HANDBOOK OF INDUSTRIAL ENGINEERING

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

SHOUT
SHARE
IT

HANDBOOK OF INDUSTRIAL ENGINEERING

Technology and Operations Management

Third Edition

Edited by

GAVRIEL SALVENDY

Purdue University

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This book is printed on acid-free paper.♾

Copyright © 2001 by John Wiley & Sons, Inc. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ @ WILEY.COM.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional person should be sought.

Disclaimer

The editor, authors, and the publisher have made every effort to provide accurate and complete inforamtion in this Handbook but the Handbook is not intended to serve as a replacement for professional advice. Any use of the information in this Handbook is at the reader's discretion. The editor, authors, and the publisher specifically disclaim any and all liability arising directly or indirectly from the use or application of any information contained in this Handbook. An appropriate professional should be consulted regarding your specific situation.

Library of Congress Cataloging-in-Publication Data:

ISBN 0471-33057-4

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

ABOUT THE EDITOR

Gavriel Salvendy is a professor of Industrial Engineering at Purdue University and is the author or coauthor of over 380 research publications, including over 190 journal papers, and the author or editor of 24 books. His publications have appeared in six languages. His main research deals with improving the usability of the network. He is the founding editor of the *International Journal on Human–Computer Interaction, International Journal on Cognitive Ergonomics,* and *Human Factors and Ergonomics in Manufacturing.* Gavriel Salvendy was the founding chair of the International Commission on Human Aspects in Computing, headquartered in Geneva, Switzerland. In 1990 he was the first member of either the Human Factors and Ergonomics Society or the International Ergonomics Association to be elected to the National Academy of Engineering. He was elected "for fundamental contributions to and professional leadership in human, physical, and cognitive aspects of engineering systems." He is the 1991 recipient of the Mikhail Vasilievich Lomonosov Medal (founder medal) of the Russian Academy of Science. This was the first time that this award was presented to a scientist outside the former USSR. In 1995 he received an Honorary Doctorate from the Chinese Academy of Sciences "for great contributions to the development of science and technology and for the great influence upon the development of science and technology in China." He is the fourth person in all fields of science and engineering in the 45-year history of the Academy ever to receive this award. He is an honorary fellow and life member of the Ergonomics Society, a Fellow of the International Ergonomics Association, Human Factors and Ergonomics Society, Institute of Industrial Engineers, and the American Psychological Association. He has advised organizations and corporations in 23 countries on the human side of effective design, implementation, and management of advanced technologies in the workplace. He earned his Ph.D. in engineering production at the University of Birmingham, United Kingdom.

ADVISORY BOARD

Hans-Jörg Bullinger
Professor and Head
Fraunhofer IAO and
Stuttgart University
Germany
Hans-Joerg.Bullinger@iao.fhg.de

John A. Buzacott
Professor
Schulich School of Business
York University
Toronto, Ontario
Canada
ibuzacot@bus.yorku.ca

Kenneth E. Case
Regents Professor
School of Industrial Engineering and Management
Oklahoma State University
Stillwater, Oklahoma
USA
kcase@okway.okstate.edu

Don B. Chaffin
The Johnson Prof. of Industrial & Operations Engr.
Director, Ergonomics Center
The University of Michigan
Ann Arbor, Michigan
USA
dchaffin@engin.umich.edu

Johnson A. Edosomwan
Chairman and Senior Executive Consultant
Johnson & Johnson Associates, Inc.
Fairfax, Virginia
USA
jedosomwan@jjaconsultants.com

Takao Enkawa
Professor
Department of Industrial Engineering
Tokyo Institute of Technology
Tokyo, Japan
enkawa@ie.me.titech.ac.jp

Shlomo Globerson
Professor
School of Business Administration
Tel Aviv University
Tel Aviv, Israel
globe@post.tau.ac.il

John J. Jarvis
Professor and Director
Dept. of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia
USA
john.jarvis@isye.gatech.edu

C. Richard Liu
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
liuch@ecn.purdue.edu

Nicolas Marmaras
Assistant Professor
Department of Mechanical Engineering
National Technical University of Athens
Sector of Industrial Management and Operational Research
Zografou, Greece
marmaras@central.ntua.gr

Aura Castillo Matias
Associate Professor and Deputy Executive Director
National Engineering Center
University of Philippines
College of Engineering, Dept. of IE
Diliman Quezon City
Philippines
matias@engg.upd.edu.ph

Barry M. Mundt
Principal
The Strategy Facilitation Group
Rowayton, Connecticut
USA
barry_mundt@earthlink.net

Gerald Nadler
IBM Chair Emeritus in Engr. Mgmt.
Industrial & Systems Engr. Dept.
University of Southern California
Los Angeles, California
USA
nadler@mizar.usc.edu

Deborah J. Nightingale
Senior Lecturer, Lean Aircraft Initiative
Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, Massachusetts
USA
dnight@mit.edu

Shimon Y. Nof
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
nof@ecn.purdue.edu

Ahmet Fahri Ozok
Professor & Head
Department of Industrial Engineering
Technical University of Istanbul
Istanbul, Turkey
isozok@tritu.bitnet

Juan R. Perez
Industrial Engineering Manager
Strategic Process Management Group
United Parcel Service
Atlanta, Georgia
USA
mla2jxp@is.ups.com

John V. Pilitsis
President and Chief Executive
Quantum Component LCC
Shrewsbury, Massachusetts
USA
stejvp@attme.att.com

John Powers
Executive Director
Institute of Industrial Engineers
Norcross, Georgia
USA

A. “Ravi” Ravindran
Professor and Head
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, Pennsylvania
USA
aravi@psu.edu

Vinod Sahney
Senior Vice President
Planning and Strategic Development
Executive Offices
Henry Ford Health System
Detroit, Michigan
USA
vsahney@hfhs.org

Laurence C. Seifert
Senior Vice President
AT&T Wireless Services
American Telephone & Telegraph Corporation
Redmon, Washington
USA
larry.seifert@attws.com

Michael J. Smith
Professor
Department of Industrial Engineering
University of Wisconsin-Madison
Madison, Wisconsin
USA
mjsmith@engr.wisc.edu

James A. Tompkins
President
Tompkins Associates, Inc.
Raleigh, North Carolina
USA
jtompkins@tompkinsinc.com

Mitchell M. Tseng
Professor and Head
Department of Industrial Engineering
Hong Kong University of Science and Technology
Hong Kong
tseng@usthk.ust.hk

Hans-Jürgen Warnecke
Professor and President
Fraunhofer Gesellschaft (Society)
Leonrodstrasse
Munich, Germany
warnecke@zv.fhg.de

Cheng Wu
Professor and Director
National CIMS Engineering Research Center
Tsinghua University
Beijing, P.R. China
wuc@tsinghua.edu.cn

CONTRIBUTORS

Mary Elizabeth A. Algeo
Computer Scientist
National Institute of Standards and Technology
Gaithersburg, Maryland
USA
algeo@cme.nist.gov

Susan Archer
Director of Operations
Micro Analysis and Design, Inc.
Pearl East Circle
Boulder, CO 80301
USA

Lajos Bálint
HUNGARNET Vice President
NIIFI Head of International Relations
Budapest, Hungary
h48bal@ella.hu

Robert M. Barker
Associate Professor
Computer Information Systems
University of Louisville
Louisville, Kentucky
USA
rmbarker@louisville.edu

Edward J. Barkmeyer
Computer Scientist
Manufacturing Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland
USA
edward.barkmeyer@nist.gov

Carl N. Belack
Principal
Oak Associates, Inc.
Maynard, Massachusetts
USA
cnb@oakinc.com

Yair Berson
Polytechnic University

Yavuz A. Bozer
Co-Director
Joel D. Tauber Manufacturing INS
Professor, Industrial-Operations Engineering
College of Engineering
University of Michigan
Ann Arbor, Michigan
yabozer@engin.umich.edu

James T. Brakefield
Professor
Department of Management
Western Illinois University
Macomb, Illinois
USA
J-Brakefield@wiu.edu

Martin Braun
Research Scientist
Competence Center Human Engineering
Fraunhofer Institute for Industrial Engineering
Stuttgart, Germany
martin.braun@iao.fhg.de

Ralf Breining
Scientist
Competence Center Virtual Reality
Fraunhofer Institute for Industrial Engineering
Stuttgart, Germany
ralf.breining@iao.fhg.de

Hans-Jörg Bullinger
Professor, Head and Director
Fraunhofer Institute of Industrial Engineering and IAT University Stuttgart
Stuttgart, Germany
Hans-Joerg.Bullinger@iao.fhg.de

Richard N. Burns
BCW Consulting Limited
Kingston, Ontario
Canada
burnsrn@attCANADA.net

John A. Buzacott
Professor
Schulich School of Business
York University
Toronto, Ontario
Canada
jbuzacot@bus.yorku.ca

Michael A. Campion
Professor
Krannert School of Management
Purdue University
West Lafayette, Indiana
USA
campionm@mgmt.purdue.edu

Pascale Carayon
Associate Professor
Department of Industrial Engineering
University of Wisconsin-Madison
Madison, Wisconsin
USA
carayon@ie.engr.wisc.edu

Tom M. Cavalier
Professor
The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering
The Pennsylvania State University
University Park, Pennsylvania
USA
tmc7@psu.edu

Thomas Cebulla
Project Leader
Fraunhofer Institute of Industrial Engineering
Stuttgart, Germany

José A. Ceroni
Professor of Industrial Engineering
Dept. of Industrial Engineering
Catholic University of Valparaí
Chile
jceroni@ucv.cl

Don B. Chaffin
The Lawton and Louise Johnson Professor
Industrial and Operations Engineering
The University of Michigan
Ann Arbor, Michigan
USA
dchaffin@engin.umich.edu

S. Chandrasekar
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
chandy@ecn.purdue.edu

Chung-Pu Chang
Professor
Eureka Consulting Company
USA

Tien-Chien Chang
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
tchang@ecn.purdue.edu

Chin I. Chiang
Institute of Traffic and Transportation
National Chiao Tung University
Taiwan
R.O. China

Soon-Yong Choi
Economist/Assistant Director
Center for Research in Electronic Commerce
Department of MSIS
University of Texas
Austin, Texas
USA

Tim Christiansen
Montana State University
txchris@montana.edu

Terry R. Collins
Assistant Professor
Industrial Engineering Department
University of Arkansas
Fayetteville, Arkansas
USA

Kevin Corker
Associate Professor
Computer Information and Systems Engineering Department
San José State University
San José, California
USA
kcorker@email.sjsu.edu

Tarsha Dargan
Professor
Department of Industrial Engineering
FAMU-FSU College of Engineering
Tallahassee, Florida
USA

Ko de Ruyter
Professor
Department of Marketing and Marketing Research
Maastricht University
Faculty of Economics & Business
Maastricht
The Netherlands
k.deruyter@mw.unimaas.nl

Xiao Deyun
Vice Director of Institute of Inspect and Electronic Technology
Department of Automation
Tsinghua University
Beijing
P.R. China
xiao@cims.tsinghua.edu.cn

Brian L. Dos Santos
Frazier Family Professor of Computer Information Systems
College of Business & Public Administration
University of Louisville
Louisville, Kentucky
USA
bldoss01@acm.org

Colin G. Drury
Professor
Department of Industrial Engineering
State University of New York at Buffalo
Buffalo, New York
USA
drury@buffalo.edu

Laura Raiman DuPont
Consultant
Quality Engineering Consultant
League City, Texas
USA
lrdupont@aol.com

Taly Dvir
Lecturer
Faculty of Management
Tel Aviv University
Tel Aviv, Israel
talyd@post.tau.ac.il

Andrea Edler
Head, Technological Planning Systems Dept.
Fraunhofer Institute for Production Systems and Design Technology
Berlin, Germany
andreas.edler@ipk.fhg.de

Johnson A. Edosomwan
Chairman and Senior Executive Consultant
Johnson & Johnson Associates, Inc.
Fairfax, Virginia
USA
jedosomwan@jjaconsultants.com

John R. English
Professor
Department of Industrial Engineering
University of Arkansas
Fayetteville, Arkansas
USA
jre@engr.uark.edu

Takao Enkawa
Professor
Industrial Engineering & Management
Tokyo Institute of Technology
Tokyo, Japan
enkawa@ie.me.titech.ac.jp

Klaus-Peter Fähnrich
Head
FHG-IAO
Stuttgart University
Stuttgart, Germany
klaus-peter.faehnrich@iao.fhg.de

Richard A. Feinberg
Professor
Consumer Sciences & Retailing
Purdue University
West Lafayette, Indiana
feinberger@cfs.purdue.edu

Klaus Feldmann
Professor and Director
Institute for Manufacturing Automation and Production Systems
University of Erlangen-Nuremberg
Erlangen, Germany
feldmann@faps.uni-erlangen.de

Martin P. Finegan
Director
KPMG LLP
Montvale New Jersey
USA
mfinegan@kpmg.com

Jeffrey H. Fischer
Director
UPS Professional Services
Atlanta, Georgia
USA
ner1jhf@ups.com

G. A. Fleischer
Professor Emeritus
Department of Industrial Engineering
University of Southern California
Los Angeles, California
USA
fleische@mizar.usc.edu

Donald Fusaro
Member of Technical Staff
Optoelectronics Quality
Lucent Technologies Bell Labs Innovations
Breinigsville, Pennsylvania
USA
fusaro@lucent.com

Alberto Garcia-Diaz
Professor
Department of Industrial Engineering
Texas A&M University
College Station, Texas
USA
agd@tamu.edu

Boaz Golany
Professor
Faculty of Industrial Engineering and Management
Technion—Israel Institute of Technology
Haifa, Israel
golany@ie.technion.ac.il

Juergen Goehringer
Scientific Assistant
Institute for Manufacturing Automation and Production Systems
Friedrich Alexander University Erlangen NurembergErlangen
Germany
goehringer@faps.un-erlangen.de

Frank Habermann
Senior Doctoral Researcher
Institute for Information Systems
Saarland University
Saarbruecken, Germany
Habermann@iwi.uni-sb.de

Michael Haischer
Fraunhofer Institute of Industrial Engineering
Stuttgart, Germany

John M. Hannon
Visiting Associate Professor
Jacobs Management Center
State University of New York-Buffalo
Buffalo, New York
USA
jmhannon@acsu.buffalo.edu

Joseph C. Hartman
Assistant Professor
Industrial and Manufacturing Systems Engineering
Lehigh University
Bethlehem, Pennsylvania
USA
jch6@lehigh.edu

On Hashida
Professor
Graduate School of Systems Management
University of Tsukuba
Tokyo, Japan
hashida@gssm.otsuka.tsukuba.ac.jp

Peter Heisig
Head
Competence Center Knowledge Management
Fraunhofer Institute for Production Systems and Design Technology
Berlin, Germany
Peter.Heisig@ipk.fhg.de

Markus Helmke
Project Leader
Institute for Machine Tools & Factory Management
Technical University Berlin
Berlin, Germany
Markus.Helmke@ipk.fhg.de

Klaus Herfurth
Professor
Technical University of Chemnitz
Germany

Ingo Hoffmann
Project Manager
Competence Center Knowledge Mgmt.
Fraunhofer Institute for Production Systems and Design Technology
Berlin, Germany
Info.Hoffmann@ipk.fhg.de

Chuck Holland
Portfolio Project Manager
United Parcel Service
Atlanta, Georgia
USA

Clyde W. Holsapple
Rosenthal Endowed Chair in MIS
Gatton College of Business and Economics
University of Kentucky
Lexington, Kentucky
USA
cwhols@pop.uky.edu

Chin-Yin Huang
Assistant Professor
Department of Industrial Engineering
Tunghai University
Taiwan

Ananth V. Iyer
Professor
Krannert School of Management
Purdue University
West Lafayette, Indiana
USA
aiyer@mgmt.purdue.edu

Robert B. Jacko
Professor
School of Civil Engineering
Purdue University
West Lafayette, Indiana
USA
jacko@ecn.purdue.edu

Jianxin Jiao
Assistant Professor
School of Mechanical & Production Engineering
Nanyang Technological University
Singapore

Albert T. Jones
Operations Research Analyst
Manufacturing Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland
USA
albert.jones@nist.gov

Swatantra K. Kachhal
Professor and Chair
Department of Industrial and Manufacturing Systems Engineering
University of Michigan-Dearborn
Dearborn, Michigan
USA
kachhal@umich.edu

Kailash C. Kapur
Professor
Industrial Engineering
The University of Washington
Seattle, Washington
USA
kkapur@u.washington.edu

Ben-Tzion Karsh
Research Scientist
Center for Quality and Productivity Improvement
University of Wisconsin-Madison
Madison, Wisconsin
USA
bkarsh@facstaff.wisc.edu

Waldemar Karwowski
Professor and Director
Center for Industrial Ergonomics
University of Louisville
Louisville, Kentucky
USA
karwowski@louisville.edu

Anton J. Kleywegt
Professor
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia
USA
Anton.Kleywegt@isye.gatech.edu

Tom Kontogiannis
Professor
Department of Production Engineering and Management
Technical University of Crete
Greece

Stephan Konz
Professor Emeritus
Department of Industrial and Manufacturing Systems Engineering
Kansas State University
Manhattan, Kansas
USA
sk@ksu.edu

Timothy M. C. LaBreche
Senior Research Engineer
Environmental & Hydraulics Engineering Area
School of Civil Engineering
Purdue University
West Lafayette, Indiana
USA

Frank-Lothar Krause
Professor
Institute for Machine Tools & Factory Management
Technical University Berlin
Berlin, Germany
Frank-L.Krause@ipk.fhg.de

Douglas M. Lambert
Mason Professor of Transportation and Logistics
Fisher College of Business
The Ohio State University
Prime F. Osborn III Eminent Scholar Chair in Transportation
University of North Florida
lambert@cob.ohio-state.edu

R. McFall Lamm, Jr.
Chief Investment Strategist
Deutsche Bank
New York, New York
USA
mac.lamm@db.com

K. Ronald Laughery, Jr.
President
Micro Analysis and Design, Inc.
Boulder, Colorado
USA
rlaughery@maad.com

Yuan-Shin Lee
Professor
Department of Industrial Engineering
North Carolina State University
Raleigh, North Carolina
USA

Mark R. Lehto
Associate Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
lehto@ecn.purdue.edu

Jens Leyh
Project Leader
Fraunhofer Institute of Industrial Engineering
Nobelstrasse 12
Stuttgart, Germany

C. Richard Liu
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
liuch@ecn.purdue.edu

Raymond P. Lutz
Professor
School of Management
The University of Texas at Dallas
Richardson, Texas
USA

Ann Majchrzak
Professor, Information Systems
Dept. of Information and Operations Management
Marshall School of Business
University of Southern California
University Park
Los Angeles, California
USA
majchrza@bus.usc.edu

Chryssi Malandraki
Lead Research Analyst
United Parcel Service
Atlanta, Georgia
USA

Tamas Maray
Associate Professor
Muegyetem Rakpart 1–3
Technical University of Budapest
Budapest H-1111
Hungary
maray@fsz.bme.hu

Nicolas Marmaras
Assistant Professor
Department of Mechanical Engineering
National Technical University of Athens
Sector of Industrial Management and Operational Research
Zografou, Greece
marmaras@central.ntua.gr

Frank O. Marrs
President/CEO
Risk Management Partners, Inc.
USA
fmarrs@wpe.com

Roy Marsten
Cutting Edge Optimization
USA

Aura Castillo Matias
Associate Professor and Deputy Executive Director
National Engineering Center
University of Philippines
College of Engineering, Dept. of IE
Diliman Quezon City
Philippines
matias@engg.upd.edu.ph

Gina J. Medsker
Senior Staff Scientist
Human Resources Research Organization
Alexandria, Virginia
USA
gmedsker@humrro.org

Emmanuel Melachrinoudis
Associate Professor
Department of Mechanical, Industrial and Manufacturing Engineering
Northeastern University
Boston, Masachusetts
USA
emelas@coe.neu.edu

Kai Mertins
Division Director Corporate Management
Fraunhofer Institute for Production Systems and Design Technology
Berlin, Germany
Kai.Mertins@ipk.fhg.de

Najmedin Meshkati
Associate Professor
Institute of Safety and Systems Management
University of Southern California
Los Angeles, California
USA
meshkati@usc.edu

George T. Milkovich
Catherwood Professor of Human Resource Studies
Center for Advanced Human Resource Studies
Cornell University
Ithaca, New York
USA
gtml@cornell.edu

Hokey Min
Executive Director
Logistics and Distribution Institute
University of Louisville
Louisville, Kentucky
USA
h0min001@gwise.louisville.edu

Kent B. Monroe
Professor
Department of Business Administration
University of Illinois at Urbana-Champaign
Champaign, Illinois
USA
k.monroe1@home.com

Barry M. Mundt
Principal
The Strategy Facilitation Group
Rowayton, Connecticut
USA
barry_mundt@earthlink.net

Kenneth Musselman
Senior Business Consultant
Frontstep, Inc.
West Lafayette, Indiana
USA

Barry L. Nelson
Professor
Department of Industrial Enginering and Management Sciences
Northwestern University
Evanston, Illinois
USA
nelsonb@iems.nwu.edu

Douglas C. Nelson
Professor
Department of Hospitality and Tourism Management
Purdue University
West Lafayette, Indiana
USA
nelsond@cfs.purdue.edu

Jens-Günter Neugebauer
Director, Automation
Fraunhofer Institute for Manufacturing Engineering and Automation
Stuttgart, Germany
jen@ipa.fhg.de

Reimund Neugebauer
Professor/Managing Director
Fraunhofer Institute for Machine Tools and Forming Technology
Chemnitz, Germany
neugebauer@iwu.fhg.de

Jerry M. Newman
Professor
School of Management
State University of New York-Buffalo
Buffalo, New York
USA
jmnewman@acsu.buffalo.edu

Abe Nisanci
Professor and Director
Research and Sponsored Programs
Industrial and Manufacturing Engineering and Technology
Bradley University
Peoria, Illinois
USA
ibo@bradley.edu

Shimon Y. Nof
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
nof@ecn.purdue.edu

Colm A. O'Cinneide
Associate Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
colm@ecn.purdue.edu

Phillip F. Ostwald
Professor Emeritus
Mechanical and Industrial Engineering
University of Colorado at Boulder
Boulder, Colarado
USA
philip.ostwald@colorado.edu

Raja M. Parvez
Vice President of Manufacturing and Quality
Fitel Technologies
Perryville Corporate Park
Perryville, Illinois
USA
rparvez@fiteltech.com

Richard B. Pearlstein
Director
Training Performance Improvement
American Red Cross
Charles Drew Biomedical Institute
Arlington, Virginia
USA
pearlstr@usa.redcross.org

Juan R. Perez
Industrial Engineering Manager
Strategic Process Management Group
United Parcel Service
Atlanta, Georgia
USA
mla2jxp@is.ups.com

Ralph W. "Pete" Peters
Principal
Tompkins Associates Inc.
Raleigh, North Carolina
USA
ppeters@tompkinsinc.com

Don T. Phillips
Professor
Department of Industrial Engineering
Texas A&M University
College Station, Texas
USA
drdon@tamu.edu

Michael Pinedo
Professor
Department of Operations Management
Stern School of Business
New York University
New York, New York
USA
mpinedo@stern.nuy.edu

David F. Poirier
Executive Vice President & CIO
Hudson's Bay Company
Toronto, Ontario
Canada

Lloyd Provost
Improvement Advisor
Associates In Process Improvement
Austin, Texas
USA
provost@fc.net

Ronald L. Rardin
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
Rardin@ecn.purdue.edu

Ulrich Raschke
Program Manager
Human Simulation Technology
Engineering Animation, Inc.
Ann Arbor, Michigan
USA
ulrich@eai.com

A. "Ravi" Ravindran
Professor and Head
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, Pennsylvania
USA
aravi@psu.edu

David Rodrick
University of Louisville
Louisville, Kentucky
USA

James R. Ross
Resource Management Systems, Inc.
USA

William B. Rouse
Chief Executive Officer
Enterprise Support Systems
Norcross, Georgia
USA
brouse@ess-advisors.com

Andrew P. Sage
Founding Dean Emeritus
University and First American Bank Professor
School of Information Technology and Engineering
Department of Systems Engineering and Operations Research
George Mason University
Fairfax, Virginia
USA
asage@gmu.edu

François Sainfort
Professor, Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive
Atlanta, Georgia
USA
sainfort@isye.gatech.edu

Hiroyuki Sakata
NTT Data Corporation
Tokyo, Japan
sakata@rd.nttdata.co.jp

Gavriel Salvendy
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
salvendy@ecn.purdue.edu

August-Wilhelm Scheer
Director
Institute of Information Systems
Saarland University
Saarbrucken, Germany
scheer@iwi.uni-sb.de

Stefan Schmid
Professor
Department Assembly Systems
Fraunhofer Institute for Manufacturing Engineering and Automation
Stuttgart, Germany
sas@ipa.fhg.de

Rolf Dieter Schraft
Director Automation
Fraunhofer Institute for Manufacturing Engineering and Automation
Stuttgart, Germany
rds@ipa.fhg.de

Lisa M. Schutte
Principal Scientist
Human Simulation Research and Development Engineering Animation Inc.
Ann Arbor, Michigan
USA
1schutte@eai.com

Shane J. Schvaneveldt
Fulbright Scholar Visiting Researcher
Tokyo Institute of Technology
Tokyo, JAPAN, and
Associate Professor of Management
Goddard School of Business and Economics
Weber State Univeristy
Ogden, Utah
USA
schvaneveldt@weber.edu

Robert E. Schwab
Engineering Manager
Chemical Products
Catepillar Inc.
Mossville, Illinois
USA
schwab_robert_e@cat.com

Sridhar Seshadri
Professor
Department of Operations Management
Stern School of Business
New York University
New York, New York
USA

J. George Shanthikumar
Professor
Haas School of Business
University of California at Berkeley
Berkeley, California
USA
shanthik@haas.berkeley.edu

Alexander Shapiro
Professor
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia
USA
ashapiro@isye.gatech.edu

Gunter P. Sharp
Professor
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia
USA
gsharp@isye.gatech.edu

Avraham Shtub
Professor
Industrial Engineering and Management
Technion—Israel Institute of Technology
Haifa, Israel
shtub@ie.technion.ac.il

Edward A. Siecienski
Executive Director
JMA Supply Chain Management
Hamilton, New Jersey
USA
eas.999@worldnet.att.net,

Wilfried Sihn
Director Corporate Management
Fraunhofer Institute for Manufacturing Engineering and Automation
Stuttgart, Germany
whs@ipa.fhg.de

D. Scott Sink
President
World Confederation of Productivity Science
Moneta, Virginia
USA
dssink@avt.edu

David Simchi-Levi
Professor
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology
77 Massachusetts Avenue, Rm. 1–171
Cambridge, Massachusetts
USA
dslevi@mit.edu

Edith Simchi-Levi
Vice President Operations
LogicTools Inc.
71 Meriam Street
Lexington, MA 02420
USA
(781)861-3777
(312)803-0448 (FAX)
edith@logic-tools.com

Douglas K. Smith
Author and Consultant
La Grangeville, New York
USA
dekaysmith@aol.com

Francis J. Smith
Principal
Francis J. Smith Management Consultants
Longmeadow, MA 01106
USA
FSmith1270@aol.com

Keith V. Smith
Professor
Krannert School of Management
Purdue University
West Lafayette, IN 47907-1310
USA
kvsmith@mgmt.purdue.edu

George L. Smith
Professor Emeritus
The Ohio State University
Columbus, Ohio
USA

Jerry D. Smith
Executive Vice President
Tompkins Associates, Inc.
Raleigh, North Carolina
USA
jsmith@tompkinsinc.com

Michael J. Smith
Professor
Department of Industrial Engineering
University of Wisconsin-Madison
Madison, Wisconsin
USA
mjsmith@engr.wisc.edu

Kay M. Stanney
Associate Professor
Industrial Engineering and Mgmt. Systems
University of Central Florida
Orlando, Florida
USA
stanney@mail.ucf.edu

Julie Ann Stuart
Assistant Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
stuart@ecn.purdue.edu

Robert W. Swezey
President
InterScience America, Inc.
Leesburg, Virginia
USA
isai@erols.com

Alvaro D. Taveira
Professor
University of Wisconsin-Whitewater
Whitewater, Wisconsin
USA

John Taylor
Coordinator, Research Programs & Service
Department of Industrial Engineering
FAMU-FSU College of Engineering
Tallahassee Florida
USA
jotaylor@.eng.fsu.edu

Oliver Thomas
Senior Doctoral Researcher
Institute for Information Systems
Saarland University
Saarbruecken, Germany
Thomas@iwi.uni-sb.de

James A. Tompkins
President
Tompkins Associates, Inc.
Raleigh, North Carolina
USA
jtompkins@tompkinsinc.com

Mitchell M. Tseng
Professor and Head
Department of Industrial Engineering
Hong Kong University of Science and Technology
Hong Kong
tseng@usthk.ust.hk

Gwo Hshiung Tzeng
Professor
College of Management
National Chiao Tung University
Taiwan
R.O. China

Reha Uzsoy
Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
uzsoy@ecn.purdue.edu

Ralf von Briel
Project Leader
Fraunhofer Institute for Manufacturing Engineering and Automation
Stuttgart, Germany

Harrison M. Wadsworth Jr.
Professor
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia
USA
harrison.wadsworth@isye.gatech.edu

William P. Wagner
Professor
Decision and Information Technologies
College of Commerce and Finance
Villanova University
Villanova, Pennsylvania
USA

Evan Wallace
Electronics Engineer
Manufacturing Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland
USA

Ben Wang
DOE Massie Chair and Professor
Department of Industrial Engineering
FAMU-FSU College of Engineering
Tallahassee, Florida
USA
indwang1@engr.fsu.edu

H. Samuel Wang
Provost
Chung Yuan Christian University
Taiwan
R.O. China

Hans-Jürgen Warnecke
Professor and President
Fraunhofer Gesellschaft (Society)
Leonrodstrasse
Munich, Germany
warnecke@zv.fhg.de

Joachim Warschat
Professor
Frunhofer Institute of Industrial Engineering
Nobelstrasse 12
Stuttgart, Germany

Martin Wetzels
Assistant Professor
Maastricht Academic Center for Research in Services
Maastricht University
Faculty of Economics & Business Administration
Maastricht
The Netherlands
m.wetzel@mw.unimaas.nl

Andrew B. Whinston
Hugh Cullen Chair Professor of Information Systems, Economics and Computer Science Director, Center for Research in Electronic Commerce
University of Texas
Graduate School of Business
Austin, Texas
USA
abw@uts.cc.utexas.edu

Richard T. Wong
Senior Systems Engineer
Telcordia Technologies, Inc.
Piscataway, New Jersey
USA
rwong1@telcordia.com

Andrew L. Wright
Assistant Professor
College of Business & Public Administration
University of Louisville
Louisville, Kentucky
USA
andrew.wright@louisville.edu

Cheng Wu
Professor and Director
National CIMS Engineering Research Center
Tsinghua University
Beijing, P.R. China
wuc@tsinghua.edu.cn

Xiaoping Yang
Doctoral Student School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
xiaoping@ecn.purdue.edu

David D. Yao
Professor
Industrial Engineering and Operations Research
Columbia University
New York, New York
USA
yao@ieor.columbia.edu

Yuehwern Yih
Associate Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana
USA
yih@ecn.purdue.edu

Po Lung Yu
Carl A. Scupin Distinguished Professor
School of Business
University of Kansas
Lawrence, Kansas
USA
pyu@bschool.wpo.ukans.edu

Fan YuShun
Professor and Vice Director
Institute of System Integration
Department of Automation
Tsinghua University
Beijing, P.R. China
fan@cims.tsinghua.edu.cn

David Zaret
Lead Research Analyst
United Parcel Service
Atlanta, Georgia
USA

FOREWORD

Many people speculated about what industrial engineering might look like in the 21st century, and now here we are. It is exciting to see how the profession's definition of the work center has broadened to embrace the information age and the global economy. Industrial engineers, with their ever-expanding toolbox, have a greater opportunity to help corporations be successful than ever before.

But while these changes and opportunities are exciting, they presented a challenge to the editor of this handbook. I met with Gavriel Salvendy as he began working to integrate all the thoughtful input he had received from his advisory committee, so I had a firsthand opportunity to observe the energy and creativity required to develop this content-rich edition. This handbook is truly an accomplishment, fully supporting industrial engineers engaged in traditional as well as new facets of the profession.

This edition has stepped up to this multidimensional challenge. The Technology section updates the previous edition, with coverage of topics such as decision support systems, but also delves into information-age topics such as electronic commerce. Look closer and you'll see that attention has been given to the growing services sector, including chapters covering specific application areas.

"Enterprise" has become a popular way to describe the total system and the broad organizational scope of problem-solving initiatives. I am pleased to see this addressed specifically through topics such as enterprise modeling and enterprise resource planning, as well as globally, as in the Management, Planning, Design, and Control section. This edition truly recognizes that IEs can and do contribute in every phase of the total product life cycle.

The mission of the Institute of Industrial Engineers is to support the professional growth of practicing industrial engineers. This third edition is an all-encompassing Handbook for the IE professional that rises to the task. Students, engineers of all types, and managers will find this a useful and insightful reference.

JOHN POWERS
Executive Director
Institute of Industrial Engineers

PREFACE

Industrial Engineering has evolved as a major engineering and management discipline, the effective utilization of which has contributed to our increased standard of living through increased productivity, quality of work and quality of services, and improvements in the working environments. The *Handbook of Industrial Engineering* provides timely and useful methodologies for achieving increased productivity and quality, competitiveness, globalization of business and for increasing the quality of working life in manufacturing and service industries. This Handbook should be of value to all industrial engineers and managers, whether they are in profit motivated operations or in other nonprofit fields of activity.

The first edition of the *Handbook of Industrial Engineering* was published in 1982. It has been translated into Japanese and published by the Japan Management Association; translated into Spanish and published by Editorial Lemusa; published in a special edition in Taiwan by Southeast Book Company; and translated into Chinese and published by Mechanical Industry Publisher; and adopted by the Macmillan book club. The Handbook was selected by the Association of American Publishers as the Best Professional and Scholarly Publication in 1982 and was widely distributed by the Institute of Industrial Engineers (IIE). The Foreword of the first edition of the Handbook was written by Donald C. Burnham, retired Chairman of Westinghouse Electric Corporation. In this Foreword Burnham wrote, "The Industrial Engineering principles that are outlines are timeless and basic and should prove useful to corporations, both large and small, both continuous process as well as discrete part manufacturing, and especially to those working in the service industries where most of the jobs are today." The second edition of the *Handbook of Industrial Engineering* maintained this thrust.

In the Foreword to the second edition, the former president of NEC Corporation wrote, "The Second Edition of the Handbook of Industrial Engineering will serve as an extremely powerful tool for both industrial engineers and managers." The many contributing authors came through magnificently. I thank them all most sincerely for agreeing so willingly to create this Handbook with me.

Each submitted chapter was carefully reviewed by experts in the field and myself. Much of the reviewing was done by the Advisory Board. In addition, the following individuals have kindly contributed to the review process: Stephan A. Konz, K. Ronald Laughery, Jack Posey, William B. Rouse, Kay M. Stanney, Mark Spearman, and Arnold L. Sweet.

For the third edition of this Handbook, 97 of the 102 chapters were completely revised and new sections added in project management (3 chapters), supply-chain management and logistics (7 chapters) and the number of chapters in service systems increased from 2 to 11 chapters. The 102 chapters of this third edition of the handbook were authored by 176 professionals with diverse training and professional affiliations from around the world. The Handbook consists of 6441 manuscript pages, 922 figures, 388 tables, and 4139 references that are cited for further in-depth coverage of all aspects of industrial engineering.

The editing of the third edition of the Handbook was made possible through the brilliant, most able, and diligent work of Kim Gilbert, my administrative assistant, who so effectively coordinated and managed all aspects of the Handbook preparation. My sincere thanks and appreciation go to her. It was a true pleasure working on this project with Bob Argentieri, the John Wiley senior editor, who is the very best there is and was a truly outstanding facilitator and editor for this Handbook.

GAVRIEL SALVENDY

West Lafayette, Indiana
September 2000

CONTENTS

I. Industrial Engineering Function and Skills **1**

1. Full Potential Utilization of Industrial and Systems Engineering in Organizations, *by D. Scott Sink, David F. Poirier, and George L. Smith* 3
2. Enterprise Concept: Business Modeling Analysis and Design, *by Frank O. Marrs and Barry M. Mundt* 26

II. Technology **61**

A. *Information Technology* 63

3. Tools for Building Information Systems, *by Robert M. Barker, Brian L. Dos Santos, Clyde W. Holsapple, William P. Wagner, and Andrew L. Wright* 65
4. Decision Support Systems, *by Andrew P. Sage* 110
5. Automation Technology, *by Chin-Yin Huang and Shimon Y. Nof* 155
6. Computer Integrated Technologies and Knowledge Management, *by Frank-Lothar Krause, Kai Mertins, Andreas Edler, Peter Heisig, Ingo Hoffmann, and Markus Helmke* 177
7. Computer Networking, *by Lajos Bálint and Tamás Máray* 227
8. Electronic Commerce, *by Soon-Yong Choi and Andrew B. Whinston* 259
9. Enterprise Modeling, *by August-Wilhelm Scheer, Frank Habermann, and Oliver Thomas* 280

B. *Manufacturing and Production Systems* 309

10. The Factory of the Future: New Structures and Methods to Enable Transformable Production, *by Hans-Jürgen Warnecke, Wilfried Sihn, and Ralf von Briel* 311
11. Enterprise Resource Planning Systems in Manufacturing, *by Mary Elizabeth A. Algeo and Edward J. Barkmeyer* 324
12. Automation and Robotics, *by Rolf Dieter Schraft, Jens-Günter Neugebauer, and Stefan Schmid* 354

13. Assembly Process, *by K. Feldmann* 401
14. Manufacturing Process Planning and Design, *by Tien-Chien Chang and Yuan-Shin Lee* 447
15. Computer Integrated Manufacturing, *by Cheng Wu, Fan YuShun, and Xiao Deyun* 484
16. Clean Manufacturing, *by Julie Ann Stuart* 530
17. Just-in-Time, Lean Production, and Complementary Paradigms, *by Takao Enkawa and Shane J. Schvaneveldt* 544
18. Near-Net-Shape Processes, *by Reimund Neugebauer and Klaus Herfurth* 562
19. Environmental Engineering: Regulation and Compliance, *by Robert B. Jacko and Timothy M. C. LaBreche* 589
20. Collaborative Manufacturing, *by José A. Ceroni and Shimon Y. Nof* 601

C. *Service Systems* 621

21. Service Industry Systems and Service Quality, *by Martin Wetzels and Ko de Ruyter* 623
22. Assessment and Design of Service Systems, *by Michael Haischer, Hans-Jörg Bullinger, and Klaus-Peter Fähnrich* 634
23. Customer Service and Service Quality, *by Richard A. Feinberg* 651
24. Pricing and Sales Promotion, *by Kent B. Monroe* 665
25. Mass Customization, *by Mitchell M. Tseng and Jianxin Jiao* 684
26. Client/Server Technology, *by On Hashida and Hiroyuki Sakata* 710
27. Industrial Engineering Applications in Health Care Systems, *by Swatantra K. Kachhal* 737
28. Industrial Engineering Applications in Financial Asset Management, *by R. McFall Lamm, Jr.* 751
29. Industrial Engineering Applications in Retailing, *by Richard A. Feinberg and Tim Christiansen* 772
30. Industrial Engineering Applications in Transportation, *by Chryssi Malandraki, David Zaret, Juan R. Perez, and Chuck Holland* 787
31. Industrial Engineering Applications in Hotels and Restaurants, *by Douglas C. Nelson* 825

III. Performance Improvement Management **837**

A. *Organization and Work Design* 839

32. Leadership, Motivation, and Strategic Human Resource Management, *by Taly Dvir and Yair Berson* 841

33. Job and Team Design, *by Gina J. Medsker and Michael A. Campion* 868
34. Job Evaluation in Organizations, *by John M. Hannon, Jerry M. Newman, George T. Milkovich, and James T. Brakefield* 899
35. Selection, Training, and Development of Personnel, *by Robert W. Swezey and Richard B. Pearlstein* 920
36. Aligning Technological and Organizational Change, *by Ann Majchrzak and Najmedin Meshkati* 948
37. Teams and Team Management and Leadership, *by François Sainfort, Alvaro D. Taveira, and Michael J. Smith* 975
38. Performance Management, *by Martin P. Finegan and Douglas K. Smith* 995

B. *Human Factors and Ergonomics* 1011

39. Cognitive Tasks, *by Nicolas Marmaras and Tom Kontogiannis* 1013
40. Physical Tasks: Analysis, Design, and Operation, *by Waldemar Karwowski and David Rodrick* 1041
41. Ergonomics in Digital Environments, *by Ulrich Raschke, Lisa M. Schutte, and Don B. Chaffin* 1111
42. Human Factors Audit, *by Colin G. Drury* 1131
43. Design for Occupational Health and Safety, *by Michael J. Smith, Pascale Carayon, and Ben-Tzion Karsh* 1156
44. Human–Computer Interaction, *by Kay M. Stanney, Michael J. Smith, Pascale Carayon, and Gavriel Salvendy* 1192

IV. Management, Planning, Design, and Control **1237**

A. *Project Management* 1239

45. Project Management Cycle: Process Used to Manage Project (Steps to Go Through), *by Avraham Shtub* 1241
46. Computer-Aided Project Management, *by Carl N. Belack* 1252
47. Work Breakdown Structure, *by Boaz Golany and Avraham Shtub* 1263

B. *Product Planning* 1281

48. Planning and Integration of Product Development, *by Hans-Jörg Bullinger, Joachim Warschat, Jens Leyh, and Thomas Cebulla* 1283
49. Human-Centered Product Planning and Design, *by William B. Rouse* 1296
50. Design for Manufacturing, *by C. Richard Liu and Xiaoping Yang* 1311

51. Managing Professional Services Projects, *by Barry M. Mundt and Francis J. Smith* 1332

C. *Manpower Resource Planning* 1351

52. Methods Engineering, *by Stephan Konz* 1353
53. Time Standards, *by Stephan Konz* 1391
54. Work Measurement: Principles and Techniques, *by Aura Castillo Matias* 1409

D. *Systems and Facilities Design* 1463

55. Facilities Size, Location, and Layout, *by James A. Tompkins* 1465
56. Material-Handling Systems, *by Yavuz A. Bozer* 1502
57. Storage and Warehousing, *by Jerry D. Smith* 1527
58. Plant and Facilities Engineering with Waste and Energy Management, *by James R. Ross* 1548
59. Maintenance Management and Control, *by Ralph W. "Pete" Peters* 1585

E. *Planning and Control* 1625

60. Queuing Models of Manufacturing and Service Systems, *by John A. Buzacott and J. George Shanthikumar* 1627
61. Production-Inventory Systems, *by David D. Yao* 1669
62. Process Design and Reengineering, *by John Taylor, Tarsha Dargan, and Ben Wang* 1695
63. Scheduling and Dispatching, *by Michael Pinedo and Sridhar Seshadri* 1718
64. Personnel Scheduling, *by Richard N. Burns* 1741
65. Monitoring and Controlling Operations, *by Albert T. Jones, Yuehwern Yih, and Evan Wallace* 1768

F. *Quality* 1791

66. Total Quality Leadership, *by Johnson A. Edosomwan* 1793
67. Quality Tools for Learning and Improvement, *by Lloyd Provost* 1808
68. Understanding Variation, *by Lloyd Provost* 1828
69. Statistical Process Control, *by John R. English and Terry R. Collins* 1856
70. Measurement Assurance, *by S. Chandrasekar* 1877
71. Human Factors and Automation in Test and Inspection, *by Colin G. Drury* 1887
72. Reliability and Maintainability, *by Kailash C. Kapur* 1921
73. Service Quality, *by Laura Raiman DuPont* 1956
74. Standardization, Certification, and Stretch Criteria, *by Harrison M. Wadsworth, Jr.* 1966

75. Design and Process Platform Characterization Methodology, *by Raja M. Parvez and Donald Fusaro* 1975

G. *Supply Chain Management and Logistics* 2005

76. Logistics Systems Modeling, *by David Simchi-Levi and Edith Simchi-Levi* 2007
77. Demand Forecasting and Planning, *by Ananth V. Iyer* 2020
78. Advanced Planning and Scheduling for Manufacturing, *by Kenneth Musselman and Reha Uzsoy* 2033
79. Transportation Management and Shipment Planning, *by Jeffrey H. Fischer* 2054
80. Restructuring a Warehouse Network: Strategies and Models, *by Hokey Min and Emanuel Melachrinoudis* 2070
81. Warehouse Management, *by Gunter P. Sharp* 2083
82. Supply Chain Planning and Management, *by Douglas M. Lambert and Edward A. Siecienski* 2110

V. Methods for Decision Making 2141

A. *Probabilistic Models and Statistics* 2143

83. Stochastic Modeling, *by Colm A. O'Cinneide* 2145
84. Decision-Making Models, *by Mark R. Lehto* 2172
85. Design of Experiments, *by H. Samuel Wang and Chung-Pu Chang* 2224
86. Statistical Inference and Hypothesis Testing, *by Don T. Phillips and Alberto Garcia-Diaz* 2241
87. Regression and Correlation, *by Raja M. Parvez and Donald Fusaro* 2264

B. *Economic Evaluation* 2295

88. Product Cost Analysis and Estimating, *by Phillip F. Ostwald* 2297
89. Activity-Based Management in Manufacturing, *by Keith V. Smith* 2317
90. Discounted Cash Flow Methods, *by Raymond P. Lutz* 2331
91. Economic Risk Analysis, *by G. A. Fleischer* 2360
92. Inflation and Price Change in Economic Analysis, *by Joseph C. Hartman* 2394

C. *Computer Simulation* 2407

93. Modeling Human Performance in Complex Systems, *by K. Ronald Laughery, Jr., Susan Archer, and Kevin Corker* 2409

94. Simulation Packages, *by Abe Nisanci and Robert E. Schwab* 2445
95. Statistical Analysis of Simulation Results, *by Barry L. Nelson* 2469
96. Virtual Reality for Industrial Engineering: Applications for Immersive Virtual Environments, *by Hans-Jörg Bullinger, Ralf Breining, and Martin Braun* 2496

D. *Optimization* 2521

97. Linear Optimization, *by A. "Ravi" Ravindran and Roy Marsten* 2523
98. Nonlinear Optimization, *by Tom M. Cavalier* 2540
99. Network Optimization, *by Richard T. Wong* 2568
100. Discrete Optimization, *by Ronald L. Rardin* 2582
101. Multicriteria Optimization, *by Po Lung Yu, Chin I. Chiang, and Gwo Hshiung Tzeng* 2602
102. Stochastic Optimization, *by Anton J. Kleywegt and Alexander Shapiro* 2625

Author Index 2651

Subject Index 2699

SECTION I

INDUSTRIAL ENGINEERING FUNCTION AND SKILLS

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 1

Full Potential Utilization of Industrial and Systems Engineering in Organizations

D. Scott Sink
Exchange Partners

David F. Poirier
Hudson's Bay Company

George L. Smith
The Ohio State University

1. **OVERVIEW** 4
    - 1.1. Full Potential Introduced 4
    - 1.2. Structure of the Chapter 4
    - 1.3. ISE Domain Defined 4
    - 1.4. Operational Definition of ISE 5
    - 1.5. Applying the Definition 5
    - 1.6. Integrating Work in Strategy and Policy, Conditions for Success, and Operations Improvement Leads to Full Potential 6
2. **THE FULL POTENTIAL MODEL FOR THE ORGANIZATION** 7
    - 2.1. Overview 7
    - 2.2. Examples of Full Potential 7
    - 2.3. Enterprise Excellence Models 8
    - 2.4. Implications for ISE 8
3. **THE FUTURE STATE VALUE PROPOSITION FOR ISE** 10
    - 3.1. Overview 10
    - 3.2. The Planning System: Position, Strategy, Implementation, and Deployment 11
        - 3.2.1. Policy Deployment 13
        - 3.2.2. Relationship Management 13
        - 3.2.3. Change Leadership: The ISE as Change Master 14
    - 3.3 Conditions for Success 15
        - 3.3.1. Culture System 15
        - 3.3.2. Infrastructure for Improvement 16
        - 3.3.3. Communication System 17
        - 3.3.4. Learning System 17
        - 3.3.5. Summary 17
    - 3.4. Operations Improvement Role 18
        - 3.4.1. Overview 18
        - 3.4.2. Business Process Improvement 18
        - 3.4.3. Building Effective Measurement Systems 20
        - 3.4.4. Organizational Systems Performance Measurement 21
4. **ORGANIZING FOR FULL-POTENTIAL ISE CONTRIBUTION** 22
    - 4.1. Overview 22
    - 4.2. Executive Sponsorship 22
    - 4.3. Business Partner Relationship Management 23
    - 4.4. Integration Role 23
5. **IIE/CIE/CIEADH RELATIONSHIP MANAGEMENT** 23
    - 5.1. Overview 23
    - 5.2. Relationship Management 23
6. **MAKING THE FULL-POTENTIAL MODEL WORK FOR YOU: LEADERSHIP AND PERSONAL MASTERY** 23

REFERENCES 24

ADDITIONAL READING 25

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. OVERVIEW

The theme of this chapter is ‘‘achieving full potential.’’ We explore how industrial and systems engineering and engineers (ISEs) can achieve full potential and how the ISE function and individual ISEs can assist their organizations in the achievement of full potential. Our fundamental premise is that organizations that desire to achieve full potential can enhance their success by more fully utilizing the potential of their ISEs. This will require holding a different definition for ISE, organizing the function differently, and having a different expectation regarding the ISE value proposition. The practicing ISE will also need to envision the role(s) he or she can play in large-scale transformation. This possibility has implications for the way ISE is organized and positioned in organizations, for higher education, and for the individual ISE.

1.1. Full Potential Introduced

Have you ever experienced being a ‘‘10’’? Perhaps you struck a perfect golf shot, had a great day when everything went perfectly, or flawlessly executed a project. We’re thinking of something that turned out even better than you expected—when you were in ‘‘flow,’’ creating the optimal experience and optimal results (Csikszentmihalyi 1990). Full potential is about realizing personal and organizational possibilities. It’s about getting into flow and staying there. It’s about creating optimal results from the organizational system.

1.2. Structure of the Chapter

In its simplest form, an organization might be modeled as a complex collection of actions that drive particular results. These actions take place in a context or environment that mediates or moderates the results. Figure 1 illustrates such a model.

This model will be our organizing frame for the chapter. We will discuss the role of ISE in corporate transformation, in the achievement of organizational full potential performance. As you see in Figure 1, there are three roles that the ISE can play in this model:

1. Strategy and positioning (e.g., what is the value proposition? Are we doing the right things? What is the strategic planning process, who is involved, what is the system, how do we ensure it works?)
2. Conditions for success (what are the conditions surrounding the actions that drive results? Is the environment right to support success?)
3. ‘‘Drivers,’’ or operations improvement (are we doing the right things right? How are we doing things?)

We will discuss these roles in that order in the chapter. Note that the third role has been the traditional focus of ISE; we will suggest an expansion of the ISE ‘‘domain.’’ Rather than cover what is well covered in the rest of this handbook in the three roles for ISE in the future state organization, we will highlight, in each of the three roles, work that we believe ISEs will migrate to in achieving full potential.

1.3. ISE Domain Defined

As James Thompson points out, a domain can be defined by the ‘‘technologies employed,’’ the ‘‘diseases treated,’’ and/or the ‘‘populations served’’ (Thompson 1967).

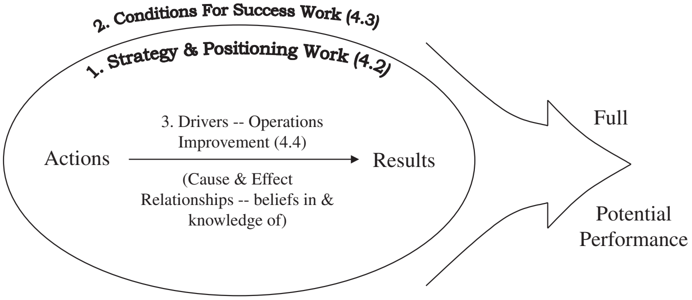

2. Conditions For Success Work (4.3)
1. Strategy & Positioning Work (4.2)
Actions
3. Drivers -- Operations Improvement (4.4)
(Cause & Effect Relationships -- beliefs in & knowledge of)
Results
Full
Potential
Performance

Figure 1 Actions → Results in Context of Conditions for Success.

Consider the remaining sections of this Handbook: Information Technology, Manufacturing and Production Systems, Service Systems, Organization and Work Design, Human Factors and Ergonomics, Project Management, Product Planning, Manpower Resource Planning, Systems and Facilities Design, Planning and Control, Quality, Supply Chain Management and Logistics, Probabilistic Models and Statistics, Economic Evaluation, Computer Simulation, and Optimization. All speak, in one way or another, to technologies employed by ISEs, diseases treated by ISEs, and/or to a lesser extent populations served by ISEs. Again, we will propose expanding the ISE role. Let's begin by examining the traditional definition of ISE and then explore a broader domain definition in the context of full potential performance.

1.4. Operational Definition of ISE

''An Industrial and Systems Engineer is one who is concerned with the design, installation, and improvement of integrated systems of people, material, information, equipment, and energy by drawing upon specialized knowledge and skills in the mathematical, physical, and social sciences, together with the principles and methods of engineering analysis and design to specify, predict, and evaluate the results to be obtained from such systems'' (Womack and Jones 1996). This is the current and fairly traditional definition for ISE.

1.5. Applying the Definition

According to Dr. W. Edward Deming, an operational definition is a definition you can ''do business with.'' Let's see if we can do business with the formal definition.

The key word in the definition is ''system.'' It prompts the question ''What system is it that ISE's work to optimize?'' Our contention is that the ultimate system of interest is the extended enterprise. Ken Wilbur (1996), a challenging author to read and understand, says growth is about ''transcending'' and ''including.'' To contribute to full potential, ISEs must include and also transcend the subsystem they are working on. ISEs must also transcend and include the roles they play and the work they do. ISEs must see how performance improvement in the target subsystem (warehouse layout, work cell configuration, display/human-equipment interface, queue design, simulation, supply chain, etc.) serves the higher good or works to optimize the performance of the larger system. Jim Tompkins (1999) speaks to this and provides an example in his lectures on the migration from warehouse management to supply chain management to supply chain synthesis. Transcending the ISEs traditional system of interest may be the most profound change facing our profession. Inability or unwillingness to address the larger system may hold a clue to the decline our professional society has experienced in the 1980s and 1990s.

Figure 2 is a portrayal of ISE extracted from the perspective of an academic program.

Notice how in this model ISE builds on a core engineering curriculum foundation and then specializes in four basic areas: human factors engineering, manufacturing systems engineering, operations research, and management systems engineering. Each of these four specialty areas dovetails

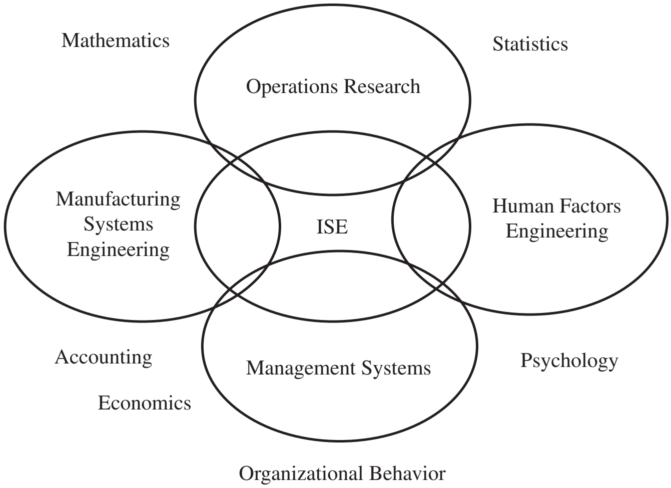


Figure 2 An Academic Portrayal of Part of the ''Domain'' Definition for ISE.

with basic knowledge areas and/or application areas such as statistics, psychology, mathematics, information sciences, accounting, and economics.

While this model is useful for portraying ISE from a curricular perspective, it is much less useful from an application perspective. Once the ISE begins reduction of theory to practice, the academic distinctions rapidly disappear. The ISE typically migrates to a setting that is defined by business processes rather than subdiscipline. The fledgling ISE is thrown into a system of people and capital that survives and thrives by continuing to enhance customer loyalty while at the same time reducing costs and improving efficiency. Fortunately, the ISE value proposition is so robust that enterprising ISEs find that they can contribute at any point and any level in the enterprise system.

ISEs at work are more accurately portrayed by the potential value contribution or offerings they bring to the enterprise as it migrates to full potential and future state. ISEs at work will increasingly find that they must transcend and include their academic training in order to contribute meaningfully to the quest for full potential. Figure 3 is an example of such a portrayal.

The specific contribution that ISEs make, the true value contribution, is the focus of this model. One example might be creating more effective measurement systems that lead to better information about the connection between improvement interventions and customer behaviors. Others might include optimized supply chain systems or increased safety and reduced lost time injuries. What is important in these examples is their impact on business results rather than their particular tools, techniques, or disciplinary focus. It is the cause-and-effect relationship between the value proposition and the business result that is the emerging emphasis. The disciplinary tool, knowledge, or technique becomes valuable when it is applied and we see its instrumentality for achieving positive business results. The ISE value proposition isn't only knowledge; it is the ability to reduce that knowledge to practice in such a way that it produces positive business results. In the coming decades, ISE practice is going to be very, very focused on creating results that move ISEs and their organizations toward full potential.

1.6. Integrating Work in Strategy and Policy, Conditions for Success, and Operations Improvement Leads to Full Potential

Our point of view is that integrating work done to achieve strategy and positioning, the management of conditions for success, and operations improvement create full potential performance (again, see Figure 1). The role of ISE needs to expand to include more involvement with this integration. ISE work in operations improvement, at minimum, needs to be seen in the context of the other two roles. At the extreme, ISE needs to be active in strategy and positioning and condition for success work. And it is value creation for the organization that is the end in the coming decades. A profession that

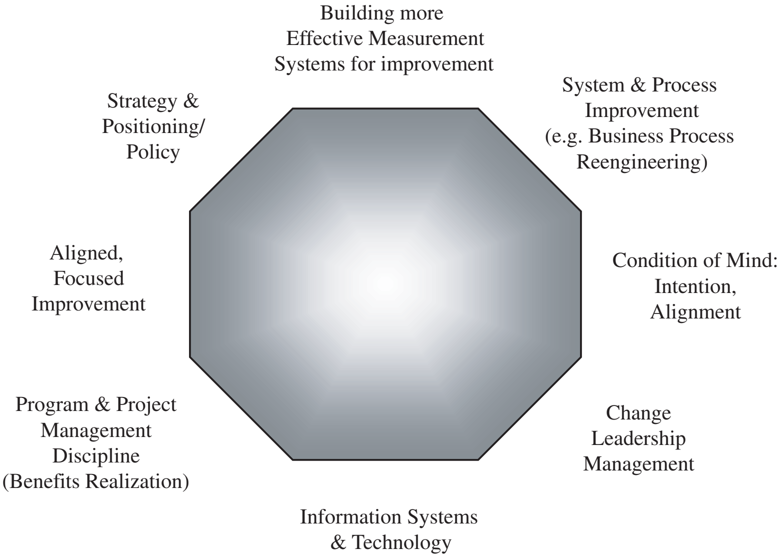

Building more
Effective Measurement
Systems for improvement
System & Process
Improvement
(e.g. Business Process
Reengineering)
Condition of Mind:
Intention,
Alignment
Change
Leadership
Management
Information Systems
& Technology
Program & Project
Management
Discipline
(Benefits Realization)
Aligned,
Focused
Improvement
Strategy &
Positioning/
Policy

Figure 3 ISE Value Proposition: A Real World Perspective.

is so constrained by the technologies it employs and doesn't focus on the end will continue to struggle. We will explain what we mean further into this chapter.

2. THE FULL POTENTIAL MODEL FOR THE ORGANIZATION

2.1. Overview

Full potential organizations are made up of full potential individuals. We will address individual full potential briefly at the end of the chapter. Our discussion of organizational full potential will be more extensive, including specific examples of full potential for organizations, discussed in the context of the enterprise or business excellence model. We will also explore the implications of the expanded role of the ISE on our profession.

2.2. Examples of Full Potential

Two seminal books come to mind when we think about full-potential organizations: *Built to Last* (Collins and Porras 1994) and *The Living Company* (DeGeus 1997). Collins and Porras portrayed eighteen ''visionary'' companies, companies that are more than successful, more than enduring. They are best of best in their industries and have been that way for decades; in fact, they performed well over some 70 years. The visionary companies were matched with 18 comparison firms* for the analysis. The DeGeus study, commissioned by Dutch Royal Shell, looked at 27 firms that were larger and older than Shell. These ''living companies'' had thrived for 100–200 years. We will review these studies to provide you with a glimpse of the ''full potential'' model (see Figure 4) and to highlight the ''causal variables'' identified by the studies. We will then explore the implications of these studies for the ISE profession.

Collins and Porras's visionary companies attained extraordinary long-term performance. Consider three $1 investments on January 1, 1926: one in a general market stock fund, one in a visionary company, and one in a comparison company. By 1996, $1 in the general market investment would have grown to $415; in the comparison firms, to $955; and in the visionary firms, to $6356 (see Figure 4).

What was full potential? What was possible? The point to be made from these numbers is that there are orders-of-magnitude differences in market performance between the visionary companies and the comparison organizations. One might even contend that ''full potential'' was in fact $10,000 or more! Surviving for 70+ years is an accomplishment in and of itself; we believe that *thriving* over that period begins to describe full potential. Collins and Porras concluded that ''visionary companies display a powerful drive for progress that enables them to change and adapt without compromising their cherished core ideals (1994, 9).

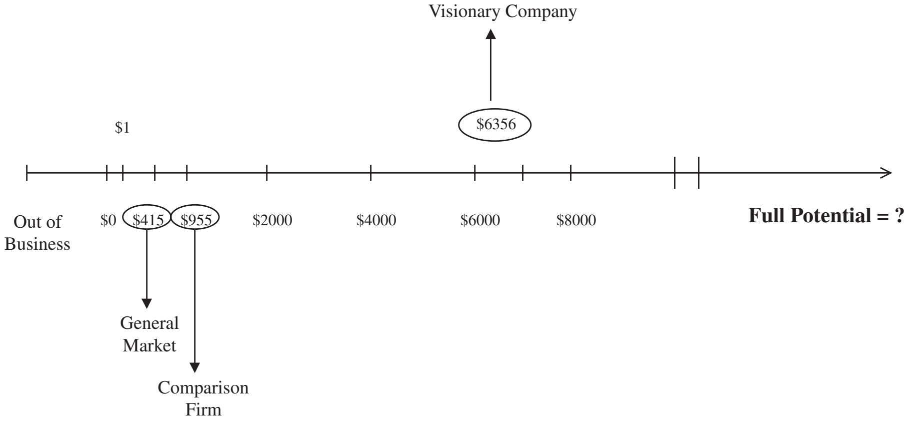


Figure 4 Relative Performance Data from Collins and Porras (1994): What Was Full Potential?

---

*E.g., 3M the visionary company, Norton the comparison firm; Boeing the visionary company, McDonnell Douglas the comparison firm; Citicorp the visionary company, Chase Manhattan the comparison firm.

What are the attributes of full potential organizations? In living companies, DeGeus found four significant factors:

1. *Sensitivity to the environment:* the ability to learn and adapt (we might add, in a timely fashion)
2. *Cohesion and identity:* the ability to build community and a distinct cultural identity (we might add, that supports full potential performance)
3. *Tolerance and decentralization:* the ability to build constructive relationships with other entities, within and without the primary organization
4. *Conservative financing:* the ability to manage growth and evolution effectively

DeGeus states, ‘‘Like all organisms, the living company exists primarily for its own survival and improvement: to fulfill its potential and to become as great as it can be’’ (1997, 11).

Collins and Porras went beyond DeGeus, identifying one distinguishing variable and six ‘‘explanatory’’ variables they attributed to performance differences between visionary companies and comparison firms. The distinguishing variable for the visionary firms was ‘‘Leadership during the formative stages.’’ The six explanatory variables that they identified were:

1. *Evidence of core ideology:* statements of ideology, historical continuity of ideology, balance in ideology (beyond profits), and consistency between ideology and actions (walk the talk)
2. *Evidence of the use of stretch goals, visioning, defining full potential for a given period of time:* ‘‘bold hairy audacious goals’’ (BHAGs); use of BHAGs, audacity of BHAGs, historical pattern of BHAGs.
3. *Evidence of ‘‘cultism’’:* building and sustaining a strong culture, seeing culture as an independent variable, not a context variable; indoctrination process, tightness of fit (alignment and attunement)
4. *Evidence of purposeful evolution:* conscious use of evolutionary progress, operational autonomy, and other mechanisms to stimulate and enable variation and innovation
5. *Evidence of management continuity:* internal vs. external CEOs, no ‘‘post-heroic-leader vacuum,’’ formal management development programs and mechanisms, careful succession planning and CEO selection mechanisms
6. *Evidence of self-improvement:* long-term investments, investments in human capabilities (recruiting, training and development), early adoption of new technologies and methods and processes, mechanisms to stimulate improvement; effective improvement cycles established and a way of doing business.

We will highlight this last finding particularly as we explore the expanding role for the ISE in the future.

2.3. Enterprise Excellence Models

We find enterprise models or business excellence models to be increasingly relevant to the central message of this chapter. They are a way to portray the lessons from the work of DeGeus and Collins and Porras. For example, the Lean Enterprise Institute is working with MIT to develop a Lean Enterprise Model (LEM) (Womack and Jones 1996). The Malcolm Baldrige Award has created a Performance Excellence Framework (National Institute of Standards and Technology 1999). We believe that the Baldrige model provides valuable insight into the variables and relationships, consistent with the lessons from Collins and Porras.

Figure 5 depicts the Baldrige Criteria for Performance Excellence: (1) leadership; (2) strategic planning; (3) customer and market focus; (4) information and analysis; (5) human resource focus; (6) process management; and (7) business results (overarching—customer- and market-focused strategy and action plans). Compare and contrast these variables to the seven identified in Collins and Porras.

Each of these models prescribes strategies for achieving full potential. Our contention is that this striving for full potential is the context within which ISE will be practiced. ISE will be challenged to present a value proposition in the context of large-scale organizational transformation. Enterprise models of excellence provide insights into how to position our profession and the work of ISE in organizations. With this base, each and every ISE has the potential to be a visionary representative of our profession. This is an excellent way to think about both personal and collective value propositions.

2.4. Implications for ISE

The context shift we describe above for ISE has several specific implications. First, each and every initiative undertaken by an ISE must be causally linked to business results. Let’s use a large retail

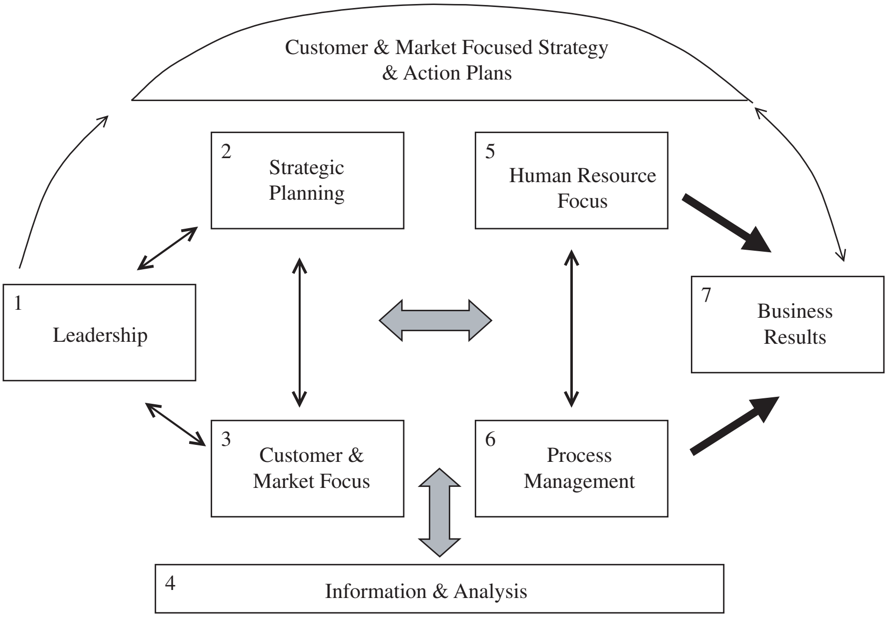

Customer & Market Focused Strategy
& Action Plans
2
Strategic Planning
5
Human Resource Focus
1
Leadership
7
Business Results
3
Customer & Market Focus
6
Process Management
4
Information & Analysis

Figure 5 Baldrige Criteria for Performance Excellence.

organization as an example. One of the major business outcomes for this type of business is to fill what we call the "treasure chest." As depicted in Figure 6, the treasure chest's three dimensions are (1) market share, (2) percent spend, and (3) length of relationship or customer loyalty.

Retail businesses want to optimize market share, get their customers to spend more of their disposable income (percent spend) in their stores, and keep their customers (optimize the value stream from customers). How can ISEs help fill the treasure chest?

Consider a profit ratio. Typically, ISEs have worked to reduce the inputs (the denominator) of the profit ratio by driving costs down and increasing efficiency. Focusing ISEs on outcomes (the nu-

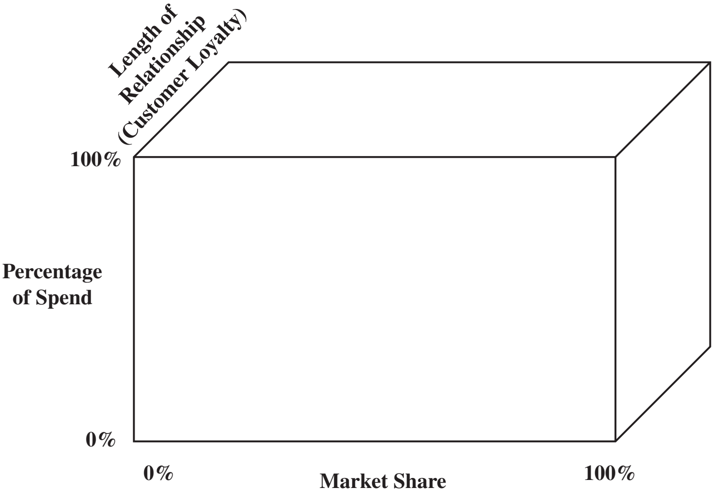

Length of Relationship (Customer Loyalty)
100%
Percentage of Spend
0%
0%
Market Share
100%

Figure 6 Filling the Treasure Chest as a Major Business Goal: What Does the Treasure Chest Model Look Like in Your Organization?

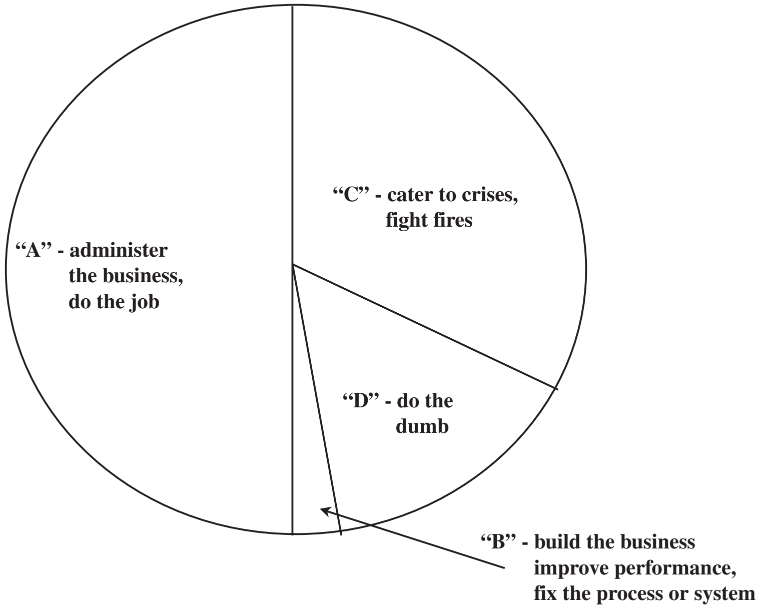


Figure 7 ABCD Model: How We Spend Our Time.

merator) shifts the paradigm about our value contribution. Our contention is that ISEs will be challenged to work on both the numerator and the denominator of the profit equation.

ISEs and the ISE function will be required to explain our contribution in terms of the profit ratio. It won't be enough to say that we improved the efficiency of a process or work cell. We will have to demonstrate how our actions lead to filling the treasure chest.

The seventh variable that Collins and Porras (1994) identified was "evidence of self-improvement." Time is the most critical resource in the knowledge-based organization. People at all levels spend their time doing four things: they **A**dminister the business, that is, do jobs ("A" work); they **B**uild the business, that is, improve performance and fix systems or processes ("B" work); they **C**ater to crises, that is, fight fires, fix problems ("C" work); and they **D**o the dumb, that is, non-value-adding things ("D" work). Figure 7 depicts how we spend our time.

Organizations that intend to achieve full potential will spend more time on B work. They will establish improvement cycles such as the Deming and Shewhart Plan-Do-Study-Act model. Rather than addressing "targets of opportunity," ISEs will deploy improvement cycles that are thought through strategically, comprehensive in scope, and well integrated. Enterprise excellence models clearly indicate this. Systems thinking will be applied at the enterprise level. This has been the clear migratory path for the past 30 years, and it will continue to be. Our profession's value proposition will focus on the special knowledge and skills the ISE brings to the quest for full potential.

In the more traditional model, ISE work is often detached from the work of transformation. ISE improvement efforts tend to be done outside the context of the enterprise improvement cycle, so they lack a clear causal connection to organizational business results (e.g., filling the treasure chest).

So the big implication for ISEs in the next several decades is that they must think enterprise, think total systems, and be connected to the enterprise improvement cycle. ISEs cannot afford to (sub)optimize targeted subsystems at the expense of the larger system, and they cannot afford to be isolated from the large-scale transformation work that characterizes the full potential organization. We might go a bit further and suggest that increasingly ISEs are going to be challenged to prescribe migration paths that move an organizational system toward full potential at a faster rate.

3. THE FUTURE STATE VALUE PROPOSITION FOR ISE

3.1. Overview

On the threshold of the 21st century, Michael Porter is perhaps the best-known researcher, author, teacher, and consultant on the subject of competitive strategy. Porter (1996) makes a distinction between operating effectiveness and efficiency and positioning and strategy; balance between the two contributes to organizational full potential. Whereas operating effectiveness and efficiency is the traditional core of ISE academic programs, positioning and strategy is rarely, if ever, addressed in ISE academic programs. We believe that our model of full potential depicts that balance. In this section of the chapter, we will approach positioning and strategy from an ISE perspective. Then we'll talk about another subject missing from most academic programs—conditions for success. Finally, we'll make some observations about a more familiar topic, the ISE's role in operations improvement. (For more on this subject, see Sink and Poirier 1999.)

3.2. The Planning System: Position, Strategy, Implementation, and Deployment

The first role in this new model for ISE contribution focuses on the planning system. The planning system includes processes by which high-level strategy and policy are determined and all processes that lead to effective implementation and benefits realization from strategy and policy deployment. In many organizations, this system is not defined, not documented, not systematic, and hence the results it creates are highly variable. We contend that the planning system needs to be reengineered, and we offer this work as a role that ISEs can and should impact. Figures 8, 9(*a*) and 9(*b*) are simple versions of what we mean by this contention.

Figure 8 is a depiction of what we call a "grand strategy system." It is a high-level picture of strategy and policy deployment. Once strategy and policy (positioning) is decided, a transformation plan needs to be developed. The vision and definition of the future state is on the right-hand side of the picture, the current reality is on the left, and the work in front of the organization is in the middle. Conceptually, it will take many improvement cycles (Plan, Do, Study, Adjust/Act) to pull off a large-scale transformation. Figure 9(*a*) depicts the improvement cycle process itself in general terms. It is not important to understand all the details of these figures; it is important that you see that ISE skills of system and process design are being applied at the strategy and policy deployment level. The premise is that more defined and explicit systems and processes for strategy and policy deployment will lead to more consistent results (benefits realization).

Positioning deals with what an organization offers and how that offering is portrayed to the customer. It has to do with whether your value proposition is clear to the customer and whether those propositions are distinctive and attractive. It has to do with whether your customers see your offering

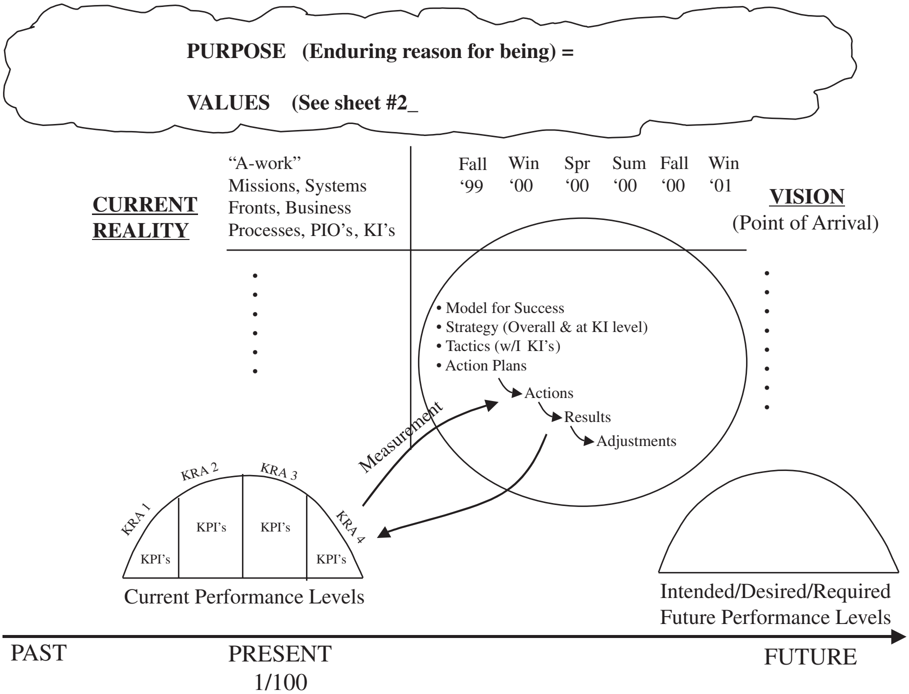
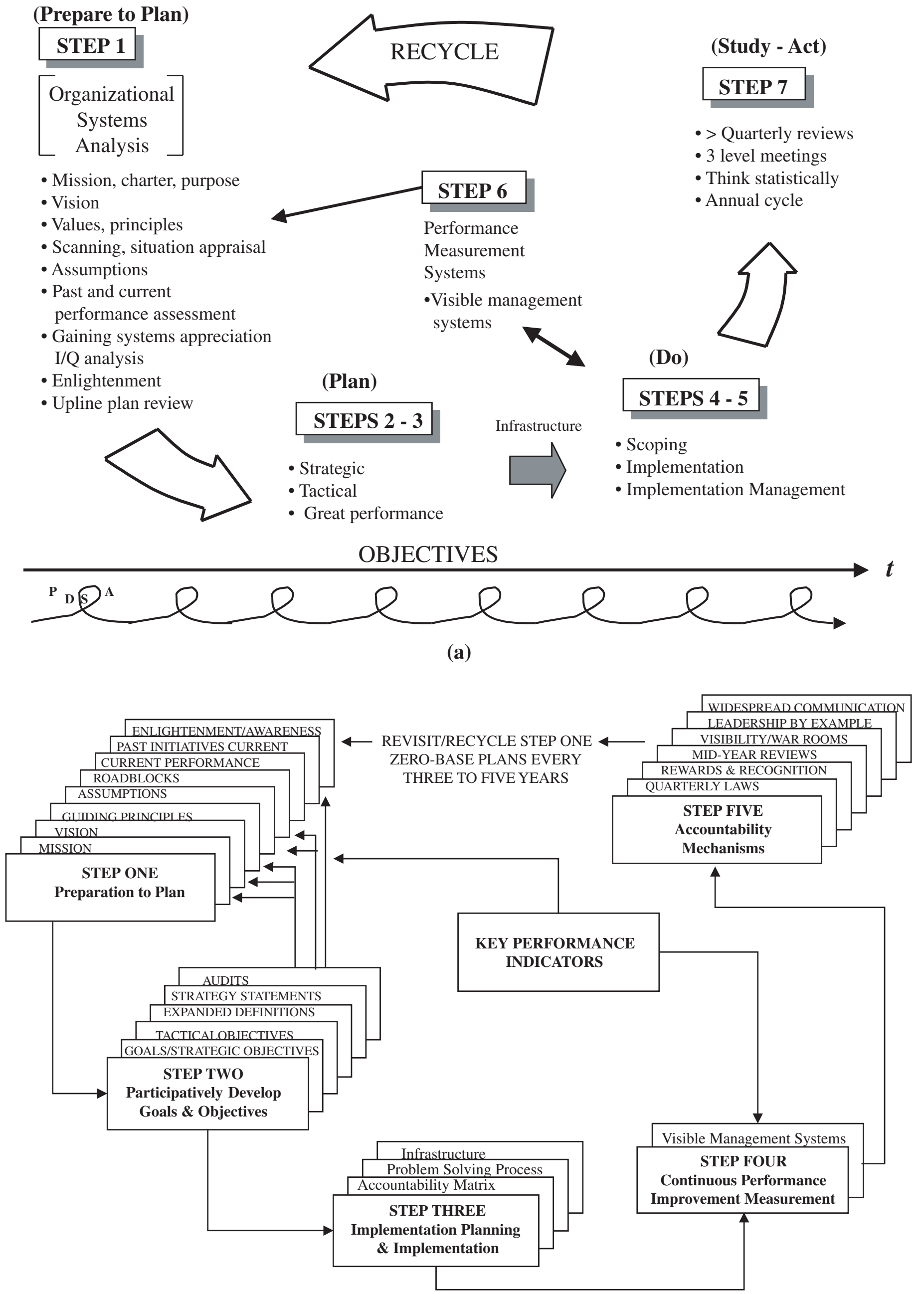

PURPOSE (Enduring reason for being) =
VALUES (See sheet #2_
"A-work"
Missions, Systems
Fronts, Business
Processes, PIO's, KI's
CURRENT
REALITY
Fall '99
Win '00
Spr '00
Sum '00
Fall '00
Win '01
VISION
(Point of Arrival)
• Model for Success
• Strategy (Overall & at KI level)
• Tactics (w/I KI's)
• Action Plans
Actions
Results
Adjustments
Measurement
KRA 1
KRA 2
KRA 3
KRA 4
KPI's
KPI's
KPI's
KPI's
Current Performance Levels
Intended/Desired/Required
Future Performance Levels
PAST
PRESENT
1/100
FUTURE

Figure 8 Grand Strategy System.


(Prepare to Plan)
STEP 1
Organizational Systems Analysis
• Mission, charter, purpose
• Vision
• Values, principles
• Scanning, situation appraisal
• Assumptions
• Past and current performance assessment
• Gaining systems appreciation I/Q analysis
• Enlightenment
• Upline plan review
RECYCLE
STEP 6
Performance Measurement Systems
•Visible management systems
(Study - Act)
STEP 7
• > Quarterly reviews
• 3 level meetings
• Think statistically
• Annual cycle
(Plan)
STEPS 2 - 3
• Strategic
• Tactical
• Great performance
Infrastructure
(Do)
STEPS 4 - 5
• Scoping
• Implementation
• Implementation Management
OBJECTIVES
t
P
D
S
A
ENLIGHTENMENT/AWARENESS
PAST INITIATIVES CURRENT
CURRENT PERFORMANCE
ROADBLOCKS
ASSUMPTIONS
GUIDING PRINCIPLES
VISION
MISSION
STEP ONE
Preparation to Plan
REVISIT/RECYCLE STEP ONE
ZERO-BASE PLANS EVERY
THREE TO FIVE YEARS
WIDESPREAD COMMUNICATION
LEADERSHIP BY EXAMPLE
VISIBILITY/WAR ROOMS
MID-YEAR REVIEWS
REWARDS & RECOGNITION
QUARTERLY LAWS
STEP FIVE
Accountability Mechanisms
KEY PERFORMANCE INDICATORS
AUDITS
STRATEGY STATEMENTS
EXPANDED DEFINITIONS
TACTICAL OBJECTIVES
GOALS/STRATEGIC OBJECTIVES
STEP TWO
Participatively Develop Goals & Objectives
Visible Management Systems
STEP FOUR
Continuous Performance Improvement Measurement
Infrastructure
Problem Solving Process
Accountability Matrix
STEP THREE
Implementation Planning & Implementation

(a)

(b)

Figure 9 Improvement Cycle Process.

as instrumental to their success. Positioning decisions occur in step 1 of the Improvement Cycle (see Figure 9(*b*)). Positioning decisions end up being articulated and portrayed on the far right side of the transformation plan (see Figure 8). *Strategy*, on the other hand, has to do with how the organization plans to deploy its offering and execute on its position. Strategy ensures that the customer takes up the offering. Success must be operationally defined; you must have a clear vision of what success looks and feels like. *Successfully executing strategy and policy* throughout the organization involves what is called policy deployment. Strategy is reflected in the middle part of the grand strategy plan (Figure 8). Strategy is fleshed out in the improvement cycle (Figures 9(*a*) and 9(*b*) in step 2.

Improvement cycles in context of positioning and strategy are similar to the relationship between a game plan, perfect practice (as Vince Lombardi used to say), and actually playing the game. Strategy and positioning are the game plan; improvement cycles are perfect practice and also playing the game. The key role we see for ISEs in the planning system is in effective implementation and deployment of strategy and policy. This may well be several levels higher in thinking and involvement than many ISEs have traditionally been.

3.2.1. Policy Deployment

Policy deployment involves the successful implementation of the strategy for success (Akao 1991). The Japanese describe it as a ‘‘waterfall-down’’ process consisting of communication and ensuing coordination of activities. Effective policy deployment processes and practices ensure alignment and attunement. It is the policy deployment process that causes improvement efforts inside the organization to be aligned with the vision for the organization. Improvement cycles become aligned as a result of the deployment processes and practices. An example of such a practice is what are called three-level meetings. Periodically, leaders and managers and employees from three levels of the business gather for an ‘‘all-hands’’ meeting to share progress and performance on key initiatives or performance improvement objectives (PIOs). During these reviews, the dialogue ensures that critical adjustments are made to refine alignment in efforts (pace, focus, resource allocation, etc.). We have found that ISEs can be effective in supporting policy deployment policy and practices; their systems thinking background makes them well suited for this activity.

3.2.2. Relationship Management

In the full potential model of the organization (see Figure 10) are at least four categories of relationships that need to be managed and optimized: (1) customers, (2) employees, (3) stakeholder/stockholders, and (4) business partners (internal and external).

At a micro level, the ISE will be expected to contribute to the conception, design, development, and implementation of improvement initiatives. The extent to which relationships are managed ef-

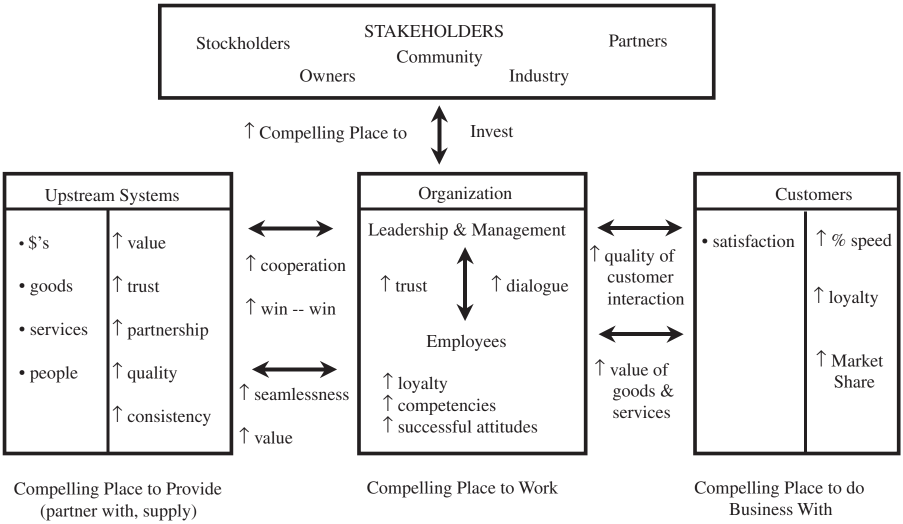

STAKEHOLDERS
Stockholders
Community
Partners
Owners
Industry
↑ Compelling Place to
Invest
Upstream Systems
• $'s
• goods
• services
• people
↑ value
↑ trust
↑ partnership
↑ quality
↑ consistency
↑ cooperation
↑ win -- win
↑ seamlessness
↑ value
Organization
Leadership & Management
↑ trust
↑ dialogue
Employees
↑ loyalty
↑ competencies
↑ successful attitudes
↑ quality of customer interaction
↑ value of goods & services
Customers
• satisfaction
↑ % speed
↑ loyalty
↑ Market Share
Compelling Place to Provide (partner with, supply)
Compelling Place to Work
Compelling Place to do Business With

Figure 10 Full-Potential Performance Requires the Optimization of Key Relationships. (↔ represents relationships (or value exchanges) we will optimize in this transformation.)

fectively will ultimately determine the effectiveness of the implementation. At the macro level and with a focus on business results, it will be the relationship with the customer that will determine how well the treasure chest is filled. Acquiring information about the customer, monitoring how that information is used, and helping the customer be successful are all key aspects of customer relationship management (CRM). The role ISE should play is, at a minimum, to be keenly aware of the importance of CRM and to align ISE project work with CRM efforts.

In that larger context and domain, value creation will derive from the ISEs ability to manage our relationship with various partner and customer segments and create results for and with them. In our view there is only one customer, the consumer. The end users or customers are the population we serve. There are no internal customers in our model. Everyone inside the organization should function as a partner on a team working to serve the customer. The ISE is in partnership with others in the enterprise, and all are working to achieve organizational full potential performance. The ISE brings technologies to the enterprise that will facilitate achieving full potential performance.

At the maximum, ISE professionals would be part of the CRM program team. This would involve designing the strategy, working on information systems, and thinking through how to deploy CRM data and information to achieve alignment and optimize the lifetime value of the customer.

Perhaps the most dramatic change for ISEs in the decades to come will be the requirement that they understand the needs of the customers and have a role in managing customer relationships. ISEs understand the process of converting data to information, using information to support decisions and actions. This understanding is sorely needed in organizations today and is a key to unlocking full potential. How will ISE solutions enhance the organization's relationships with its customers? This will be a key area of emphasis for the ISE in the future.

The relationship management aspect of strategy and policy deployment goes well beyond just improved customer relationship management systems and processes. As Figure 10 shows, full potential performance requires that all relationships be managed differently. It has not been uncommon for supplier and vendor relationships to be in transition in the past; what has the role of ISE been in that transition? We contend that relationships with employees must also be managed differently; what is the role of ISE in that? Sears has a well-publicized transformation effort that focuses on being a compelling place to invest, a compelling place to shop, and a compelling place to work. This concept will show up in Figure 10. We would argue that ISEs can and should play a role in the relationship management associated with making the organization compelling from all constituent vantage points. This will lead to full potential performance.

Managing these relationships differently will take different levels of personal mastery. Listening skills will be even more important. How we hold ourselves in relationship to others in the system is foundational; do we see ourselves as partners or as competitors/adversaries to be negotiated with? First we need a different model for full potential. Then we need to develop the skills to work the model effectively.

We have discussed grand strategy, improvement cycle, policy deployment, and relationship management as key elements of planning system (strategy and policy development and execution) transformation. Other changes are required in the planning system in order for an organization to migrate to full potential performance, which we will either mention here and not elaborate on or mention in upcoming sections. The transformation to full potential is a stream of improvement cycles. There are improvement cycles embedded in improvement cycles. There is one for the overall enterprise and then embedded improvement cycles (aligned) for subsystems. Coordinating all these improvement cycles is critical to overall success—yet another natural role for ISEs to play in the future.

Improvement cycles, at all levels, are conceptually a process of Planning, Doing, Studying (progress and performance), and then Adjusting plans based on results. We've addressed the Plan and Do steps. In the third role, operations effectiveness and efficiency, we'll discuss the role of measurement in the Study process and outline how to build more effective measurement systems—yet another key role for ISEs in the work to achieve full potential performance.

All the Planning and Doing and Studying and Adjusting is being done in a context or organizational environment. Figure 1 depicted this. Culture shift is central to creating conditions that will support migration to full potential performance. We would once again contend that ISEs can and should play a role in designing and executing the culture shift to create conditions that will fully support transformation to full potential. This might be a support role to the HR function, team member/collaborator, and so on. We will flesh this out in the next section.

3.2.3. Change Leadership: The IE as Change Master

Pulling off a major shift in the improvement cycle process in organizations will require many, if not most, to get outside their comfort zones. People are going to be asked to do different things differently. One key function of ISE in the context of improvement cycles will be bringing solutions to the business that help fill the treasure chest. Ensuring that the benefits of improvement are actually realized is a three-ball juggling challenge. The first ball is solution design and development. Solutions must be thought through in the context of strategy and policy of the larger system. The second ball is project leadership and management: delivering the solution. The third ball, and often the most critical, is change leadership and management (Sink 1998). It is the ISEs smooth juggling of these three balls that creates effective change. Clearly there is a science and an art to this, and ISEs can become its foremost practitioners. In large-scale transformation, the three-ball challenge becomes the three$^{\text{nth}}$-ball challenge. Full potential will require that many (n) solutions be delivered concurrently. The challenge is to manage multiple projects rather than a single project. We find that as ISEs gain experience they become excellent candidates for this challenge due to their training in dealing with system interactions and complexity.

Here is a model that is especially useful for understanding change leadership and management:

$$\text{Readiness for change } (R) = V \times BP \times A$$

Change will occur if $R > C$, where:

V = vision, i.e., people's connection to success, their clarity on what success looks like and feels like, an operationally defined desired point of arrival

BP = burning platform, i.e., sense of urgency, level of dissatisfaction with the status quo

A = approach, i.e., clear and pragmatic first steps, knowing what to do next to move in direction of vision

C = Perceived cost and/or risk associated with change, i.e., level of discomfort associated with moving outside the comfort zone

We believe that in the new environment the power of creation will supplant problem solving. It is easier to get people motivated to deal with an uncertain future if they are clear on the vision and they choose for it. (However, in fairness to problem solving, you don't need a vision to fix a broken window.) So we suggest strongly that ISEs work on developing creation skills to complement their highly developed problem-solving skills (Fritz 1991). Once people are in the creation mode, there will be little resistance to change and change leadership will be quite natural and almost easy.

3.3. Conditions for Success

We will now turn from the planning system to look at what we call *conditions for success.* Recall the action-results and conditions for success model (see Figure 1). Much of the work of ISE tends to be in the domain of action to results; working on what we call drivers of change that directly result in improved business results. Often the effectiveness of efforts to influence drivers is diminished because the conditions for enabling the efforts to be effective are not present. So we identify "enablers" that are collectively constitute conditions for success. We will introduce four enablers that are central to conditions for success:

1. *Culture system:* shaping values/attitudes/behaviors that will support full potential performance
2. *Infrastructure:* how we are organized to do B and implications that B work will have on the A work infrastructure
3. *Communication system:* sharing information necessary to ensure that people are informed as we migrate to full potential performance
4. *Learning system:* ensuring that knowledge and skills keep pace to support the transformation

3.3.1. Culture System

Culture is "a pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems" (Schein 1992).

The fundamental question for organizations aspiring to reach full potential (or should we say leaders who wish to lead groups to full potential?) is, "What is the full potential culture?" Figure 11 is a pictorial we use to describe our answer. The descriptors on the right side reflect values, attitudes, and behaviors that we believe are consistent with achieving full potential. Listed on the left are commonly experienced values, attitudes, and behaviors typical of underperforming organizations. We believe that most people inherently have core values that match those on the right: serving, learning, integrity, and excellence. Unfortunately, we have found that when individuals come together in organizations, things conspire to cause their attitudes and behaviors to migrate to the left. By raising consciousness about personal choices and what motivates those choices, it is possible to induce a shift to the right. We might more accurately say it is possible to create conditions that will naturally bring people back to their core values and to the attitudes and behaviors that are so critical to full

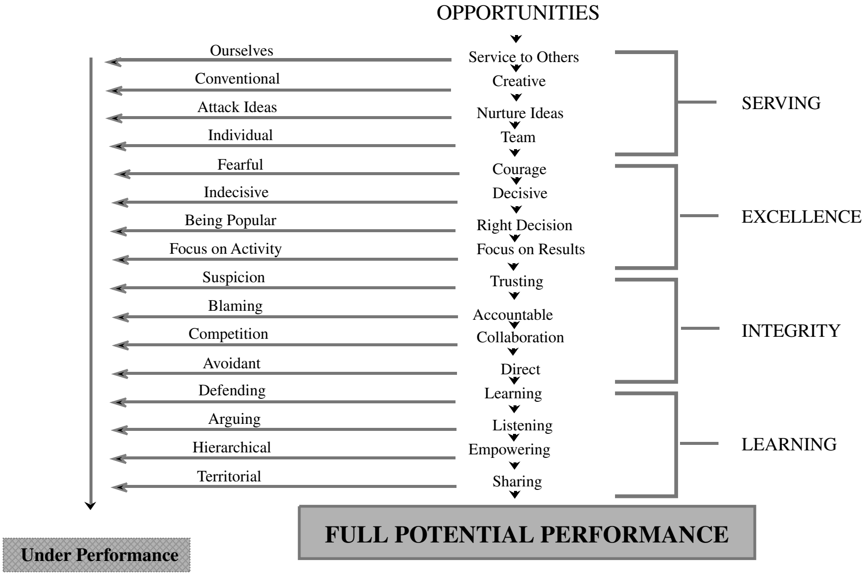

OPPORTUNITIES
Ourselves
Conventional
Attack Ideas
Individual
Fearful
Indecisive
Being Popular
Focus on Activity
Suspicion
Blaming
Competition
Avoidant
Defending
Arguing
Hierarchical
Territorial
Service to Others
Creative
Nurture Ideas
Team
Courage
Decisive
Right Decision
Focus on Results
Trusting
Accountable
Collaboration
Direct
Learning
Listening
Empowering
Sharing
SERVING
EXCELLENCE
INTEGRITY
LEARNING
FULL POTENTIAL PERFORMANCE
Under Performance

Figure 11 Leadership Values Model.

potential performance. Whether the ISE is an active or a passive player in this process, there still must be an awareness of the role culture plays in performance improvement. Having said this, we are convinced that there can be no passive players. Everyone's behavior must reflect full-potential values. (i.e., ''walk the talk.'') The ISE function is too important to be ignored. The ISEs must model these core values in order to be effective as individuals and as a collective force for improvement.

Note that we entitled the section ''Culture System.'' The implication is that culture is the product of a system. If we want/need the culture to be different in order to achieve full potential, then we need to change the system that shapes culture. This is an active rather than passive approach to culture. We are consciously shifting culture toward what is natural (in our view) and toward values-attitudes-behaviors that support achievement of full potential. Because ISEs are about integrating systems to optimize performance, recognizing that culture is a subsystem that needs to be led and managed is integral to our work—yet another area of potential work for ISEs in the future.

3.3.2. Infrastructure for Improvement

Earlier we alluded to the ABCD model (Figure 7) depicting how we spend our time. Spear and Bowen (1999) highlight the importance of integrating administering the business (A work) with building the business. What are the roles and rules in the context of A and ''B?'' Toyota manages to demand rigid performance specifications while at the same time ensuring flexibility and creativity. This seeming paradox is the ''DNA'' or essence of their success. Disciplined execution of A makes it possible to also do B. The infrastructure for this is supported by culture, discipline, learning, role clarity, and rules.

The potential ISE role in this regard is tremendous. Many organizations lack discipline and specification relative to A work, leading to a great deal of variation and unpredictability. This variability in performance is a source of D and ineffective management of C. Established methods and procedures are central to gaining control of A-work and freeing up more time for B. It is important for ISEs to reestablish their role in the continued ''rationalization'' of A. As Spear and Bowen (1999) mention, rationalizing A involves using experimentation and the scientific method to improve things. A word of caution: things will get worse before they get better. If the experiments are well designed and carefully thought through, the loss in performance will be manageable and the final results will be well worth the cost. Ideally, one could create a system that would perform and improve simultaneously. How to organize this and make it happen is clearly an important domain of activity for ISEs.

How is your organization organized to do B, and what role should ISE play in the infrastructure piece? We have found that the infrastructure for B will determine how effective and efficient the B work is. ISEs can play an important role in this design variable.

3.3.3. Communication System

Noted organizational authority Dr. Edward Lawler (1986) recommends that organizations first share information, then knowledge, then power, and only then share rewards. Communication is sharing information. The theory is that better-informed people perform better. Our experience and a host of research results confirm that hypothesis. Sharing information leads to better coordination, and better coordination creates the conditions for realizing full potential. The communication system needs to be thought through and well designed for effective and efficient implementation and deployment. We have found that ISEs can help improve the information sharing system. We see this as a key area of ISE contribution in the future.

Remember the sequence Lawler recommends: Reengineer the communication system before addressing the learning system, then create the involvement system, and finally, put into place the reward and recognition system. Taken in order, these four steps produce an integrated system that can benefit from key capabilities of the ISE. Taken randomly, as is too often the case, valuable resources are wasted and the value proposition to the enterprise gets diluted and disappears.

3.3.4. Learning System

Full-potential organizations are also learning organizations. Learning organizations must be made up of learning individuals. Each of us is a product of our experience. For the better part of 16–20+ years we participate in a particular educational process. The curriculum and the syllabi are provided for us. The goals, the things to be learned, the sequence, the practice, and so on are all prescribed. And then, suddenly, teachers are no longer part of the process, and we must make choices: either drift away from learning and growth or take charge of our own learning. Unfortunately, most people drift, and as a result their value proposition—their ability to continue to add value—diminishes over time. In the 1990s, as technology, competition, and complexity increased, organizational leaders began to stress the importance of learning and growth as part of the personnel development process. However, stressing importance typically has little or no effect. Lifelong learning is something that must be ‘‘chosen for’’ if we are to continue to contribute.

Clearly, the ISEs can play at least two roles in the learning process. First, ISEs of the future will need to be more systematic and more disciplined about continuing education and development. Our professional society—in this case, the Institute of Industrial Engineers—will play an increasingly important role through its conferences and seminars. Additionally, many companies and agencies provide formal programs to support the learning process. Second, the ISE can and should play a role in assisting the organization with the design and execution of the learning (education, training, and development) process. Learning is a key business process. Improving the competencies and capabilities of the human resource requires a well-designed and executed process. Business process reengineering is often required if an organization is to move beyond the minimally effective training systems of the past to the learning systems that support the creation of positive business results and keep employees vital.

We are proposing that learning is the result of a process. The education, training, development process in most organizations is not effective. A work knowledge and skills as well as B work knowledge and skills are needed. This will require being a team member with the HR function. Our experience is that the HR personnel have welcomed the contribution that ISEs can make to rethinking how to spark and sustain learning. The migration to full potential performance will require a tremendous amount of new learning. This will require, first, recognition that this is important, and second, willingness and desire to learn. We believe that creating a learning orientation is part of the culture shift work. People have a natural tendency and desire to learn and grow. When you change the condition within which people are working, this natural tendency will reappear. Then it is just a matter of directing the learning toward the knowledge and skills that support performance improvement. Much of this will be quality and productivity and improvement related, and this is where the ISE comes in. Changing the culture and then changing the system for learning are central to creating conditions that support full potential performance. ISEs can and should play a role in this shift.

3.3.5. Summary

The involvement of ISE in strategy and positioning and in conditions for success occurs too infrequently. Full potential, as we have posited, requires an integration of work in the three areas. We asserted earlier that for ISE to be positioned differently, to adopt a broader domain, a senior champion or sponsor is often required. This will be someone who is an ISE or understands the value proposition of ISE and is constantly looking for effective pieces of work to involve them in, ways to use ISE to migrate to full potential. Each month, *Industrial Management* features ISEs who are in these roles.*

---

* An outstanding example of an executive champion of ISE is our coauthor and colleague, David Poirier. His story appears in the May/June 1998 issue of *Industrial Management*.

The true impact of leadership in achieving and sustaining full potential performance is most often seen in the work they do in establishing conditions for success. All ISEs would be well served to consider the role they play in establishing and sustaining the proper conditions or environment to support full potential performance.

We believe that ISE modes of thinking will be valuable in the challenging strategy and positioning work of the organizations of the future and in creating optimal conditions for success. A perspective built on systems thinking is ideally suited for the type of planning and analysis required as organizations work to enhance positioning and strategy and creating positive internal conditions in the ever-changing markets.

The challenge to practicing ISEs is to understand the connection between their work and organizational attainment of its full potential. Seminal questions include "How does my work support the overall positioning and strategies of the organization?" "What are the cause-and-effect linkages between my work and 'filling the treasure chest'?" "How is my work connected to other efforts in the organization, and how can I strengthen the connection and create synergies where they don't currently exist?" "What is full potential performance and how is my work moving the organization toward that goal?" "How can the ISE role work on key condition for success issues in a way that enables or enhances the success we achieve over time?"

We have offered some examples of what we call "strategy and positioning" and "conditions for success" and described the roles that ISE can play in these categories of endeavor. Too often the work of ISE gets over- or underemphasized due to a lack of understanding of where they fit in the organization. ISEs need to be aware of the big picture and the leverage points to ensure that their work is contributing in an optimal fashion to the greater good.

To ensure that we are not misunderstood, we would offer the following. We are not calling for ISEs to be all things to all people. We understand the core value proposition of ISE, what it has been, and we have our views on what it will be. We are simply suggesting that if ISEs are to live up to their lofty definition as systems integrators, the future will require us to continue to transcend and include more traditional roles and migrate to enterprise level contributions.

3.4. Operations Improvement Role

3.4.1. Overview

We will now turn to the traditional role of ISE, that of problem solving and operations improvement. We place improvements in efficiency, quality, and technology (e.g., methods, hardware, software, procedures, processes) in this category. Regardless of whether the ISE specializes in operations research, human factors, manufacturing systems, or management systems methodology, the challenge will be to change how things are done such that total system performance is improved.

Traditionally, ISEs have operated at the work center level. Over the past 30+ years the scope of the system of interest has broadened. The word "system" in industrial and systems engineering has taken on increased importance. This migration has occurred naturally; it is what was required for the ISE to be more successful. Unfortunately, industrial engineering tended to lag behind other disciplines in the evolutionary process. Our profession has missed opportunities to continue to add value over the past 30 years. The more striking examples include abandoning the quality movement, and missing both the tactic of business process reengineering and the emergence of the balanced scorecard in the measurement domain. We believe that the decades to come will provide opportunities for ISE to reintegrate and reposition itself as a leader and doer in the field of performance improvement. At the risk of being repetitive, this requires rethinking the role and relationship among positioning and strategy, conditions for success, and operations improvement.

We will not duplicate material in the remaining chapters of this Handbook. However, we do want to fill in some blanks by highlighting a couple of areas of operations improvement that we feel might not be represented: systems and process improvement (more specifically, business process reengineering) and measurement systems. Again, our contention is that the bulk of this Handbook focuses on the traditional ISE role in achieving operations effectiveness.

3.4.2. Business Process Improvement

The term *unit of analysis* applies to the scope of the system of interest. When IE first began, the unit of analysis—the scope of the system of interest—was confined to the worker, the work cell, and individual work methods. Over time, the scope of the system of interest to the ISE has increased—for example, from economic lot size to inventory control to warehouse management system to supply chain optimization to enterprise or supply chain synthesis. It is important to keep in mind that this expansion in focus is an "and," not an "or." Attending to larger units of analysis is a "transcend and include" strategy. Inventory control remains in the ISE toolkit, but now it is being applied in the context of larger systems. This causes the ISE to rethink what it means to optimize the system of interest. If a system's performance objective is simply to fill customer orders, one might employ a ‘‘just-in-case’’ inventory policy that relies on very high inventory levels to ensure satisfied customers. Vendor relationships of one type might drive inventory even higher. If the unit of analysis is shifted to the next-more inclusive system, so that it includes different relationships with vendors, high weighting on holding costs, preference for short lead times, high values placed on shortage costs, and so on, then the outcome could very well be an entirely different strategy for inventory management. It is interesting to note that the ‘‘dot com’’ organizations have to reinvent supply chain models in order to thrive in the high-paced and fast-moving environment. The real challenge will be for more traditional organizations with legacy systems to migrate to a dot com pace.

This migration to a larger system of interest also happened to process improvement. Total Quality Management (TQM) revitalized the interest of organizations in quality issues.* The focus on improving embedded processes has shifted to an enterprise-level focus on improving business processes. Once again we see the transcend-and-include principle being required. A business process begins and ends with the customer. Organizations can no longer afford to optimize a subsystem (embedded process) at the expense of the performance of the larger system (business process). So the techniques and methods that were employed to improve smaller processes are being adapted and enhanced with some expertise borrowed from organizational leadership and management of change, creating a whole new line of business for a large group of professionals.

At the time we go to press, our experience is that few ISE undergraduates are being introduced to business process engineering. Our plea is that ISEs embrace this technology. For example, business process reengineering, we believe, is a context within which most ISE improvement efforts will be done in the near future.

Figure 12 is an example of a business process reengineering roadmap to provide insight as to how the ISE might contribute. First, systems thinking is required to do BPR. ISEs are trained to

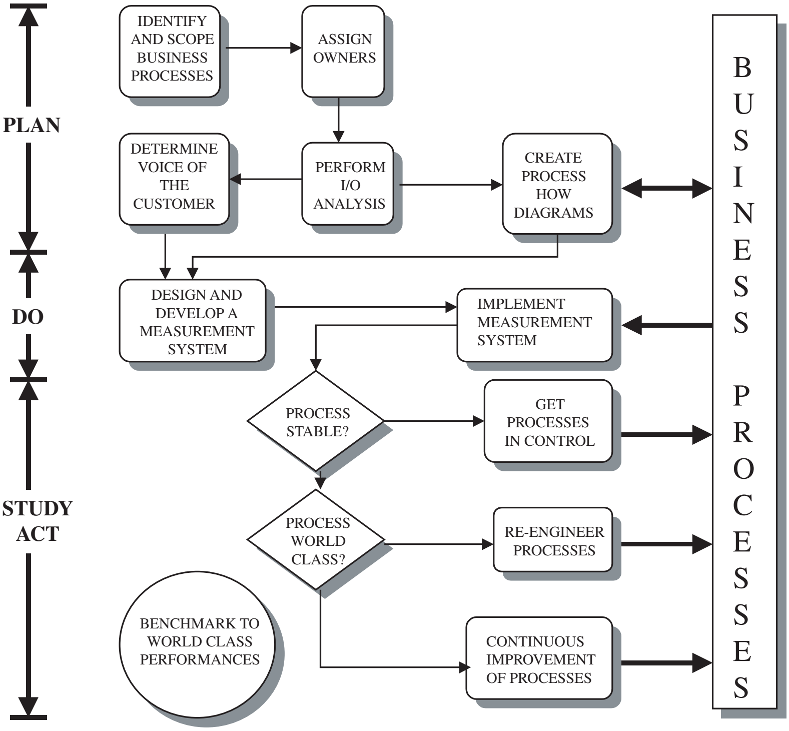

PLAN
DO
STUDY
ACT
IDENTIFY AND SCOPE BUSINESS PROCESSES
ASSIGN OWNERS
DETERMINE VOICE OF THE CUSTOMER
PERFORM I/O ANALYSIS
CREATE PROCESS HOW DIAGRAMS
DESIGN AND DEVELOP A MEASUREMENT SYSTEM
IMPLEMENT MEASUREMENT SYSTEM
PROCESS STABLE?
GET PROCESSES IN CONTROL
PROCESS WORLD CLASS?
RE-ENGINEER PROCESSES
BENCHMARK TO WORLD CLASS PERFORMANCES
CONTINUOUS IMPROVEMENT OF PROCESSES
BUSINESS PROCESSES

Figure 12 Business Process Reengineering Roadmap: How the ISE Might Contribute.

---

* Interestingly, quality assurance, control, and sampling, etc. were taught almost exclusively in ISE programs before the 1980s. Now it would be surprising to find an MBA program without TQM coursework.

think systems. Second, many of the substeps in the process (see Figure 12) require specific knowledge and skills that ISEs are grounded in. We understand that ISEs are involved in many BPR efforts in organizations today. We also know that many BPR efforts are information technology-driven and many exclude ISE involvement. We believe this is an error that leads to underperformance.

At several points in the chapter we have mentioned the importance of measurement in transforming to full potential performance. Let's turn our attention to a specific focus on this role and see if we can highlight differences we see in the future in terms of ISE contribution.

3.4.3. Building Effective Measurement Systems

Work measurement and *methods engineering* are terms that have traditionally described actions taken to quantify the performance of a worker or a work unit and to improve the performance of the individual and the unit. The relevant body of knowledge has been developed and preserved through the entire history of ISE. Throughout that history, one principle has served as its unifying force: *an appreciation for and the applied knowledge of systems.* In fact, Frederick Taylor (1856–1915), who is generally recognized as the father of industrial engineering, wrote in 1911, "The system must be first." Methods engineering was pioneered by Frank Gilbreth and his wife Lillian, whose lives were memorialized in the Hollywood film *Cheaper by the Dozen.* Soon after methods engineering began to be practiced, the need for measurement technology became clear. This was inevitable. Once analysts proposed improvements to the way a job was done, natural curiosity led to the question "How much better is it?" The ISEs translation of and response to that question led to the development of tools and techniques for measuring work.

More recently, the importance of methods engineering and work measurement was underscored by the late W. Edwards Deming in his legendary seminars. Dr. Deming, when questioned about a system performance deficit, would confront his audience with the challenge "By what method?" He virtually browbeat thousands of paying customers into the realization that insufficient thought and planning went into the design of work systems. Dr. Deming believed that workers, by and large, strive to meet our expectations of them. The lack of sufficient high-quality output, he taught, stemmed not from poor worker attitude, but from poor management and poor design of the methods, tools, and systems we provide to the otherwise willing worker.

Dr. Deming also promoted the ISEs contribution through work measurement with an additional challenge. Once a solution to the "By what method" question was offered, Deming would ask, "How would you know?" This query highlighted his insistence that decisions regarding process improvements be data driven. In practice, this means that effective systems improvement activities require evidence as to whether the changes make any difference. The requirement is that we use our measurement expertise to quantify the results of our efforts to design and implement better systems. The Deming questions—"By what method?" and "How would you know?"—articulate the defining concerns of the early ISEs, concerns that continue to this very day.

The ability to measure individual and group performance allowed organizations to anticipate work cycle times, which led to more control over costs and ultimately more profitability and better positioning in the marketplace. Understanding how long it *actually* takes to do a task led to inquiry about how long it *should* take to do work through the application of scientific methods. Standard times became prescriptive rather than descriptive. The next step in the evolution was the integration of production standards into incentive pay systems that encouraged workers to exceed prescribed levels of output. Application of extrinsic rewards became an additional instrument in the ISE toolbox, vestiges of which linger on.

So much for the evolution of the work measurement and methods aspects of traditional ISE practice. The following are some evolutionary enhancements that have become part of the ISE measurement value proposition:

- Statistical thinking plays a more critical role in understanding work performance. Variation is inherent in all processes and systems. Discovering the underlying nature of variation and managing the key variables has become more critical than just establishing a standard.
- The role of production quotas is being reexamined. Should we throw out all standards, quotas, and targets, as Dr. Deming suggested? We think not. We contend that the effective approach is to establish a system within which teams of employees hold themselves and each other accountable for system performance and are encouraged to reduce variation and improve performance on their own. Standards that control and limit employee creativity should be eliminated. The key is understanding performance variation and what causes it and creating a partnership with employees so that they are integral members of the team working to improve it.
- Work measurement and methods improvement became detached, to some extent, from the larger system of improvement efforts. Today, efforts to improve what workers do and how they do it is being tied to overall business strategy and actions. This means that measures of performance at the work unit level will have to be tied to and integrated with measures of performance for

larger units of analysis. Linkages between individual worker and team performance and system-level measures of performance are becoming better understood and managed. Here again, our message is ‘‘transcend and include.’’ Efforts to understand how long something does or should take at the employee or work center level will expand to include understanding of how the unit-level systems need to perform in order to fill the treasure chest of the organization.

- Time and quality are no longer the only indicators of organizational performance for which ISEs are being held responsible. This is manifest in a vector of performance indicators, including efficiency, effectiveness, productivity, financial performance, quality of work life, customer satisfaction, and innovation, which are are being made elements in a rich and balanced scorecard on which the organization is being graded (Kaplan and Norton 1996).
- Visibility of measurement systems and portrayal of performance data are being recognized as critical elements in the deployment of enhanced measurement systems. Traditionally, a worker knew the standard and that was it. In the future, employees at every level of the organization will have continual access to their scorecard indicators. Furthermore, they will be aware of how these measures are linked to the performance of more inclusive systems. For example, employees at the checkout counter will understand that their behaviors and attitudes influence whether the store is a compelling place to shop and will have data to tell them how this helps to fill the treasure chest and how that affects them. Store managers will understand that a five-point increase in employee attitudes translates into a 1.3% increase in customer satisfaction, which translates into a 0.5% increase in gross store revenues, and their chart book will portray the data that portray how their unit is performing. As a result, daily, hourly, moment-to-moment decisions will be made on the basis of visible data and facts.

The ISE will be integral to the design, implementation, and maintenance of these comprehensive and very visible measurement systems.

3.4.4. Organizational Systems Performance Measurement

An organizational system is two or more people whose activities are coordinated to accomplish a common purpose. Examples of organizational systems are a work unit, a section, branch, plant, division, company, enterprise.

In Figure 13, the management system is portrayed as consisting of three elements: *who* manages, *what* is managed, and *how* managing is accomplished. In a traditional organization, ‘‘who’’ may be

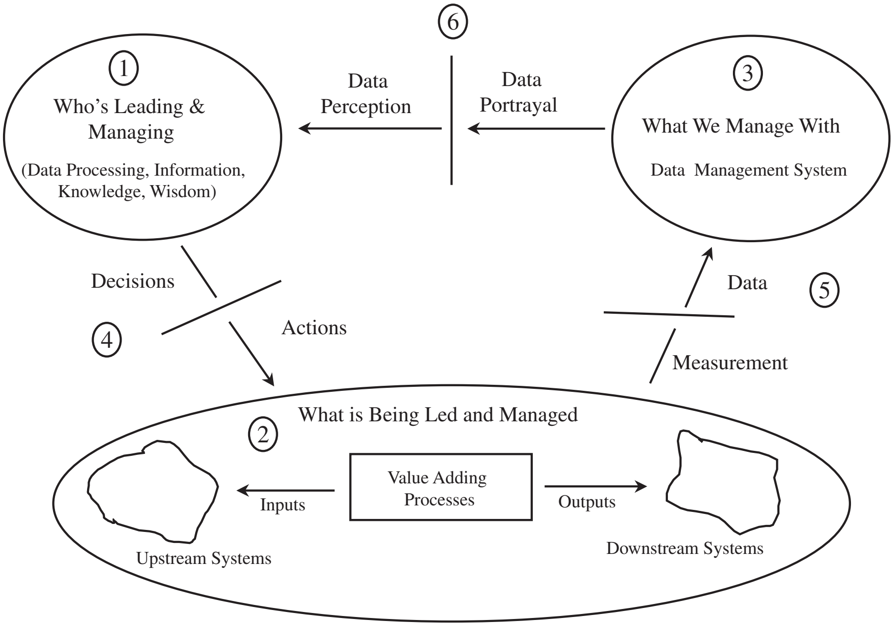

⑥
①
Who’s Leading & Managing
(Data Processing, Information, Knowledge, Wisdom)
Data Perception
Data Portrayal
③
What We Manage With
Data Management System
Decisions
④
Actions
Data
⑤
Measurement
What is Being Led and Managed
②
Value Adding Processes
Inputs
Outputs
Upstream Systems
Downstream Systems

Figure 13 Management System Model.

the manager; but in the contemporary organization, ‘‘Who’’ refers to a management team. For the long-range planning horizon, the management team establishes the goals and objectives—the vision—of what the organization is to become. In the short time horizon, levels of system performance must be specified. The ‘‘What’’ is the system that is being managed; the organization, system, or process of interest, the object of the management efforts. ‘‘How’’ refers to managerial processes and procedures and more specifically to the transformation of data about the performance of the organization (‘‘What’’) into information regarding the need for actions or interventions.

The management system model can also be characterized as a feedback or closed-loop control system. In this version, the management team is the controller (who), the process is the system being controlled (what), and the instrumentation (how) monitors the system states and feeds these back to the controller so that deviations between the actual and the desired states can be nulled. The interfaces between each of the elements also represent the management process. Between the ‘‘what’’ and the ‘‘how’’ elements is the measurement-to-data interface. Between the ‘‘how’’ and ‘‘who’’ elements is the information portrayal/information perception interface. And between the ‘‘who’’ and the ‘‘what’’ elements is the decision-to-action interface. Viewed from the perspective of this model, the management of a function would entail:

1. Determining what performance is expected from the system
2. Monitoring the system to determine how well it is performing in light of what is expected
3. Deciding what corrective action is necessary
4. Putting the correction into place

Note that any embedded organizational system is operating in the context of a larger system and the linkages are critical to total system optimization (Sink and Smith 1994).

This model provides the frame and the outline to be followed in building an effective measurement system. The steps one takes to do this are covered in other references we have provided. We will simply restate that understanding what constitutes success for an embedded system must be clearly understood and operationally defined. The model for success must also be operationalized and understood. Senge (1990) and others devote much of their work to helping us understand the tools of systems modeling in this regard. Without a clear, specific, focused understanding of what the key result areas and their related key performance indicators are, it is difficult for workers or managers to assess how they are doing and on what basis to make those assessments. The measurement system allows for effective Study (S) in the Shewhart/Deming PDSA improvement cycle. A modern measurement system will have to be comprehensive, well integrated, and strategic as well as operational. It needs to portray causal linkages from the system of interest to the next-larger system. This assists in ensuring that we do not fall into the trap of optimizing the subsystem at the expense of the larger system.

In our experience, ISEs understand measurement perhaps better than other disciplines, and yet the traditional approach is often reductionist in its orientation. The key to ISE being better integrated to enterprise improvement is that we apply our strengths in a way that avoids suboptimization and clearly ties to the higher organizational good. This is especially true with measurement.

4. ORGANIZING FOR FULL-POTENTIAL ISE CONTRIBUTION

4.1. Overview

We’ve portrayed what we believe is an extended and expanded value proposition for ISE. To summarize: the ISE is playing a key role in strategy and positioning, planning, change leadership and management, culture, measurement, learning, reengineering, and infrastructure. When you couple this view with the more traditional focus that is discussed in the bulk of the Handbook, we think you will get a flavor of the full potential for ISE and how exciting ISE will be in the decades to come. The issue here is how to position and organize the ISE role such that the potential contribution can be realized. The traditional view is that the ISE function should be located in a dedicated unit that is well connected to positions of power.

An alternative view is emerging. We believe that the environment that made the ISE department or function effective has changed. The challenge is to position the ISE value proposition, not the function. Once again, this transformation requires that there be strong motivation and enlightened executive sponsorship for deploying ISE skills throughout the organization. ISEs will be positioned in ways that were impossible to envision in the past.

4.2. Executive Sponsorship

At the risk of becoming too repetitive, we will reassert that full-potential organizations will have a senior leader (CEO, VP) who champions the ISE contribution. This sponsor will also have a clear mental model of what needs to be done for the organization to migrate toward full potential and how

ISEs can support that. ISEs will be deployed throughout the business in positions of influence and support relative to key initiatives.

This may not seem too different from the current reality, but the major difference will be the extent to which oversight exists. The executive sponsor will own the process of positioning ISEs as their talents are expanded and applied across the enterprise. Review the readiness for change model presented in Section 3.2.3. All the requirements for successful change are present. We've seen executive sponsorship work, and we believe it is the wave of the future.

4.3. Business Partner Relationship Management

Earlier in the chapter we noted that in our conceptual framework, organizations have only one customer. We have abandoned the term *internal customer.* Everyone inside the organization is a partner working on a team dedicated to serving the customer and filling the treasure chest. In this respect the executive sponsor for ISE is in partnership with other key executives working together to migrate the business toward full potential. ISEs are deployed as resources that can support functional and cross-functional efforts to ensure that improvements are well managed and produce the desired results. The model is value based and relationship based. To achieve full potential, it is necessary (but not sufficient) that relationships among employees, internal business partners, stockholders and stakeholders, customers, and suppliers/vendors be effectively managed. This is true for the executive sponsor of the ISE role, for the ISE, and for the partners of ISEs within the enterprise.

4.4. Integration Role

The word ''integration'' is central to the operational definition of ISE. It takes a lot of integration to achieve full potential. It should be clear by now that we view the ISE as a key integrator. Businesses are decentralized or differentiated to reduce span of control and enhance agility, responsiveness, and ownership. Differentiation escalates the need for integration, and integration requires information sharing and cooperation to achieve the higher good. The executive sponsor for ISE should be in a position to understand the higher good, be an integral contributor to organization's senior leadership team, and passionately promote integration. ISEs throughout the organization will have to understand the big picture so that they too can promote integration. To do so, they will be in touch with what is going on across the organization, understand the high level strategy and actions, and share this knowledge when working on their projects throughout the business. This ensures that ISE and their colleagues don't lose sight of the forest for the trees.

5. IIE/CIE/CIEADH RELATIONSHIP MANAGEMENT

5.1. Overview

The leadership in the extended ISE community can be portrayed as consisting of three key constituencies. First is our professional society, the Institute of Industrial Engineers (IIE). Second is the academic institutions that teach and certify competency in the knowledge and skills of their graduates, who then become practitioners. This important group is represented by The Council of Industrial Engineering Academic Department Heads (CIEADH). Third is the Council of Industrial Engineering (CIE), a group of ''executive sponsors'' of ISE in major organizations from around the world. There are other groups in the ISE community; these are three we have chosen for this discussion.

Communication and coordination among these three groups will be essential to our profession achieving full potential. Ideally, from our point of view, CIEADH would view CIE as a focus group of its most influential customers and CIE would view CIEADH as a leadership body representing their supplier organizations. Both CIE and CIEADH would be viewed by IIE as essential customers, suppliers, and stakeholders. IIE would serve both CIE and CIEADH as the organization that encourages and supports lifelong learning and continued personal and professional development for ISE practitioners *and* as their professional society for life. The leaders of all three organizations would accept their roles as stewards of our profession and our professional society.

5.2. Relationship Management

As we've said so many times to this point, much depends on relationships in the context of core values. ISE will achieve full potential in the years to come if and only if those in leadership positions in the three constituencies acknowledge their mutual dependencies and manage their interrelationships. If we truly want our profession to realize its full potential and make the contribution it is capable of, we will opt for improved relationships.

6. MAKING THE FULL-POTENTIAL MODEL WORK FOR YOU: LEADERSHIP AND PERSONAL MASTERY

At the outset we mentioned that it would take full-potential ISEs to create full-potential organizations. We have deferred addressing the issue of personal mastery and making the full-potential model work for the individual ISE. If ISEs accept personal responsibility for their own growth and development, full potential will be within the grasp of ISEs, our customers, and our organizations. The fundamental questions that make up the process are straightforward:

- What is our purpose in life (our enduring reason for being)? What is our vision (full potential, the look and the feel of success)?
- What will it take for us to be successful?
- What work do we have to do to migrate toward our particular visions?
- What conditions do we have to create for ourselves in order to be successful?
- How will we know how we are doing, on what basis will we make this assessment, and how will we manage our performance over time?

So much good work has been done today to achieve personal and professional mastery that it is difficult to suggest a starting point. Each of us has found Senge's and Covey's work on personal and professional mastery a good place to begin (Senge, 1990; Covey, 1994).

As Dr. Deming used to say, the journey to full potential performance will take a lifetime and that's good because that's all you've got *and* you can begin anytime you want as long as you start now!

REFERENCES

Akao, Y., Ed. (1991), *Hoshin Kanri: Policy Deployment for Successful TQM,* Productivity Press, Cambridge, MA.

Collins, J. C., and Porras, J. I., (1994), *Built to Last: Successful Habits of Visionary Companies,* HarperCollins, New York.

Covey, S. R. (1994), *First Things First,* Simon & Schuster, New York.

Csikszentmihalyi, M. (1990), *Flow: The Psychology of Optimal Experience*, Harper & Row, New York.

DeGeus, A. (1997), *The Living Company: Habits for Survival in a Turbulent Business Environment*, Harvard Business School Press, Boston.

Fritz, R. (1991), *Creating: A Practical Guide to the Creative Process and How to Use It to Create Everything,* Fawcett Columbine, New York.

Kaplan, R. S., and Norton, D. P. (1996), *Translating Strategy into Action: The Balanced Scorecard,* Harvard Business School Press, Boston.

Lawler, E. E., III, (1986), *High Involvement Management,* Jossey-Bass, San Francisco.

National Institute of Standards and Technology (1999), "Criteria for Performance Excellence," in *Malcolm Baldrige Business Criteria*, NIST, Gaithersburg, MD.

Porter, M. E. (1996), "What Is Strategy," *Harvard Business Review,* Volume 74, November–December, pp. 61–78.

Schein, E. H. (1992), *Organizational Culture and Leadership,* Jossey-Bass, San Francisco.

Senge, P. M. (1990), *The Fifth Discipline,* Doubleday-Currency, New York.

Sink, D. S. (1998), "The IE as Change Master," *IIE Solutions,* Vol. 30, No. 10, pp. 36–40.

Sink, D. S., and Poirier, D. F. (1999), "Get Better Results: For Performance Improvement, Integrate Strategy and Operations Effectiveness," *IIE Solutions,* Vol. 31, No. 10, pp. 22–28.

Sink, D. S., and Smith, G. L. (1994), "The Influence of Organizational Linkages and Measurement Practices on Productivity and Management," in *Organizational Linkages: Understanding the Productivity Paradox,* D. H. Harris, Ed., National Academy Press, Washington, DC.

Spears, S., and Bowen, H. K. (1999), "Decoding the DNA of the Toyota Production System," *Harvard Business Review,* Vol. 77, No. 5, pp. 96–106.

Thompson, J. D. (1967), *Organizations in Action,* McGraw-Hill, New York.

Tompkins, J. (1999), *No Boundaries: Moving Beyond Supply Chain Management,* Tompkins Press, Raleigh, NC.

Wilbur, K. (1996), *A Brief History of Everything,* Shambala Press, Boston.

Womack, J. T., and Jones, D. T. (1996), *Lean Thinking: Banish Waste and Create Wealth in Your Corporation,* Simon & Schuster, New York.

ADDITIONAL READING

Miller, D. “Profitability = Productivity × Price Recovery,” *Harvard Business Review,* Vol. 62, May–June, 1984.

Sink, D. S., and Tuttle, T. C., *Planning and Measurement in Your Organization of the Future,* IEM Press, Norcross, GA, 1989.

Sink, D. S., and Morris, W. T., *By What Method?,* IEM Press, Norcross, GA, 1995.

Vaill, P., *Learning as a Way of Being,* Jossey-Bass, San Francisco, 1996.

CHAPTER 2

Enterprise Concept: Business Modeling Analysis and Design

Frank O. Marrs
Risk Management Partners, Inc.

Barry M. Mundt
The Strategy Facilitation Group

1. DEFINING THE ENTERPRISE 27
- 1.1. The Enterprise as a Complex, Living System 28
- 1.2. The Impact of the Global Business Environment 28
- 1.3. Increasing and Changing Business Risks 28
- 1.4. The Business Model and Its Purpose 28
- 1.5. The Business Model as the IE's Context 28

2. A COMPREHENSIVE BUSINESS MODEL FRAMEWORK 29
- 2.1. Looking at the Entire Enterprise System 29
- 2.2. Overview of Business Model Elements and Structure 29
- 2.3. The IE's Use of the Business Model 30
  - 2.3.1. Gaining an Understanding of the Whole Business 30
  - 2.3.2. Facilitating a Common Understanding of the Business by Others 30
  - 2.3.3. Identifying Opportunities for Process Improvements 30
  - 2.3.4. Identifying and Mitigating Business Risks 30
  - 2.3.5. Developing the Basis for Process Performance Measurement 30
  - 2.3.6. Facilitating the Development of the Enterprise's Directional Course 30

3. BUILDING THE BUSINESS MODEL 31
- 3.1. Obtain a Committed Sponsor at the Appropriate Level of the Enterprise 31
- 3.2. Set out the Purpose and Scope of the Model 31
- 3.3. Gather and Orient the Model-Building Team 31
- 3.4. Determine Information Requirements for Each Element of the Model 31
- 3.5. Construct the Business Model 31
- 3.6. Build Consensus for the Model 31

4. BUSINESS MODEL CONTENT 31
- 4.1. Developing the Business Model Content 31
- 4.2. Model Element 1: External Forces and Agents 32
- 4.3. Model Element 2: Markets 34
- 4.4. Model Element 3: Business Processes 34
- 4.5. Model Element 4: Alliances and Relationships 34
- 4.6. Model Element 5: Core Products and Services 34
- 4.7. Model Element 6: Customers 34
- 4.8. Summary 35

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

5. ELEMENT 1: EXTERNAL FORCES AND AGENTS 35
- 5.1. The Multiplicity of External Forces and Their Relative Impact on the Business 35
  - 5.1.1. Globalization of Business 36
  - 5.1.2. Revolution in Information Technology 36
  - 5.1.3. Growth of Knowledge Work 36
  - 5.1.4. Data Access Transforms Business Reporting 37
  - 5.1.5. Other Social, Demographic, and Political Changes 37
- 5.2. Customers 38
- 5.3. Competitors 39
- 5.4. Regulators 39
- 5.5. The Community 39
- 5.6. Alliances 39
- 5.7. Stakeholders and Owners 39
- 5.8. Suppliers 39
- 5.9. Capital Markets 39
- 5.10. The Economy 40

6. ELEMENT 2: MARKETS 40
- 6.1. "Market" Defined 40
- 6.2. Market Domains Served 40

7. ELEMENT 3: BUSINESS PROCESSES 40
- 7.1. Categories of Business Processes 40
  - 7.1.1. Strategic Management Process 41
  - 7.1.2. Core Business Processes 43
  - 7.1.3. Resource Management Processes 43
- 7.2. Process Analysis Components 43
  - 7.2.1. Process Objectives 43
  - 7.2.2. Inputs 44
  - 7.2.3. Activities 44
  - 7.2.4. Outputs 45
  - 7.2.5. Supporting Systems 45
  - 7.2.6. Risks That Threaten Objectives 45
  - 7.2.7. Controls Linked to Risks 45

8. ELEMENT 4: ALLIANCES AND RELATIONSHIPS 46
- 8.1. "Alliance" Defined 46
- 8.2. Strategic Alliances in the Value/Supply Chain 48
- 8.3. Performance Management 48

9. ELEMENT 5: CORE PRODUCTS AND SERVICES 49
- 9.1. "Core Product and Service" Defined 49
- 9.2. Categories of Products and Services 49
- 9.3. Measuring Product Performance 49

10. ELEMENT 6: CUSTOMERS 50
- 10.1. "Customer" Defined 50
- 10.2. Categories of Customers 50
- 10.3. Products and Services and Customer Linkages 50
- 10.4. Relationship of Customers to Markets 50

11. APPLYING THE BUSINESS MODEL 51
- 11.1. Communicating the Nature of the Business 51
- 11.2. Improving the Business 51
  - 11.2.1. Strategic Analysis 51
  - 11.2.2. Business Process Analysis 52
  - 11.2.3. Business Performance Measurement 54
  - 11.2.4. Risk Assessment 56
- 11.3. Continuous Improvement 57

ADDITIONAL READING 57

APPENDIX: LIST OF GENERIC BUSINESS PROCESSES AND SUBPROCESSES 58

1. DEFINING THE ENTERPRISE

In this chapter we use the term *enterprise* in its classical sense: an undertaking, especially one of some scope, complication, and risk. Thus, an enterprise could be a business corporation or partnership, a government agency, or a not-for-profit organization. The business modeling concepts described herein can be applied to any kind of enterprise.

1.1. The Enterprise as a Complex, Living System

Defining any given enterprise is a difficult endeavor because the enterprise is perceived differently by each individual or group that views it. Furthermore, each enterprise is a complex, living system that is continually changing, so today's view may be very different from yesterday's.

Often people attempt to define an enterprise by its organizational structure and the executives who occupy key positions. But this is only a small part of the picture. The enterprise actually operates as a complex system, with many parts that interact to function as a whole. In addition to organizational structure, an enterprise's system includes its economic and social environment; the customers it serves; other enterprises with which it cooperates to achieve its objectives; and the internal processes that are designed to set strategic direction, identify and satisfy the customers, and acquire and provide the resources necessary to keep the enterprise running. Thus, to define an enterprise properly one must define the system within which it operates. Ultimately, the success of an enterprise depends on the strength of its intra- and interconnections—the couplings among the organization's internal processes and between the organization and its external economic agents.

1.2. The Impact of the Global Business Environment

In recent years, technological advances in communications have paved the way for enterprises to operate effectively in a global, rather than just a local, environment. The foundation for this globalization was set by the technological advances in transportation experienced during the twentieth century. As a result, global expansion—often through mergers, acquisitions, and alliances—is now commonplace. Indeed, in some industries globalization has become a requisite for survival.

But globalization brings a whole new level of complexity to the enterprise. When an enterprise seeks to operate in a new environment, markets, competition, regulations, economies, and human resources can be very different from what an enterprise has experienced. Accommodating such differences requires understanding them and how they will affect the strategies and processes of the enterprise.

1.3. Increasing and Changing Business Risks

Another aspect of globalization is that it significantly increases the enterprise's business risks—that is, risks that threaten achievement of the enterprise's objectives. Traditionally, management of risks has been focused on the local external environment, including such areas as the nature and size of direct competition, the labor market, the cost of capital, customer and supplier relationships, and competitor innovations.

But in a global operation the business risks become greater and are not always well defined. For example, regulatory environments in foreign countries may favor local enterprises; vying for limited resources—both natural and human—may be problematic; and the foreign work ethic may not be conducive to productive operation and delivery of quality products and services. The business risks in a foreign environment need to be identified, defined, and managed if the enterprise is to be successful.

Even the business risks of local enterprises are affected by globalization. For example, new market entrants from foreign countries can provide unexpected, lower-priced competition, or product innovations originating in another country can become direct product substitutes. As a result, even local enterprises must anticipate new business risks brought on by the strategies of global organizations.

1.4. The Business Model and Its Purpose

An enterprise business model is designed to compile, integrate, and convey information about an enterprise's business and industry. Ideally, it depicts the entire system within which the enterprise operates—both internal and external to the organization. Not only does the construction of a model help enterprise management better understand the structure, nature, and direction of their business, but it provides the basis for communicating such information to employees and other interested stakeholders. The model can be the catalyst for developing a shared understanding of what the business is today and what needs to be done to move the enterprise to some desired future state.

A business model can be as detailed as the users deem necessary to fit their needs. Other factors regarding level of detail include the availability of information and the capabilities and availability of the business analysts who will "build" the model.

1.5. The Business Model as the IE's Context

The business model is a tool that helps the industrial engineer develop an understanding of the effectiveness of the design and management of the enterprise's business, as well as the critical performance-related issues it faces, to evaluate opportunities and manage risk better.

One of the industrial engineer's key roles is to improve the productivity of enterprise business processes. Often he or she is assigned to analyze a particular process or subprocess and make recommendations for changes that will enhance product/service quality, increase throughput, and/or reduce cycle time and cost. But in today's environment, a given process is not a stand-alone operation; rather, it is an integral part of an entire enterprise system. Changes made to one process may very well affect the performance of other processes in the system—sometimes adversely.

A comprehensive business model can provide the enterprise context within which the engineer conducts his or her process analysis. Specifically, the model displays how and where the process fits into the enterprise system, what other processes are affected by it, and what business and information systems must support it. This context helps the engineer make sure that proposed process changes will not degrade the performance of other processes and systems.

2. A COMPREHENSIVE BUSINESS MODEL FRAMEWORK

2.1. Looking at the Entire Enterprise System

The comprehensive business model paints a picture of the entire enterprise system. It depicts not only the internal operations of the enterprise, but the external forces that act upon it. Strategies, business objectives, business risks, and management controls are reflected, as well. Because of its comprehensiveness, the model helps management address key questions about the enterprise:

- Do the enterprise's strategy and the business relationships it has formed address the external forces in the industry?
- Does the design of the business processes established by the enterprise support its strategic objectives?
- Has management gained a complete perception of the business risks that could affect achievement of the strategic and business process objectives?
- Does the design of the management control framework adequately address the business risks?
- Does management monitor and measure those factors that are critical to the achievement of its significant business objectives?

An industrial engineer who is responsible for answering and resolving such questions within an enterprise clearly is in a very strategic position.

2.2. Overview of Business Model Elements and Structure

Figure 1 shows the basic framework for the comprehensive enterprise business model. Six components comprise the model:

- *External forces:* political, economic, social, and technological factors, pressures, and forces from outside the entity that threaten the attainment of the entity's business objectives
- *Markets:* the domains in which the enterprise may choose to operate

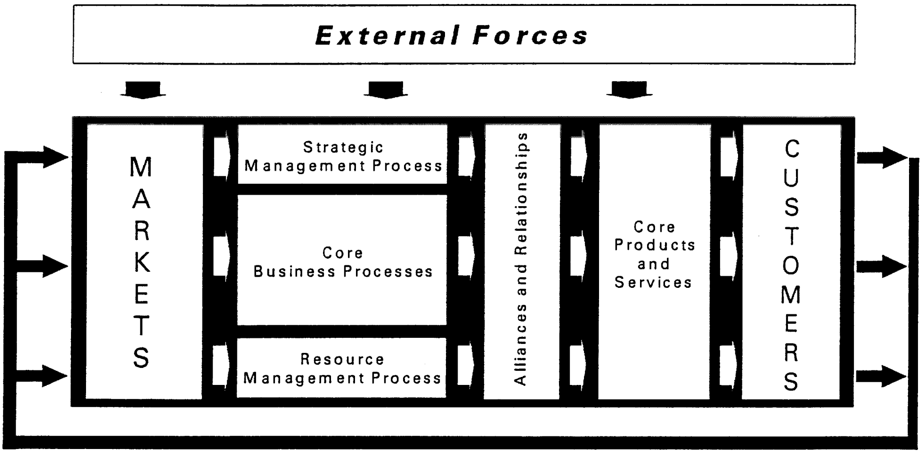

External Forces
MARKETS
Strategic Management Process
Core Business Processes
Resource Management Process
Alliances and Relationships
Core Products and Services
CUSTOMERS

Figure 1 Enterprise Business Model Framework. (From Bell et al. 1997)

- *Business processes*, including:
  - Strategic management processes: the processes by which
    - the enterprise's mission is developed
    - business objectives are defined
    - business risks that threaten attainment of the business objectives are identified
    - business risk management processes are established
    - progress toward meeting business objectives is monitored
  - Core business processes: the processes that develop, produce, market, and distribute the enterprise's products and services.
  - Resource management processes: the processes by which resources are acquired, developed, and allocated to the core business activities
- *Alliances:* the relationships established by the enterprise to attain business objectives, expand business opportunities, and/or reduce or transfer business risk
- *Core products and services:* The value that the enterprise brings to the market
- *Customers:* the individuals and organizations who purchase the enterprise's output

Each of these components is discussed more fully later in this chapter.

2.3. The IE's Use of the Business Model

The business model can be used by the industrial engineer in a number of ways, depending on his or her placement and role in the enterprise. Following are several important potential uses. A more detailed discussion of how the industrial engineer might apply the business model is provided later in the chapter.

2.3.1. Gaining an Understanding of the Whole Business

Regardless of a person's placement and role, a good understanding of the whole business will help the individual see how he or she fits in, both operationally and strategically. This should help the person ensure that his or her objectives are consistent with and supportive of those of the enterprise, improving the potential for advancement within the organization.

2.3.2. Facilitating a Common Understanding of the Business by Others

Each person in an enterprise has an individual view of the business, and this view is often limited in scope and parochial in nature. Variations among viewpoints can cause considerable misunderstanding among individuals and groups about the purposes and direction of the enterprise. A well-documented, comprehensive business model can facilitate a common understanding of the business, both internally and externally.

2.3.3. Identifying Opportunities for Process Improvements

As noted earlier, process improvement is at the heart of the industrial engineer's purpose, regardless of where and at what level he or she is placed in the enterprise. The business model provides the framework for assessing process performance and interrelationships with other processes. Such assessment can lead directly to the identification and design of process changes that will improve the performance of the process as well as the enterprise as a whole.

2.3.4. Identifying and Mitigating Business Risks

Critical to the success of any enterprise is effective management of business risk. The business modeling process provides the basis for identifying the most significant risks, both internal and external, and developing means for mitigating those risks.

2.3.5. Developing the Basis for Process Performance Measurement

Performance measurement is fundamental to continuous improvement. Development of a comprehensive business model includes the identification or establishment of specific performance objectives for each business process. The performance objectives then provide the basis for an ongoing process performance measurement program.

2.3.6. Facilitating the Development of the Enterprise's Directional Course

The comprehensive business model can provide the basis for painting a future picture of the enterprise and determining the course of action to get there. This is done by developing a vision of what leadership wants the enterprise to be at some future point—for example, in three years. This vision is translated into a model of what the enterprise needs to look like to support the vision (the "to be" model). The "to be" model then is compared with today's "as is" model, and a migration plan is developed to transform the business to the new vision. This use of the model is for the industrial engineer who is placed at the highest levels of the enterprise.

3. BUILDING THE BUSINESS MODEL

The enterprise business model depicts a complex system, and building it can be a major effort. Accordingly, the effort should be planned and managed like any complex business project. The major steps in a model-building project plan follow.

3.1. Obtain a Committed Sponsor at the Appropriate Level of the Enterprise

Developing a comprehensive business model can involve considerable information gathering and analysis. Substantial time and cost may be involved. Accordingly, there must be a sponsor for the effort, at the appropriate level of the enterprise, who will back the model design team and ensure that the necessary funding and other resources are provided. In essence, the sponsor will legitimize the effort and be responsible for making sure that the finished model meets its design objectives.

3.2. Set out the Purpose and Scope of the Model

The sponsor and others, as appropriate, clearly articulate the purposes and expected uses of the model. The purpose and use statement provides the basis for determining the scope of the model (e.g., in terms of geographic coverage, business unit coverage, and "as is" vs. "to be" views of the enterprise). The purpose, use, and scope statements then are translated into model-development time and cost objectives for the design team.

3.3. Gather and Orient the Model-Building Team

The team that will develop the enterprise business model is assembled and briefed on the purpose, scope, and framework of the model. The team members may be from various parts of the enterprise, typically including representatives from the key business processes (strategic, core, and resource management). The internal team may be supplemented by outside resources, as necessary (e.g., information specialists, process facilitators, and the like). A skilled project manager is appointed whose role is to ensure that the model-development effort is properly planned and completed on time and within budget.

3.4. Determine Information Requirements for Each Element of the Model

Each element of the model requires the development of information. In most cases the information will be readily available within the enterprise, but, particularly in the external forces area, the information may have to be developed with the assistance of outside information providers. Determining the information requirements for each element will highlight those areas that will be problematic and require special attention.

3.5. Construct the Business Model

The team gathers, compiles, and integrates the information to develop a draft of the business model. The model is depicted graphically and supported by textual material, as necessary. Reviews are conducted to ensure that the model is developed to an appropriate level of detail and that the various elements are properly integrated.

3.6. Build Consensus for the Model

Consensus for the model is built best through involvement. Such involvement can come through participating directly as a member of the design team, participating as a member of a project "steering committee," or acting as a reviewer of the draft business model. The key is knowing who needs to be involved and in what ways they can participate most effectively.

4. BUSINESS MODEL CONTENT

4.1. Developing the Business Model Content

As previously mentioned, the business model is a tool that helps the industrial engineer develop an understanding of the effectiveness of the design and management of the enterprise's business, as well as the critical performance-related issues it faces, to evaluate opportunities and manage risk better. When completed, the business model is a strategic-systems decision frame that describes (a) the interlinking activities carried out within a business entity, (b) the external forces that bear upon the entity, and (c) the business relationships with persons and other organizations outside of the entity.

In the initial stage of developing the business model, pertinent background information is gathered to gain a full understanding of the industry structure, profitability, and operating environment. This industry background information and preliminary analysis then is used to determine the impact on the enterprise's business. Each of the elements of the business model provides a summary of information that is pertinent to developing an understanding of the competitive environment and the enterprise's relative strengths and weaknesses in the marketplace.

The processes the engineer uses to assimilate the acquired knowledge will be unique for each enterprise and each engineer and therefore cannot and should not be reduced to highly structured formats, such as templates, checklists, and mathematical models.

A thorough understanding of five key business principles—strategic analysis, business process analysis, business measurement, risk management, and continuous improvement—will be necessary as the engineer seeks to acquire knowledge about the company's business and industry for the purpose of developing the full business model. These business principles and their interrelationships are depicted in Figure 2.

Throughout the model-building process, the engineer is working toward the ultimate goal of integrating the knowledge he or she obtains about the enterprise's systems dynamics and the congruence between strategy and the environment. He or she may use mental processes or more formal business simulation and systems thinking tools, or some combination of both, to structure his or her thinking about the dynamics of the enterprise's strategic systems.

4.2. Model Element 1: External Forces and Agents

External forces and agents encompass the environment in which an enterprise operates. They are the forces that shape the enterprise's competitive marketplace and provide new opportunities, as well as areas of risk to be managed.

Continuous monitoring and assessment of external forces is critical to the future of any business. The environment plays a critical role in shaping the destinies of entire industries, as well as those of individual enterprises. Perhaps the most basic tenet of strategic management is that managers must adjust their strategies to reflect the environment in which their businesses operate.

To begin understanding what makes a successful business, one must first consider the environment in which the enterprise operates and the alignment of its strategy with that environment. "Environment" covers a lot of territory—essentially everything outside the organization's control. The analysis of external forces and agents includes an assessment of both the general environment and the competitive environment. To be practical, one must focus attention on those parts of the general and competitive environments that will most affect the business. Figure 3 provides an example of a framework that can be used in assessing the general environment.

The general environment consists of factors external to the industry that may have a significant impact on the enterprise's strategies. These factors often overlap, and developments in one area may influence those in another. The general environment usually holds both opportunities for and threats to expansion.

The competitive environment, generally referred to as the "industry environment," is the situation facing an organization within its specific competitive arena. The competitive environment combines

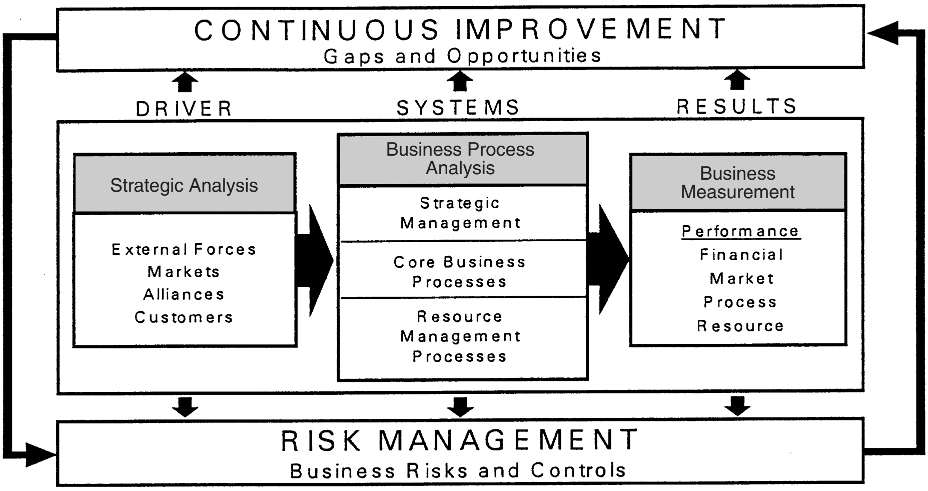


Figure 2 Business Improvement Principles. (From Bell et al. 1997)

Political-Legal Forces	***Social Forces***
■ Tax Laws ■ Trade Regulations ■ Lending Regulations ■ Environmental Laws ■ Workforce Laws ■ Etc.	■ Attitudes ■ Lifestyles ■ Life Expectations ■ Shifts in Workforce ■ Population Shifts ■ Etc.
Economic Forces	***Technological Forces***
■ Money Supply ■ Monetary Policy ■ Unemployment Rates ■ Stage of Business Cycle ■ Globalization ■ Etc.	■ R&D Expenditures ■ Rate of new-products ■ Automation ■ E-commerce ■ Etc.

Figure 3 Dimensions in General Environment Assessment. (From Risk Management Partners, Inc.)

forces that are particularly relevant to an enterprise's strategy, including competitors (existing and potential), customers, and suppliers. The five forces model developed by Michael Porter, probably the most commonly utilized analytical tool for examining the competitive environment, broadens thinking about how forces in the competitive environment shape strategies and affect performance. Figure 4 is a graphic depiction of the five basic forces.

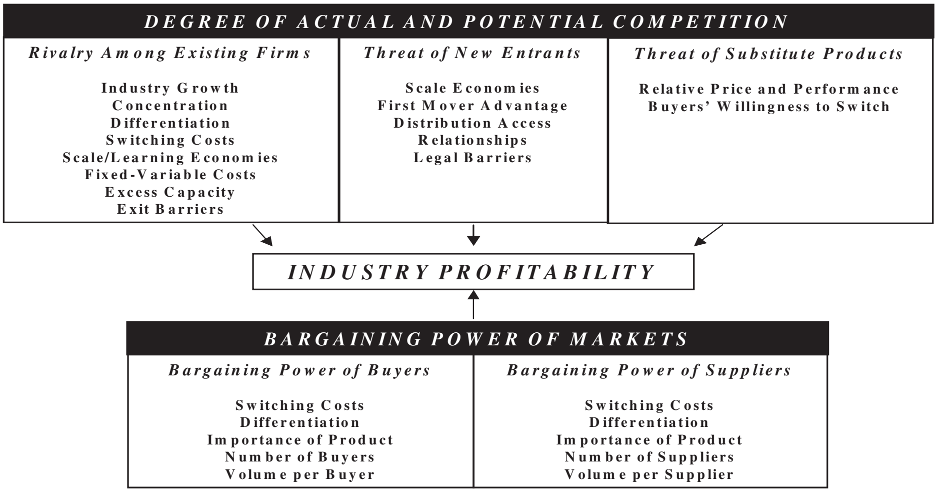

DEGREE OF ACTUAL AND POTENTIAL COMPETITION
Rivalry Among Existing Firms
Industry Growth
Concentration
Differentiation
Switching Costs
Scale/Learning Economies
Fixed-Variable Costs
Excess Capacity
Exit Barriers
Threat of New Entrants
Scale Economies
First Mover Advantage
Distribution Access
Relationships
Legal Barriers
Threat of Substitute Products
Relative Price and Performance
Buyers' Willingness to Switch
INDUSTRY PROFITABILITY
BARGAINING POWER OF MARKETS
Bargaining Power of Buyers
Switching Costs
Differentiation
Importance of Product
Number of Buyers
Volume per Buyer
Bargaining Power of Suppliers
Switching Costs
Differentiation
Importance of Product
Number of Suppliers
Volume per Supplier

Figure 4 Five Forces Model of Competition. (Adapted from Porter 1985)

4.3. Model Element 2: Markets

Understanding the markets in which the enterprise competes is critical in developing the knowledge base for the business model. The extent to which an enterprise concentrates on a narrowly defined niche or segment of the market is referred to as *focus*. The engineer should understand the relevant advantages and disadvantages of particular levels of focus in terms of their impact on competitive advantage. For example, a differentiation strategy is often associated with focusing on a narrowly defined market niche. In contrast, a cost leadership strategy is often associated with a broadly defined target market.

Markets are not static—they emerge, grow, mature, and decline. As a market moves from one life cycle stage to another, changes occur in its strategic considerations, from innovation rates to customer price-sensitivity to intensity of competitive rivalry and beyond. The market life cycle provides a useful framework for studying markets and their impact on the enterprise's value proposition.

4.4. Model Element 3: Business Processes

In the enterprise, work gets done through a complex network of business processes. Work processes are the vehicles of business life. If properly configured and aligned and if properly coordinated by an integrated set of goals and measures, they produce a constant flow of value creation.

Process view of the business involves elements of structure, focus, measurement, ownership, and customers. A process is a set of activities designed to produce a specified output for a particular customer or market. It implies a strong emphasis on *how* work is done rather than *what* is done. Thus, a process is a structured set of work activities with clearly defined inputs and outputs. Understanding the structural elements of the process is key to understanding workflow, measuring process performance, and recommending process improvements.

4.5. Model Element 4: Alliances and Relationships

Financial pressures and time constraints continually squeeze managers who do not have the resources to fill the resource gaps through internal development. Acquisitions have not always been the most effective way to fill these resource gaps. They have proved expensive and brought not only the capabilities needed, but also many that were not desired. As a result, an increasing number of global enterprises recognize that strategic alliances can provide growth at a fraction of the cost of going it alone. In addition to sharing risks and investment, a well-structured, well-managed approach to alliance formation can support other goals, such as quality and productivity improvement. Alliances provide a way for organizations to leverage resources.

The rapid emergence of strategic collaborations as alternatives to the usual go-it-alone entrepreneurial ventures is evident everywhere, from the growing collaborative efforts of large multinationals to the continuing use of alliances to help maintain competitive advantage.

4.6. Model Element 5: Core Products and Services

Intense global competition, rapid technological change, and shifting patterns of world market opportunities compel firms to develop new products and services continually. Superior and differentiated products—those that deliver unique benefits and superior value to the customer—are the key to business success. Understanding the enterprise's core products and services and the value they bring to the customer is essential in developing a business model.

The importance and benefits of measuring new product success and failure cannot be overstated. Measuring new product performance has several benefits. Measurement (a) facilitates organizational learning and process improvements, (b) fulfills the need for consensus on new product outcomes and determinants, and (c) leads to observable benefits, such as improved cycle times, improved new product success rates, and an enhanced ability to assess changes to the new product development process.

4.7. Model Element 6: Customers

An organization cannot survive and prosper in today's world without customers. Customers allow organizations to exist, and yet customer capital is a mismanaged intangible asset. In many cases, companies are providing the wrong products and services for the markets they serve. Such organizations focus on pushing products onto the customer, rather than involving the customer in the design and development activities. Only mismanagement of customer capital can explain why U.S. companies, on average, lose half their customers every five years.

In a knowledge economy, information is more valuable than ever, and generally speaking, customers have more knowledge than the enterprise. As information and the economic power it conveys move downstream, it is vital that businesses manage customer relationships in new ways. Enterprises that learn with their customers, simultaneously teaching them and learning from them, form dependencies with them. Their people and systems—human and structural capital—mesh better than before.

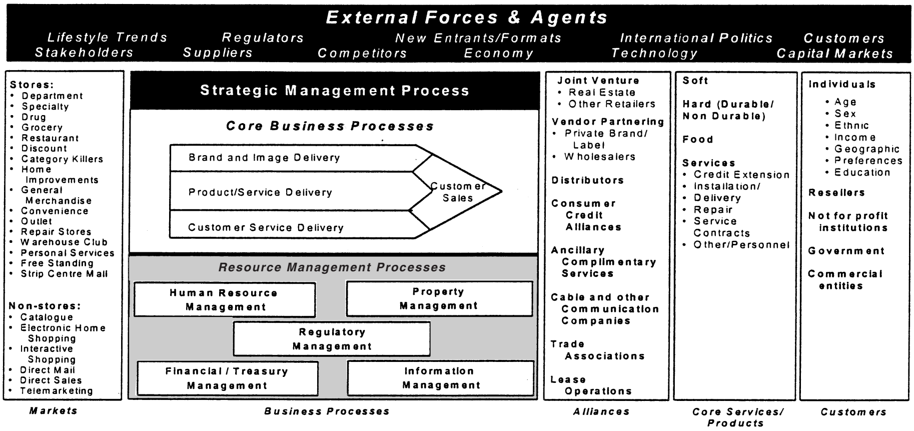

External Forces & Agents
Lifestyle Trends
Regulators
New Entrants/Formats
International Politics
Customers
Stakeholders
Suppliers
Competitors
Economy
Technology
Capital Markets
Stores:
• Department
• Specialty
• Drug
• Grocery
• Restaurant
• Discount
• Category Killers
• Home Improvements
• General Merchandise
• Convenience
• Outlet
• Repair Stores
• Warehouse Club
• Personal Services
• Free Standing
• Strip Centre Mall
Non-stores:
• Catalogue
• Electronic Home Shopping
• Interactive Shopping
• Direct Mail
• Direct Sales
• Telemarketing
Strategic Management Process
Core Business Processes
Brand and Image Delivery
Product/Service Delivery
Customer Service Delivery
Customer Sales
Resource Management Processes
Human Resource Management
Property Management
Regulatory Management
Financial / Treasury Management
Information Management
Joint Venture
• Real Estate
• Other Retailers
Vendor Partnering
• Private Brand/ Label
• Wholesalers
Distributors
Consumer Credit Alliances
Ancillary Complimentary Services
Cable and other Communication Companies
Trade Associations
Lease Operations
Soft
Hard (Durable/ Non Durable)
Food
Services
• Credit Extension
• Installation/
• Delivery
• Repair
• Service Contracts
• Other/Personnel
Individuals
• Age
• Sex
• Ethnic
• Income
• Geographic
• Preferences
• Education
Resellers
Not for profit institutions
Government
Commercial entities
Markets
Business Processes
Alliances
Core Services/ Products
Customers

Figure 5 Example Entity-Level Model—Retail Company. (From Bell et al. 1997)

Understanding the power of the enterprise's customers, their needs and expectations, and the manner in which they are integrated into the enterprise's value proposition is critical in developing a knowledge-based business model.

4.8. Summary

Figure 5 provides an example of the major elements of a hypothetical retail company's business model. Each of the six business model elements is discussed more fully in the following sections.

5. ELEMENT 1: EXTERNAL FORCES AND AGENTS

5.1. The Multiplicity of External Forces and Their Relative Impact on the Business

The general environment consists of many factors external to the industry that may have a significant impact on the strategies of the enterprise. Systematic analysis of the factors making up the general environment can identify major trends in various industry segments. The framework shown in Figure 6 provides insight into the factors that should be understood when the enterprise engages in an analysis of its general, competitive, and internal environments.

In today's world, distance is no longer a barrier to market entry, technologies are rapidly replicated by competitors, and information and communications technologies are shaping a new economic order. To manage their business operations effectively, organizations must now view their playing field as

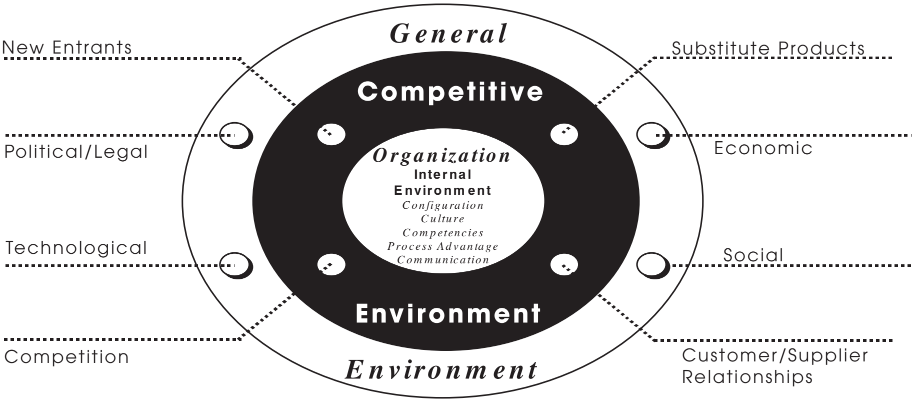

General
New Entrants
Substitute Products
Competitive
Political/Legal
Economic
Organization
Internal
Environment
Configuration
Culture
Competencies
Process Advantage
Communication
Technological
Social
Environment
Competition
Customer/Supplier Relationships
Environment

Figure 6 External/Internal Analysis—A Global Perspective. (Adapted from Bell et al. 1997)

the whole global economy. Prudent managers continually scan these environments for indications of the emergence of business opportunities and risks. They understand current trends and the relevant implications for their business.

Many trends are impacting the way business will be conducted in the future, but the New Economy is taking shape at the intersection of three very significant long-term trends that will continue to gather momentum in the decades ahead: the *globalization of business,* the *revolution in information technology*, and the *growth of knowledge work.* These trends are undermining the old order, forcing businesses to restructure and dramatically change their business models. In the following paragraphs, we discuss these trends and other significant social, demographic, political, and business reporting trends.

5.1.1. Globalization of Business

Simply put, capitalism is spreading around the world—if not full-blown capitalism, at least the introduction of market forces, freer trade, and widespread deregulation. It's happening in the former Communist countries, in the developing world of Latin American and Asia, and even in the industrialized West, with economic union in Europe and the Free Trade agreement in North America.

The number of foreign companies operating in the United States is growing rapidly, at about 2% per year. They now total more than 40,000 and account for about 7% of all corporate assets in the United States. These foreign firms bring to bear the financial and production resources of their home countries on almost any emerging market opportunity in the United States and can quickly take market share with products manufactured less expensively abroad. Foreign competition has arisen not just in large industrial enterprises—automobiles, trucks, aircraft, and computers—but even in some traditional natural monopolies, such as telecommunications and satellite broadcasting.

International trade and investment will play a much larger role in the U.S. economy in the future. Exports and imports already constitute over 25% of the economy.

Increasingly porous national borders, changing corporate cultures, and continuing labor shortages are contributing to the emergence of a global workforce. As regions develop into pockets of specific talent, more workers will relocate to them. In other cases, companies will go to workers, hiring them where they live. Technology will enable the efficient execution of tasks regardless of proximity to home office. This fluid interchange of individuals and information will bring together people of disparate backgrounds. Individuals with dual nationalities will be commonplace.

5.1.2. Revolution in Information Technology

The foundation of the New Economy is the revolutionary explosion of computer processing power. Computing power doubles every 18 months. In addition, we have seen a 22-fold increase in the speed of data transmission obtained over ordinary telephone lines during the past two decades. This wave of innovation in data communications has promoted the extraordinary build-out of the world's computer networks. Over 60 million computers are connected via the Internet. The network phenomenon is just as remarkable as the explosion in computing power. As information technology reduces the trade-off between the depth and quality of information and its access, the economics that underlie industry and organizational structures will be transformed. As more individuals and businesses connect electronically, the economics of scope will change organizational relationships, competition, and make vs. buy decisions.

Technology is likely to continue on its path of being smaller, faster, cheaper, and less visible in the everyday world. The intersection of computing and telecommunications will bring about a fundamental change in the perception of distance and time. At the same time, the traditional interface that has existed between humans and technology will rapidly disappear. Remote sensing, data collection systems, cameras, and adjuncts to sensing abilities are among the major new applications in this field.

More significantly, information technology is transforming the global business environment. Housing and autos used to drive the U.S. economy. Now information technology accounts for a quarter to a third of economic growth.

5.1.3. Growth of Knowledge Work

Increasingly, knowledge has become more valuable and more powerful than natural resources and physical facilities. Value propositions are based on information and ideas, rather than on the mere physical attributes of products. This phenomenon is true in both service businesses and in businesses that have historically focused heavily on tangible products. Knowledge is being used to enhance greatly the value of all physical products. The result is that much of our economic growth in recent years has been intangible. As value creation shifts from the mere economics of physical products to the economics of information, managing information, knowledge, and innovation will become a business imperative.

Information and knowledge have become the sources of competitive advantage. In industry after industry, success comes to those enterprises that manage knowledge and information more effectively than their competitors. Companies like Wal-Mart, Microsoft, and Toyota became great companies because they had intellectual capital—knowledge, information, intellectual property, and experience—and used it to achieve competitive advantage. Knowledge and information have become the New Economy's primary raw materials and the source of its most important products.

5.1.4. Data Access Transforms Business Reporting

The demand for data from users, including employees, capital suppliers, customers, regulators, and the like, will continue to accelerate rapidly. This will require the use of massive databases and network technologies that will provide customized information to users—anyplace, anytime.

In addition, highly networked environments will increase user needs for instantaneous information. Information on demand will be a requirement of all network participants. Users won't be satisfied with paging through static report images; they will need to analyze data—within the scope of the reporting system—from multiple perspectives in order to answer critical business questions involving all aspects of the value chain.

Data access will replace traditional forms of reporting. Companies will employ complex, multifaceted, client/server report-processing environments that include intelligent report viewers, tiered report processors, flexible report schedulers, report distribution schemes, user flexibility, and administration and control mechanisms. Content agents that allow users to customize their reports and receive data in real time will support these databases.

5.1.5. Other Social, Demographic, and Political Changes

The following social, demographic, and political trends have been obtained from various research studies. They are illustrative examples of current trends and their related business implications. Many government agencies and private research firms continuously analyze and monitor trends that affect the way business is conducted. The enterprise and the engineer should cultivate and become familiar with these resources and monitor current trends during the design and development of the full business model.

5.1.5.1. Birth Rates The significant changes in birth rates during the past half century continue to affect the marketplace in subtle ways. New births have now become part of a much larger population base and are not as likely to have the major impact that the "baby boomers" have had. Markets will be smaller and enterprises will have to spend more time and resources to tap into them.

5.1.5.2. Immigration As births fall in the next decade, immigration will have even more of an effect on the composition of the market. Immigrants bring a diverse set of skills and attitudes to the United States, the most prominent being their enthusiasm and desire to partake in the U.S. experience.

Immigrants expand the labor pool in the workplace at both ends, as unskilled workers and as high-end workers with specialized talent and training. Business will find that the foreign-born make up an increasing share of their markets, especially in the four largest and most dynamic states: California, Texas, New York, and Florida. Companies will have to learn about the cultures of these groups as they become both customers and employees.

5.1.5.3. Household Growth Households are of particular importance to future markets because they are the fundamental purchasing unit. Virtually all consumer-spending decisions are made within the context of household needs and budgets. Thus, new households create new sales opportunities.

One of the most important longer-term consequences of the aging baby boomers is that household formation is slowing down. For companies that sell products and services to new households (housing, home furnishings, new cars, insurance), the days of growing with the market are over. Opportunities will now have to come from understanding the composition of the household, more than sheer growth in numbers. More and more sales will be driven by increasing the provision of value-added goods and services to households.

5.1.5.4. Families Continue to Change The composition of the household is changing. The share of households made up of married couples with children declined from 40% in 1970 to 25% in 1995. That share will continue to go down. The households with the most dynamic growth rates will be married couples without children.

Business will make more sales to smaller households, and the sales will demand more one-on-one interaction. This is because needs and tastes will no longer be driven by family imperatives, which tend to be similar to each other, but by those of adults, which tend to be more personal. This means much wider swings in the purchasing patterns of households.

5.1.5.5. Income Mobility of Workers Increasing immigration rates, the growing importance of education, and changing systems of compensation that reward high performers have contributed to a disturbing trend in the United States: a growing inequality in income between the rich and the poor. In the last three decades, the number of households with incomes under \$15,000 (in constant 1996 dollars) has grown from 14 million to 21 million, while the number with incomes over \$75,000 (in constant 1996 dollars) has grown from 4 million to 17 million. The good news is that every year about one third of adults of working age move out of their income quintile. In five years, almost half move.

Consumers' purchasing behavior is driven by household resources. But access to credit means that consumer purchases may not always be limited by current earnings. Many household purchases are based on expected changes in future income. The fact that as many as 50% of adults can expect to find themselves in a different income quintile in the next five years suggests that payment flexibility will be a critical tool for enterprises trying to meet the ever-more elusive needs of the 21st century consumer.

5.1.5.6. The Virtual Workforce Technology and changing organizational cultures are enabling more people to work wherever they choose to live. In the next five years, the number of telecommuters is expected to reach 20 million or more. Some predict that half the workforce will be telecommuting from home, office, or shared facilities within the next decade.

At the same time, the number of temporary workers, freelancers, independent contractors, and the like exceeds 25 million by some estimates. Virtual partnerships/alliances between independent contractors are expected to flourish as sophisticated telecommunications capabilities enable people to link up with anyone, anywhere.

5.1.5.7. Political and Regulatory Changes The regulation of commerce and industry is being adapted to meet the needs of the new consumer. The United States spent almost a century building a regulatory network to protect citizens from the complex risks of a rich, urbanized, industrialized society. Now a more sophisticated consumer, new technologies, and a much more competitive global market are gradually creating an environment more self-regulating and open to consumer discretion, in which it is easier to spread throughout the marketplace the risk that the government formerly took on. As a result, regulatory barriers are coming down.

Sophisticated consumers, one of the key drivers of the move toward deregulation, will become a roadblock if they feel that their fundamental birthrights are threatened: (a) affordable and accessible health care choices, (b) safety and security of the financial system, (c) quality of life that environmental regulations protect; and (4) any issue that seems to lead to an unwanted invasion of privacy.

5.1.5.8. Privatization The transformation of state-owned or controlled enterprises into privately owned or managed enterprises is sweeping the world for the second decade. This phenomenon indicates expanding confidence in the benefits of market forces the world over.

While privatization initiatives have been common all over the world, Europe has felt the largest impact. Almost 60% of the private money raised in the last few years has been in Europe. The flow of public enterprises into the private sector will continue in the 21st century, though probably at a slower rate. Privatization has increased competition and lowered prices, given consumers more choices, increased the rate of innovation, ended subsidies to state-run enterprises, provided new investment opportunities, and replaced monopolies with new players.

5.1.5.9. Key Trends to Follow The New Economy will be affected largely by future developments in three key areas:

1. *Information technology*, providing refinements of computer technologies to optimize the possibilities of electronic commerce
2. *Biotechnology*, where the manipulation of genes will allow new control over diseases and health possibilities
3. *Nanotechnology*, the development of miniaturized systems so that everyday instruments such as mobile phones and calculators can be used in extraordinary ways

5.2. Customers

In the New Economy, the companies with the smartest customers win. The richness and reach of information created by the network economy has moved the power downstream to the customer. With more timely, accurate, and relevant information, customers will be in the driver's seat as existing products are improved and new products are introduced. Their power in the New Economy cannot be overemphasized. The extent to which customers are integrated into the design, development, and improvement of products and services will determine competitive advantage. Knowledge is the driver of the New Economy, and customers generally have much more of it than the producers of products and services.

5.3. Competitors

As the richness and reach of information improve, traditional barriers to entry in most markets will be reduced substantially. Enterprises can no longer focus most of their efforts on analyzing existing competition. New market entrants will appear at an ever-increasing rate, and value propositions will be challenged continuously.

5.4. Regulators

The regulation of commerce and industry is being adapted to meet the needs of the new consumer. The competitive global market is gradually creating an environment that is more self-regulating and open to consumer discretion. As a result, sophisticated consumers are driving deregulation; however, markets will deregulate at varying speeds around the world, some much more slowly than others. In-depth understanding of local regulations will be important as enterprises expand into an ever-increasing number of new markets.

While many regulatory barriers will disappear over time, new regulations will emerge regarding the environment, safety of the financial system, access to health care, and other areas that are considered important to the new generation.

5.5. The Community

The networked economy will ensure that an enterprise's stakeholders, including the local communities that it serves, have a breadth of information to help them make choices. Managing the brand image will become even more important as companies realize the importance of providing information to their stakeholders about social and environmental performance.

5.6. Alliances

Alliances are fraught with uncertain risks and opportunities. Uncertainty brings ambiguity, and ambiguity can lead to business failures. Alliances are most advisable when conditions are right within both the enterprise and the target industry. When alliances are considered, there is a range of strategic options that should be measured in terms of their related risks and rewards.

Rewards can be measured in a variety of ways: market share, cash flow, depth of product line, and growth, to name a few. Risks generally include political, monetary, technological, partner, and market risks, among others.

The architecture of alliances is composed of a number of critical elements, including common language; principles and practices; strategies; structure, roles, and responsibilities; processes and systems design; interrelationships and interfaces; early warning signals; and performance management.

5.7. Stakeholders and Owners

Stockholders are only one of several possible stakeholder groups. Obligations to a firm's stockholders are generally referred to as the firm's fiscal responsibility, while obligations to its stakeholders—parties that have an interest, or stake, in the success or performance of the company—are referred to as the firm's social responsibility.

Sweeping trends in corporate governance are placing more oversight responsibility on boards of directors, audit committees, and other corporate committees to improve fiscal and social performance, as well as stakeholder communication. In addition, stockholders are demanding board of director independence and accountability.

5.8. Suppliers

In many industries, the cost of purchased supplies accounts for 60–80% of production costs, so suppliers can have an important impact on an industry's profit potential. When the number of suppliers in a market is limited and substitute products are lacking, suppliers can exercise considerable power over their customers because the switching costs can be problematic and costly.

The relative importance of the suppliers' products to the buyer, and conversely the relative lack of importance of the buyer to the supplier group, give significant power to suppliers. Other factors that increase the bargaining power of suppliers include high switching costs and a credible threat of suppliers to move into various stages of the value chain as direct competitors.

5.9. Capital Markets

The changes in capital markets in Europe and North America are spreading throughout the world. China is already entering the commercial capital markets in Europe and the United States, and Russia will follow. The West is investing in rising and maturing stock exchanges throughout the world as it seeks market positions and greater return on capital. Major new enterprises in developing economies, created through privatization, are entering the world's capital markets. But as companies tap capital sources outside their home countries, capital suppliers are likely to demand representation on their boards of directors.

5.10. The Economy

Macroeconomic developments, such as interest rate fluctuations, the rate of inflation, and exchange rate variations, are extremely difficult to predict on a medium- or long-term basis. Unpredictable movements of these macroeconomic indicators cannot only affect a company's reported quarterly earnings, but even determine whether a company survives. There is general agreement that the financial environment, characterized by increased volatility in financial markets, is more risky today than in the past. Growing uncertainty about inflation has been followed quickly by uncertainty about foreign exchange rates, interest rates, and commodity prices.

The increased economic uncertainty has altered the way financial markets function. Companies have discovered that their value is subject to various financial price risks in addition to the risk inherent in their core business. New risk management instruments and hybrid securities have proliferated in the market, enabling companies to manage financial risk actively rather than try to predict price movements.

6. ELEMENT 2: MARKETS

6.1. "Market" Defined

In simple terms, a market is a group of customers who have a specific unsatisfied need or want and are able to purchase a product or service to satisfy that need. For example, the market for automobiles consists of anyone older than the legal driving age with a need for transportation, access to roads, and enough money to purchase or make a payment on a car.

Markets can be broken down in numerous ways as marketers try to find distinctive groups of consumers within the total market. Market segmentation allows marketers to allocate promotional expenses to the most profitable segments within the total market and develop specific ad campaigns for each one. Segmentation produces a better match between what a marketer offers and what the consumer wants, so customers don't have to make compromises when they purchase a product.

6.2. Market Domains Served

The served market is the portion of a market that the enterprise decides to pursue. For example, a company that manufactures video games defines the market as anyone who owns a television. The *potential market* is defined as households with children and a television. The *available market* is limited to households with children, a television, enough income to make the purchase, and a store nearby that carries the game. The *served market* consists of households with a television, access to a toy store, sufficient income to buy the product, and children within a specific age range.

7. ELEMENT 3: BUSINESS PROCESSES

7.1. Categories of Business Processes

Considerable controversy revolves around the number of processes appropriate to a given organization. The difficulty derives from the fact that processes are almost infinitely divisible; the activities involved in taking and fulfilling a customer order, for example, can be viewed as one process or hundreds. Process identification is key to making process definitions and determining their implications. If the objective is incremental improvement, working with many narrowly defined processes is sufficient because the risk of failure is relatively low, particularly if those responsible for improving a process are also responsible for managing and executing it. But when the objective is radical process change, a process must be defined as broadly as possible.

Before we explain the major business process categories, the following definitions may be useful:

- A *business process* is a logical, related, sequential—connected—set of activities that takes an input from a supplier, adds value to it, and produces an output to a customer.
- A *key business process* is a process that usually involves more than one function within the organizational structure, and its operation has a significant impact on the way the organization functions.
- A *subprocess* is a portion of a major process that accomplishes a specific objective in support of that major process.
- *Activities* are work elements that go on within a process or subprocess. They may be performed by one person or a team of people.
- *Tasks* are individual elements and/or subsets of an activity. Normally, tasks relate to how an individual performs a specific assignment.

For purposes of understanding the business and documenting the business model, it is useful to organize the business processes into categories:

1. *Strategic management processes:* those processes that develop the value proposition of the enterprise and define the business objectives
2. *Core business processes:* those processes that develop, produce, sell, and distribute products and services (i.e., the value chain)
3. *Resource management processes:* those processes that support and provide appropriate resources to the value-creating processes of the enterprise

Most enterprises can define their business in terms of 10–15 key processes: one to three strategic management processes, 5–7 core business processes, and 3–5 resource management processes. Figure 7 is an illustrative example of process definitions in the industrial products industry.

Identifying and selecting processes for the business model is an important prerequisite to understanding the enterprise and its competitive strengths and weaknesses. Without some focus on critical processes, an organization's energies, resources, and time will not be focused appropriately. The appendix to this chapter provides a list of key generic business processes and their related subprocesses that should be useful as a guide in identifying and selecting processes in any organization.

7.1.1. Strategic Management Processes

Strategic management is the name given to the most important, difficult, and encompassing challenge that confronts any private or public organization. The conflict between the demands of the present and the requirements of the future lies at the heart of strategic management. Change is the central concern and focus of strategic management: change in the environment, change inside the enterprise, and change in how the enterprise links strategy and structure. Strategic management is the process of identifying opportunities to achieve tangible and sustainable success in the marketplace and understanding the risks that threaten achievement of that success. Figure 8 shows one view of the major components of a strategic management process.

Information technology (IT) has introduced new challenges and opportunities for business. Businesses are forced to adopt IT strategies that provide for connectivity to everyone in the business network—stakeholders, suppliers, customers, alliance partners, and others—and react strategically and operationally in real time. The external forces affecting business are changing so rapidly that organizational structures must be designed so they can respond quickly to changes in the business environment. In addition, IT investments must be leveraged by the adoption of open systems and standards that allow rapid changes in those systems that support the execution of key business processes.

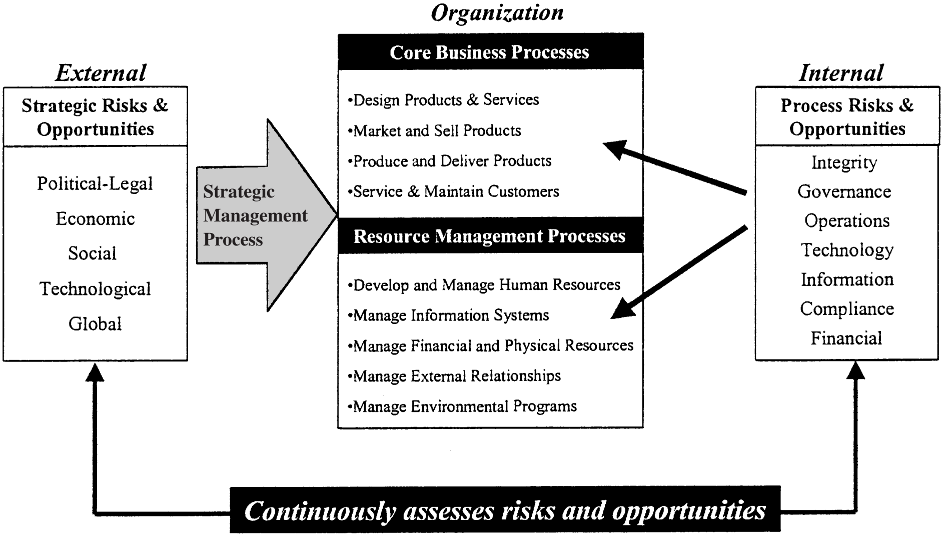

Organization
External
Strategic Risks & Opportunities
Political-Legal
Economic
Social
Technological
Global
Strategic Management Process
Core Business Processes
•Design Products & Services
•Market and Sell Products
•Produce and Deliver Products
•Service & Maintain Customers
Resource Management Processes
•Develop and Manage Human Resources
•Manage Information Systems
•Manage Financial and Physical Resources
•Manage External Relationships
•Manage Environmental Programs
Internal
Process Risks & Opportunities
Integrity
Governance
Operations
Technology
Information
Compliance
Financial
Continuously assesses risks and opportunities

Figure 7 Process Definitions—Industrial Products Example. (From Risk Management Partners, Inc.)

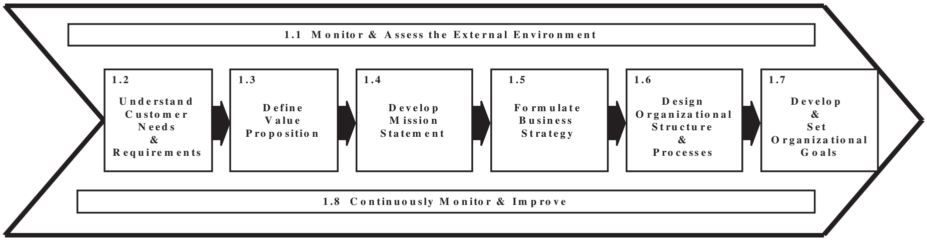

1.1 Monitor & Assess the External Environment
1.2 Understand Customer Needs & Requirements
1.3 Define Value Proposition
1.4 Develop Mission Statement
1.5 Formulate Business Strategy
1.6 Design Organizational Structure & Processes
1.7 Develop & Set Organizational Goals
1.8 Continuously Monitor & Improve

Figure 8 Illustrative Strategic Management Process (From Risk Management Partners, Inc.)

Strategic management principles in the New Economy must recognize the importance of knowledge and information to all value propositions and the need for structures that can adapt quickly and continuously improve in real time. Figure 9 provides an example of the important components of strategy in a connected world.

In the ''Enterprise Business Model'' block in Figure 9:

- ''Knowledge Networks'' means capturing and managing knowledge as a strategic asset.
- ''Process Excellence'' means designing and managing processes to achieve competitive advantage.
- ''Core Competencies'' means focusing on those things that the enterprise does best and using alliance partners to supplement those skills.

In particular, strategy involves continuously reconceiving the business and the role that business can play in the marketplace. The new business model is a real-time structure that can change continually and adapt more quickly and better than the competition.

The traditional approach to strategy development relied upon a set of powerful analytic tools that allowed executives to make fact-based decisions about strategic alternatives. The goal of such analysis was to discuss and test alternative scenarios to find the most likely outcome and create a strategy based on it. This approach served companies well in relatively stable business environments; however, fact-based decisions in the rapidly changing New Economy will be largely replaced by imagination and vision.

Rapidly changing business environments with ever-increasing uncertainty require new approaches to strategy development and deployment. While traditional approaches are at best marginally helpful and at worst very dangerous, misjudging uncertainty can lead to strategies that do not identify the business opportunities and business risks. However, making systematically sound strategic decisions will continue to be important in the future. Even in the most uncertain environments, executives can generally identify a range of potential scenarios. The key will be to design business models that can

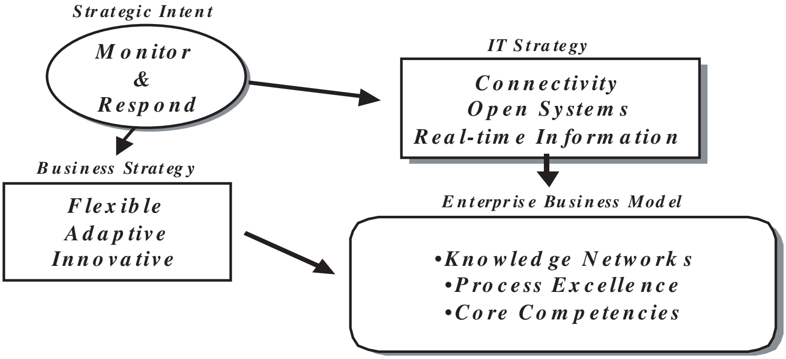

Strategic Intent
Monitor & Respond
IT Strategy
Connectivity
Open Systems
Real-time Information
Business Strategy
Flexible
Adaptive
Innovative
Enterprise Business Model
•Knowledge Networks
•Process Excellence
•Core Competencies

Figure 9 A New Economy Strategy. (From Risk Management Partners, Inc.)

respond quickly as economic realities change. Understanding competitive environments, external forces, and the levels of uncertainty in the marketplace will lead to more informed and confident strategic decisions. The design and development of a real-time business model can provide an invaluable tool to support the strategic management process.

7.1.2. Core Business Processes

Core business processes develop, produce, sell, and distribute an entity's products and services; they are the entity's value chain. These processes do not follow traditional organizational or functional lines, but reflect the grouping of related business activities.

The management of core business processes is about execution of the enterprise's value proposition. Significant changes are taking place in the New Economy that will have profound impacts on core business processes and the way they are managed in the future.

IT is driving the transformation of core business processes by creating a new competitive dynamic that rewards institutional agility. Business processes will no longer be viewed within the boundaries of the organization. Multiple partners will be involved in key activities of many organizations' key business processes.

Historically, the fundamental disposition of core business processes is to prepackage and shrinkwrap as much product as possible, take advantage of economies of scale, and then persuasively offer these products. Service is viewed as a way of enhancing the attractiveness of products. The New Economy enterprise, on the other hand, will design core processes that concentrate on assembling modular elements into a customized response to a specific customer's specific request.

The evolution of extranets and content standards alters the bases of competitive advantage by increasing the quality of information and its access. These technologies diminish the value of established business processes and relationships because most buyers can easily find suppliers—worldwide—who offer the best product. Higher-quality information will allow companies to improve business processes by supplementing internal capabilities with the needed skills and technologies from others—worldwide.

The opportunities to radically change traditional business processes and take advantage of network externalities, multiple strategic and alliance partners, and the richness and reach of information cannot be overstated. Using business process excellence to obtain competitive advantage will be more important than ever before. Deconstructing value/supply chains for the purpose of exploiting market opportunities and reducing transaction costs will be the norm for most successful businesses in the New Economy.

The ultimate challenge posed by deconstructing value chains will be to the traditional hierarchical organization, with its static business processes. Core business process improvement will be both challenging and exciting as opportunities are unleashed to create value in new ways. Outsourcing will flourish as organizations design core processes that emphasize their strengths and supplement them with the outstanding capabilities of other firms that can add significant value to their products and services.

7.1.3. Resource Management Processes

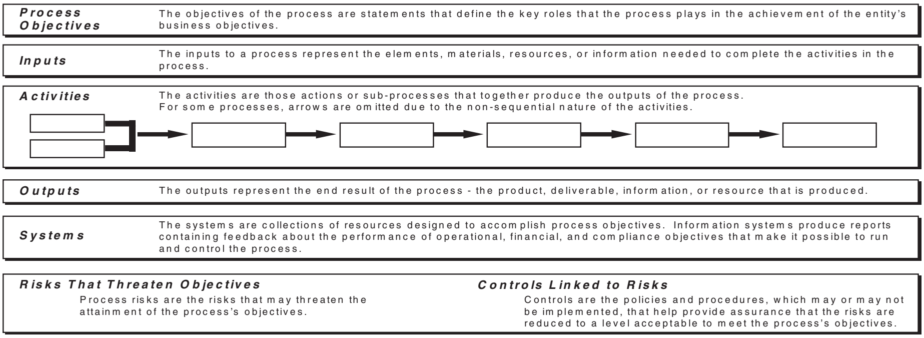
Resource management processes are the processes by which organizations allocate resources and monitor their use. They provide appropriate resources to support the other business processes. Resource management processes can be placed into three basic categories: information, people, and capital. They are focused on serving the needs of internal customers, principally those in the enterprise's core business processes who are externally focused on serving customers outside the organization. Without appropriate resources—information, people, and capital—the core processes cannot offer the value customers need and will cease to provide an effective source of competitive advantage.

7.2. Process Analysis Components

Figure 10 displays a framework for process analysis. The process analysis components in this framework are discussed in this section.

7.2.1. Process Objectives

Processes are established to serve specific customer needs. The customers may be internal, such as another process, or external to the enterprise. The process objectives define what value will be supplied to the customer. One can look at them as the whole purpose for which the organization has put together this set of resources and activities. Process objectives need to be specific, measurable, attainable, and realistic and to have a sense of time. Business process objectives may differ significantly between enterprises within an industry or industry segment, being shaped by the organization's strategic objectives and related critical success factors. For example, the business objectives for the "materials procurement process" in a consumer products company might be as follows:



Process Objectives The objectives of the process are statements that define the key roles that the process plays in the achievement of the entity's business objectives.

Inputs The inputs to a process represent the elements, materials, resources, or information needed to complete the activities in the process.

Activities The activities are those actions or sub-processes that together produce the outputs of the process. For some processes, arrows are omitted due to the non-sequential nature of the activities.

Outputs The outputs represent the end result of the process - the product, deliverable, information, or resource that is produced.

Systems The systems are collections of resources designed to accomplish process objectives. Information systems produce reports containing feedback about the performance of operational, financial, and compliance objectives that make it possible to run and control the process.

Risks That Threaten Objectives Process risks are the risks that may threaten the attainment of the process's objectives.

Controls Linked to Risks Controls are the policies and procedures, which may or may not be implemented, that help provide assurance that the risks are reduced to a level acceptable to meet the process's objectives.

Figure 10 Process Analysis Framework. (Adapted from Bell et al. 1997)

- Effectively manage supplier relationships to ensure the highest quality materials at the lowest cost.
- Provide accurate, real-time production information to suppliers to minimize inventory levels.

7.2.2. Inputs

The inputs to a process represent the elements, materials, resources, or information needed to complete the activities in the process. Examples of inputs for the aforementioned materials procurement process could be the following:

- Material requirements, supply requisitions, negotiated prices;
- Material specifications, bills of material, capital authorizations
- Supply base, production reports, lead times

7.2.3. Activities

Almost everything that we do or are involved in is a process. Some processes are highly complex, involving thousands of people, and some are very simple, requiring only seconds of time. Therefore, a process hierarchy is necessary to understand processes and their key activities. From a macro view, processes are the key activities required to manage and/or run an organization. Any key business process—a strategic management process, a core business process, or a resource management process—can be subdivided into subprocesses that are logically related and contribute to the objectives of the key business process. For example, Figure 11 provides an example of the subprocess components of a new product planning process for a consulting firm.

Every key business process or subprocess is made up of a number of activities. As the name implies, activities are the actions required to produce a particular result. Furthermore, each activity is made up of a number of tasks that are performed by an individual or by small teams. Taken together, tasks form a microview of the process.

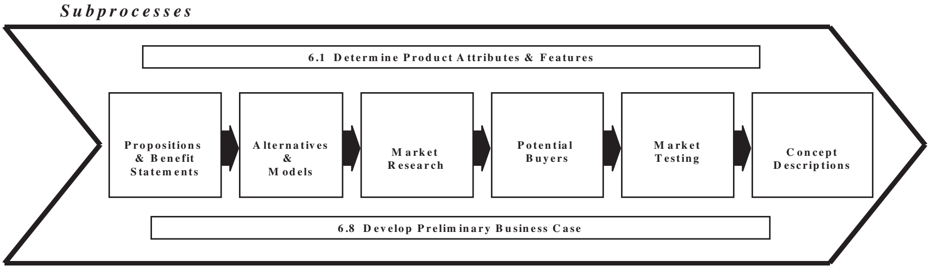

Subprocesses
6.1 Determine Product Attributes & Features
Propositions & Benefit Statements
Alternatives & Models
Market Research
Potential Buyers
Market Testing
Concept Descriptions
6.8 Develop Preliminary Business Case

Figure 11 Example: New Product Planning Process—Consulting. (From Risk Management Partners, Inc.)

7.2.4. Outputs

Outputs represent the end result of a process; they are the products, deliverables, information, or resources that are produced.

7.2.5. Supporting Systems

Supporting systems include the hardware, software, information, and communications capabilities that the organization requires to fulfill its mission and achieve its business objectives.

7.2.6. Risks That Threaten Objectives

Business risk is the threat that an event or action will adversely affect an entity's ability to achieve its business objectives and execute its strategies successfully. A business risk is always related to one or more business objectives and can be described as the antithesis of those objectives.

Business risks can be categorized as:

- *External:* strategic risks that threaten an enterprise's marketplace objectives and are mitigated by an effective strategic management process
- *Internal:* process risks that threaten an enterprise's ability to execute its strategy effectively and are mitigated by effective process controls

Figure 12 shows examples of generic risks that could affect the achievement of core process objectives. Similarly, Figure 13 provides examples of generic risks that could affect the achievement of resource management process objectives.

7.2.7. Controls Linked to Risks

A new business risk control paradigm is changing the way organizations manage their business risks. Over the years, businesses have used a variety of practices to control risk. In many organizations, however, control is a misunderstood and misapplied concept. Control all too often means inflexible and unimaginative budgets, seemingly endless management reporting requirements, and an overburdening and often irrelevant stream of information up and down the corporate hierarchy. The new control paradigm is illustrated in Figure 14.

Fast-moving markets, flattening corporate hierarchies, and the need for an expanding scope of authority at the local level are making salient the costs of misunderstanding and misapplying control. In response, managers are rethinking fundamental definitions of control and how business risks should be identified and mitigated.

Empowerment will increasingly blur lines of authority, and increasingly flat organizations will provide fewer opportunities for segregation of duties. Traditional notions of control (such as segregation of duties and functions, proper authorization for expenditures, controlled access to assets, and proper recording of transactions) that define the procedural checks and balances that safeguard assets

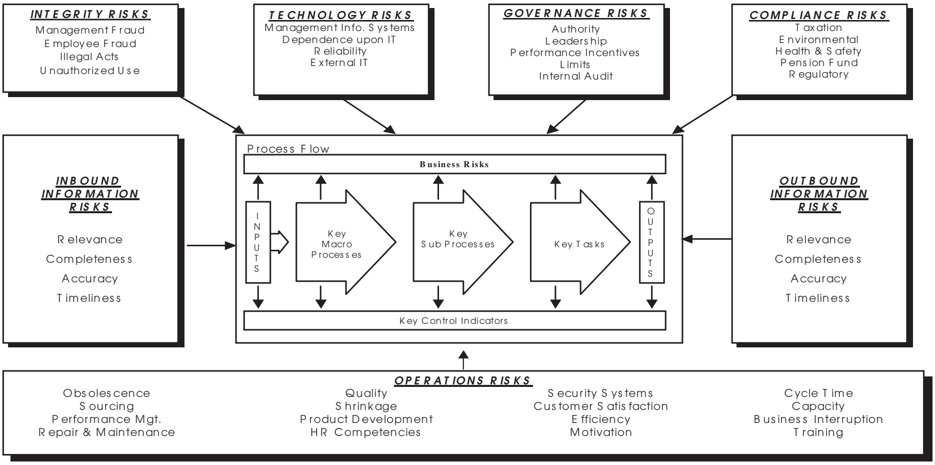

INTEGRITY RISKS
Management Fraud
Employee Fraud
Illegal Acts
Unauthorized Use
TECHNOLOGY RISKS
Management Info. Systems
Dependence upon IT
Reliability
External IT
GOVERNANCE RISKS
Authority
Leadership
Performance Incentives
Limits
Internal Audit
COMPLIANCE RISKS
Taxation
Environmental
Health & Safety
Pension Fund
Regulatory
INBOUND INFORMATION RISKS
Relevance
Completeness
Accuracy
Timeliness
Process Flow
Business Risks
INPUTS
Key Macro Processes
Key Sub Processes
Key Tasks
OUTPUTS
Key Control Indicators
OUTBOUND INFORMATION RISKS
Relevance
Completeness
Accuracy
Timeliness
OPERATIONS RISKS
Obsolescence
Sourcing
Performance Mgt.
Repair & Maintenance
Quality
Shrinkage
Product Development
HR Competencies
Security Systems
Customer Satisfaction
Efficiency
Motivation
Cycle Time
Capacity
Business Interruption
Training

Figure 12 Illustrative Process Risks for Core Processes. (From Risk Management Partners, Inc.)

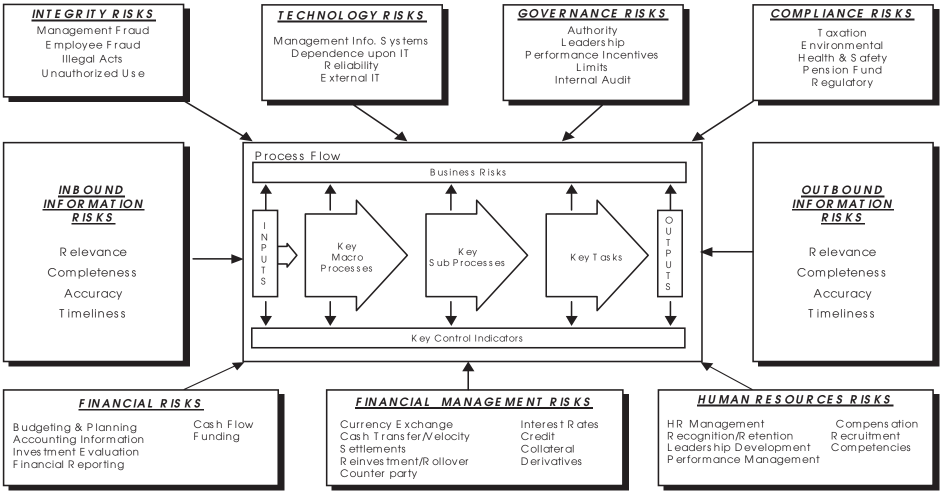

INTEGRITY RISKS
Management Fraud
Employee Fraud
Illegal Acts
Unauthorized Use
TECHNOLOGY RISKS
Management Info. Systems
Dependence upon IT
Reliability
External IT
GOVERNANCE RISKS
Authority
Leadership
Performance Incentives
Limits
Internal Audit
COMPLIANCE RISKS
Taxation
Environmental
Health & Safety
Pension Fund
Regulatory
INBOUND INFORMATION RISKS
Relevance
Completeness
Accuracy
Timeliness
Process Flow
Business Risks
INPUTS
Key Macro Processes
Key Sub Processes
Key Tasks
OUTPUTS
Key Control Indicators
OUTBOUND INFORMATION RISKS
Relevance
Completeness
Accuracy
Timeliness
FINANCIAL RISKS
Budgeting & Planning
Accounting Information
Investment Evaluation
Financial Reporting
Cash Flow
Funding
FINANCIAL MANAGEMENT RISKS
Currency Exchange
Cash Transfer/Velocity
Settlements
Reinvestment/Rollover
Counter party
Interest Rates
Credit
Collateral
Derivatives
HUMAN RESOURCES RISKS
HR Management
Recognition/Retention
Leadership Development
Performance Management
Compensation
Recruitment
Competencies

Figure 13 Illustrative Process Risks for Resource Management Processes. (From Risk Management Partners, Inc.)

and assure integrity of data must be recognized as only one aspect of the contemporary organization's control structure.

Other components of the control structure include diagnostic control systems, belief systems, boundary systems, and incentives. Diagnostic control systems, for example, recognize that empowerment requires a change in what is controlled. Consistently, empowered individuals are being asked to take risks and there must be commensurate rewards for the risk taking and achievement of superior performance. Such rewards, which can be either monetary or nonmonetary, are made on the basis of tangible performance consistent with the organization's mission.

The evolving organizational controls structure consists of strategic controls, management controls, and business process controls. A brief description of these elements follows:

- *Strategic controls* are designed to assess continuously the effect of changes in environment risks on the business, formulate business risk control strategies, and align the organization with those strategies.
- *Management controls* drive business risk assessment and control throughout the organization.
- *Process controls* are designed to assess continuously the risk that business processes do not achieve what they were designed to achieve. Embedded in process risk is information processing/technology risk, which arises when the information technologies used in the process are not operating as intended or are compromising the availability, security, integrity, relevance, and credibility of information produced.

Figure 15 provides examples of these types of controls.

8. ELEMENT 4: ALLIANCES AND RELATIONSHIPS

8.1. "Alliance" Defined

Several types of alliances exist, each with a specific purpose. The following are three of the more common types:

- *Transactional alliances* are established for a specific purpose, typically to improve each participant's ability to conduct its business. Cross-licensing in the pharmaceutical industry is an example. Open-ended purchase orders for specific products would be another.
- *Strategic sourcing* involves a longer-term commitment. It is a partnership between buyer and seller that can reduce the cost and friction between supplier and buyer by sharing product development plans, jointly programming production, sharing confidential information, or otherwise working together much more closely than do typical suppliers and customers. Wal-Mart and Home Depot are examples of companies with substantial capabilities in strategic sourcing.

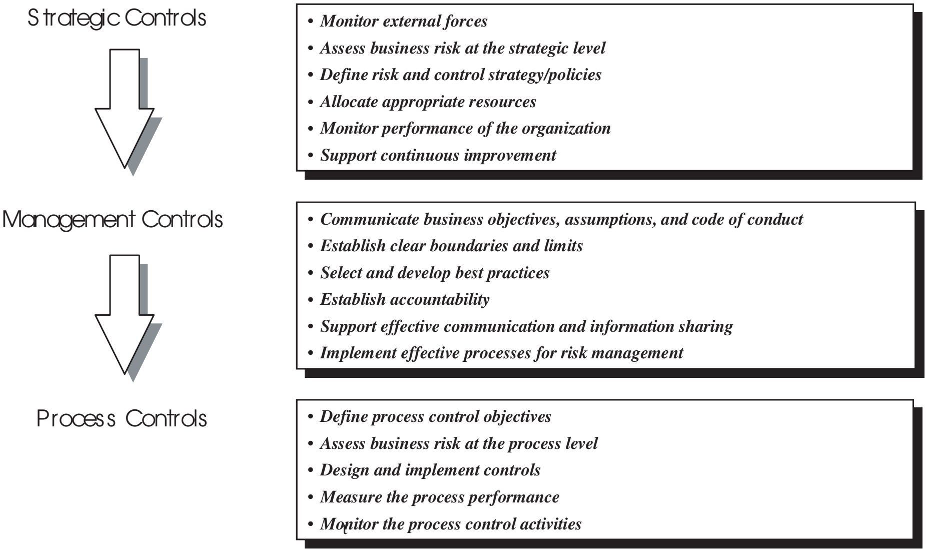
Old Paradigm

- Risk assessment occurs periodically
- Accounting, treasury, and internal audit responsible for identifying risks and managing controls
- Fragmentation—every function behaves independently
- Control is focused on financial risk avoidance
- Business risk controls policies, if established, generally do not have the full support of upper management or are inadequately communicated throughout the company
- Inspect and detect business risk, then react at the source
- Ineffective people are the primary source of business risk

New Paradigm

- Risk assessment is a continuous process
- Business risk identification and control management are the responsibility of all members of the organization
- Connections—Business risk assessment and control are focused and coordinated with senior-level oversight
- Control is focused on the avoidance of unacceptable business risk, followed closely by management of other unavoidable business risks to reduce them to an acceptable level
- A formal business risk controls policy is approved by management and board and communicated throughout the company
- Anticipate and prevent business risk, and monitor business risk controls continuously
- Ineffective processes are the primary source of business risk

Figure 14 The Old and New Business Risk Control Paradigms. (From Risk Management Partners, Inc.)



Strategic Controls

- *Monitor external forces*
- *Assess business risk at the strategic level*
- *Define risk and control strategy/policies*
- *Allocate appropriate resources*
- *Monitor performance of the organization*
- *Support continuous improvement*

Management Controls

- *Communicate business objectives, assumptions, and code of conduct*
- *Establish clear boundaries and limits*
- *Select and develop best practices*
- *Establish accountability*
- *Support effective communication and information sharing*
- *Implement effective processes for risk management*

Process Controls

- *Define process control objectives*
- *Assess business risk at the process level*
- *Design and implement controls*
- *Measure the process performance*
- *Monitor the process control activities*

Figure 15 Illustrative Examples of Controls. (From Risk Management Partners, Inc.)

- *Strategic alliances* involve two enterprises pulling together to share resources, funding, or even equity in a new enterprise on a long-term basis. For example, Motorola, Apple, and IBM pooled research programs and made financial commitments to develop a new-generation engine for personal computers—the Power PC.

8.2. Strategic Alliances in the Value/Supply Chain

The move toward strategic alliances is very strong because of increased global competition and industry convergence. For example, in banking, insurance, mutual funds, financial planning, and credit cards, industry boundaries are beginning to blur.

Strategic alliances form when one enterprise alone can’t fill the gap in serving the needs of the marketplace. Financial pressures and time constraints have squeezed managers without resources to fill the gaps through internal development. Also, acquisitions have proved expensive and have not always brought the needed capabilities.

An increasing number of global enterprises recognize that strategic alliances can provide growth at a fraction of the cost of going it alone. In addition to sharing risks and investment, a well-structured, well-managed approach to alliance formation can support other goals, such as quality and productivity improvement. Alliances provide a way for organizations to leverage resources. In the future, many organizations will be nothing more than ‘‘boxes of contracts,’’ with substantially all of the traditional value/supply chain components outsourced to business partners in the form of alliances or other strategic relationships.

In addition to the more obvious reasons to focus on a company’s ‘‘core competencies,’’ the new economics of information will dramatically dismantle traditional business structures and processes. New business structures will reform based on the separate economics of information and physical products. Over the next few years, many relationships throughout the business world will change. The dismantling and reformulation of traditional business structures will include value chains, supply chains, and business models. This will result from the separation of the economics of information from the economics of physical products. In addition, the explosion of networks and content standards will allow informational value chains to separate from the physical chain. This will create enormous opportunities to use information innovatively and create new knowledge for competitive advantage. As value chains deconstruct to create new market opportunities, more and more alliances and new relationships will be formed to maximize opportunities and fill the gaps that will exist in traditional business processes.

8.3. Performance Management

While the promise of alliances is very bright, managing a myriad of different and sometimes very complex relationships will be a significant challenge for most enterprises. Improper guidelines, poor communications between the enterprise and its partners (suppliers, customers, manufacturers, outsourcers, etc.), and unrealistic expectations can lay the groundwork for failure.

The significant shift from hierarchies to alliance networks will require a high level of competency in IT and relationship management. The enterprise and its partners must develop relationships at multiple levels throughout the organization to gain the trust and understanding required for long-term success. That includes making senior managers of the network well acquainted with one another.

Although multiple relationships are beneficial, consistent communication must take place between the partners at various levels within the respective organizations. Partners can receive confusing and conflicting direction from multiple sources within the enterprise. All parties must be connected to a database that tracks commitments and instructions made for their respective staffs.

To achieve success in these complex relationships, enterprises and their partners should speak a common business language. All parties will need to coordinate standards (such as software and network protocols) and business process elements (such as procurement, production, logistics, and customer service). In addition, all parties should participate in joint budgeting in order to understand the key cost drivers inherent in the network's infrastructure.

It is critical for the parties in the network to have access to detailed cost and performance information associated with each partner's respective services. Alliances typically are described by agreements that represent long-term relationships in which unforeseen opportunities, technologies, and business conditions will need to be addressed. The agreements should be flexible in order to accommodate changes in technologies and technology costs. Without careful planning and management, the network can get out of control, producing disastrous results.

The needs of all parties should be addressed in the contractual agreements. These contracts may be inexpensive to enter but extremely expensive to exit. Contracts should define intermediate consequences—generally penalties—for poor behavior and performance, with termination of the contract only as a last resort. Additional payments should be provided to reward superior performance that provides a measurable benefit.

Measurement becomes a major issue in managing alliance relationships. If performance cannot be measured effectively on a real-time basis, failure will likely occur. Service-level agreements must be defined in detail and allow for the addition of service levels that reflect changing business requirements and service levels. Partners should agree to measure and benchmark their cost structures against others on a regular basis for performance goal-setting and ongoing performance evaluation.

Effective day-to-day management of a complex network of relationships begins with planning. Management on both sides of an alliance must be committed to the communication and flexibility required for the relationship to shape and reshape itself as needs and opportunities arise. Well-designed management structures and processes enhance the probability of success for all parties.

9. ELEMENT 5: CORE PRODUCTS AND SERVICES

9.1. "Core Product and Service" Defined

The general goal of a product is to fill customers' specific needs. Core products are the product offerings that are most closely associated with the enterprise's brand image and generate the largest share of product revenue. The enterprise's core products and services are generally the basis for its marketplace strategy and value proposition.

9.2. Categories of Products and Services

A product family is a set of products based on a common platform. A platform includes design components that can be shared by products in the family. These shared design components allow for variations in product functions and features to meet different customer needs. As product strategies change and new technologies are introduced, new platforms are created. The automobile industry, for example, uses product platforms extensively to reduce costs and leverage technologies.

Derivative products are the products that are designed around a platform product. For example, Sony designed over 200 Walkman products, based on three platforms, to meet the specific needs of global markets. The economic proposition for derivative products is to reduce cost by reducing features (low-cost products), or to add features without changing the price significantly (line extensions), or to add features to increase the value to the customer (superior products).

9.3. Measuring Product Performance

Measurements of product performance should exist along four dimensions:

- *Market performance:* How does the marketplace view the company's products?
- *Process performance:* How well does the product development process work?
- *Resource performance:* How well do cross-functional teams perform?
- *Financial performance:* How profitable are the products/services?

Key Performance Indicator	Short-Term	Long-Term
Market Performance		
Customer Acceptance	X	X
Customer Satisfaction	X	X
Market Share		X
Unit Sales Growth		X
Brand Image		X
Process Performance		
Time-to-market	X	X
Quality Standards	X	X
Unique Benefits	X	X
Technology Enablers	X	X
Resource Performance		
Cross-functional Teaming	X	X
Performance Against Goals	X	X
Financial Performance		
Margin Goals		X
Profitability Goals		X
Return on Investment		X
New Product Sales/total Sales	X	X

Figure 16 Core Products/Services Measurement Framework. (From Risk Management Partners, Inc.)

The measurement framework in Figure 16 provides an example of the product dimensions that should be continuously monitored and improved.

10. ELEMENT 6: CUSTOMERS

10.1. "Customer" Defined

Customers are the reason that organizations exist—they are the most valuable assets. They are consumers or other businesses that utilize the enterprise's products and services. Large customers and/or groups of customers can exert significant influence over an organization. The business model is a framework for analyzing the contributions of individual activities in a business to the overall level of customer value that an enterprise produces, and ultimately to its financial performance.

10.2. Categories of Customers

An organization's customer base is made up of customers with many different attributes. They may be categorized along many different dimensions, depending on the purpose for which the segmentation is performed. Possible segmentation criteria include size, market, profitability, geographic location, customer preferences, influence or bargaining power, and intellectual capital. Segmentation gets the enterprise closer to the customer and allows the enterprise to understand customer needs in a very deep way. This closeness gives the enterprise access to information that is critical to strategy formulation and implementation. In addition, the enterprise and/or engineer can utilize customer segmentation techniques for various improvement initiatives.

10.3. Product and Services and Customer Linkages

Products and services and customers are inextricably linked. An enterprise's value proposition is expressed in the value—the products and services—it delivers to its customers. In the New Economy, these linkages will become more formalized as organizations innovate and produce new products *with* their customers. Customer capital will grow when the enterprise and its customers learn from each other. Collaborative innovation will be in everyone's best interest.

10.4. Relationship of Customers to Markets

Markets are made up of customers that have some common interests. Markets can be divided into finer and finer segments (customer groups), with each segment having its own issues. Market segmentation allows an organization to pursue the acquisition of those customers that are most attractive to its value proposition. Segmenting markets also provides the basis for tailoring products more specifically to the customers' needs.

11. APPLYING THE BUSINESS MODEL

The engineer can apply the enterprise business model to his or her work in a number of ways, including:

- Communication of the nature of the business
- Strategic analysis
- Business process analysis
- Business performance measurement
- Risk assessment

In this section, we will discuss each of the above applications.

11.1. Communicating the Nature of the Business

Throughout the design and development of the business model, the engineer should review findings and conclusions with management. The review could take the form of overhead slides or other visual aids, discussions, or a written document. During a review with the enterprise's management, the engineer should confirm that his or her understanding of the business is accurate and complete and provide management with potential new perspectives that may assist in organizational improvement. Once adopted by management, the model can be used as the basis for communicating the nature and structure of the business to employees and other interested stakeholders.

11.2. Improving the Business

As discussed in prior sections, the business model should be designed and developed with an objective of improving the business. The improvement model shown earlier in Figure 2 includes five key business principles that will guide the engineer as he or she achieves that stated objective.

11.2.1. Strategic Analysis

The business model focuses the engineer's attention on whether managers have designed effective strategies for reshaping patterns of behavior. The strategic analysis is intended to provide the engineer with a deep understanding of the broad environment in which the enterprise operates, and it focuses on the organization's strategic orientation and potential for reorientation. Included therein are both the industry and global environs of the organization. Also included is the engineer's understanding of the enterprise's strategy for achieving a sustainable competitive advantage within the industry context. The business risks that threaten achievement of this strategy are consistently identified, along with the enterprise's responses to such risks.

As part of the strategic analysis, the engineer will obtain or update an understanding of the organization's history, management's business strategy and objectives, the business risks faced by the organization, management's planned responses to such business risks, and the business processes that management has implemented. The strategic analysis is also focused on the articulation between the business strategy and the supporting business processes, as well as the articulation between the identified business risks and management's responses or controls.

During strategic analysis, the engineer may first obtain general industry information, including that which is available from trade associations, periodicals, and the like. Then he or she will consider obtaining information about the structure of the industry, including its segmentation, the dynamics among the various organizations that comprise the industry, the critical business issues facing entities in the industry, and significant industry risks.

At the conclusion of the strategic analysis, the engineer will have learned the "directional course" the management has set in response to the environment, taking into consideration:

- The relationship between the broad economic environment and the industry segment(s) in which the enterprise competes
- The enterprise's position and role within its respective industry segment(s)
- Threats to maintaining or improving the current position
- The needs and wants of the enterprise's chosen market segment(s)
- The total productive capacity of the enterprise and its competitors for each niche

- Management's vision of how to satisfy the market needs better than its rivals
- Management's specific strategies and plans for achieving that vision

Also, the engineer will have obtained an understanding of how and to what extent management steers the business and attains a fit between its strategy and the range of environmental forces acting on it. This will have been done through review of:

- The enterprise's strategic management process
- The formalized strategic plan
- The enterprise's approach to ''environmental scanning'' to monitor emerging or changing external threats
- Management's methods for communicating strategies throughout the organization, as well as the clarity of such communications
- The methods and measures used to monitor entity-level performance in terms of the strategic goals

The strategic analysis will provide the engineer with in-depth knowledge of the enterprise's value proposition and insight into opportunities to improve business performance and mitigate the risks that threaten achievement of the established objectives.

11.2.2. Business Process Analysis

Business process analysis is designed to provide the engineer with an in-depth understanding of the key business processes identified earlier during strategic analysis. Through this analysis, the engineer learns how the organization creates value. Specifically, each core business process is studied in depth to discern significant process objectives, the business risks related to these objectives, the controls established to mitigate the risks, and the financial implications of the risks and controls. Likewise, each significant resource management process is examined with the same foci.

Business process analysis adopts a ''value chain'' approach to analyzing the interconnected activities in the business, both domestically and globally. It is consistent with W. Edward Deming's views of business processes and the role of total quality management in monitoring the value of these processes. Core business processes represent the main customer-facing activities of the business. It is the successful combination and execution of the core business processes that creates value in the eyes of customers and therefore results in profitable customer sales. During business process analysis, the engineer recognizes the cross-functional nature of activities in the enterprise's business, that not all activities within and across processes are sequential, and that important linkages exist between processes.

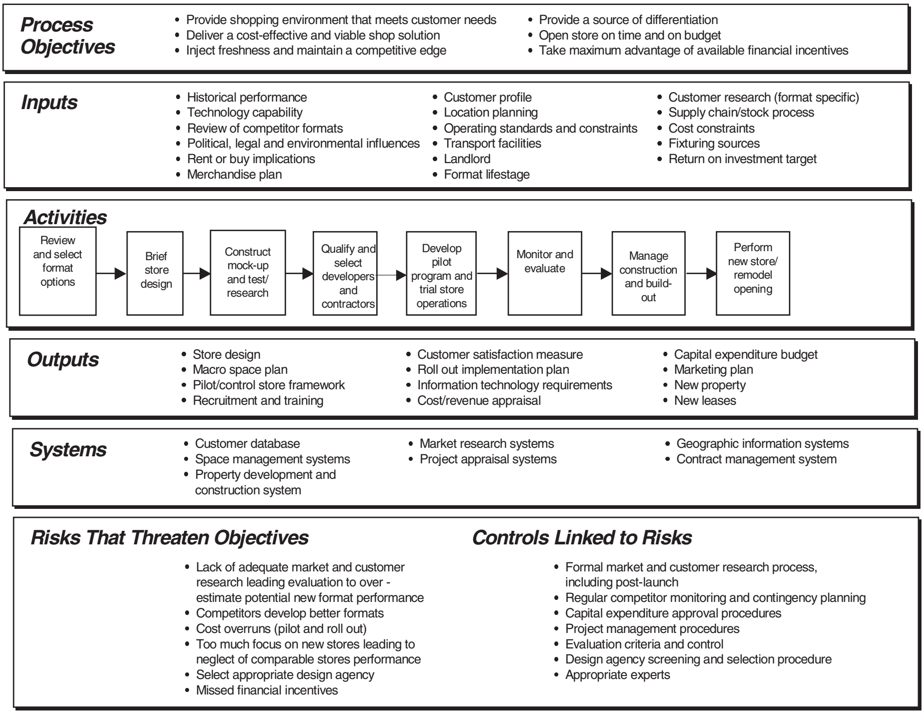
Figure 5, above, provides the context for a process analysis example. Specifically, it depicts the four core business processes of a hypothetical retail company: brand and image delivery, product/service delivery, customer service delivery, and customer sales. Consider the brand and image delivery core business process, which might include the following subprocesses: format development and site selection, brand management, advertising and promotion, visual merchandising, and proprietary credit. Figure 17 presents an example of a completed process analysis template for the format development and site selection subprocess. Such a process analysis template can be used by the engineer to analyze his or her enterprise's core business processes and significant resource management processes. The template is a framework that guides the engineer's collection and integration of information about business processes, using eight components: process objectives, inputs, activities, outputs, systems, risks that threaten objectives, and management controls linked to risks. Refer to Section 7.2 for descriptions of each of these components.

In the retail company, the engineer would address each of the following objectives, which are typical for this process:

1. Provide an environment in which the customer's needs can be met.
2. Deliver a cost-effective and viable shop solution.
3. Inject freshness and maintain a competitive edge.
4. Use the store as a vehicle for differentiation.
5. Open the store on time and stay on budget.
6. Take maximum advantage of available financial incentives.

From a value-chain perspective, and focusing first on process inputs, among the key considerations are historical performance, technology capability, competitor formats, customer profile, and cost constraints.

Continuing the value-chain perspective, the engineer will gather information about process activities such as:



Process Objectives

- Provide shopping environment that meets customer needs
- Deliver a cost-effective and viable shop solution
- Inject freshness and maintain a competitive edge
- Provide a source of differentiation
- Open store on time and on budget
- Take maximum advantage of available financial incentives

Inputs

- Historical performance
- Technology capability
- Review of competitor formats
- Political, legal and environmental influences
- Rent or buy implications
- Merchandise plan
- Customer profile
- Location planning
- Operating standards and constraints
- Transport facilities
- Landlord
- Format lifestage
- Customer research (format specific)
- Supply chain/stock process
- Cost constraints
- Fixturing sources
- Return on investment target

Activities

Review and select format options
Brief store design
Construct mock-up and test/ research
Qualify and select developers and contractors
Develop pilot program and trial store operations
Monitor and evaluate
Manage construction and build-out
Perform new store/ remodel opening

Outputs

- Store design
- Macro space plan
- Pilot/control store framework
- Recruitment and training
- Customer satisfaction measure
- Roll out implementation plan
- Information technology requirements
- Cost/revenue appraisal
- Capital expenditure budget
- Marketing plan
- New property
- New leases

Systems

- Customer database
- Space management systems
- Property development and construction system
- Market research systems
- Project appraisal systems
- Geographic information systems
- Contract management system

Risks That Threaten Objectives

- Lack of adequate market and customer research leading evaluation to over - estimate potential new format performance
- Competitors develop better formats
- Cost overruns (pilot and roll out)
- Too much focus on new stores leading to neglect of comparable stores performance
- Select appropriate design agency
- Missed financial incentives

Controls Linked to Risks

- Formal market and customer research process, including post-launch
- Regular competitor monitoring and contingency planning
- Capital expenditure approval procedures
- Project management procedures
- Evaluation criteria and control
- Design agency screening and selection procedure
- Appropriate experts

Figure 17 Format Development and Site Selection Subprocess—Retail Company. (Adapted from Bell et al. 1997)

1. Review and selection of format options
2. Store design
3. Store mock-up and research
4. Developer and contract selection

And the engineer will consider outputs like the following:

1. Store design
2. The macrospace plan
3. Recruitment and training
4. Customer satisfaction measures
5. The roll-out implementation plan

The engineer also will recognize that various systems are germane to the format development and site selection subprocess. These systems include the customer database, space management, property development and construction, market research, project appraisal, and contract management systems.

The engineer next considers the risks that threaten achievement of the process objectives and the controls that have been implemented to mitigate such risks. Continuing with the focus on the format development and site selection subprocess, such risks may include the possibility that competitors will develop better store formats, or an overemphasis on new stores relative to existing stores. Controls that could mitigate such risks are regular monitoring of competitors, in concert with contingency planning and usage of appropriate evaluation criteria.

A similar approach is taken by the engineer for the significant resource management processes, which were identified for the retail company in Figure 5 to be financial/treasury management, information management, human resource management, property management, and regulatory management. Figure 18 presents an example of a completed process analysis template for a retail company's human resource management process.

As shown in Figure 18, the following are among the process objectives of relevance: attract and retain a skilled and motivated workforce; control employee costs while maintaining morale and productivity; comply with regulatory/tax filing requirements; and adhere to the organization's code of conduct. Maintaining a value-chain perspective, the engineer next considers inputs to this process, including the organization's strategic plan, its operating plan, employee regulations, tax regulations, union contracts, industry statistics and market data, and training goals. Activities are then considered, such as developing and maintaining human resource policies and procedures; establishing and maintaining compensation and benefit policies and programs; identifying resource requirements; recruitment and hiring; training and development; performance reviews; compensation and benefit administration; monitoring of union contracts and grievances; and monitoring compliance with regulations.

The engineer then will consider outputs, such as regulatory filings, personnel files, tax filings, and performance reviews. Of course, various systems will be recognized as keys to successful human resource management, such as those related to compensation and benefits, tax compliance, and regulatory compliance.

Subsequently, the engineer considers risks related to the human resource management function, including high levels of staff turnover, noncompliance with regulations, and noncompetitive compensation packages. In turn, the engineer considers the controls that can mitigate the risks, such as implementing growth and opportunity plans for employees; regulatory monitoring; and benchmarking salary costs against industry and other norms.

At the conclusion of business process analysis, the engineer will have updated his/her understanding of (a) how the enterprise creates value, (b) whether the enterprise has effectively aligned the business process activities with the business strategy, (c) what the significant process risks are that threaten the achievement of the enterprise's business objectives, and (d) how effective the processes are at controlling the significant strategic and process risks. This detailed and updated knowledge about the business provides a basis for the engineer's development of recommendations about improvement opportunities and risk management.

11.2.3. Business Performance Measurement

Information-age enterprises succeed by investing in and managing their intellectual assets as well as integrating functional specialization into customer-based business processes. As organizations acquire these new capabilities, their measure of success should not depend solely on a traditional, historical

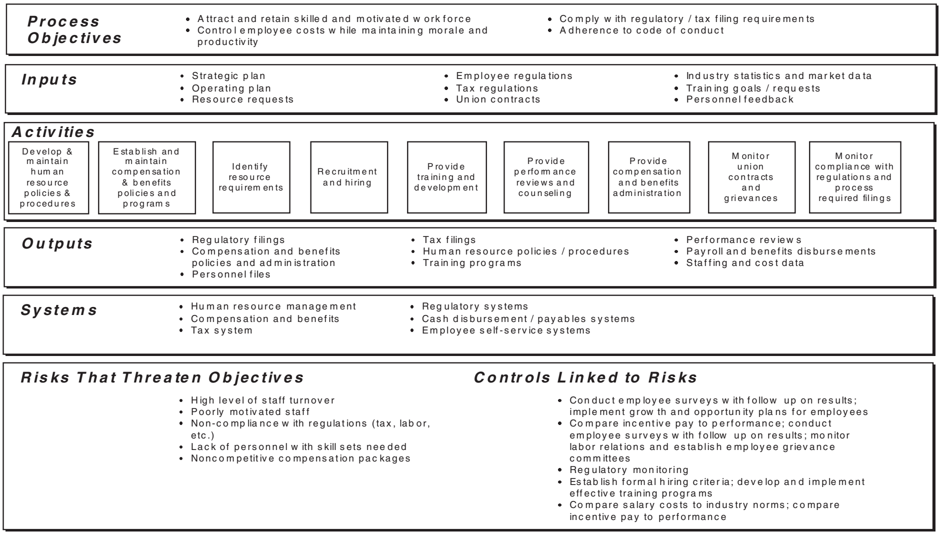

Process Objectives
• Attract and retain skilled and motivated workforce
• Control employee costs while maintaining morale and productivity
• Comply with regulatory / tax filing requirements
• Adherence to code of conduct
Inputs
• Strategic plan
• Operating plan
• Resource requests
• Employee regulations
• Tax regulations
• Union contracts
• Industry statistics and market data
• Training goals / requests
• Personnel feedback
Activities
Develop & maintain human resource policies & procedures
Establish and maintain compensation & benefits policies and programs
Identify resource requirements
Recruitment and hiring
Provide training and development
Provide performance reviews and counseling
Provide compensation and benefits administration
Monitor union contracts and grievances
Monitor compliance with regulations and process required filings
Outputs
• Regulatory filings
• Compensation and benefits policies and administration
• Personnel files
• Tax filings
• Human resource policies / procedures
• Training programs
• Performance reviews
• Payroll and benefits disbursements
• Staffing and cost data
Systems
• Human resource management
• Compensation and benefits
• Tax system
• Regulatory systems
• Cash disbursement / payables systems
• Employee self-service systems
Risks That Threaten Objectives
• High level of staff turnover
• Poorly motivated staff
• Non-compliance with regulations (tax, labor, etc.)
• Lack of personnel with skill sets needed
• Noncompetitive compensation packages
Controls Linked to Risks
• Conduct employee surveys with follow up on results; implement growth and opportunity plans for employees
• Compare incentive pay to performance; conduct employee surveys with follow up on results; monitor labor relations and establish employee grievance committees
• Regulatory monitoring
• Establish formal hiring criteria; develop and implement effective training programs
• Compare salary costs to industry norms; compare incentive pay to performance

Figure 18 Example Human Resource Management Process—Retail Company. (Adapted from Bell et al. 1997)

financial accounting model. Rather, measurement and management in the information age requires enterprises to become much more competent at identifying and monitoring measures that drive their business performance.

To develop a credible business model, the engineer must gain an understanding of how the enterprise measures and monitors business performance. Financial performance measures are valuable for summarizing the readily measurable economic consequences of actions already taken; however, outcome measures without performance drivers do not communicate how the outcomes have been achieved.

Performance drivers are key indicators of an enterprise's future financial performance. Understanding the cause-and-effect relationships between resource, process, market, and financial performances is essential to understanding the enterprise's strengths and weaknesses.

Figure 19 shows the cause and effect relationships between financial and nonfinancial performance.

Figure 20 provides some illustrative examples of financial and nonfinancial measures for a large management consulting firm, using the measurement framework shown in Figure 19. The performance measures are designed to provide feedback regarding implementation of strategic initiatives.

The strategic analysis provides the engineer with a basis to judge the effectiveness of the enterprise's performance management system. The business measurement approach should include the perspectives mentioned in Figure 19. During the development of the business model, the engineer will be in a unique position to evaluate the cause-and-effect relationships of the major elements of the performance management system. The engineer will review measures of resource performance, process performance, market performance, and financial performance; he or she will determine the business processes and variables that appear to have the greatest impact on the organization. In addition, the engineer will analyze interrelated key performance measures, both financial and nonfinancial, over time and relative to similar organizations. These measurements and assessments are combined with the engineer's knowledge about the business opportunities/risks that are documented in the business model. The updated business model, as well as the mental or more formal simulations performed by the engineer to better understand the organization's strategic-systems dynamics, provide a knowledge-base for development of expectations about the entity's achieved level of overall performance.

During business measurement, the engineer also evaluates the performance of the entity taken as a whole and its key business processes, using key performance indicators (KPIs) and the collective knowledge contained in the business model. KPIs are quantitative measurements, both financial and nonfinancial, collected by an entity or by the engineer, either continuously or periodically, and used by management and the engineer to evaluate performance in terms of the entity's defined business

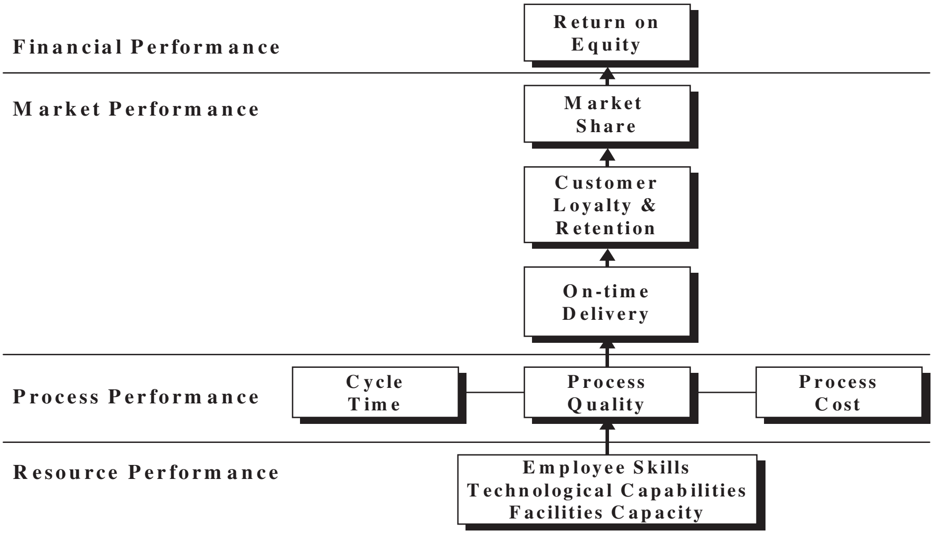


Figure 19 Financial and Nonfinancial Performance Relationships. (From Risk Management Partners, Inc.)

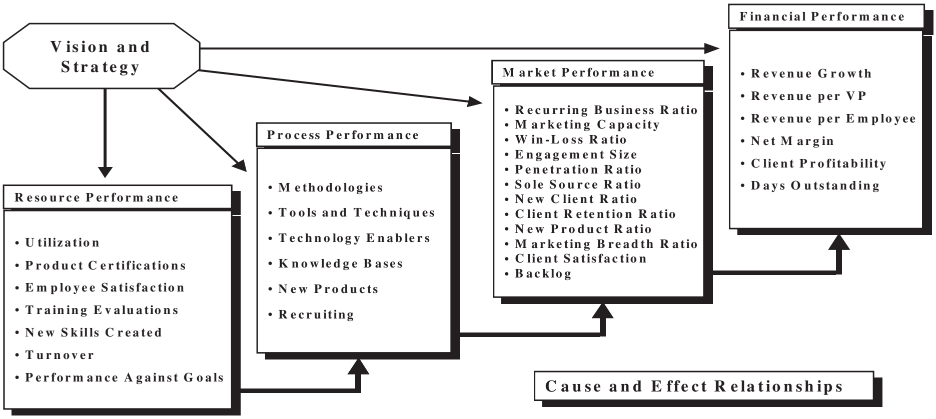

Vision and Strategy
Financial Performance
• Revenue Growth
• Revenue per VP
• Revenue per Employee
• Net Margin
• Client Profitability
• Days Outstanding
Market Performance
• Recurring Business Ratio
• Marketing Capacity
• Win-Loss Ratio
• Engagement Size
• Penetration Ratio
• Sole Source Ratio
• New Client Ratio
• Client Retention Ratio
• New Product Ratio
• Marketing Breadth Ratio
• Client Satisfaction
• Backlog
Process Performance
• Methodologies
• Tools and Techniques
• Technology Enablers
• Knowledge Bases
• New Products
• Recruiting
Resource Performance
• Utilization
• Product Certifications
• Employee Satisfaction
• Training Evaluations
• New Skills Created
• Turnover
• Performance Against Goals
Cause and Effect Relationships

Figure 20 Sample Financial and Nonfinancial Measures—Consulting Firm. (From Risk Management Partners, Inc.)

objectives. KPIs at the process level typically focus on three dimensions of process performance: cycle time, process quality, and process cost. More specifically, management might monitor and control process performance using one or more of the following types of KPIs:

- Waste, rework, and other indicators of process inefficiency
- Backlog of work in process
- Customer response time
- Number of times work is recycled between subprocesses and departments
- Number of document errors
- Customer satisfaction ratings
- Number of routing errors
- Value-adding processing time
- Information processing errors

An integrated performance management system, with the appropriate KPIs, can provide evidence to the engineer that the organization is maintaining the level of process quality required to sustain product demand.

11.2.4. Risk Assessment

Risk assessment is a continuous process performed throughout the design and development of the business model. During strategic analysis and business process analysis, the engineer reviews the processes and procedures that the enterprise has established to identify and manage strategic and process risks.

During the engineer's review of the enterprise's risk management activities, he or she develops an understanding of management's perceptions of business risk, both strategic risks and business process risks, and considers the reasonableness of the assumptions that underlie management's assessments of the potential impacts of these risks. These underlying assumptions may be viewed as a combination of assumptions about the probability of occurrence and assumptions about the magnitude of impact. Also, the engineer uses other information obtained during the strategic and business process analyses to make judgments about coverage (i.e., whether management has considered all significant business risks). And he or she uses this information to make judgments about the extent to which strategic and process risks remain uncontrolled (i.e., to determine the level of residual risk).

Next, the engineer further integrates information about residual business risks by grouping risks based on the particular business model elements to which they relate. He or she will also consider possible interactions among these groups of risks and develop expectations about how they might be manifested in the performance of the business. This integrated knowledge, together with the appropriate business measurements, provides the engineer with a basis for performing a diagnosis of the business. Furthermore, it guides tactical planning about the type and extent of additional information he or she should obtain in order to make recommendations for improving risk management activities.

By this point, the engineer will have developed a business risk profile of the organization. In the business risk profile, residual business risks are classified as either strategic or process risks. Also, interactions among risks are identified, and the risk classifications and identified interactions are cross-matched with related business performance attributes.

11.3. Continuous Improvement

At the conclusion of the business model design and development effort, the engineer will have constructed a fully integrated business model containing all of the information he or she has collected and integrated through the application of the five business principles shown earlier in Figure 2: strategic analysis, business process analysis, risk assessment, business measurement, and continuous improvement. The engineer will use the completed model as the basis for final review of the recommendations for improving business performance. But it must be remembered that the business model is a living document that must be updated and maintained on a continuous basis to reflect changing market conditions, new or improved value propositions, changes in organization structures, and the like. Continuous improvement applies just as much to the business model as it does to the business itself.

Acknowledgement

The enterprise business modeling concept described in this chapter was developed by KPMG LLP as an integral and fundamental part of its proprietary audit approach called the Business Measurement Process (BMP). The authors, both KPMG LLP retired partners, were involved in the development of the BMP: Frank Marrs led the entire development effort as National Managing Partner, Assurance Services; Barry Mundt was involved in the design of the business modeling concept and facilitated the development of generic enterprise business models for six industries. A large part of this chapter is based on a research monograph published by KPMG LLP, entitled *Auditing Organizations Through a Strategic-Systems Lens.*

ADDITIONAL READING

Bell, T., Marrs, F., Solomon, I., and Thomas, H., *Auditing Organizations Through a Strategic-Systems Lens: The KPMG Business Measurement Process,* KPMG, New York, 1997.

Boyett, J. H. and Boyett, J. T., *Beyond Workplace 2000: Essential Strategies for the New American Corporation,* Penguin, Harmondsworth, Middlesex, England, 1995.

Davenport, T. H., *Process Innovation: Reengineering Work Through Information Technology,* Harvard Business School Press, Boston, 1993.

Davis, S. M., *Future Perfect,* Addison-Wesley, Reading, MA, 1987.

Deschamps, J., and Nayak, P. R., *Product Juggernauts: How Companies Mobilize to Generate a Stream of Market Winners,* Harvard Business School Press, Boston, 1995.

Evans, P., and Wurster, T. S., *Blown to Bits: How the New Economics of Information Transforms Strategy,* Harvard Business School Press, Boston, 2000.

Fahey, L., and Randall, R. M., *The Portable MBA in Strategy,* John Wiley & Sons, New York, 1994.

Fisher, K., and Fisher, M. D., *The Distributed Mind: Achieving High Performance Through the Collective Intelligence of Knowledge Work Teams,* AMACOM, New York, 1998.

Gleick, J., *Faster: The Acceleration of Just About Everything,* Pantheon, New York, 1999.

Grantham, C., *The Future of Work: The Promise of the New Digital Work Society,* McGraw-Hill, New York, 2000.

Harrington, H. J., *Business Process Improvement: The Breakthrough Strategy for Total Quality, Productivity, and Competitiveness,* McGraw-Hill, New York, 1991.

Hesselbein, F., Goldsmith, M., and Beckhard, R., Eds., *The Organization of the Future,* Josey-Bass, San Francisco, 1997.

Kelly, K., *New Rules for the New Economy:* 10 *Radical Strategies for a Connected World,* Penguin, Harmondsworth, Middlesex, England, 1998.

Lynch, R. P., *Business Alliances Guide: The Hidden Competitive Weapon,* John Wiley & Sons, New York, 1993.

Magretta, J., Ed., *Managing in the New Economy,* Harvard Business School Press, Boston, 1999.

Marquardt, M., and Reynolds, A., *The Global Learning Organization: Gaining Competitive Advantage Through Continuous Learning,* Irwin Professional Publishing, New York, 1994.

Miller, A., and Dess, G. G., *Strategic Management,* McGraw-Hill, New York, 1996.

Negroponte, N., *Being Digital,* Alfred A. Knopf, New York, 1995.

Porter, M. E., *Competitive Advantage: Creating and Sustaining Superior Performance,* The Free Press, New York, 1985.

Rosenau, M. D., Jr., Ed., *The PDMA Handbook of New Product Development,* John Wiley & Sons, New York, 1996.

Shapiro, C., and Varian, H. R., *Information Rules: A Strategic Guide to the Network Economy,* Harvard Business School Press, Boston, 1999.

Sherman, H., and Schultz, R., *Open Boundaries: Creating Business Innovation Through Complexity,* Perseus Books, Reading, MA, 1998.

Stewart, T. A., *Intellectual Capital: The New Wealth of Organizations,* Doubleday, New York, 1997.

Thomas, R. J., *New Product Development: Managing and Forecasting for Strategic Success,* John Wiley & Sons, New York, 1993.

APPENDIX
List of Generic Business Processes and Subprocesses

Strategic Management Processes

1.0 Understand Markets and Customers:

- **1.1** Determine customer needs and wants.
- **1.2** Monitor changes in market and customer expectations.

2.0 Develop Vision and Strategy:

- **2.1** Monitor the external environment.
- **2.2** Define value proposition and organizational strategy.
- **2.3** Design organizational structure/processes/relationships.
- **2.4** Develop and set organizational goals.

3.0 Manage Improvement and Change:

- **3.1** Measure organizational performance.
- **3.2** Conduct quality assessments.
- **3.3** Benchmark performance.
- **3.4** Improve processes and systems.

Core Business Processes

4.0 Design Products and Services:

- **4.1** Develop new product/service concepts and plans.
- **4.2** Design, build, and evaluate prototype products/services.
- **4.3** Refine existing products/services.
- **4.4** Test effectiveness of new/revised products or services.
- **4.5** Prepare for production.

5.0 Market and Sell Products/Services:

- **5.1** Market products/services to relevant customers.
- **5.2** Process customer orders.

6.0 Produce and Deliver Goods:

- **6.1** Plan for and acquire necessary resources.
- **6.2** Convert resources or inputs into products.
- **6.3** Deliver products.
- **6.4** Manage production and delivery process.

7.0 Produce and Deliver Services:

- **7.1** Plan for and acquire necessary resources.
- **7.2** Develop human resource skills.

- **7.3** Deliver service to the customer.
- **7.4** Ensure quality of service.

8.0 Invoice and Service Customers:

- **8.1** Bill the customer.
- **8.2** Provide after-sales service.
- **8.3** Respond to customer inquiries.

Resource Management Processes

9.0 Develop and Manage Human Resources:

- **9.1** Create and manage human resource strategies.
- **9.2** Perform work level analysis and planning.
- **9.3** Manage deployment of personnel.
- **9.4** Develop and train employees.
- **9.5** Manage employee performance, reward, and recognition.
- **9.6** Ensure employee well-being and satisfaction.
- **9.7** Ensure employee involvement.
- **9.8** Manage labor/management relationships.
- **9.9** Develop human resource information systems.

10.0 Manage Information Resources:

- **10.1** Plan for information resource management.
- **10.2** Develop and deploy enterprise support systems.
- **10.3** Implement systems security and controls.
- **10.4** Manage information storage and retrieval.
- **10.5** Manage facilities and network operations.
- **10.6** Manage information resources.
- **10.7** Facilitate information sharing and communication.
- **10.8** Evaluate and audit information quality.

11.0 Manage Financial and Physical Resources:

- **11.1** Manage financial resources.
- **11.2** Process finance and accounting transactions.
- **11.3** Report information.
- **11.4** Conduct internal audits.
- **11.5** Manage the tax function.
- **11.6** Manage physical resources.

12.0 Execute Environmental Management Program:

- **12.1** Formulate environmental management strategy.
- **12.2** Ensure compliance with regulations.
- **12.3** Train and educate employees.
- **12.4** Implement pollution-prevention program.
- **12.5** Manage remediation efforts.
- **12.6** Implement emergency response program.
- **12.7** Manage government agency and public relations.
- **12.8** Manage acquisition/divestiture environmental issues.
- **12.9** Develop/manage environmental information systems.

13.0 Manage External Relationships:

- **13.1** Establish communication networks and requirements.
- **13.2** Communicate with stakeholders.
- **13.3** Manage government relationships.

13.4 Build relationships with network participants.
13.5 Develop public relations program.
13.6 Interface with board of directors.
13.7 Develop community relations.
13.8 Manage legal and ethical issues.

SECTION II

TECHNOLOGY

A. Information Technology
B. Manufacturing and Production Systems
C. Service Systems

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

II.A

Information Technology

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 3

Tools for Building Information Systems

ROBERT M. BARKER
University of Louisville

BRIAN L. DOS SANTOS
University of Louisville

CLYDE W. HOLSAPPLE
University of Kentucky

WILLIAM P. WAGNER
Villanova University

ANDREW L. WRIGHT
University of Louisville

1. INTRODUCTION 66
- 1.1. The Nature of Information Systems 66
- 1.2. Classes of Information Systems 67
  - 1.2.1. Local Information Systems 68
  - 1.2.2. Functional Information Systems 68
  - 1.2.3. Enterprise Information Systems 69
  - 1.2.4. Transorganizational Information Systems 69
- 1.3. Overview of Information System Development and Tools 70

2. PROGRAMMING LANGUAGES 70
- 2.1. Overview 70
- 2.2. Historical Review 70
- 2.3. C++ 72
- 2.4 Visual Basic 73
- 2.5. Web-based Programming 76
  - 2.5.1. HTML 76
  - 2.5.2. CGI 77
  - 2.5.3. Java 78
  - 2.5.4. ColdFusion 78
  - 2.5.5. ASP 79

3. DATABASE MANAGEMENT TOOLS 79
- 3.1 DBM Concepts 79
  - 3.1.1. Relational DBMS 80
  - 3.1.2. Processing 81
- 3.2. Object-Oriented Databases 82
- 3.3. Data Warehouses 83

4. ENTERPRISE TOOLS 83
- 4.1. Enterprise Tools Defined 85
- 4.2. Evolution of Enterprise Resource Planning 85
  - 4.2.1. The Appearance of Enterprise Resource Planning 86
  - 4.2.2. The Enterprise Tools Market 87
- 4.3. Basic Concepts of Enterprise Tools 88
  - 4.3.1. ERP and the “Process View” 88
  - 4.3.2. ERP and the Open Systems Architecture 88
  - 4.3.3. ERP Application and Data Integration 89
- 4.4. Standard Enterprise Tool Applications 89

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

4.4.1. Sales and Distribution Applications 90
4.4.2. Manufacturing and Procurement Applications 90
4.4.3. Accounting and Finance Applications 91
4.4.4. Human Resource Applications 91
4.5. Implementing an Enterprise System 92
4.5.1. Choosing an Enterprise Tool 92
4.5.2. Enterprise System Implementation Strategies 93
4.5.3. Critical Success Factors for ERP Implementations 93
4.6. Future of Enterprise Tools 94
4.6.1. Enterprise Tools and Supply Chain Management (SCM) 94
4.6.2. Enterprise Tools and Customer Relationship Management (CRM) 95
4.6.3. Enterprise Tools and Electronic Commerce (EC) 96
5. TOOLS FOR ANALYSIS AND DESIGN OF IS 96
5.1. The Systems Development Life Cycle 96
5.2. Systems Development Tools 97
5.2.1. Feasibility Analysis 98
5.2.2. Process Modeling Using Data Flow Diagrams 99
5.2.3. Structured English 100
5.2.4. ERD/Data Dictionaries 102
5.2.5. Gantt/PERT Diagram 103
5.2.6. JAD/RAD 104
5.2.7. CASE 105
6. CONCLUSION 105
REFERENCES 107

1. INTRODUCTION

Spanning the series of revolutionary changes in computer technology and hardware has been an ongoing evolution of software and methodology. This evolution has been motivated by the intent of harnessing raw computing power as a means for increasing human productivity and enhancing organizations' competitiveness. The result is a vast array of computer systems in today's organizations. These systems are being used on a daily basis to assist managers and operating personnel in all kinds of work environments, ranging from manufacturing to service tasks. Two of the most important types of these computer systems are *information systems* and *decision support systems*. In this chapter we will focus on prominent software tools that can be used in building an information system (IS). Decision support systems are dealt with in Chapter 4.

Progress in the development of information systems has been greatly affected by ongoing advances in hardware, tools, and methods. Many information systems in use today would have been economically or technically infeasible a mere decade ago. Continuous decreases in hardware costs per unit of storage and processing power have increased the ubiquity and size of ISs, making them essential components of even modest-sized organizations. Software tools that greatly improve a developer's leverage in implementing information systems have led to ISs of greater sophistication, having more features and better interfaces. For example, software that makes it easy to store, manipulate, and update data (i.e., database management systems) and that allows development of programs in high-level languages (i.e., fourth-generation languages) provides tools that greatly affect the costs of developing information systems and the capabilities that these systems can provide.

In addition to improvements in tools available for IS development, advances have been made in methodologies for developing information systems. In some instances, automated tools support methodology. A structured development method, commonly referred to as the systems life cycle, is routinely used by professional IS developers when building information systems. In addition, various structured development techniques have been devised to support different phases of the systems life cycle. For example, techniques such as data flow diagrams, structure charts, and decision tables have had positive impacts on systems development. One class of software tools, commonly referred to as computer-aided software engineering tools, facilitates the use of these methods and techniques. This chapter examines both the tools that are used in the actual building of information systems (e.g., database management software) and tools that support methodologies for system development.

1.1. The Nature of Information Systems

The core function of an information system is record keeping. Record keeping is able to represent and process information about some domain of interest such as a job shop, an inventory, suppliers of parts for a manufacturing operation, customer accounts, and so forth. Regardless of its application domain, an information system is able to store records describing one or more states of that domain (e.g., current state, historical states, and hypothetical states). The manner in which this descriptive knowledge is organized in a computer is a problem of information representation. The way in which information is represented strongly influences how it can be processed (e.g., updated, retrieved for presentation).

The principal value of an information system being able to represent and process large volumes of descriptive knowledge is that it can give engineers, managers, and operating personnel a way of recording, viewing, and monitoring states of an application domain. An information system provides its users with means for invoking any of a variety of standard reports. When an IS is designed, the developer must determine what reports will be allowed. Each type of report has a particular layout by which it presents information gleaned from the stored data. Thus, in implementing an IS, the developer must be concerned not only with record keeping issues, but with devising means for extracting information represented in the records and mapping it into corresponding report layouts.

Often the predefined reports are produced at predefined times. These may be periodic, as in the case of a weekly sales report or monthly expense report, or they may be event-triggered, as in the case of a stockout report or shipment report. In either case, the reports that an IS can produce may enable users to keep abreast of the current (or past) state of affairs pertaining to some aspect of an organization's operations. Such awareness is essential for a manager's ability to control those operations and can furnish a useful background for decision making. However, the ability of a traditional management information system to support decision making is quite limited because its reports are predefined, tend to be unavailable on the spur of the moment, and are based only on descriptive knowledge (Holsapple and Whinston, 1996).

Except in the most structured cases, the situation a decision maker faces can be very dynamic, with information needs arising unexpectedly and changing more rapidly than ISs can be economically revised to produce new kinds of reports. Even when the newly needed information exists in a set of predefined reports, it may be a needle in the haystack, distributed across several reports, or presented in a way not particularly helpful to the decision maker. It may be incomplete or unfocused or require some value-added processing. In contrast, the ideal of a decision support system (DSS) is to provide focused, complete, fully processed knowledge in the desired presentation format.

Periodic or event-triggered reports from an IS can be useful. However, in dynamic or unstructured decision situations, it is important for a decision maker to control the timing of reporting. With an IS, one must wait for the next periodic or triggered report. Beyond scheduled or event-triggered reporting, a DSS tends to allow desired knowledge to be easily and readily requested on an as-needed basis.

Keeping sufficiently up-to-date records of relevant descriptive knowledge is important as a prerequisite for IS report production. However, other kinds of knowledge enter into decision making, such as procedural knowledge and reasoning (Holsapple 1995). Procedural knowledge involves step-by-step specifications of how to do something. For instance, a solver is an algorithm that typically operates on existing descriptive knowledge to yield new descriptive knowledge in the guise of expectations, beliefs, facts, or model solutions. Reasoning knowledge is concerned with what conclusions are valid under some set of circumstances. For instance, a rule indicates what conclusion to draw when a premise is satisfied; a set of rules can be subjected to inference in order to produce diagnoses or recommendations. Traditional ISs have little concern with storing, updating, and manipulating such knowledge representations as solver libraries or rule sets. DSSs are not only concerned with them, but with how to integrate them as well.

Aside from being able to produce any standard report from a predefined portfolio of report types, an IS may be equipped with a query facility to provide greater flexibility in the kinds of reports that can be produced. After an IS has been implemented, it is not unusual for its users to want reports other than those in the predefined portfolio. The usefulness of these kinds of reports may have been overlooked in the original design of the system, or new reports may be needed because of changing conditions in a user's work environment. A query facility provides a nonprocedural way to state what should be in the desired report, allowing a user to specify what is desired without having to specify how to extract data from records in order to produce the report. An information system equipped with a query facility is a step in the direction of a decision support system. It gives the user a way to address some of the ad hoc knowledge needs that arise in the course of decision making.

1.2. Classes of Information Systems

Having characterized information systems and distinguished them from decision support systems, we can now look at classes of ISs. One approach to classification is based on functional application domains: ISs for finance applications, manufacturing applications, human resource applications, accounting applications, sales applications, and so forth. Another classification approach is to differentiate ISs in terms of the underlying technologies used to build them: file management, database management, data warehouses, Visual Basic, C++, COBOL, HTML, XML, artificial intelligence, and so on. However, it is not uncommon for multiple technologies to be employed in a single IS.

Here, we classify ISs in terms of their intended scopes. Scope can be thought of in two complementary ways: the scope of the records that are kept by the IS and the scope of IS usage. The IS classes form a progression from those of a local scope, to ISs with a functional scope to enterprise-wide ISs to ISs that are transorganizational in scope. Some tools for building information systems are universal in that they can be used for any of these classes. Examples include common programming languages and database management tools. Others are more specialized, being oriented toward a particular class. Representative examples of both universal and specialized tools are presented in this chapter.

1.2.1. Local Information Systems

Local ISs are in wide use today. This is due to such factors as the tremendous rise in computer literacy over the past two decades, the economical availability of increasingly powerful computing devices (desktop and handheld), and the appearance of more convenient, inexpensive development tools. Local ISs are used in both small and large organizations. They are also prominent outside the organizational setting, ranging from electronic calendars and address books to personal finance systems to genealogy systems.

A local IS does record keeping and reporting that relate to a specific task performed by an individual. For instance, a salesperson may have an IS for tracking sales leads. The IS would keep records of the salesperson's current, past, and prospective customers, including contact information, customer characteristics (e.g., size, needs, tastes), history of prior interactions, and schedules of planned interactions. The IS would allow such records to be updated as needed. Reports produced by this IS might include reports showing who to contact at a customer location and how to contact them, a list of all customers having a particular characteristic (e.g., having a need that will be addressed by a new product offering), or a list of prospects in a given locale and a schedule showing who to contact today.

Operation of a local IS tends to be under its user's personal control. The user typically operates the IS directly, assuming responsibility for creating and updating its records and requesting the production of reports. The behavior of a local IS, in terms of what records it can keep and what reports it can produce, may also be under the user's control. This is the case when the IS software is custom-built to meet the particular user's requirements. The user either personally builds the IS software or has it built by a professional IS developer according to his or her specifications. At the opposite extreme, the user may have little control over the behavior of the local IS's software. This is the case when the user obtains off-the-shelf, packaged software that has been designed by a vendor to suit most of the needs of a large class of IS users.

Thus, the user of a local IS has a choice between creating or obtaining custom-built IS software and acquiring ready-made software. Customized construction has the advantage of being tailor-made to suit the user's needs exactly, but it tends to be time-consuming and expensive. Ready-made software is quickly available and relatively inexpensive, but the behavior of an off-the-shelf software package may not match the IS user's needs or desires. Such mismatches range from crippling omissions that render a package unworkable for a particular user to situations where the behavior is adequate in light of cost and time savings.

A middle ground between custom-made and ready-made software is package software that can be configured to behave in a certain way. In principle, this is packaged software that yields a variety of local ISs. The user selects the particular IS behavior that most closely suits his or her needs and, through an interactive process, configures the packaged software to exhibit this behavior. This approach may be somewhat more expensive and time-consuming than strictly ready-made packages, but it also permits some tailoring without the degree of effort required in the custom-made approach.

1.2.2. Functional Information Systems

A functional IS is one that performs record keeping and reporting related to some function of an organization such as production, finance, accounting, marketing, sales, purchasing, logistics, personnel, or research. Such systems include those that give reports on production status, inventory levels, financial positions, accounts (e.g., payable, receivable), sales levels, orders, shipments, fringe benefits, and so forth. These ISs are much broader in scope than local ISs and may be thought of as being more for departmental use than individual use. A functional IS tends to be larger and more complex in terms of records it possesses and processes. Often there are many individuals who use a given functional IS.

Functional ISs are typically administered by IS professionals rather than individual users. It would be even more unusual for an individual user to build his or her own functional IS, as the size and complexity of these systems usually calls for formal system development methodologies. The options for building a functional IS are using customized development, purchasing ready-made software, and purchasing configurable software. Customized development can be performed by an organization's IS department or contracted out to an external developer. In either case, the developer works closely with prospective users of the system, as well as a functional sponsor (e.g., a department head), to ascertain exactly what the system needs to do. The customized development approach is the most flexible for building a system that closely fits users' needs and departmental requirements.

With ready-made software, building a functional IS becomes largely a matter of using the software to enter the system's initial records (and to update them over time). This software allows production of a predefined set of reports from the IS's records. The resultant functional IS can be constructed much more quickly than its custom-made counterpart, but the users will need to conform to the system rather than having a system specifically designed to conform to their needs.

Configurable software forms a middle ground, offering some options in the nature of records to be stored and processed for a given functional application, as well as various reporting options. After a developer sets up a configuration that sufficiently approximates the desired system's behavior, records are loaded into the system for subsequent updating and reporting.

As functional ISs proliferate within a department and across departments, issues of integration and consistency arise. For instance, information held in two functional ISs may be needed in a single report, which cannot be produced (even with an ad hoc query facility) from either IS on its own. Or two functional ISs may store some of the same kinds of records. If updates to these records in the two systems are not made simultaneously, the inconsistencies between their records can lead to inconsistencies between the reports produced by the two systems. One way to try to address integration and consistency issues is to devise additional systems that recognize interrelationships between functional ISs and serve to link them for update or reporting purposes. A second way is to build information systems at an enterprise scale, encompassing multiple functionality in a single IS.

1.2.3. Enterprise Information Systems

Enterprise ISs cross traditional functional boundaries. They keep records pertinent to multiple organizational functions, maintaining them in a consistent fashion and producing both functional and cross-functional reports. The magnitude and complexity of enterprise ISs far exceed those of functional ISs. The creation and operation of enterprise ISs are strictly in the domain of IS professionals. The earliest examples of these systems were custom-made via very large-scale projects. At the time, no off-the-shelf software solutions were available on an enterprise-wide scale. However, over the past decade several vendors have successfully offered configurable off-the-shelf software that functions as a tool for building enterprise ISs. These are known by such names as enterprise resource planning (ERP) and enterprise asset management (EAM) software. For simplicity, the former term is used here.

In some cases, ERP offerings are industry-specific. For instance, one might be oriented toward record keeping and reporting needs of the oil industry, another toward the automotive industry, and yet another toward retailers. ERP offerings also tend to be functionally modular. For instance, an organization may decide to start with implementations of accounting, finance, and production functions. Later, an integral human resources module may be added. In any case, building an enterprise IS with ERP tools is a matter of configuring the record storage mechanisms and configuring the behavior of ERP software to try to fit specific enterprises. This is by no means a trivial task and often causes a redesign of the organization to fit the IS.

The main attraction of an enterprise IS is the potential for consistency of information and integration of information usage. As an example, when a salesperson enters an order, the records describing it are immediately available to others in the enterprise. The factory receives a report of it and starts production. The logistics function receives reports on production progress, allowing it to schedule shipments. Inventory and procurement receive reports of production, leading to replenishment of raw materials and parts. Accounting receives reports on orders, production, and shipments; accounts records are updated accordingly. The enterprise's senior management receives reports allowing it to monitor sales activity, production progress, and inventory levels.

1.2.4. Transorganizational Information Systems

Transorganizational ISs are those that involve information flows to and from entities beyond an enterprise's boundaries. Information for updating records comes directly from customers, suppliers, partners, regulators, and others. Reports are issued directly to these same entities. All of this transorganizational activity is either linked to various internal ISs (local, functional, or enterprise) or designed as an extension of enterprise ISs. Especially notable examples are so-called supply chain management systems and customer relationship management systems. Configurable off-the-shelf software for building each of these is available, in some cases allowing them to be treated as additional ERP modules.

More broadly, transorganizational ISs form a major element of most organizations' electronic commerce initiatives. The Internet and the World Wide Web have been tremendous facilitators of transorganizational ISs, which take two main forms: those that involve accepting and reporting information from and to other businesses (B2B electronic commerce) and those that involve accepting and reporting information from and to consumers (B2C electronic commerce). As companions to transorganizational ISs, many examples of Web-oriented decision support systems exist (Holsapple et al. 2000).

1.3. Overview of Information System Development and Tools

Given an application domain and a class of potential users, we are confronted with the problem of how to create a useful information system. The act of creation spans such activities as analysis, design, and implementation. System analysis is concerned with determining what the potential users want the system to do. System design involves transforming analysis results into a plan for achieving those results. Implementation consists of carrying out the plan, transforming the design into a working information system.

IS implementation may involve software tools such as programming-language compilers and database management software. Or it may involve configuring off-the-shelf software packages. The activity of design can be strongly influenced by what tools are to be used for implementation. Both the analysis and design activities can themselves be supported by tools such as data flow diagrams, data dictionaries, HIPO (hierarchical input, process, output) charts, structure charts, and tools for computer-assisted software engineering (CASE).

In the early days of IS development, the principal software tool used by developers was a programming language with its attendant compiler or interpreter. This tool, together with text-editing software, was used to specify all aspects of an information system's behavior in terms of programs. When executed, these programs governed the overall flow of the information system's actions. They accomplished information storage and processing tasks. They also accomplished user interaction tasks, including the interpretation of user requests and production of reports for users. Section 2 provides an overview of programming, accompanied by highlights of two languages widely used for IS implementation: Visual Basic and C++.

Although programming is a valuable way for developers to specify the flow of control (i.e., what should happen and when) in an information system's behavior, its use for specifying the system's data representation and processing behaviors has steadily diminished in concert with the proliferation of database management tools. Today, database management systems are a cornerstone of information system development. Most popular among the various approaches to database management is the relational, which is described in Section 3. In addition to packaging these as separate tools from conventional (so-called third-generation) languages such as COBOL, C, and FORTRAN, various efforts have been made to integrate database management and programming facilities into single facility. The resultant tools are examples of fourth-generation languages.

Programming and database management tools can be used in implementing ISs in any of the four classes discussed in Section 2. There is great variation in off-the-shelf packages available for IS implementation both within and across the IS classes. Section 2.5 considers prominent Web-based tools used for implementing transorganizational information systems. Section 3 provides a brief overview of database management, which forms the foundation for most information systems today. Section 4 focuses on tools for the class of enterprise ISs. Finally, Section 5 considers ancillary tools that can be valuable in the activity of developing information systems. These are examined in the context of the system development life cycle of analysis, design, implementation, and maintenance. We will focus on tools often used by IS professionals during the analysis and design phases.

2. PROGRAMMING LANGUAGES

2.1. Overview

This section describes how programming languages may be used to build information systems. First, a brief historical review of programming languages helps explain how programming tools have become more powerful and easier to use over the years. We characterize today's modern programming languages, such as C++ and Visual Basic, by considering the advantages and disadvantages of each. Then we review some basic programming concepts by showing some typical examples from these languages. Finally, we consider how program development is affected when the World Wide Web is targeted as a platform. This includes a comparison of Internet programming tools, such as HTML, Java, and CGI scripting.

2.2. Historical Review

In the earliest days of program development, programmers worked directly with the computer's own machine language, using a sequence of binary digits. This process was tedious and error-prone, so programmers quickly started to develop tools to assist in making programming easier for the humans involved. The first innovation allowed programmers to use mnemonic codes instead of actual binary digits. These codes, such as *LOAD, ADD,* and *STORE,* correspond directly to operations in the computer's instruction set but are much easier to remember and use than the sequence of binary digits that the machines required. A programming tool, called an *assembler,* translates mnemonics into machine codes for the computer to execute. This extremely low-level approach to programming is hence referred to as *assembly language* programming. It forces the programmer to think in very small steps because the operations supported by most computers (called *instruction sets*) are very simple tasks, such as reading from a memory location or adding two numbers. A greater problem with this approach is that each computer architecture has its own assembly language, due to differences in each CPU's instruction set. This meant that early programmers had to recreate entire information systems from scratch each time they needed to move a program from one computer to another that had a different architecture.

The next major innovation in computer programming was the introduction of *high-level languages* in the late 1950s. These languages, such as COBOL and FORTRAN, allow a programmer to think in terms of larger, more complex steps than the computer's instructions set allows. In addition, these languages were designed to be portable from machine to machine without regard to the underlying computer architecture. A tool called a *compiler* translates the high-level statements into machine instructions for the computer to execute. The compiler's job is more complex than an assembler's simple task of one-to-one translation. As a result, the machine code that is generated is often less efficient than that produced by a well-trained assembly programmer. For this reason, programming in assembly language is still done today when execution time is absolutely critical. The cost of developing and maintaining high-level code is significantly reduced, however. This is especially important given another trend in computing, towards cheaper and more powerful hardware and more expensive and harder-to-find programmers.

As high-level languages became more prevalent, programmers changed to a new development paradigm, called *structured programming*. This approach, embodied in languages such as Pascal and C from the 1970s and 1980s, emphasizes functional decomposition, that is, breaking large programming tasks into smaller and more manageable blocks, called functions. Data used within a function block is local to that function and may not generally be seen or modified by other blocks of code. This style of programming is better suited to the development of large-scale information systems because different functions can be assigned to different members of a large development team. Each function can be thought of as a black box whose behavior can be described without revealing how the work is actually being done but rather in terms of input and output. This principle of *information hiding* is fundamental to the structured programming approach. Another advantage is *code reuse,* achieved by using the more general-purpose functions developed for one project over again in other systems.

As the information systems being developed became more complex, development with structured programming languages became increasingly difficult. A new paradigm was required to overcome the problem, and it arrived in the late 1980s in the form of *object-oriented programming* (OOP). This continues to play an important role today. In structured programming, functions are the dominant element and data are passed around from one function to another, with each function having a dependency on the structure of the data. If the way the data are represented changes, each function that manipulates the data must, in turn, be modified. In OOP, data are elevated to the same level of importance as the functions. The first principle of object-oriented programming is called *encapsulation*. It involves joining data together with the functions that manipulate that data into an inseparable unit, usually referred to as a *class*. A class is a blueprint for actual *objects* that exist in a program. For example, the class *Clock* would describe how all clocks (objects or instances of the class) in a program will behave. It does not create any clocks; it just describes what one would be like if you built one, much as an architect's blueprint describes what a building would look like if you built one.

The importance of encapsulation is seen when one considers the well-known Year 2000 problem encountered in many programs written with structured languages, such as COBOL. In many of these information systems, literally thousands of functions passed dates around as data, with only two digits reserved for storing the year. When the structure of the data had to be changed to use four digits for the year, each of these functions had to be changed as well. Of course, each function had to be identified as having a dependency on the date. In an OOP language, a single class would exist where the structure of the data representing a date is stored along with the only functions that may manipulate that data directly. Encapsulation, then, makes it easier to isolate the changes required when the structure of data must be modified. It strengthens information hiding by making it difficult to create a data dependency within a function outside a class.

OOP also makes code reuse easier than structured programming allows. In a structured environment, if you need to perform a task that is similar but not identical to an existing function, you must create a new function from scratch. You might copy the code from the original function and use it as a foundation, but the functions will generally be independent. This approach is error-prone and makes long-term program maintenance difficult. In OOP, one may use a principle called *inheritance* to simplify this process. With this approach, an existing class, called the *base class* or *superclass,* is used to create a new class, called the *derived class* or *subclass*. This new derived class is like its parent base class in all respects except what the programmer chooses to make different. Only the differences are programmed in the new class. Those aspects that remain the same need not be developed from scratch or even through copying and pasting the parent's code. This saves time and reduces errors.

The first OOP languages, such as SmallTalk and Eiffel, were rarely used in real world projects, however, and were relegated to university research settings. As is often the case with easing system development and maintenance, program execution speed suffered. With the advent of C++, an OOP hybrid language, however, performance improved dramatically and OOP took off. Soon thereafter, new development tools appeared that simplified development further. *Rapid Application Development* (RAD) tools speed development further by incorporating a more visual development environment for creating graphical user interface (GUI) programs. These tools rely on wizards and code generators to create frameworks based on a programmer's screen layout, which may be easily modified. Examples include Borland's Delphi (a visual OOP language based on Pascal), Borland's C++ Builder (a visual C++ language), and Microsoft's Visual Basic.

2.3. C++

The C++ language was originally developed by Bjarne Stroustrup (Stroustrup 1994) but is now controlled and standardized by the American National Standards Institute (ANSI). It is an extension (literally, an increment) of the C programming language. C is well known for the speed of its compiled executable code, and Stroustrup strove to make C++ into a similarly efficient object-oriented language (see Stroustrup 1994 for a description of the development of the language). Technically, C++ is a hybrid language in that it supports both structured (from C) and object-oriented development. In order to achieve the speed that earlier OOP languages could not, C++ makes some sacrifices to the purity of the OO model. The tradeoffs required, however, helped make OOP popular with real-world developers and sparked a revolution in modern program development.

Like C before it, C++ is a language with a rich abundance of operators and a large standard library of code for a programmer's use. In its current form, C++ includes the Standard Template Library, which offers most of the important data structures and algorithms required for program development, including stacks, queues, vectors, lists, sets, sorts, and searches (Stroustrup 1997). Programs written using C++'s standard components are easily ported from one platform to another. The language lacks, however, a standard library of graphics routines for creating a graphical user interface program under different operating systems. Instead, each compiler vendor tends to offer its own classes and functions for interacting with a specific operating system's windowing routines. For example, Microsoft offers the Microsoft Foundation Classes (MFC) for developing Windows applications with its compiler, while Borland offers its Object Windows Library (OWL) for the same purpose. This makes it difficult to port GUI programs from one vendor's compiler on one operating system to another vendor's compiler on another operating system.

A simple example (Main and Savitch 1997) is shown below that declares a basic *Clock* class. It is an abstraction of a real-world clock used for telling time.

```
class Clock {
 public:
  Clock();
  void set_time(int hour, int minute, bool morning);
  void advance(int minutes);
  int get_hour() const;
  int get_minute() const;
  bool is_morning() const;
 private:
  int hour_24, // Stores the current hour
  minute_24; // Stores the current minute
};
```

This class is typical of the structure of most classes in C++. It is divided into two sections, *public* and *private*. In keeping with information hiding, the data are normally kept private so that others outside of the class may not examine or modify the values stored. The private data are exclusively manipulated by the class functions, normally referred to as *methods* in OOP. The functions are declared in the public section of the class, so that others may use them. For example, one could ask a clock object its minute by using its *get_minute*() function. This might be written as in the following code fragment:

```
Clock c; // Creates an actual object, c
cout << ''The minute is '' << c.get_minute() ;<< endl;
```

Of course, this simple class provides a method for modifying the time stored. The public method *set_time*() is used for this purpose. By forcing the use of this method, the class designer can ensure that the data are manipulated in accordance with appropriate rules. For example, the *set_time*() method would not allow the storage of an illegal time value. This could not be guaranteed if the data were public and directly available for modification. The implementation of this method is shown below:

```
void Clock::set_time(int hour, int minute, bool morning) {
 assert((hour >= 1) && (hour <= 12));
 assert((minute >= 0) && (minute <= 59));
 minute_24 = minute;
 if ((morning)&&(hour == 12))
  hour_24 = 0;
 else if ((!morning) && (hour < 12))
  hour_24 = hour + 12;
 else
  hour_24 = hour;}
```

Note how the values passed to the method are tested via an *assert*() statement before they are used. Only if the assertion proves true are the values used; otherwise a runtime error message is produced.

To show an example of inheritance, consider extending the *Clock* class to create a new class, *CuckooClock*. The only difference is that this type of clock has a bird that chirps on the hour. In all other respects it is identical to a regular clock.

```
class CuckooClock : public Clock {
 public:
  bool is_cuckooing( ) const;
};

bool CuckooClock::is_cuckooing( ) const {
 return (get_minute( ) == 0);}
```

Note, we declare the new *CuckooClock* class in terms of the existing *Clock* class. The only methods we need to describe and implement are those that are new to or different from the base class. Here there is only one new function, called *is_cuckooing*(), which returns true on the hour. It is important to understand that *CuckooClock* is a *Clock*. In important respects, an instance of *CuckooClock* can do anything that a *Clock* object can do. In fact, the compiler will allow us to send a *CuckooClock* object anywhere a *Clock* object is expected. For example, we might have a function written to compare two *Clock* objects to see if one is "equal to" the other.

```
bool operator ==(const Clock& c1, const Clock& c2) {
 return ((c1.get_hour() == c2.get_hour()) &&
 (c1.get_minute() == c2.get_minute()) &&
 (c1.is_morning() == c2.is_morning())); }
```

We may confidently send a *CuckooClock* to this function even though it is written explicitly to expect *Clock* objects. Because of inheritance, we do not need to write a different version of the function for each class derived from *Clock*—*CuckooClock* is a *Clock.* This simplifies things considerably for the programmer. Another interesting note about this code segment is that C++ allows us to provide definitions for most of its built-in operators in the context of a class. This is referred to as *operator overloading,* and C++ is rare among programming languages in allowing this kind of access to operators. Here we define a meaning for the equality operator (==) for *Clock* objects.

In summary, C++ is a powerful and rich object-oriented programming language. Although widely recognized as difficult to work with, it offers efficient execution times to programmers that can master its ways. If you want portable code, you must stick to creating console applications. If this is not a concern, modern compilers assist in creating GUI applications through wizards, such as in Microsoft's Visual C++. This is still more difficult to accomplish using C++ than a RAD tool such as Visual Basic, which will be discussed next.

2.4. Visual Basic

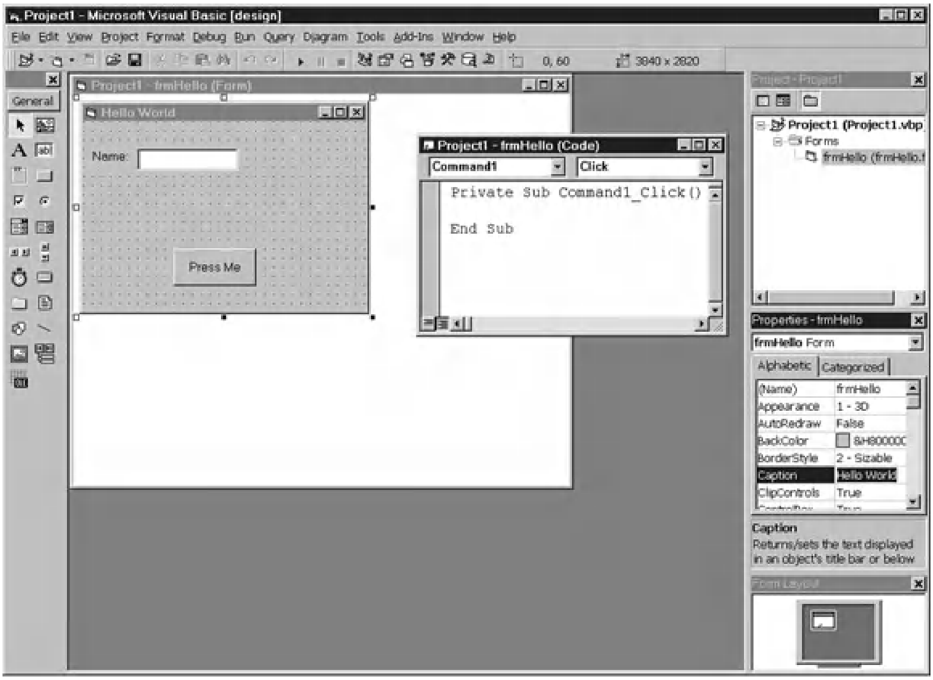
Visual Basic (VB) is a programming language and development tool from Microsoft designed primarily for rapidly creating graphical user interface applications for Microsoft's Windows operating system. First introduced in 1991, the language is an extension of BASIC, the Beginners' All-Purpose Symbolic Instruction Code (Eliason and Malarkey, 1999). BASIC has been in use since John Kemeny and Thomas Kurta introduced it in 1965 (Brookshear 1999). Over the years, Visual Basic has evolved more and more towards an object-oriented programming model. In its current version, 6.0, VB provides programmers with the ability to create their own classes using encapsulation, but it does not yet support inheritance. It simplifies the creation of Windows applications by making the enormously complex Windows Application Program Interface (consisting of over 800 functions) available through easy-to-use objects such as forms, labels, command buttons, and menus (Eliason and Malarkey 1999).

As its name suggests, Visual Basic is a highly *visual* tool. This implies not only that the development environment is GUI-based but also that the tool allows you to design a program's user interface by placing components directly onto windows and forms. This significantly reduces development time. Figure 1 provides a snapshot of the Visual Basic development environment.

Another feature that makes Visual Basic into a Rapid Application Development tool is that it supports both interpreted and compiled execution. When VB is used in *interpreted* mode, the tool allows the programmer to quickly see the effects of their code changes without a lengthy compilation directly to machine code. Of course, run-time performance is better when the code is actually compiled, and this can be done before distributing the application.

In Visual Basic, objects encapsulate properties, methods, and events. *Properties* are an object's data, such as a label's caption or a form's background color. *Methods* are an object's functions, as in C++ and other OOP languages. *Events* are typically user-initiated actions, such as clicking a form's button with the mouse or making a selection from a menu, that require a response from the object. Figure 1 shows some of the properties of the highlighted command button object, *Command1*. Here the button's text is set to "Press Me" via the *Caption* property. In the code window for the form, the *Click* event's code is displayed for this command button. This is where you would add the code to respond when the user clicks on the form's button. The tool only provides the outline for the code, as seen here. The programmer must add actual VB statements to perform the required action. Visual Basic does provide many wizards, however, to assist in creating different kinds of applications, including forms that are connected back to a database. These wizards can significantly reduce development time and improve reliability.

The fundamental objects that are important to understand in Visual Basic development are *forms* and *controls*. A form serves as a container for controls. It is the "canvas" upon which the programmer


Project1 - Microsoft Visual Basic [design]
File Edit View Project Format Debug Run Query Diagram Tools Add-Ins Window Help
0, 60
3840 x 2820
General
Project1 - frmHello (Form)
Hello World
Name:
Press Me
Project1 - frmHello (Code)
Command1
Click
Project - Project1
Project1 (Project1.vbp
Forms
frmHello (frmHello.f
Properties - frmHello
frmHello Form
Alphabetic
Categorized
Form Layout

```
Private Sub Command1_Click()

End Sub
```

(Name)	frmHello
Appearance	1 - 3D
AutoRedraw	False
BackColor	&H80000C
BorderStyle	2 - Sizable
Caption	Hello World
ClipControls	True

Caption
Returns/sets the text displayed in an object's title bar or below

Figure 1 The Visual Basic Development Environment.

“paints” the application’s user interface. This is accomplished by placing controls onto the form. *Controls* are primarily GUI components such as command buttons, labels, text boxes, and images. In Figure 1, the standard toolbox of controls is docked on the far left of the screen. These controls are what the user interacts with when using the application. Each control has appropriate properties and events associated with it that affect its appearance and behavior. For example, consider the form shown in Figure 2.

This form shows the ease with which you can create a database-driven application. There are three controls on this form. At the bottom of the form is a *data* control. Its properties allow the programmer to specify a database and table to connect to. In this example, it is attached to a simple customer database. At the top are two *textbox* controls, used for displaying and editing textual information. In this case, these controls are *bound* to the data control, indicating that their text will be coming from the associated database table. The data control allows the user to step through each record in the table by using the left and right arrows. Each time the user advances to another record from the table, the bound controls update the user’s text to display that customer’s name and address. In the code fragment below, you can see some of the important properties of these controls.

```
Begin VB.Data Data1
 Caption = ``Customer''
 Connect = ``Access''
 DatabaseName = ``C:\VB Example\custdata''
 DefaultCursorType= 0 `DefaultCursor
 DefaultType = 2 `UseODBC
 Exclusive = 0 `False
 ReadOnly = 0 `False
 RecordsetType = 1 `Dynaset
 RecordSource = ``Customer''
End

Begin VB.TextBox Text1
 DataField = ``customer name''
 DataSource = ``Data1''
End

Begin VB.TextBox Text2
 DataField = ``street address''
 DataSource = ``Data1''
End
```

In the data control object, *Data1,* there are several properties to consider. The *Connect* property specifies the type of database being attached, here an Access database. The *DatabaseName* property specifies the name and location of the database. *RecordSource* specifies the table or stored query that will be used by the data control. In this example, we are connecting to the *Customer* table of the example database. For the bound text controls, the programmer need only specify the name of the data control and the field to be used for the text to be displayed. The properties that need to be set for this are *DataSource* and *DataField,* respectively.

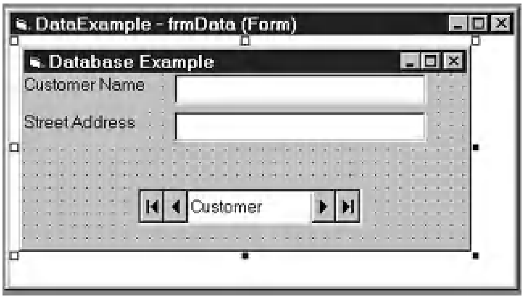


Figure 2 Database Example.

As these examples demonstrate, Visual Basic makes it relatively easy to create complex GUI applications that run exclusively under the Windows operating system. When these programs are compiled into machine code, the performance of these applications is quite acceptable, although not yet as fast as typical C++ programs. The object-oriented model makes development easier, especially since most of the properties can be set visually, with the tool itself writing the necessary code. Unfortunately, VB still does not support inheritance, which limits a programmer's ability to reuse code. Recent additions in the language, however, make it easier to target the Internet as an application platform. By using *Active X controls,* which run under Windows and within Microsoft's Internet Explorer (IE) web browser, and *VBScript* (a subset of VB that runs in IE), you can create applications that may be ported to the World Wide Web. Your choice of controls is somewhat limited and the user must have Microsoft's own web browser, but in a corporate Intranet environment (where the use of IE can be ensured) this might be feasible. For other situations a more flexible solution is required. The next section will explore some of the tools for achieving this.

2.5. Web-Based Programming

Developing applications where the user interface appears in a web browser is an important new skill for programmers. The tools that enable a programmer to accomplish this type of development include HTML, Java, CGI, Perl, ColdFusion, ASP, etc.—a veritable cornucopia of new acronyms. This section explains these terms and shows how these new tools may be used to create web pages that behave more like traditional software applications.

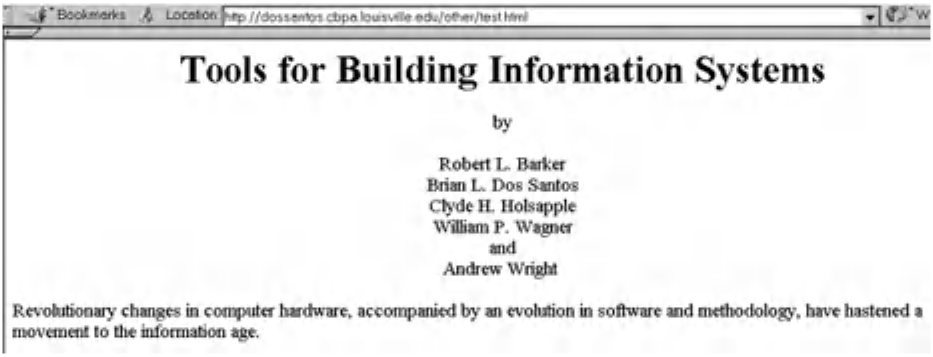
2.5.1. HTML

The HyperText Markup Language (HTML) is *the* current language of the World Wide Web. HTML is a markup language, not a programming language. Very simply, a document is "marked up" to define its appearance. This involves placing markup tags (or commands) around text and pictures, sound, and video in a document. The general syntax for HTML tags is:

```
<Tagname [options]>your text here</tagname>
```

These tags indicate how to display portions of the document. Opening tags, along with available options, are enclosed in < and > and closing tags include the / before the tagname. In some instances both opening and closing tags are required to indicate what parts of the documents the tags should affect. In other instances no closing tag is required. The contents of the file (including HTML tags and web page contents) that generates the simple web page shown in Figure 3 are as follows:

```
<html>
<head>
<title>Tools for Building Information Systems</title>
</head>
<body>
<center><h1>Tools for Building Information Systems</h1>
by<p>
Robert L. Barker<br>Brian L. Dos Santos<br>Clyde H.
```



http://dossantos.cbpa.louisville.edu/other/test.html

Tools for Building Information Systems

by

Robert L. Barker
Brian L. Dos Santos
Clyde H. Holsapple
William P. Wagner
and
Andrew Wright

Revolutionary changes in computer hardware, accompanied by an evolution in software and methodology, have hastened a movement to the information age.

Figure 3 Web Page and HTML Tags.

```
Holsapple<br>William P. Wagner<bR>and<br>Andrew Wright
</center><p>
Revolutionary changes in computer hardware, accompanied by
an evolution in software and methodology, have hastened a
movement to the information age.
</body>
</html>
```

HTML documents may then be processed and displayed using a browser that runs on any number of computer platforms, from UNIX workstations down to cellular telephones. These documents are simple text files that can be created with WYSIWYG tools (such as Microsoft's FrontPage or Netscape's Composer) or even a simple text editor such as the Windows Notepad. HTML relies on *tags,* keywords enclosed in angle brackets, like `<h1>`, `<b>`, `<li>`, etc., to specify what type of content or formatting is intended. HTML standards are controlled by an organization called the World Wide Web Consortium (W3C), and their website (http://www.w3.org/) has the most accurate and up-to-date information about HTML.

The latest version of HTML is 4.0 (http://www.w3.org/MarkUp/), and several of today's browsers support most of its features. In addition, the leading producers of web browsers, Microsoft and Netscape, frequently include support for proprietary features not specified in the W3C's version of HTML. These extensions can make it difficult to create extremely complicated documents that will work on both browsers. This is especially true for *dynamic* HTML documents, or documents whose contents change even after the page is loaded into a browser (Castro, 1999). HTML 4.0 includes support for Cascading Style Sheets (http://www.w3.org/Style/), which gives page designers easier control of large and complicated websites. HTML is essentially a system for formatting documents. It includes both structuring and formatting tags, which makes it difficult to maintain complex websites. The introduction of cascading style sheets allows a designer to separate structuring and formatting tags. Formatting tags are maintained in cascading style sheets so that HTML is used only for structuring documents. Style sheets are very powerful but not as easy to use as basic HTML. The future, however, lies in Extensible Markup Language (XML) (http://www.w3.org/XML/). XML is designed to allow groups to create their own markup languages, like HTML, specifically suited to their needs (Castro, 1999). It is especially important for its ability to put structured data into a text format. This might make it easy to view and edit a spreadsheet or database via the Web, for example. In fact, the W3C is rewriting HTML (as XHTML) and its Cascading Style Sheets (as XSL) using XML (see http://www.w3.org/TR/xhtml1/ and http://www.w3.org/Style/XSL/).

HTML, by itself, does not provide the capabilities required for electronic commerce. For e-commerce, it is necessary to develop applications that are interactive in nature, permitting visitors to dynamically obtain information they need (e.g., search a product catalog) or complete some task (e.g., submit an order). Such capabilities typically require that user-provided data be processed in real time and require interaction with databases. This type of processing may be accomplished via programming languages such as Perl, Visual Basic, and C++. Recently, several tools have emerged that reduce the need for traditional programming while essentially achieving the same results (e.g., ColdFusion and Tango). These capabilities are often achieved via the Common Gateway Interface (CGI).

2.5.2. CGI

To create truly powerful websites, a programmer must be able to access current data, typically from a corporate database. For example, commerce sites need to be able to tell a user whether a product is in stock and record how many items a customer would like to purchase. This cannot be done with static HTML files. Instead, processing is required on the server-side before the contents of a web page are delivered to a user's web browser. One widely used approach is the Common Gateway Interface (CGI). CGI specifies a common method for allowing Web pages to communicate with programs running on a Web server.

A typical CGI transaction begins when a user submits a form that he or she has filled out on a web page. The user may be searching for a book by a certain author at an online bookstore and have just entered the author's name on a search page. The HTML used for displaying the form also specifies where to send the data when the user clicks on the submit button. In this case, the data is sent to a CGI program for processing. Now the bookstore's CGI program searches its database of authors and creates a list of books. In order to return this information to the user, the CGI program must create a response in HTML. Once the HTML is generated, control may once again be returned to the web server to deliver the dynamically created web page. CGI programs may be written in nearly any programming language supported by the operating system but are often written in specialized scripting languages like Perl.

CGI provides a flexible, but inefficient mechanism for creating dynamic responses to Web queries. The major problem is that the web server must start a new program for each request, and that incurs significant overhead processing costs. Several more efficient alternatives are available, such as writing programs directly to the Web server's application program interface. These server APIs, such as Microsoft's ISAPI and Netscape's NSAPI, allow more efficient transactions by reducing overhead through the use of dynamic link libraries, integrated databases, and so on. Writing programs with the server APIs is not always an easy task, however. Other alternatives include using products like Allaire's ColdFusion or Microsoft's Active Server Pages (ASP).

2.5.3. Java

Java is a programming language from Sun Microsystems that is similar to C/C++ in syntax and may be used for general-purpose applications and, more significantly, for embedding programs in web pages (Hall 1998). Although Java is an extremely young language, it has gained widespread acceptance because of its powerful and easy-to-use features. It is not yet a mature product, however, and experiences some stability problems as the language continues to evolve.

Java has a lot in common with C++, so much so that it is often referred to as "C++ Lite" by experienced programmers. It is an object-oriented programming language that was built from the ground up, unlike the hybrid C++. It is designed to be a cross-platform development tool; in other words, the code is easy to port from one computer architecture and operating system to another. It does a better job of this than C++ does by including a standard set of classes for creating graphical user interfaces. This allows a GUI application designed on a Macintosh to be run under Windows or Linux, for example. Extremely complex interfaces may still require tweaking to achieve complete interchangeability of the interface, but new additions to the language (the Swing classes) make this less necessary. Java is widely considered a simpler language to use than C++ because it forgoes some of the more troublesome C++ (though powerful) features. For example, Java does not explicitly use pointers and has automatic memory management, areas where C++ programmers often introduce bugs into their programs. This makes Java much easier to work with and somewhat less error-prone. Java only supports *single inheritance,* which means that a derived class may only have a single parent base class. C++ allows the use of the more powerful form, *multiple inheritance,* which allows several parents for a derived class. Java does not support *templates,* which allow C++ programmers to create functions and classes that work for many different types of data without needing to be implemented separately for each. Java also does not support operator overloading. Like C++, however, Java does have a rich and powerful set of standard libraries. Java's libraries even include components for network programming and database access (through JDBC), which standard C++ lacks. Even with all these improvements to C++, Java would be a footnote in the annals of programming if it were not for its web features.

Java allows the creation of specialized applications called *applets,* which are designed to be run within a web browser. Most modern web browsers support Java in some form, and Sun offers plugins that provides the most up-to-date language features. Early applets focused on creating more dynamic web pages through animations and other graphical components easily created in Java but hard to reproduce in pure HTML. Today, applets are used for delivering entire software applications via the Web. This has many advantages over a traditional development and distribution model. For example, programmers do not have to worry about the end-user platform because applets run in a browser. Users don't have to worry about applying patches and bug fixes because they always get the latest version from the Web. Security is an obvious concern for applets because it might be possible to write a malicious Java applet and place it on a web page for unsuspecting users. Luckily, the Java community has created a robust and ever-improving security model to prevent these situations. For example, web browsers generally restrict applets from accessing memory locations outside their own programs and prevent writing to the user's hard drive. For trusted applets, such as those being run from a company's Intranet, the security restrictions can be relaxed by the end user to allow more powerful functionality to be included.

2.5.4. ColdFusion

ColdFusion is a tool for easily creating web pages that connect to a database. Developed by Allaire (http://www.allaire.com), ColdFusion is being used in thousands of Web applications by leading companies, including Reebok, DHL, Casio, and Siemens. The ColdFusion Markup Language (CFML) allows a user to create Web applications that are interactive and interface with databases. The syntax of CFML is similar to HTML, which is what makes it relatively easy to use; it makes ColdFusion appealing to Web designers who do not have a background in programming. ColdFusion encapsulates in a single tag what might take ten or a hundred lines of code in a CGI or ASP program. This allows for more rapid development of applications than competing methods provide. However, if a tag does not do exactly what you need it to, then you must either change your application or resort to a programmatic approach. Although ColdFusion is not a programming language per se, knowledge of fundamental programming concepts and SQL is essential for sophisticated applications. CFML allows a user to display output based on the results of a query to a database and update databases based on user input. In order to use CFML, a ColdFusion Application Server (CFAS) must be installed. CFAS works in conjunction with a web server. The CFAS can do much more than query/update databases; it interfaces with many different Internet services to provide the functionality that a website designer desires. Figure 4 shows how a ColdFusion Application Server interacts with a web server, databases, and other services it provides.

The following sequence of actions illustrates how ColdFusion functions to produce dynamic results.

1. A user submits a form via a browser. The form is designed to be processed by a file containing CFML code. This file must have a ‘‘cfm’’ extension.
2. The web server recognizes the cfm extension and hands the file to the CFAS for processing.
3. The CFAS executes the instructions per the CFML code. This typically results in information from a database to be included in the web page. In addition, it may interact with other Internet services, such as other web servers, e-mail, etc.
4. The web page generated by the CFAS is then returned to the web server.
5. The web server sends this page to the browser.

Allaire offers versions that run under Microsoft Windows NT and the Unix-based operating systems Solaris and HP-UX. ColdFusion programs interface with many different Web server APIs to provide greater efficiency than traditional CGI programs.

2.5.5. ASP

One of Microsoft’s methods for creating more dynamic web pages is through its web server’s support of Active Server Pages. ASP lets programmers use the VBScript programming language on the server-side to create HTML content on-the-fly. Because ASP is built into the Web server, it is more efficient than other CGI programs. It offers greater flexibility and control than a product like ColdFusion but requires greater programming knowledge to achieve results.

3. DATABASE MANAGEMENT TOOLS

3.1. DBM Concepts

One of the most important functions that an information system must perform is data management (i.e., record keeping). Prior to the advent of true database software, computer data management was characterized by excessive data redundancy (or duplication), large data dependence (or coupling),

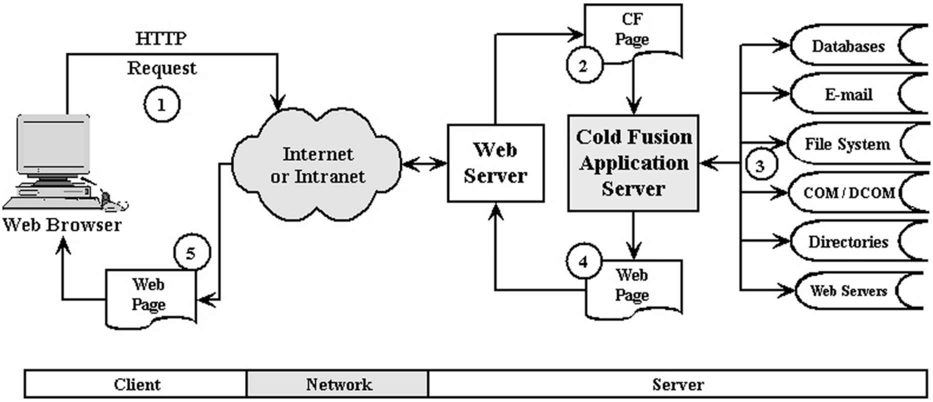


Figure 4 ColdFusion Application Server.

and diffused data ownership (McLeod 1998). In early data processing systems, the data files were often created with "little thought as to how those files affected other systems. Perhaps much, or even all, of the new data in a new file was already contained in an existing file" (McLeod 1998). This resulted in significant data redundancy and several associated problems. Obviously, data duplication results in wasted space, but worse than this are synchronization problems. "Multiple copies of data were often updated on different frequencies. One file would be updated daily, another weekly, and a third monthly. Reports based on one file could conflict with reports based on another, with the user being unaware of the differences" (McLeod 1998). These early systems also exhibited a high degree of dependence between the data specifications and the programs written to manipulate the data. Every time the representation of the data or the file storage format changed, all of the programs that depended on that data would also need to be changed. Other problems resulted from a lack of agreement on data ownership and standardization. "This situation of the 1950s and early 1960s may have been the reason why many of the first management information systems failed in the late 1960s. Information specialists appeared unable to deliver systems that provided consistent, accurate, and reliable information" (McLeod 1998).

Database systems sought to address these issues by separating the *logical* organization of the data from the *physical* organization. The logical organization is how the *user* of the system might view the data. The physical organization, however, is how the *computer* views the data. A database, then, may be defined as "an integrated collection of computer data, organized and stored in a manner that facilitates easy retrieval" by the user (McLeod 1998). It seeks to minimize data redundancy and data dependence. A database is often considered to consist of a hierarchy of data: files, records, and fields. A *field* is the smallest organized element of data, consisting of a group of characters that has a specific meaning. For example, a person's last name and telephone number would be common fields. Related fields are collected and stored in a structure referred to as a *record*. A specific person's record might consist of his or her name, address, telephone number, and so on (all values associated with that individual). Related records are grouped into a *file,* such as a file of employees or customers. Keep in mind that the physical representation of the database will be shielded from the user and should not limit the user's ability retrieve information from the system.

A database management system (DBMS) is a "collection of programs that manage the database structure and controls access to the data stored in the database. The DBMS makes it possible to share the data in the database among multiple applications or users" (Rob and Coronel 1997). In a sense, it sits between the user and the data, translating the users' requests into the necessary program code needed to accomplish the desired tasks. General Electric's Integrated Data Store (IDS) system, introduced in 1964, is considered by most to be the first generalized DBMS (Elmasri and Navathe 2000).

Database systems are often characterized by their method of data organization. Several popular models have been proposed over the years. Conceptual models focus on the logical nature of the data organization, while implementation models focus on how the data are represented in the database itself. Popular conceptual models include the entity-relationship (E-R) model and the object-oriented model discussed below. These models typically describe relationships or associations among data as one-to-one, one-to-many, or many-to-many. For example, in most organizations an employee is associated with one department only but a department may have several employees. This is a one-to-many relationship between department and employee (*one* department has *many* employees). But the relationship between employee and skills would be many-to-many, as a single employee might possess several skills and a single skill might be possessed by several employees (*many* employees have *many* skills).

Implementation models include the hierarchical model, network model, and relational model. Each model has advantages and disadvantages that made it popular in its day. The relational model is easily understood and has gained widespread acceptance and use. It is currently the dominant model in use by today's database management systems.

3.1.1. Relational DBMS

The relational database model, introduced by E. F. Cod in 1970, is considered a major advance for users and database designers alike. At the time, however, many considered the model impractical because its "conceptual simplicity was bought at the expense of computer overhead" (Rob and Coronel 1997). As processing power has increased, the cost of implementing the relational model has dropped rapidly. Now you will find relational DBMS implementations from mainframes down to microcomputers in products like DB2, Oracle, and even Microsoft's Access.

To the user, the relational database consists of a collection of *tables* (or *relations* in the formal language of the model). Each table may be thought of as a matrix of values divided into rows and columns. Each row represents a record and is referred to as a *tuple* in the model. Each column is a field and is called an *attribute* in the model. Below are some simple tables for Employee and Department.

Employee	EmployeeID	EmpName	EmpPhone	DepartmentID
	1234	Jane Smith	5025551234	11
	2111	Robert Adams	8125551212	20
	2525	Karen Johnson	6065553333	11
	3222	Alan Shreve	5025558521	32
	3536	Mary Roberts	8125554131	20

Department	DepartmentID	DeptName
	11	Marketing
	20	Engineering
	32	Accounting

Pay special attention to how relationships between tables are established by sharing a common field value. In this example, there is a one-to-many relationship between the Employee and Department tables established by the common *DepartmentID* field. Here, Robert Adams and Mary Roberts are in the Engineering department (*DepartmentID* = 20). The tables represented here are independent of one another yet easily connected in this manner. It is also important to recognize that the table structure is only a logical structure. How the relational DBMS chooses to represent the data physically is of little concern to the user or database designer. They are shielded from these details by the DBMS.

3.1.2. Processing

One of the main reasons for the relational model's wide acceptance and use is its powerful and flexible ad hoc query capability. Most relational DBMS products on the market today use the same query language, Structured Query Language (SQL). SQL is considered a fourth-generation language (4GL) that "allows the user to specify *what* must be done *without* specifying *how* it must be done" (Rob and Coronel 1997). The relational DBMS then translates the SQL request into whatever program actions are needed to fulfill the request. This means far less programming for the user than with other database models or earlier file systems.

SQL statements exist for defining the structure of the database as well as manipulating its data. You may create and drop tables from the database, although many people use tools provided by the DBMS software itself for these tasks. SQL may also be used to insert, update, and delete data from existing tables. The most common task performed by users, however, is data retrieval. For this, the *SELECT* statement is used. The typical syntax for *SELECT* is as follows:

```
SELECT <field(s)>
FROM <table(s)>
WHERE <conditions>;
```

For example, to produce a list of employee names and telephone numbers from the Engineering department using the tables introduced earlier, you would write:

```
SELECT EmpName, EmpPhone
FROM Employee
WHERE DepartmentID = 20;
```

Of course, you may use the typical complement of logical operators, such as *AND, OR,* and *NOT,* to control the information returned by the query.

The *SELECT* statement may also be used to join data from different tables based on their common field values. For example, in order to produce a list of employee names and their departments, you would write:

```
SELECT Employee.EmpName, Department.DeptName
FROM Employee, Department
WHERE Employee.DepartmentID = Department.DepartmentID;
```

In this example, the rows returned by the query are established by matching the common values of *DepartmentID* stored in both the *Employee* and *Department* tables. These joins may be performed across several tables at a time, creating meaningful reports without requiring excessive data duplication.

It is also possible to sort the results returned by a query by adding the *ORDER BY* clause to the *SELECT* statement. For example, to create an alphabetical listing of department names and IDs, you would write:

```
SELECT DeptName, DepartmentID
FROM Department
ORDER BY DeptName;
```

These examples demonstrate the ease with which SQL may be used to perform meaningful ad hoc queries of a relational database. Of course, SQL supports many more operators and statements than those shown here, but these simple examples give a hint of what is possible with the language.

3.2. Object-Oriented Databases

The evolution of database models (from hierarchical to network to relational) is driven by the need to represent and manipulate increasingly complex real-world data (Rob and Coronel 1997). Today's systems must deal with more complex applications that interact with multimedia data. The relational approach now faces a challenge from new systems based on an object-oriented data model (OODM). Just as OO concepts have influenced programming languages, so too are they gaining in popularity with database researchers and vendors.

There is not yet a uniformly accepted definition of what an OODM should consist of. Different systems support different aspects of object orientation. This section will explore some of the similar characteristics shared by most of these new systems. Rob and Coronel (1997) suggest that, at the very least, an OODM

1. Must support the representation of *complex objects*
2. Must be *extensible,* or capable of defining new data types and operations
3. Must support *encapsulation* (as described in Section 2)
4. Must support *inheritance* (as described in Section 2)
5. Must support the concept of *object identity,* which is similar to the notion of a primary key from the relational model that uniquely identifies an object. This object identity is used to relate objects and thus does not need to use *JOIN*s for this purpose.

Consider an example involving data about a *Person* (from Rob and Coronel 1997). In a typical relational system, *Person* might become a table with fields for storing name, address, date of birth, and so on as strings. In an OODM, *Person* would be a class, describing how all instances of the class (*Person* objects) will be represented. Here, additional classes might be used to represent name, address, date of birth, and so on. The *Address* class might store city, state, street, and so on as separate attributes. *Age* might be an attribute of a *Person* object but whose value is controlled by a method, or function of the class. This inclusion of methods makes objects an *active* component in the system. In traditional systems, data are considered *passive* components waiting to be acted upon. In an OODM, we can use inheritance to create specialized classes based on the common characteristics of an existing class, such as the *Person* class. For example, we could create a new class called *Employee* by declaring *Person* as its superclass. An *Employee* object might have additional attributes of salary and social security number. Because of inheritance, an *Employee* is a *Person* and, hence, would also gain the attributes that all *Person* objects share—address, date of birth, name, and so on. *Employee* could be specialized even further to create categories of employees, such as manager, secretary, cashier, and the like, each with its own additional data and methods for manipulating that data.

Rob and Coronel (1997) summarize the advantages and disadvantages of object-oriented database systems well. On the positive side, they state that OO systems allow the inclusion of more semantic information than a traditional database, providing a more natural representation of real-world objects. OO systems are better at representing complex data, such as required in multimedia systems, making them especially popular in CAD/CAM applications. They can provide dramatic performance improvements over relational systems under certain circumstances. Faster application development time can be achieved through the use of inheritance and reuse of classes. These OO systems are not without problems, however. Most of them are based on immature technology, which can lead to product instability. This is exacerbated by the lack of standards and an agreed-upon theoretical foundation. OO database systems especially lack a standard ad hoc query language (like SQL for relational systems). In addition, the initial learning curve is steep, particularly for individuals who are well accustomed to the relational model. On the whole, it seems clear that object-oriented databases are worth consideration. As standards develop over time and people become more familiar with object-oriented systems in general, OO databases could prove to be the successor to the relational model.

3.3. Data Warehouses

One of the major changes in the information systems area within the past decade is a recognition that there are two fundamentally different types of information systems in all enterprises: *operational systems* and *informational systems*. Operational systems are just what their name implies: they are the systems that help us run the enterprise day-to-day. These are the backbone systems of any enterprise: the order entry, inventory, manufacturing, payroll, and accounting systems. Because of their importance to the organization, operational systems were almost always the first parts of the enterprise to be computerized. Over the years, these operational systems have been enhanced and maintained to the point that they are completely integrated into the organization. Indeed, most large enterprises around the world today couldn't operate without their operational systems and the data that these systems maintain. Operational data normally is current (the present) and is focused on an area within the enterprise.

On the other hand, other functions go on within the enterprise that have to do with planning, forecasting, and managing the organization. These functions are also critical to the survival of the organization, especially in our fast-paced world. Functions like marketing planning, engineering planning, and financial analysis also require information systems to support them. But these functions are different from operational ones, and the types of systems and information required are also different. The information needed to aid such major decisions that will affect how the enterprise will operate, now and in the future, typically is obtained by analyzing operational data over a long time period and covering many different areas within an enterprise.

A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data to support management decision making. The data in a data warehouse are stored in a manner that will support the information and analytical processing needs of the organization over a long historical time period. They are typically physically separated from the operational databases that an organization must maintain and update to function effectively on a day-to-day basis (Singh 1998; Barquín and Edelstein 1997).

There are many data flows from the operational databases to the data warehouse and from the data warehouse to users. Data must flow from the enterprise's legacy systems to a more compact form, then to the data warehouse storage, and finally out to consumers. The flows necessary to manage the warehouse itself (metaflow) must also be considered (Hackathorn 1995). For each flow, data warehouses require tools to make these processes function effectively (Mattison 1996).

The tools for developing data warehouses can be grouped into three categories, based on their activities: acquisition tools (for inflow), storage tools (for upflow and downflow), and access products (for outflow) (Mattison 1996). Acquisition tools are necessary to perform tasks such as modeling, designing, and populating data warehouses. These tools extract data from various sources (operational databases and external sources) and transform it (i.e., condition it, clean it up, and denormalize it) to make the data usable in the data warehouse. They also establish the metadata, where information about the data in the warehouse is stored (Francett 1994).

Storage is usually managed by relational DBMSs and other special tools in ways that enable the use of data for effective decision support (Mattison 1996). Access to warehouse contents is provided by data mining tools using such methods as neural networks and data discovery mechanisms (Mattison 1996), plus multidimensional analysis software that supports end users in accessing and analyzing the data in the warehouse in various ways (Francett 1994). Data mining is the process of making a discovery from large amounts of detailed data (Barry 1995) by drilling down through detailed data in the warehouse. Data mining tools sift through large volumes of data to find patterns or similarities in the data. Data mining and online analytical processing tools translate warehouse contents into business intelligence by means of a variety of statistical analyses and data visualization methods (Brown 1995; Fogarty 1994). Table 1 provides a list of some of the common terms used within the data warehouse community.

A number of commercial products attempt to fulfill warehousing needs, including the SAS Institute's SAS/Warehouse Administrator, IBM's Visual Warehouse, Cognos's PowerPlay, Red Brick Systems' Red Brick Warehouse, and Information Builders' FOCUS. A discussion of these tools is beyond the scope of this chapter.

4. ENTERPRISE TOOLS

Since the early days of business computing, system designers have envisioned information system architectures that would allow the seamless sharing of information throughout the corporation. Each major advance in information technology spawned a new generation of applications such as CIM or MRP II that claimed to integrate large parts of the typical manufacturing enterprise. With the recent

TABLE 1 Common Terms Used by the Data Warehousing Community

Term	Description
Aggregates	Precalculated and prestored summaries that are stored in the data warehouse to improve query performance. For example, for every VAR there might be a prestored summary detailing the total number of licenses, purchased per VAR, each month.
Business Intelligence Tools	Those client products, which typically reside on a PC, that are the decision support systems (DSS) to the warehouse. These products provide the user with a method of looking and manipulating the data.
Data Extraction, Acquisition	The process of copying data from a legacy or production system in order to load it into a warehouse.
Data Mart	Separate, smaller warehouses typically defined along organization's departmental needs. This selectivity of information results in greater query performance and manageability of data. A collection of data marts (functional warehouses) for each of the organization's business functions can be considered an enterprise warehousing solution.
Data Mining	A collection of powerful analysis techniques for making sense out of very large datasets.
Data Modeling	The process of changing the format of production data to make it usable for heuristic business reporting. It also serves as a roadmap for the integration of data sources into the data warehouse.
Data Staging	The data staging area is the data warehouse workbench. It is the place where raw data are brought in, cleaned, combined, archived, and eventually exported to one or more data marts.
Data Transformation	Performed when data is extracted from the operational systems, including integrating dissimilar data types and processing calculations.
Data Warehouse	Architecture for putting data within reach of business intelligence systems. These are data from a production system that now resides on a different machine, to be used strictly for business analysis and querying, allowing the production machine to handle mostly data input.
Database Gateway	Used to extract or pass data between dissimilar databases or systems. This middleware component is the front-end component prior to the transformation tools.
Drill Down	The process of navigating from a top-level view of overall sales down through the sales territories, down to the individual sales person level. This is a more intuitive way to obtain information at the detail level.
DSS (Decision Support Systems)	Business intelligence tools that utilize data to form the systems that support the business decision-making process of the organization.
EIS (Executive Information Systems)	Business intelligence tools that are aimed at less sophisticated users, who want to look at complex information without the need to have complete manipulative control of the information presented.
Metadata	Data that define or describes the warehouse data. These are separate from the actual warehouse data, which are used to maintain, manage, and support the actual warehouse data.
OLAP (Online Analytical Processing)	Describes the systems used not for application delivery, but for analyzing the business, e.g., sales forecasting, market trends analysis, etc. These systems are also move conducive to heuristic reporting and often involve multidimensional data analysis capabilities.
OLTP (Online Transactional Processing)	Describes the activities and systems associated with a company's day-to-day operational processing and data (order entry, invoicing, general ledger, etc.).
Query Tools and Queries	Queries can either browse the contents of a single table or, using the databases SQL engine, they can perform join conditioned queries that produce result sets involving data from multiple tables that meet specific selection criteria.

TABLE 1 (*Continued*)

Term	Description
Scrubbing/Transformation	The processes of altering data from its original form into a format suitable for business analysis by nontechnical staff.
Star Schema	A method of database design used by relational databases to model multidimensional data. A star schema usually contains two types of tables: fact and dimension. The fact table contains the measurement data, e.g., the salary paid, vacation earned, etc. The dimensions hold descriptive data, e.g., name, address, etc.

growth in client/server applications and more powerful relational databases, a new breed of ''enterprise'' tools has emerged in the last 10 years. These enterprise tools enable the design of information systems, commonly called ''enterprise resource planning'' or ''ERP systems,'' that are truly enterprise-wide. They have become popular in large part because they are replacing older legacy systems that are often complex, redundant, and notoriously inflexible, with a suite of integrated application modules and integrated data that is accessible throughout the enterprise. Another common feature among the various enterprise tools is that they are intended to develop a set of integrated applications, typically for large business enterprises. These business enterprises may be global in nature, and these systems are typically running on a distributed client/server platform with a common, integrated database.

4.1. Enterprise Tools Defined

Today, when software vendors and consultants talk about enterprise tools they are referring to a class of tools that are frequently referred to as enterprise resource planning (ERP) software. *ERP* is a term used by the industry to refer to a collection of applications or modules that can be used to manage the whole business. ERP software describes the class of prepackaged software applications that can be configured and modified for a specific business to create an ERP system. Here we use the term *ERP* interchangeably with the term *enterprise tools*. ERP software or packages are defined here as:

An integrated set of software modules designed to support and automate core business processes that may include logistics, sales, manufacturing, finance, and human resources.

A key point to note in this definition is that the software is integrated around a common user interface and around a shared database so that data can easily be shared between applications such as production and finance. Also of importance is the notion that the software is designed to be modular, so users can implement whichever modules they choose, and that it is oriented towards supporting business processes rather than the traditional functional view of the organization. It is this process view that is often the most misunderstood part of implementing an ERP system, since it can entail substantial reengineering of core business processes.

4.2. Evolution of Enterprise Resource Planning

The idea that a large corporate enterprise could be supported by a completely integrated information system is certainly not new. It has been envisioned since the early days of computers. These visions began to become more concrete in the manufacturing area when computer integrated manufacturing (CIM) systems became popular in the early 1980s. The goal of CIM was to link up the CAD output from the engineering function to marketing personnel so it could be reviewed and also be shared with the production planning and production departments. A part of the product's design was the bill of materials (BOM), which contained the material specifications and quantities to produce a single unit of the particular product. In later versions of CIM, this BOM might also be used create an electronic order and send it to approved vendors via an electronic data interchange (EDI) connection. The basic benefits were that it would decrease product development time and production costs while allowing designs to be changed rapidly and smaller production runs to be made. Because of its inherent complexity, few firms developed these in-house and most chose to purchase generic CIM software. However, CIM did not achieve all the anticipated gains in integration and gave way to another manufacturing integration concept, MRP II.

The lessons learned in trying to implement CIM in the manufacturing environment were used in the development of the newer material requirements planning (MRP) systems in the mid-1980s. The functionality of these new planning systems continued its expansion as companies tried to take a more holistic view of manufacturing processes. Improvements in new relational database technology made it possible to integrate the different departments involved in the manufacturing processed around a central database. MRP II systems extended the integrated manufacturing concept by linking the production schedule to other organizational systems in finance, marketing, and human resources. For instance, changes in customer demand would have a ripple effect on cash flow planning, and personnel planning, in addition to purchasing. The goals again were to cut costs of production, reduce inventory, improve customer service, and integrate production decisions better with finance and marketing functions.

4.2.1. The Appearance of Enterprise Resource Planning

While MRP II remained focused on shortening the ordering and production cycles in a manufacturing environment, several advances in technology in the late 1980s led to the development of ERP tools and applications. Beyond the manufacturing processes, ERP encompasses business planning, sales planning, forecasting, production planning, master scheduling, material planning, shop floor control, purchasing and financial planning, routing and bill of material management, inventory control, and other functions. As with MRP II, a centralized relational database is key to the success of an ERP application. Other new technologies that enabled the spread of ERP tools include the rapidly maturing client/server distributed architecture and object-oriented programming (OOP) development practices. Both of these factors make the ERP system more scalable both in terms of the hardware and also the modularity of the software. This scalability in turn lends itself to continuous expansion of their functionality, as can be seen by the rise of new ERP modules that extend to supply chain and customer relationship concepts, generating tighter linkages with both the customer's and supplier's own planning systems.

As the scope of these integrated applications has grown, it is clear that the complexity has also grown. Horror stories about runaway ERP implementations appear frequently in the popular press. Two examples occurred in 1999 with the much-publicized problems of Hershey's ERP implementation and also a similar set of problems at Whirlpool. After more than $100 million had been spent on Hershey's ERP implementation, Hershey's inventory system did not work and they reportedly lost more than $100 million during the busy Halloween season (Stedman 1999b). ERP costs for large multinational firms can easily take several years and cost several hundred million dollars. Corning Inc. reported in 1998 that they expected their ERP project to take between five and eight years to roll out to all 10 of its manufacturing divisions (Deutsch 1998). Common complaints about ERP products are that they are inflexible and if anything special is desired, expensive consultants must be hired to design "tack on" modules that often must be programmed in some obscure proprietary language. This "one size fits all" mentality makes it difficult for smaller companies to justify the expense of an ERP systems. It may be that ERP will not fit with certain types of organizations that require extensive flexibility. Experience has shown that ERP systems seem to work best in those organizations with a strong top-down structure. Also, the very complexity of an ERP implementation frightens off many potential adopters because it may involve shifting to a client/server environment combined with an often painful process of reengineering.

With all the problems and risks associated with ERP, why would anyone even want to consider an ERP system? Despite these problems, from 1995 to 1999 there was a tremendous growth in corporate implementation of ERP systems. Over 70% of the Fortune 500 firms (including computer giants Microsoft, IBM, and Apple) have implemented an ERP system. Though market growth slowed a bit in 1999 due to corporate focus on the Y2K problem, during this same five-year period, ERP vendors averaged about 35% annual growth (Curran and Ladd, 2000). Globally, the ERP market is projected to grow at a compound annual rate of 37% over the next few years, reaching $52 billion by 2002.

The reality is that if a company today wants an enterprise-wide IS, there are few alternatives to ERP tools. Though the risks are very high, the potential benefits include Y2K and Euro compliance, standardization of systems and business processes, and improved ability to analyze data (due to improved data integration and consistency). A vice president at AlliedSignal Turbocharging Systems said the company was replacing 110 old applications with an ERP system and this would help the $1 billion unit do a much better job of filling orders and meeting delivery commitments (Davenport 1998). Prior to the year 2000, the most common expected benefits of implementing an ERP solution cited in interviews of information technology executives included:

- Solved potential Y2K and Euro problems
- Helped standardize their systems across the enterprise
- Improved business processes
- Improved efficiency and lowered costs
- Needed to support growth and market demands
- Global system capabilities
- Ability to integrate data (Bancroft et al., 1998)

4.2.2. The Enterprise Tools Market

Next to electronic commerce, ERP is one of the most dynamic and rapidly growing areas of business computing. The market for ERP tools has grown at an average annual rate of 30–40% since 1994. ERP revenues for 1998 were approximately $23 billion and were expected to grow to $52B by 2002. Because it grew out of the MRP II market, the early focus of ERP tools was on the manufacturing complex, and it expanded from there, to service and even government and educational institutions. Penetration of this market by ERP tools has reached 70%, and it is said that almost all of the remaining companies are considering installing an ERP solution. The ERP tools market was really defined and is still dominated by a German software company, SAP AG, which currently controls 33% of the ERP market. SAP had revenues in 1998 of $5.05 billion, a 41% increase from the year before (Curran and Ladd 2000). Oracle is the next-leading ERP tools vendor, at 10% of the market, then J.D. Edwards Inc. at 7%, PeopleSoft, Inc. at 6%, and Baan, NV rounding out the top 5 ERP vendors at 5%.

Begun in 1972 by four IBM German engineers, SAP virtually defined the ERP tools market. Its early product, R/2, was a mainframe-based extension to MRP II. However, its top leadership made the strategic decision in 1988 to focus all R&D efforts on developing enterprise solutions for the client/server platform. SAP was one of the earliest software vendors to recognize this fundamental shift and is now the leading client/server application software vendor and the fourth-largest software company in the world. The list of their clients includes 9 of the 10 companies with the top market value in the world and includes IBM, Microsoft, and Apple within the computer industry itself. Their product, R/3, is available in 14 languages and is used by over 10,000 customers in more than 90 countries (Curran and Ladd 2000). SAP has expanded beyond its traditional base in manufacturing industries into service industries, education, and governmental agencies. Service industry clients include Burger King, Sothebys, Lufthansa, and Deutsche Banke. Among governments, the Canadian government standardized on SAP and has one of the largest public installations. MIT and several other institutions of higher learning have also implemented SAP to run their fundamental operations.

The cost of SAP's R/3 for a mid-sized company averages about $750,000, but the implementation costs can be many times the software costs. Typical costs for a full-fledged Fortune 500 firm can typically run about $30 million in license fees and $100–200 million in consulting fees (Kirkpatrick 1998). SAP was the first to institute a University Alliance, which fosters the teaching of business courses that use R/3. SAP invests about 20–25% of its revenues in R&D. This represents an amount greater than all the other ERP vendors combined and ensures that it should be able to respond rapidly to changes in the ERP market.

Of the other ERP vendors, Oracle is in the odd position of being the second-largest software company in the world next to Microsoft and being second to SAP in the area of ERP tools. One reason for this is the fact that Oracle is the preferred database engine for SAP and so is often even shipped with R/3 to new clients. However, Oracle has begun to compete more aggressively with SAP with respect to its ERP tools. Initially, it focused on developing financial application modules and targeted the smaller to mid-sized corporate market. In the last several years, it has either developed application modules in-house or purchased smaller software development firms, so that they now can compete with SAP on the entire ERP spectrum. In addition to its enterprise-wide financial applications, Oracle has recently rolled out an Internet Procurement Suite, Oracle Exchange Suite, and the Oracle Customer Relationship Management (CRM) package. Of all the ERP vendors, it was the first to aggressively extend the ERP functions to e-commerce. As such, 19 of the top 20 e-businesses (e.g., Amazon.com, eBay, Barnes & Noble, and CDNow) use Oracle ERP software. One indicator of this strategy is that in 1988 it was the first to allow users to access their enterprise ISs from a standard web browser. Many venture capitalists reportedly refuse to fund an Internet start-up unless Oracle is a part of the business plan (Stroud, 1999).

PeopleSoft Inc. made its niche in the ERP market by being the company that disaffected SAP users could turn to as an alternative. When it was founded in 1987, it was a vendor of software for the human resource (HR) function. It has since focused on this application in the ERP market, while SAP has focused on inventory management and logistics processes. The quality of PeopleSoft's HR modules was such that some companies would adopt SAP for part of their processes and PeopleSoft specifically for their HR processes. Because it was more flexible in its implementations, PeopleSoft won over customers who just wanted a partial ERP installation that they could integrate more easily than with their existing systems. PeopleSoft was perceived as being more flexible and cost-effective because it developed its whole ERP software suite by collaborating with firms like Hyperion, Inc. for data warehousing, data mining, and workflow management applications. Once PeopleSoft got into a company with its HR applications, it then began offering the financial and other ERP software modules. This led to rapid growth; 1992 revenues of $33 million grew to $1.4 billion in 1998 (Kirkpatrick 1998). Along with the rest of the ERP software industry, PeopleSoft has been scrambling to extend its ERP offerings to the Internet. In 1999, it began development of eStore and eProcurement, and it acquired CRM vendor Vantive (Davis 1999).

Baan Software, a Dutch firm, was an early player in the ERP software market, shipping its first product in 1982. It was the first ERP vendor to focus on an open UNIX computing environment in 1987. Its early strengths were in the areas of procurement and supply chain management. Through internal development and acquisitions, Baan offers one of the most complete suites of ERP tools, including accounting, financial, and HR applications in one comprehensive package or separately for functional IS development. Shifting toward an Internet focus, Baan has recently acquired a number of tool development firms and entered into an alliance with Sun for its Java-based software. It is also moving aggressively to integrate XML-based middleware into its next release of BaanERP (Stedman 1999a).

J.D. Edwards began in 1977 as a CASE tool vendor but developed its own suite of ERP applications, primarily for the AS/400 platform. Its modules now include manufacturing, distribution, finance, and human resources. The company emphasizes the flexibility of its tools and has risen quickly to be fourth-largest ERP vendor, with nearly $1 billion in 1998 sales (Kirkpatrick 1998). It has been an aggressive ERP vendor in moving to the ERP Outsourcing model, where firms access the J.D. Edwards software over a secure network link to an offsite computer. This means firms' pay a fixed monthly fee per user and don't have large up-front investments in expensive implementations and software and skills that are rapidly out of date. This network-distribution-based outsourcing model for ERP vendors has yet to be validated, though, and presents many potential technical problems.

4.3. Basic Concepts of Enterprise Tools

As one reads the literature on ERP tools, it quickly becomes apparent that very little academic research has been performed to in this rapidly emerging area. The field, for the most part, is defined by a few case studies, and research consists primarily of industry publications (Brown and Vessey 1999). Articles appearing in industry publications generate many new terms, which the authors seldom define, and they also tend to over-hype new technology. Various ERP-related terms have been put forth, including *ERP tools, ERP integration tools, ERP network tools, ERP security tools, ERP software,* and *ERP system*. Some vendors are also trying to move away from the term *ERP* to the term *enterprise planning* (EP) and even *extended resource planning* (XRP) systems (Jetly, 1999). As mentioned earlier, *enterprise resource planning* (ERP) is a term used by the industry to refer to prepackaged software tools that can be configured and modified for a specific business to create an enterprise IS. Common enterprise-wide processes are integrated around a shared database. Although mainframe versions of ERP tools exist, for the most part ERP tools assume a client/server platform organized around a centralized database or data warehouse. The essential concepts that generally define ERP tools include integration of application modules, standard user interfaces, "process view" support, integrated database with real-time access, and use of an "open" system architecture usually built around a client/server platform.

4.3.1. ERP and the "Process View"

Another common feature of ERP tools/systems is that they usually involve the extensive application of business process reengineering (Curran and Ladd 2000). This process is often very painful because it may require an organization to reexamine all of its core processes and redesign whole departments. The ERP tool chosen actively supports this "process view." In this sense, a process can be thought of specified series of activities or a way of working to accomplish an organizational goal. To be properly supported by an ERP tool, processes have a defined structure and order and identifiable inputs and outputs. This process view is in contrast to the more traditional functional view, where processes are broken down into activities that lack integration. ERP tools offer further support for the process view of an organization by building into the system standard business processes such as order management or shipping scenarios. These scenarios are then used as the basis for configuring and customizing an ERP system for the specific company. Certain vendors go so far as to try and capture the "best practices" within whole industries, such as the pharmaceutical industry, and use these as models for implementing ERP systems within that industry. Benefits include streamlined business processes, better integration among business units, and greater access to real-time information by organizational members. For many organizations, movement to an ERP-based IS is seen as one way to use computer technology to provide substantial gains in productivity and speed of operations, leading to competitive advantage (Davenport 1998).

4.3.2. ERP and the Open Systems Architecture

It is clear that one of the reasons ERP tools became so successful was that they were able to take advantage of improvements in the client/server architecture and in middleware. The general movement to the client/server platform in the late 1980s meant that firms could reduce the cost of a highly centralized configuration in favor of a more flexible client/server setup. The C/S configuration preferred by most ERP vendors is that of the standard three-tier C/S platform, which consists of a presentation server, connected to the application server, which is in turn connected to the database server. The “3” in SAP’s R/3 product name even refers to this three-tier platform, as opposed to the old R/2 mainframe product.

ERP systems take advantage of new developments in middleware that make it feasible to distribute applications across multiple, heterogeneous platforms. This means that ERP applications can be running on different operating systems and can be accessing data stored in various formats on different database servers worldwide. The middle-ware built into such ERP packages such as SAP and PeopleSoft contains all the communication and transaction protocols needed to move the data back and forth from the database server to the presentation server. The ultimate goal of this is that the movement of data across the varied platforms be transparent to the user, who comes into the ERP applications via a common graphical user interface or even a standard browser interface for e-commerce extensions to the enterprise ISs.

4.3.3. ERP Application and Data Integration

ERP packages include software for a wide range of applications, ranging from purchase orders to accounting and procurement to warehousing. They grew out of the need to plan, manage, and account for resources in manufacturing environments. In recent years, their functionality has grown in both breadth and depth. As companies integrate business units through consolidation or global operations, their information technology’s ability to support these changes is often challenged. The broad functionality of the ERP applications allows companies to replace much of their systems with ERPs, providing better support for new business structures and strategies. The great appeal of ERPs is that employees enter information only once and that information is then available to all applications company-wide. Activities tracked or triggered by the ERP software are also automatically captured in the enterprise’s accounting general ledger, without the risk of reentry errors. This means that IS reports throughout a company are based on accurate, real-time information. The value of this is borne out by the large amounts of money that multinational firms are willing to pay to implement ERPs.

An ERP tool provides companies with a standard interface for all of a firm’s IS applications, a feature that may be lacking with the older legacy application systems. This has become even more important with rise of CRM and the movement of ERP toward Web-enabled enterprise applications. Although standard, SAP’s R/3 interface is famous for some confusing and complicated aspects. Research on an improved user interface is continuing, and the new versions are much improved and more intuitive. For e-commerce extensions to ERP-based ISs, users will soon be able to navigate using a standard web browser such as Microsoft’s Internet Explorer. Soon users will also have the option of accessing their ERP-based ISs using wireless, pen-based, hand-held computers (Marion 1999b).

4.4. Standard Enterprise Tool Applications

When companies look to implement an IS with an ERP tool, they typically analyze their functional requirements, technical integration issues, costs, and strategic issues. With respect to analyzing the functional requirements of a company and comparing them to the functionality of the various ERP tools available, this task is clouded somewhat by several factors. First, with the ERP emphasis on supporting core business processes, designers must be more knowledgeable about how all the corporate processes flow into each other, from creating a sales document (i.e., report) all the way to maintaining records of employee profiles. Beyond this, with the new emphasis on business plans that integrate alliance partners and customers more closely, a company must move its focus away from a strictly intraorganizational one to an increasing awareness of transorganizational needs. Both of these needs are further complicated by the fact that each vendor defines the standard business processes in different ways and hence the component architectures of their software reflect a different set of configurable components. Also, within the different vendor packages there are varying degrees of overlap in the functionality of each of the modules. This makes it very difficult to compare the functionality of specific modules from different vendors and how much of them will be able to be integrated into existing systems. The functionality varies, making it difficult to do a proper comparison of the purchase cost and implementation cost of the modules. This problem is further complicated when one considers the different sets of “industry solutions” that ERP vendors push as ways of smoothing the implementation process.

One way to get around some of the problems in making comparisons of ERP tool functionality is to try to translate a particular company’s needs into a generic set of core processes, such as order management, inventory management, and so forth. Companies may then go through several iterations of attempting to map each vendor’s modules to those requirements until the desired level of functionality has been met. In this way, some of the gaps and overlaps will become more apparent. Further requirements, such as those related to technical, cost, support, and data integration can be overlaid onto the matrix to refine the comparison further. This section examines the four standard core business processes that almost all ERP tools are designed to support. These processes include the sales and distribution processes, manufacturing and procurement, financial management, and human resources. Each ERP package defines these processes slightly differently and has varying degrees of functionality, but all tend to support the basic processes presented here with various combinations of modules.

4.4.1. Sales and Distribution Applications

With the advent of newer, more customer-oriented business models such as Dell's "build to order" computer manufacturing model, it is increasingly important for companies to take customer needs into account and to do it in a timely fashion. ERP tools for supporting sales and distribution are designed to give customers a tighter linkage to internal business operations and help to optimize the order management process. The processes that are most typically supported by the sales and distribution (sales logistics) ERP modules include order entry, sales support, shipping, quotation generation, contract management, pricing, delivery, and inquiry processing. Because the data and functional applications are integrated in a single real-time environment, all the downstream processes, such as inventory management and production planning, will also see similar improvements in timeliness and efficiency.

When companies adopt a particular ERP vendor's package, they often begin with the sales and distribution model because this is the beginning of the business cycle. ERP customers are given a variety of different sales scenarios and can choose the ones that most closely mirror how they would like to run their sales processes. Then they can configure that module to reflect more accurately the way they want to support their sales processes. If they need to modify the module more substantially, they will have to employ programmers to customize the module and possibly link it to existing custom packages that they wish to keep. Whichever approach they choose, the generic sales processes that are chosen by the ERP customer can be used as a good starting point from which to conduct user requirements interview sessions with the end users within the company.

One example of a typical sales process that would be supported by an ERP sales and distribution module is the "standard order processing" scenario. In the standard scenario, the company must respond to inquiries, enter and modify orders, set up delivery, and generate a customer invoice (Curran and Ladd 2000). The system should also be able to help generate and manage quotes and contracts, do credit and inventory availability checks, and plan shipping and generate optimal transportation routes. One area that does distinguish some of the ERP packages is their level of support for international commerce. For example, SAP is very good in its support for multiple currencies and also international freight regulations. Their sales and distribution module will automatically perform a check of trade embargoes to see if any particular items are blocked or are under other trade limits for different countries. They will also maintain international calendars to check for national holidays when planning shipping schedules. This application is also where customer-related data are entered and corresponding records maintained. Pricing data that is specific to each customer's negotiated terms may also be a part of this module in some ERP packages (Welti 1999).

Other sales processes that might be considered a part of the sales and distribution module (depending on the particular ERP product) include support for direct mailing campaigns, maintenance of customer contact information, and special order scenarios such as third party and consignment orders. With the rapid growth of the newer CRM software packages, much of the support for these processes is now being moved to the Internet. E-mail and Web-based forms enabling customer input mean that ERP sales and distribution modules will have to evolve rapidly along CRM lines. Currently, the most advanced with respect to this type of web integration is the ERP package from Oracle, but the other vendors are moving quickly to enhance their current sales and distribution module functionality (Stroud 1999).

4.4.2. Manufacturing and Procurement Applications

Certainly the most complex of the core business processes supported is the large area generally called manufacturing and procurement. Because ERP tools grew out of MRP II and its precursors, these modules have been around the longest and have been fine-tuned the most. The suite of ERP modules for this application is designed to support processes involved in everything from the procurement of goods used in production to the accounting for the costs of production. These modules are closely integrated with the financial modules for cash flow and asset management and with the sales and distribution modules for demand flow and inventory management. Key records maintained by these application modules include material and vendor information.

Among the supply chain activities, the manufacturing and procurement applications handle material management, purchasing, inventory and warehouse management, vendor evaluation, and invoice verification (Bancroft et al. 1998). These activities are so complex that they are typically divided up among a whole suite of manufacturing modules, such as in SAP's R/3, which consists of materials management (MM), production planning (PP), quality management (QM), plant maintenance (PM), and project management (PM). These various modules can be configured in a wide variety of ways and can accommodate job shop, process-oriented, repetitive, and build-to-order manufacturing environments. The PM module is designed to support the management of large, complex projects, such as shipbuilding or special construction projects. Also, despite being oriented more towards large manufacturing enterprises, ERP vendors have made great strides in modifying these modules to accommodate the special needs of service industries such as education and government.

With respect to procurement processes, ERP tools can be used to configure the desired IS in many ways, but they are generally built around a standard procurement scenario. Whether a company must procure subcontract work, consumable goods, or material for their day-to-day production stock, the chain of procurement activities is somewhat similar. These activities involve recognizing a material need, generating a purchase requisition, sending out RFQs to approved vendors, choosing a vendor, receiving and inspecting the goods, and verifying the vendor invoice (Curran and Ladd, 2000). In this area, many companies already use EDI effectively to support their procurement processes, and so most ERP tools are equipped to support standard EDI transactions. The growth of the Internet and extended supply chain management means that ERP tools are increasingly linked directly to their preapproved vendors to further shorten the procurement cycles.

4.4.3. Accounting and Finance Applications

Like the other core business activities, accounting has been undergoing a total reorientation towards a more process-oriented view of the enterprise. Most ERP tools support the accounting and financial processes with another suite of modules and submodules. These include general ledger, AR/AP, asset management, controlling, taxation, and cash flow projection. The more robust ERP packages support accounting for multiple divisions, across multiple currencies, and also support consolidation of financial statements across countries and divisions. Again, because of the integrated nature of all the ERP applications and data, the output from these modules is generated in a much more efficient and timely manner than in the older, legacy systems. The key, of course, is to make sure that the data are highly accurate when it is entered into the system. The downside of data integration is that inaccurate data can be spread through many different applications' reports.

A large part of the accounting and financial activities in an enterprise is the efficient processing of vendor payments and the posting of payments in the correct order into the various ledger accounts. This can be done automatically via EFT or manually in an ERP system, depending on the preferences of the vendor or customer. One of the outputs from the procurement process is a vendor-produced invoice, and this input is the usual starting point for the external accounting activities. Once the invoice is entered either via EDI, manual entry, or even scanning, the invoice is posted to the appropriate account and the amount is verified by matching it with a purchase order and any other partial payments. Once verified, a payment release is issued and payment is issued either electronically or manually. On the customer side, the accounting module also maintains the customer credit records and generates dunning notices (i.e., reports) according to a predefined dunning procedure for each individual customer. The ERP application system will also automatically add on any interest or processing charges that are defined as part of the sales contract for that particular customer. As payments are received and entered into the system, it applies them to the appropriate invoice and moves them to the appropriate GL account.

Most ERP tools also support a variety of management accounting activities. The largest part of these internal accounting activities center on cost management and profitability analysis. Depending on the type of organization, costs can be collected as a byproduct of procurement and production and the system can be configured to generate cost estimates, simultaneous production costing, and also final costs for a period or a project. These costs can be further analyzed and broken down by organizational unit and by specific product or project. Profitability analysis can likewise be conducted at various points in the process and be broken down by organizational unit and product or project. One must bear in mind that with a properly configured suite of ERP financials, the data that are collected and analyzed for these controlling reports are very timely and generating the reports can be done much more efficiently than previously. Analysis can also be conducted across international divisions of the company and whole product lines.

4.4.4. Human Resource Applications

Typically the last set of business processes to be converted to an ERP system, HR applications have recently become very popular as companies look to areas where they might be able to gain a competitive advantage within their respective industries. Because of this, there has been substantial improvement in the HR modules of the leading ERP vendors. This improvement in turn has led to a substantial second wave of ERP implementations among those early adopters, who may have only adopted the financial and manufacturing or sales modules in the first wave of ERP implementations (Caldwell and Stein 1998).

In most organizations, HR too often has become a catchall for those functions that do not seem to fit elsewhere. As a result, HR typically houses such disparate functions as personnel management, recruiting, public relations, benefit plan management, training, and regulatory compliance. In ERP applications, the HR modules house the employee data and also data about the current organizational structure and the various standard job descriptions. The latest wave of ERP tools now supports most of these activities in a growing suite of HR modules and submodules. For example, personnel management includes the recruitment, training and development, and compensation of the corporate workforce. Typical recruitment activities include realizing the internal need for new personnel, generating a job description, posting advertisements, processing applicants, screening, and selecting the final choice for the position. Once the position has been filled, then training assessments can be done and a benefit and development plan can be implemented. Other core processes in the HR module include payroll processing, benefits, and salary administration; sometimes they include time and travel management to help employees streamline reimbursement for travel expenses.

4.5. Implementing an Enterprise System

Because of the large investment in time and money, the selection of the right enterprise tool has a tremendous impact on the firm's strategy for many years afterward. Choosing a particular ERP package means that a company is entrusting its most intimate core processes to that vendor's software. This is a decision that cannot be easily reversed after it has been made. Some firms have chosen to implement an ERP package not realizing the extent of the commitment involved and then had to stop the project in the middle (Stedman 1999b). This is often because these companies are unwilling to reexamine their core processes to take advantage of the integrated nature of an ERP package.

What makes choosing an ERP package different from other types of organizational computing decisions? As discussed above, the large scale and scope of an ERP system make it even more critical than local or functional IS applications for the well-being of the enterprise, apart from the high-cost issue. It is also different in that it is oriented towards supporting business processes rather than separate islands of automation and so the typical implementation involves business experts more than technology experts as opposed to traditional system implementations. The process-oriented nature of an ERP implementation also means that cross-functional teams are a large part of the design process. Because of the high potential impact on an organization due to the redesign of core business processes and reduction of the labor force needed for these processes, one of the more subtle differences is that end users may fight the changes brought about by an ERP system. Therefore, project leaders must also be skilled in "change management" (Davenport 1998).

4.5.1. Choosing an Enterprise Tool

Given that an ERP implementation is so risky to an enterprise, what should a firm consider when choosing an ERP tool? There are currently a handful of case studies in the academic literature and numerous anecdotal references in the trade literature. From these we can derive some factors that a firm should take into account. Probably the most important criterion is how well the software functionality fits with the requirements of the firm. A firm could develop a matrix of its core business process requirements and then try to map the modules from the various ERP packages to the matrix to see where obvious gaps and overlaps might exist (Jetly 1999). A byproduct of this analysis will be a better understanding of the level of data and application integration that the package offers. Increasingly, the ERP package's support for electronic commerce functions will be a major factor in choosing the right package.

Beyond an analysis of the ERP package functionality, a firm must also take technical factors into consideration. Each package should be evaluated for the flexibility of the architectures that it supports and its scalability. How well will this architecture fit with the existing corporate systems? How many concurrent users will it support and what is the maximum number of transactions? Will the module require extensive modifications to make it fit with our business processes and how difficult will it be to change them again if my business changes? Given the potential complexity of an ERP implementation, potential users should be concerned about the ease of use and amount of training that will be required for end users to get the most out of the new system. ERP vendor-related questions should center on the level of support that is offered and how much of that support can be found locally. They should also look into the overall financial health and stability of their prospective ERP vendors and the scope and frequency of their new product releases and updates.

With respect to cost, several factors should be kept in mind when choosing an ERP package. First is the cost of the software and hardware itself that will be required. Besides the required number of new database and application servers, hardware for backup and archiving may also be required. Software costs are typically on a per-module basis and may vary between modules, but a firm may also have to purchase the submodules in order to get the desired functionality. If these modules need to be customized or special interfaces with existing systems need to be programmed, firms should consider the amount of customization required. This cost is exaggerated for some ERP packages, such as SAP's R/3, which require knowledge of the rather exotic proprietary language ABAP/4 in order do extensive customization. Training and ongoing support for these packages will be a continuing expense that should also be considered.

4.5.2. Enterprise System Implementation Strategies

With the high cost of implementing an enterprise information system, developers are under pressure to build them as quickly as possible. Given that the actual system development costs can typically range from 2–10 times the actual cost of the software and hardware (Doane 1997), many ERP consulting firms have sprung up to service this market. Of course, each of them has its own methodology which it claims to be faster and more efficient than the other's. Most of these methodologies represent some variation of the traditional phased systems implementation, so this will be discussed here. Another current ERP debate is the relative merits of the ''big bang'' conversion strategy as opposed to a slower, phased conversion (Marion 1999a).

Because of the strategic nature of ERP decisions, they are usually made at the top executive level. To help coordinate an ERP implementation, firms often create a steering committee with some top executives and high-level management from each of the affected divisions of the firm. The committee will determine the initial scope of the ERP implementation and also which area of the firm will act as a test bed for debugging the initial IS, assuming the firm does not choose to do a big bang implementation. As a result, a firm may choose to begin with the sales and distribution module in its marketing department and expand from there. A project team consisting of IS professionals and end users should then be formed. This team will work in conjunction with the steering committee. Studies show that the importance of the project leader cannot be underestimated and that this person should have strong leadership skills and some experience in implementing an ERP-based IS (Bancroft et al. 1998).

It is the project team's job to analyze the current business processes that will be the focus of the ERP-based IS and look at the existing hardware and software along with the existing data and interfaces. In order to explore the ERP software functionality in more detail, firms sometimes map the initial business processes to the chosen ERP package and do a limited test installation. After exploring the ERP package, the project team will come up with a detailed new design of the business processes in question and a prototype system may be developed and tested by users. Once the prototype has been accepted by users, then the complete ERP-based IS application will be developed and implemented and the data migration will take place. Once the IS has been rolled out for the initial application, it will be expanded to include the rest of the enterprise applications that have been identified in the initial scope of the project. It should be noted that most ERP packages support this development process with built-in CASE tools, standard process scenarios and scripts, and possibly a set of recommended ''best'' practices for specific industries.

As an alternative to this phased approach to ERP implementation, some firms choose the big bang approach, which is riskier but less expensive, in which a firm shifts its entire application for the whole firm from its old legacy system to its new ERP application. This is a high-risk implementation method because the resulting system may have so many bugs that chaos may result, after which it may take months for the firm to recover. An example of the potential problems occurred in the fall of 1999 when Hershey's tried to implement its ERP inventory application using the big bang approach and lost over $100 million due to its failure to manage inventory properly (Stedman 1999b). However, there have been big bang success stories at McDonalds Corp. and Rayovac, and perhaps the benefits of this approach can outweigh the risks (Marion 1999a). Some of the benefits include lower cost because it requires less expensive consultant time. It also reduces the amount of interface development with the old legacy systems because there is no temporary connection to maintain with the new ERP applications. The other problem with the long phased approach is that when one application is ready, updated versions of the ERP software are available for other applications when they are ready to be rolled out, and then developers must deal with a whole new set of compatibility and conversion issues.

4.5.3. Critical Success Factors for ERP Implementations

From the various case studies reported in the trade literature, several key factors have been reported that seem critical for the success of an ERP implementation (Bancroft et al. 1998). These include:

1. Management support for making the hard decisions
2. Capability and readiness of the organization for making changes
3. A balanced project team with strong leadership
4. Reasonable project scope
5. The right ERP package
6. The right implementation method
7. A strong training plan for users and the project team

The most commonly cited of these factors is the importance of management support for the success of the project. This factor has been widely documented for other types of business computing projects, but it is especially important in implementing an ERP system because this may involve a radical redesign of some core processes and resulting major changes in the organization. In a project as complex as an ERP system, problems will occur. Leaders must maintain the strong vision to persevere through all the changes. One study showed that nearly 75% of new implementations resulted in some organizational change, and these changes could be very severe (Boudreau and Robey 1999). Project teams should also be balanced with respect to business area experts and IS professionals. As the software becomes easier to configure and has more complete functionality, it seems clear that the trend of shifting more responsibility away from the IS professionals to the business experts will continue (Bancroft et al., 1998). Because of the cross-functional nature of ERP-based ISs, it is also important that team members be drawn from multiple business units and assigned full-time to the project (Norris et al., 1998).

4.6 Future of Enterprise Tools

The new generation of enterprise tools is beginning to enable large corporations to reap significant gains in efficiency and productivity, but the investment costs of implementing an ERP system can be a huge hurdle for smaller companies. Enterprise tools will continue to improve in overall functionality and ease of use as modules are developed and refined. The functionality will continue to be extended to the customer, supplier, and other business partners in what some call the Extended Resource Planning (XRP) software generation (Jetly 1999). Most of these extensions are based on some form of e-business model using the Internet. A description of some of the principal newer enhancements to ERP systems follows. These are taking enterprise ISs into the transorganizational realm.

4.6.1. Enterprise Tools and Supply Chain Management (SCM)

One logical outgrowth of the increased integration of enterprise-wide data and support for core business processes is the current interest in supply chain management. In the 1980s, much of the focus in manufacturing was on increasing the efficiency of manufacturing processes using JIT, TQM, Kanban, and other such techniques. Given the large investment in these improvements in manufacturing strategies, companies managed to lower the cost of manufacturing goods as far as was practical. These same companies are finding that they can take a system view of the whole supply chain, from suppliers to customers, and leverage their investment in ERP systems to begin to optimize the efficiency of the whole logistics network. In this context, supply chain management is defined as follows:

> A set of approaches utilized to efficiently integrate suppliers, manufacturers, warehouses, and stores, so that merchandise is produced and distributed at the right quantities, to the right locations, and at the right time, in order to minimize system-wide costs while satisfying service level requirements (Simchi-Levi et al. 2000).

Until recently, ERP vendors have focused their efforts on providing real-time support for all the manufacturing transactions that form the backbone of the supply chain and left the more focused decision support applications to software companies such as i^2 Technologies and Manugistics. These companies have produced whole families of decision support systems to help analyze the data produced by the ERP-based IS and make better decisions with respect to the supply chain. The interfaces between the standard ERP tools and the new generation of DSSs is where some of the most interesting and valuable research and development is being conducted. As companies improve their support of specific supply chain activities, they can expand the level of integration to include supporting activities from financial operations and human resources.

When taking a supply chain view of its operations, a company must begin with the notion that there are three basic flows within their enterprise: materials, financials, and information. At defined points in the supply chain, data can be collected and stored. The goal of ERP-based ISs in this view is to enable the enterprise to collect and access this information in a timely fashion. This reduces inventory levels, reduces costs, and improves customer service and satisfaction, and each product produced by the enterprise will also have an accompanying information trail that can be tracked and used for planning and estimating lead times. It must be kept in mind that this integrated information flow is available in real time as a basis for decision making.

Much of the recent interest in supply chain management and ERP stems from the e-business models that have been emerging. These models now call for tighter linkages between suppliers and customers, and the ERP tools must support these linkages. This extended supply chain management makes use of the Internet to link with suppliers and customers via standard browsers and TCP/IP protocols. This means that not only will customers be able to track their orders, but companies will be sharing inventory and BOM information with their suppliers. In fact, one of the major unresolved issues in ERP research is the question of inter-ERP linkage. Most vendors have their own standard format, and until there is general agreement within the industry, inter-ERP data sharing will require the development of sophisticated middleware to share data between ISs at companies using different vendors' ERP tools (Simchi-Levi et al. 2000).

4.6.2. Enterprise Tools and Customer Relationship Management (CRM)

As ERP vendors continue to extend the functionality of their packages, one area where there has been tremendous growth has been in the ability of the software to support a company's links to its clients. Interest in better supporting these linkages has led to the development of a niche software market called customer relationship management (CRM) software. CRM supports business functions typically performed by marketing, sales, and service. CRM software is designed to support these functions by helping to identify, select, develop and retain customers. With the recent move to e-business, the traditional view of CRM has changed to look to how CRM software can be combined with electronic commerce technology to gain a competitive advantage. This means that new CRM software helps to move existing processes such as order entry to the Internet and supports new processes such as email-based customer support. Companies are now revising their business strategies to take advantage of Web-enabled CRM software to support Web marketing and customer interaction. Besides supporting sales staff, the new breed of CRM software supports a "contact center," which is a focal point of phone, Web, e-mail, ATM, and kiosk customer interactions. These front-office activities then must be linked in real time to back office functions that fall into the domain of the traditional ERP package.

While the predictions for the maturing ERP market indicate that its growth rate is slowing, the market for CRM software is just starting to explode. One industry analyst has predicted that the market will grow at an annual growth rate of 49% over the period from 1999–2003, when it will reach $16.8 billion in sales (Bowen, 1999). Of this market, the top vendors in 1998 were Siebel Systems, Vantive, and Clarify, who together represent 40% of the CRM tool market. Siebel became the market leader by changing its focus from client/server sales force automation applications to developing an integrated, Web-based CRM suite, with sales, marketing, and customer modules. It has generated more than 50 partnerships with industry giants, such as the recent alliance with IBM. With strategy, their revenues increased from $36 million in 1996 to $329 million in 1998 (Zerega 1999).

Because of the immense potential of the CRM market and because it fits well with their attempts to turn their product lines toward an e-business orientation, all the major ERP vendors are now looking to expand into the front-office domain. ERP vendors are doing this by acquiring CRM vendors, developing alliances with pure CRM vendors, developing their own packages, or as in the case of SAP, a combination of these. In October of 1999, Peoplesoft announced that it was acquiring one of the leading CRM vendors, Vantive. Similarly, Baan has already purchased Aurum Software and has formed an alliance with Cap Gemini to integrate its CRM applications into BaanERP (Goldbaum 1999). Of the top five ERP vendors, Oracle is the only one currently offering a front-office system. It is ahead of the ERP market in developing a suite of Web-enabled CRM applications that integrate easily with its tools for back-office enterprise systems. It has aggressively pursued alliances with hardware vendors such as HP to allow Oracle and HP salespeople to collaborate and share information over the Internet (Stroud 1999). In this way, it cosells its CRM solutions and shares customer data. Oracle also announced in March of 1999 that it would integrate its CRM suite with SAP's ERP tools (Sykes 1999). In contrast, SAP does not currently offer a CRM module for its product, R/3.

SAP is the ERP market leader with the most comprehensive and integrated package of ERP applications. Yet it is seen as being late to move its focus to Web-enabled enterprise application systems. In the fall of 1999, it unveiled its long-awaited MySAP.com and is attempting to use its large client base of active installations for leverage into the CRM market. Given its huge investments in R&D, SAP has never been inclined to purchase niche players in related software markets but has preferred to develop its own tools in-house. With respect to CRM tools, its eventual plan is to package CRM modules as a part of its new MySAP.com architecture for e-business portals (Bowen 1999). It will be browser-based, using Microsoft's Internet Explorer, and will provide seamless access to R/3's enterprise-wide applications. Although SAP has yet to deliver its own CRM modules, it is predicting that CRM-related revenues will eventually outpace revenues from sales of "pure" ERP packages (Bowen 1999).

Because of its huge growth potential, the CRM market is very dynamic, with a stream of new entrants and the expansion of current CRM vendors' tools. With respect to ERP packages, the trend is toward the eventual blending of ERP with CRM modules. This will occur as the major ERP vendors continue to turn their focus toward e-business. It is expected that this new amalgam of front-office with back-office applications will lead to a seamless new set of IS development tools called "e-business suites" (Goldbaum 1999).

4.6.3. Enterprise Tools and Electronic Commerce (EC)

As the ERP market has matured, ERP vendors have rushed to be the first to redesign their products so that they Web-enable their enterprise tools. The rapidly evolving model of ERP and e-commerce means that users anywhere in the world will be able to access their specific enterprise-wide ISs via a standard, browser-based interface. In this model, the ERP platform defines a new suite of e-business modules that are distributed to suppliers and clients using the standards of the Internet. The ERP tool can thus be used to build an e-business "portal" that is designed specifically for everything from small "e-tailers" to large multinational corporations. This portal will streamline links with customers, suppliers, employees, and business partners. It will extend the linkage between the enterprise and transorganizational classes of ISs.

One way that this model is emerging is as a turnkey supplier of e-business function support. For example, SAP's new web portal model, MySAP.com, can be used to help a new e-tailer design and build its website. Using an innovative natural language interface, a business person can describe the features he or she wants for the company website and a basic web page will be automatically generated. The website will be further linked with whatever functionality is desired from the back-office enterprise IS, via the Web-enabled suite of modules in MySAP.com. These modules might include processing customer orders, creating and maintaining a corporate catalog for customers to browse, processing bank transactions, monitoring customer satisfaction, and allowing customers to check the status of their orders. The same platform can be used to maintain the company's business-to-business applications, such as eProcurement and information sharing with business partners and clients. For a fee, MySAP.com users can "rent" back-office decision support software for forecasting, data mining, and inventory planning and management. These services are delivered over the Internet and users pay on a per-transaction basis. This web portal also serves as the platform for business-employee transactions, helping employees communicate better, monitor their benefit plans, plan for and receive training, and possibly provide more advanced knowledge management support.

5. TOOLS FOR ANALYSIS AND DESIGN OF IS

Managing the development and implementation of new IS is vital to the long-term viability of the organization. In this section, we present and discuss some of the tools and methodologies that have emerged over the last 40 years to help facilitate and manage the development of information systems. The systems development life cycle (SDLC) is introduced, and the tools used within each step of the life cycle are explained. It is important to realize that although professional systems developers have employed these tools and methodologies for the last 20 years with some success, the tools do not guarantee a successful outcome. Utilization of the SDLC will increase the likelihood of development success, but other forces may affect the outcome. Careful attention must also be paid to the needs of users of the system and to the business context within which the system will be implemented. This attention will increase the likelihood of system acceptance and viability.

5.1. The Systems Development Life Cycle

The SDLC was developed as a methodology to aid in the development of information systems. By defining the major activities in the process, the SDLC ensures that all the required tasks are completed, and the time and resources necessary in the different phases can be more precisely estimated. The SDLC separates the tasks that have to be performed when developing an IS into five phases:

1. Systems planning
2. Systems analysis
3. Systems design
4. Systems implementation/testing
5. Systems use and maintenance

Some have referred to this process as a waterfall because one phase naturally leads to the next, in a descending cascade. In business, all information systems become obsolete over time, either functionally or operationally. When the system fails to achieve the desired results, a change in that system is necessary. The existing system may be in the use and maintenance phase; however, it is necessary to move into the systems planning phase in order to initiate a change in the existing system.

During the systems planning phase, an investigation is conducted into the desirability and feasibility of changing the existing system. At this point, the major purpose is to understand the problem and ascertain whether it is cost effective to change it. At the outset of this phase, a systems development team is formed, made up of members of user groups, members of the information technology group, and some members of management. This team typically works together through the entire SDLC. At the end of this phase, a feasibility report is produced. This report should outline the major business needs to be met by the new system, the system's major inputs and outputs, the scope of the actual problem, and an estimate of the time and resources necessary to solve the problem.

The tools used during this phase usually include automated tools that enable team members to model the existing system using a context diagram and an economic analysis using tools that enable the team to conduct traditional financial analysis, such as net present value and return on investment. A determination is made at the end of this phase whether to continue on with the development or terminate the project. Often the project is either too trivial or cannot be solved in a cost-effective manner to warrant continuation. In such cases, the project is terminated at the end of this phase without major expenditures. If the team decides that the project is worth pursuing, they will proceed to the system analysis phase.

In the system analysis phase, some of the planning activities are repeated in order to flush out the details. The analysts must fully understand the entire problem and begin to create a guide to solve it. This involves a much more detailed study of the existing system. The analysts will question users in much more depth to gather data on their exact requirements. The existing system is typically modeled using tools such as data flow diagrams, data dictionaries, and entity relationship diagrams. Each of these tools will be discussed in greater detail in a following section. The modeling tools allow the analyst team to gain a thorough understanding of the existing system. This understanding, when compared with the new system requirements, will generate the new system proposal. This proposal spells out the changes necessary to correct the deficiencies in the existing system. The proposal is typically presented to a management group that is responsible for funding the next step in the process, system design. At this point, management determines whether the project should continue. They will either provide funds for the project to move on to the system design phase or terminate the project. If the project is terminated at this point, only labor costs are incurred because no information technology expenditures have yet been expanded.

In the design phase, the team will determine how user requirements will be met, working from the outputs backward. A new physical model of the proposed system is created. If the analysts are able to determine precisely what outputs the users require, they can then work back through the processes needed to produce those outputs and the input data needed in the process. This provides a blueprint for the design: the formulation of the data files, the specifications for the software, and the configuration of the needed hardware. The data configuration and software and hardware requirements are determined during this phase. The rule of thumb in systems has been to work in this manner because the data and software design drive the purchase of the actual computing equipment. If the hardware were designed first, later it might either prove inadequate for the long-term requirements of the new system or overshoot those requirements so much as to be less cost effective. At this point, nothing has actually been purchased. The new physical design is summarized in a new system proposal, which is presented to management for funding. If management declines to fund the proposal, the design team disbands. If management funds the new proposal, then the project moves on to system implementation.

If the project reaches this phase, it has reached the point of no return. During the implementation phase, the new system design is used to produce a working physical system. The needed pieces of the system are either created or acquired, then installed and tested on the hardware that has been earmarked for this system. Generally, the team begins by creating the test data needed for the new system. The software specifications are then converted into computing code and tested with the test data. The users are trained in the procedures required to use the new system. The site where the system will be installed is prepared, and the new equipment is installed and tested. The entire system typically undergoes final tests before being released into the production environment.

After the system has been installed, the system enters the use and maintenance phase. In order for the system to deliver value to the organization, minor maintenance typically is necessary to correct errors, ensure system currency, enhance features, or adapt hardware. The system will be periodically tested and audited to increase the reliability of the system outputs. When the system ages to a point where it can no longer be easily modified, the SDLC will shift back to the planning phase and the life cycle will begin again.

During each phase of the SDLC, the organization should maintain detailed documentation of the system. This documentation is the accumulation of all the models, analysis and design specifications, and records during system development. Such documentation is invaluable to the next team charged with a development project because it explains all of the steps and assumptions made during the system's development. The team must keep meticulous records of all meetings, findings, and results, which together make up the system document. This document is maintained at the organization as long as the system is in production.

5.2. Systems Development Tools

This section examines the models and methodologies that are used during the different phases of the SDLC.

5.2.1. Feasibility Analysis

Information systems development projects are almost always done under tight budget and time constraints. This means that during the systems planning phase of the SDLC, a feasibility assessment of the project must be performed to ensure that the project is worth the resources invested and will generate positive returns for the organization. Typically, feasibility is assessed on several different dimensions, such as technical, operational, schedule, and economic. Technical feasibility ascertains whether the organization has the technical skills to develop the proposed system. It assesses the development group's understanding of the possible target technologies, such as hardware, software, and operating environments. During this assessment, the size of the project is considered, along with the systems complexity and the group's experience with the required technology. From these three dimensions, the proposed project's risk of failure can be estimated. All systems development projects involve some degree of failure risk; some are riskier than others. If system risk is not assessed and managed, the organization may fail to receive anticipated benefits from the technology, fail to complete the development in a timely manner, fail to achieve desired system performance levels, or fail to integrate the new system properly with the existing information system architecture. Reducing the scope of the system, acquiring the necessary development skills from outside, or unbundling some of the system components to make the development project smaller can reduce these risks.

Operational feasibility examines the manner in which the system will achieve desired objectives. The system must solve some business problem or take advantage of opportunities. The system should make an impact on value chain activities in such a way that it supports the overall mission of the organization or addresses new regulations or laws. The system should be consistent with the operational environment of the organization and should, at worst, be minimally disruptive to organizational procedures and policies.

Another dimension of feasibility, projected schedule feasibility, relates to project duration. The purpose of this assessment is to determine whether the timelines and due dates can be met and that meeting those dates will be sufficient to meet the objectives of the organization. For example, it could be that the system must be developed in time to meet a regulatory deadline or must be completed at a certain point in a business cycle, such as a holiday or a point in the year when new products are typically introduced to market.

Usually, the economic feasibility assessment is the one of overriding importance. In this case, the system must be able to generate sufficient benefits in either cost reductions or revenue enhancements to offset the cost of developing and operating the system. The benefits of the system could include reducing the occurrence of errors associated with a certain procedure; reducing the time involved in processing transactions; reducing the amount of human labor associated with a procedure; or providing a new sales opportunity. The benefits that can be directly measured in dollars are referred to as tangible benefits. The system may also have some benefits that, though not directly quantifiable, assist the organization in reaching its goals. These are referred to as intangible benefits. Only tangible benefits will be used in the final economic feasibility assessment, however. After benefits have been calculated, the tangible costs of the systems are estimated. From a development perspective, these costs include hardware costs, labor costs, and operational costs such as site renovation and employee training. The costs can be divided into two classes: one-time (or sunk) costs and those costs associated with the operation of the new system, or recurring costs. The benefits and costs are determined for a prespecified period of time, usually the depreciable life of the system. This time period could be 1 year, 5 years, or 10 years, depending on the extent of the investment and the financial guidelines used within the organization.

The tangible costs and benefits are then used to determine the viability of the project. Typically, this involves determining the net present value (NPV) of the system. Since the cost of developing the system is incurred before the systems begins to generate value for the organization over the life of the system, it is necessary to discount the future revenue streams to determine whether the project is worth undertaking. The present value of anticipated costs at a future point in time is determined by:

$$\mathrm{PV}_n = \left(\frac{Y_n}{(1 + i)^n}\right)$$

where PV_n is the present value of an anticipated cash inflow or outflow Y_n (Y_n is the monetary value (\$) of the cost or benefit) in the nth year into the project's life and i is the interest rate at which future cash flows are to be discounted. The net present value (NPV) of the project is:

$$\mathrm{NPV} = -C_0 + \sum_{t=1}^{T} PV_t$$

where C_0 is the cost of developing the system and putting it into production and T is the useful life of the system.

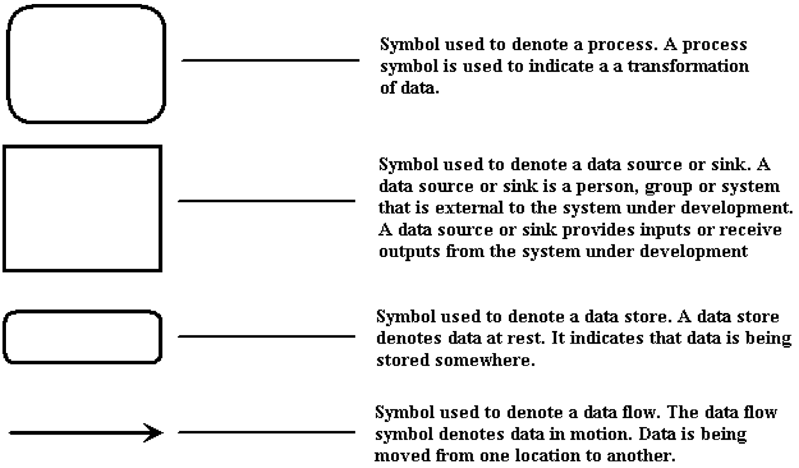
Other financial analyses that provide useful information as one seeks to determine whether the organization should proceed with the project include return on investment (ROI) and break-even (BE) analysis. ROI is the ratio of the net returns from the project divided by the cash outlays for the project. BE analysis seeks to determine how long it will take before the early cost outlays can be recouped. If the ROI is high and the BE point is small (i.e., reached quickly), the project becomes more desirable because cash flow estimates far into the future are much more difficult to forecast accurately. Table 2 is a simple spreadsheet that was used to determine economic feasibility.

5.2.2. Process Modeling Using Data Flow Diagrams

During the analysis phase, it is critical that the development team gain a good understanding of the existing system so that they can be in a position to know how the current system produces information. This understanding must be at a very low level of detail, at the level of the processes used to produce information. This level of understanding is best obtained by developing models of the existing system. Typically, the existing system is first graphically represented using a diagramming method known as a data flow diagram (DFD). This method provides a simple way to describe the system graphically. The diagrams can be produced quickly by the development team yet can provide a rich level of detail as they are expanded on lower levels. They are also easy for novices to read, which means they can be powerful communication tools among the development team, managers, and users. Finally, they can be decomposed, that is, logically broken down, showing increasing levels of detail without losing their ability to convey meaning to users and management.

DFDs are drawn with just four symbols: data sources and sinks, processes, data flows and data stores (see Figure 5). Data sources are entities that provide data to the system (a source) or receive information from the system (a sink). These sources lie outside the boundary of the system and are not active participants in the processing that occurs within the system. A process is any activity that converts data into information. Generally, processes are named with a verb of the action in the process and are numbered at the top of the symbol. Numbers are useful for cross-referencing processes on lower-level diagrams to the activities in higher-level diagrams. Processes are interconnected with other processes by data flows. Data flows are symbols that represent data in motion. Each data flow is named with a noun that describes the data represented in the flow. A data flow must be attached to a process at some point, either as a source or as a sink. A data store is a representation of the data at rest. Examples of data stores could include electronic data files, a database table, or a physical file folder.

When the symbols are used in combination, they trace the movement of data from its origin, the source, through all of the steps that transform that data into information for the users, to its eventual destination, a sink. The feature of DFDs that make it so useful in the context of system development is that the diagrams are decomposable. Every system must be understood at a high level of detail.



Symbol used to denote a process. A process symbol is used to indicate a a transformation of data.

Symbol used to denote a data source or sink. A data source or sink is a person, group or system that is external to the system under development. A data source or sink provides inputs or receive outputs from the system under development

Symbol used to denote a data store. A data store denotes data at rest. It indicates that data is being stored somewhere.

Symbol used to denote a data flow. The data flow symbol denotes data in motion. Data is being moved from one location to another.

Figure 5 Data Flow Diagram Symbols.

DFDs allow the same system to be examined at many levels of detail, from the context diagram level through whatever level the development team chooses to stop at.

The lowest level of detail in a DFD is called a context diagram, also called a "black box" diagram, because it shows only the sources and sinks and their connection to the system through aggregate data flows. The system itself is represented by only one process. That central process will be decomposed into many subsystems in the Level 0 diagram, which shows the system's major processes, data flows, and data stores at a higher level of detail. Figure 6 shows a level 0 diagram for a simple payroll processing system. Each process is numbered, and each represents a major information-processing activity. When the diagram moves to Level 1, each subsystem from Level 0 is individually decomposed. Thus, process 1.0 on the DFD may be functionally subdivided into processes 1.1, 1.2, 1.3, and so on for as many steps as necessary. In Figure 7, we show a Level 1 diagram that decomposes the third process from the Level 0 diagram in Figure 6. As the diagrams are decomposed further, at Level 2 the steps in process 1.1 may be decomposed into subprocesses 1.1.1, 1.1.2, 1.1.3, etc. Each Level 2 DFD will expand on a single process from Level 1. These diagrams can continue to be leveled down until the steps are so simple that they cannot be decomposed any further. This set of DFDs would be referred to as primitive diagrams because no further detail can be gained by decomposing further. The power of this tool is that while it is conceptually simple, with only four symbols, it is able to convey a great degree of detail from the nesting of the decompositions. For systems analysis, this detail is invaluable for an understanding of existing system functionality and to understand the processes.

5.2.3. *Structured English*

The next step in process modeling, usually performed after the creation of the DFDs, is to model process logic with Structured English, a modified form of the English language used to express information system process procedures. This tool uses the numbering convention from the DFDs to designate the process under examination, then semantically describes the steps in that DFD process. Structured English is used to represent the logic of the process in a manner that is easy for a programmer to turn into computer code. In fact, similar tools were called "pseudo-code" to indicate that they were closely related to the design of computer programs. At that point they were used

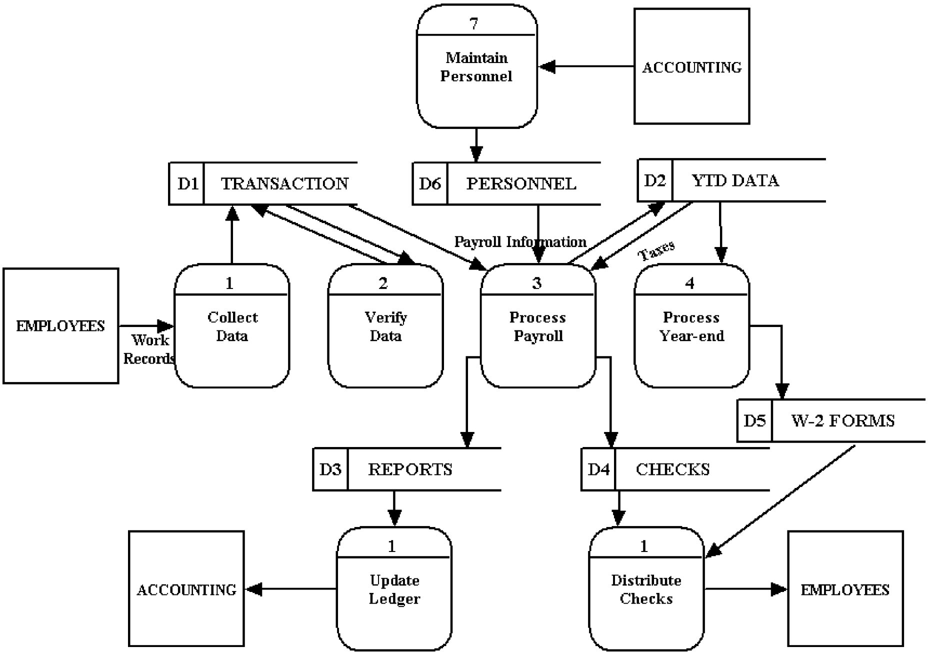


Figure 6 Data Flow Diagram for a Simple Payroll System.

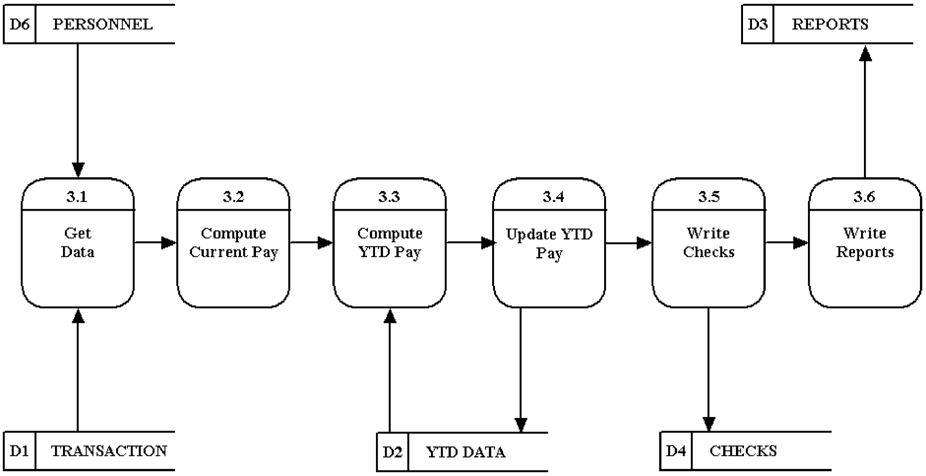


Figure 7 An Explosion of the Process Payroll Process.

to describe individual programs, not entire information systems. As a tool, Structured English representations of processes help in the implementation of the system, making it much easier to develop specifications that can be handed off to those who will develop the programs. Generally, Structured English is generic enough that it can used to develop programs in most programming languages.

The steps in developing Structured English versions of a process are fairly straightforward. First, describe the steps in the process being examined, beginning with a strong verb. The verb modifies a noun, which is generally a data structure, data element, or data flow. Any given statement should include no adjectives or adverbs. While there is no standard for Structured English, it represents basically three types of processes: sequence, conditional statements, and iterations. Sequences are lists of statements where one statement simply follows another. An example of a sequence is:

```
Process 1.1: Update Master File
DO
     OPEN Master File
     READ Record
     MATCH Key to Master File Key
     REPLACE Master record with New Master Record
     CLOSE Master File
END DO
```

The second type of Structured English process is a conditional statement. A case statement is declared, followed by a group of possible actions that will be executed after the evaluation of the case. Programmers refer to this as an "if/then" statement. An example of a conditional statement is:

```
Process 2.3: Order Inventory
BEGIN IF
     IF Quantity on hand is less than Reorder quantity
     THEN generate New order
     ELSE do nothing
END IF
```

The last type of Structured English process is an iteration. Iterations are special cases of the conditional statement, where steps are repeated until a stated condition is satisfied. This type of process is also sometimes referred to as a repetition. An example of an iteration is:

```
Read Order File
WHILE NOT End Of File DO
     BEGIN IF
          READ Order Address
          GENERATE Mailing Label
     END IF
END DO
```

Structured English has been found to be a good supplement to DFDs because, in addition to helping developers at the implementation stage, they have been found to be a useful way to communicate processing requirements with users and managers. They are also a good way to document user requirements for use in future SDLC activities.

5.2.4. ERD/Data Dictionaries

During the analysis phase of the SDLC, data are represented in the DFDs in the form of data flows and data stores. The data flows and data stores are depicted as DFD components in a processing context that will enable the system to do what is required. To produce a system, however, it is necessary to focus on the data, independent of the processes. Conceptual data modeling is used to examine the data "at rest" in the DFD to provide more precise knowledge about the data, such as definitions, structure, and relationships within the data. Without this knowledge, little is actually known about how the system will have to manipulate the data in the processes. Two conceptual data modeling tools are commonly used: entity relationship diagrams (ERDs), and data dictionaries (DDs). Each of these tools provides critical information to the development effort. ERDs and DDs also provide important documentation of the system.

An ERD is a graphical modeling tool that enables the team to depict the data and the relationships that exist among the data in the system under study. Entities are anything about which the organization stores data. Entities are not necessarily people; they can also be places, objects, events, or concepts. Examples of entities would be customers, regions, bills, orders, and divisions. Entity types are collections of entities that share a common property or characteristic, and entity instances are single occurrences of an entity type. For example, a CUSTOMER is an entity type, while Bill Smith is an entity instance. For each entity type, one needs to list all the attributes of the entity that one needs to store in the system. An attribute is a named property or characteristic of an entity type. For example, the attributes for a CUSTOMER could be customer number, customer name, customer address, and customer phone number. One or a combination of the attributes uniquely identifies an instance of the entity. Such an attribute is referred to as a *Candidate key,* or *identifier*. The candidate key must be isolated in the data: it is the means by which an entity instance is identified.

Associations between entity types are referred to as *relationships*. These are the links that exist between the data entities and tie together the system's data. An association usually means that an event has occurred or that some natural linkage exists between the entity types. For example, an order is associated with an invoice, or a customer is associated with accounts receivable instances. Usually the relationship is named with a meaningful verb phase that describes the relationship between the entities. The relationship can then be "read" from one entity to the other across the relationship. Figure 8 shows a simple ERD for a system in a university environment.

Once an ERD has been drawn and the attributes defined, the diagram is refined with the degree of the relationships between the entities. A unary relationship is a one-to-one or one-to-many relationship within the entity type—for example, a person is married to another person, or an employee manages many other employees. A binary relationship is a one-to-one, one-to-many, or many-to-many relationship between two entity types—for example, an accounts receivable instance is associated with one customer, or one order contains many items. A tenary relationship is a simultaneous relationship between three entity types—for example, a store, vendor, or part, all sharing stock-on-hand. Once the degree of the relationships have been determined, the ERD can be used to begin the process of refining the data model via normalization of the data. Normalization is necessary to increase the stability of the data so that there are fewer data updating and concurrency anomalies. The normalization process is described in detail in every database management textbook (e.g., McFadden et al., 1999).

The ERD is an excellent tool to help understand and explore the conceptual data model. It is less useful, however, for documentation purposes. To address the documentation requirements, a data dictionary (DD) is used. The DD is a repository of all data definitions for all the data in the system. A DD entry is required for each data flow and data store on every DFD and for each entity in the ERDs. Developing a DD entry typically begins with naming the data structure being described, either by flow name, file name, or entity type. The list of attributes follow, with a description of the attribute, the attribute type, and the maximum size of the attribute, in characters. The attribute type is specified (e.g., integer, text, date, or binary). In the case of a data file, the average size of the file is indicated.

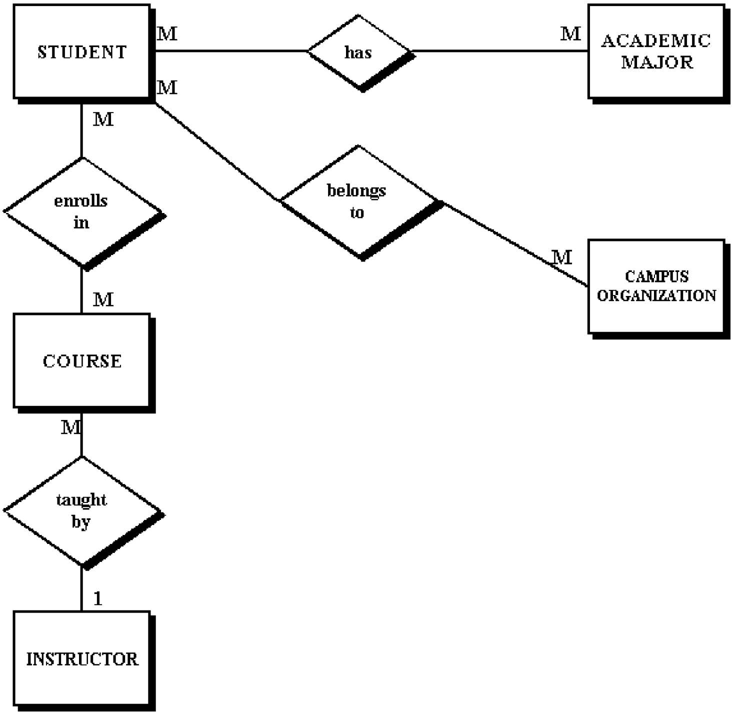


Figure 8 An Entity-Relationship Diagram.

For a data flow, the minimum, maximum, and average number of occurrences of that flow in a business day is determined and listed with the attributes. For the entity type, the entry will note the various associations that the entity type has with the other entities. An example of a DD entry is:

```
        Data Stores

Name:   Customer Mailing List

Key     Customer Number

Size    250,000
```

	Name	Definition	Type	Size
Attributes:	Cust_Num	Unique customer Number	Integer	5
	Cust_Name	Customer name	Text	25
	Cust_Add	Customer Address	Text	30

5.2.5. Gantt/PERT Diagram

Gantt charts and PERT diagrams are used in the planning phase of the SDLC to schedule and allocate time to key activities in the development process. Gantt charts are two-dimensional charts that show major activities on one axis and a time line on the other. The duration of the activities listed is designated with a solid bar under the time line, next to the name for that activity. The time lines for the bars can overlap because certain activities may overlap during the development process. An

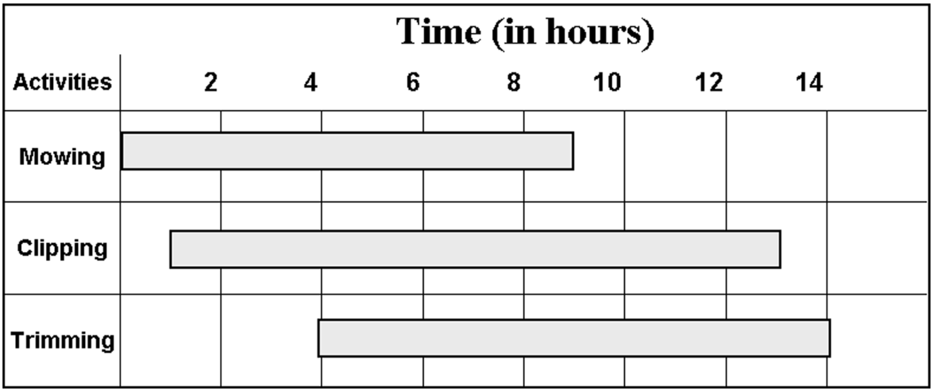


Figure 9 A Gantt Chart for a Simple Lawn-Maintenance Task.

example of this would be program code development and site preparation during system implementation. Figure 9 shows a Gantt chart for a lawn maintenance job with two workers assigned to each activity (Wu and Wu 1994).

PERT diagrams serve the same function as Gantt charts but are more complex and more detailed. They are essential tools for managing the development process. PERT is an acronym for Program Evaluation and Review Technique and was developed by the United States military during World War II to schedule the activities involved in the construction of naval vessels. Since then, PERT has found its way into production environments and structured system development because it allows one to show the causal relationships across activities, which a Gantt chart does not depict. PERT diagrams show the activities and then the relationships between the activities. They also show which activities must be completed before other activities can begin and which activities can be done in parallel. Figure 10 shows an example of a PERT diagram for the lawn maintenance job (Wu and Wu 1994).

Which planning tool is chosen depends upon the complexity of the system being developed. In the case of a systems development effort that is relatively small and simple and where the design can be completed quickly, the Gantt chart is preferable. Where the design is large, complex, and involves many activities that must be carefully coordinated, the PERT diagram is best.

5.2.6. JAD/RAD

Rapid application deployment (RAD), joint application deployment (JAD), and prototyping are approaches used in system development that bypass parts of the SDLC in an effort to speed up the

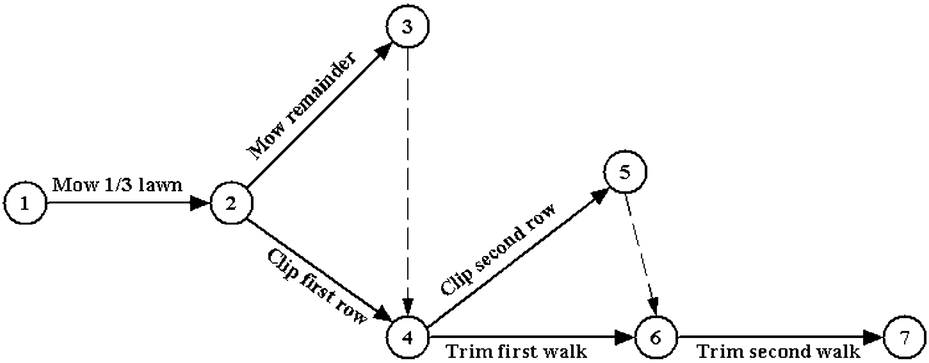


Figure 10 A PERT Diagram for a Simple Lawn-Maintenance Task.

development of the system. Traditionally, the development team will use the analysis phase of the SDLC to determine how the existing system functions and assess what the users of the new system would like to see in that system. RAD and JAD methods have emerged in an effort to remove or greatly reduce the time spent in the analysis phase of development. They are designed for development efforts that jump from planning directly to implementation, such as when software is acquired from a third party. RAD and JAD eliminate the detailed study required during the analysis phase and in doing so speed up user requirements analysis.

JAD is a structured approach whereby users, managers, and analysts work together for several days of intensive meetings to specify or review systems requirements. The coordination of these groups during these meetings determines whether the effort is successful. The groups must agree to work together to iron out any differences and have to work closely together to elicit requirements that might take months to emerge using traditional methods. RAD works by compressing the SDLC steps into four stages: planning, design, development, and cutover. Working only on the systems functions and user interface characteristics shortens planning and design. Then design and development work in an iterative fashion as the model is designed and implemented, then refined through further design and implementation until it produces the features that users require. RAD entails the use of a prototyping tool, which is a software tool that allows the rapid development of a new physical model that can be used and tested by the users. The distinction between prototyping and RAD as development approaches is that in many cases prototypes are never utilized in production environments. They are used to elicit requirements, but the actual prototype is then discarded in favor of a physical model produced with traditional SDLC methods. In RAD, the designed model will be implemented once it has reached a stage where no further refinement is necessary. That change would occur in the implementation phase.

The advantage of these approaches is in the speed with which systems can be developed. For smaller systems, or systems whose failure would not significantly affect the ability of the organization to engage in commerce, these approaches work very well. For systems that handle a significant volume of data, or where security and controls are a concern, these approaches are not advisable. The time saved using these approaches is usually gained at the expense of the attention paid to security, controls, and testing when the traditional development approach is used. Time is the usual hedge against the risk of systems failure: time in analysis and time in design. That time is sacrificed in these approaches.

5.2.7. CASE

Another way to speed systems development is to use computer aided software engineering (CASE) tools. In contrast to RAD or JAD, CASE tools are used with the traditional SDLC methodology. CASE aids the SDLC by using automation in place of manual methods. CASE tools support the activities in the SDLC by increasing productivity and improving the overall quality of the finished product. There are basically three different types of CASE tools. Upper CASE supports the planning, analysis and design phases of the SDLC. Upper CASE usually has planning and scheduling tools as well as DFD generators and data dictionary support. Lower CASE supports the latter parts of the SDLC, such as maintenance and implementation. Last come cross-life cycle CASE toolboxes, which support all the SDLC activities, from planning through maintenance.

There are many compelling reasons to use CASE tools. CASE tools can greatly shorten development times without sacrificing system quality. The tools usually enforce consistency across different phases by providing automated consistency checking between DFDs, ERDs, and the data dictionary to ensure that the attributes and data structures have been named and defined consistently across all of them. This, in turn, can increase the productivity of the SDLC team because a data store defined in a DFD can be automatically inserted as an entry in the data dictionary. This, in turn, can increase the consistency of the development effort across projects.

The disadvantages of CASE mostly revolve around cost. CASE tools tend to be very expensive, both in monetary and in resource terms. CASE tools cost a lot, and learning to use them can be very time consuming because of the number of features and the functionality that they provide. They also place great demands on the hardware in terms of processing needs and storage requirements at the server and workstation level. Essentially, all they do is automate manual processes, so the decision to use CASE hinges on the traditional tradeoff between labor costs and automation costs.

6. CONCLUSION

Over the past 35 years, tremendous strides have been made in management information systems, contributing greatly to organizations' abilities to grow, provide quality outputs, and operate efficiently. These strides have been fostered by continually improving hardware, advances in software tools for building ISs, formal methods for IS development, and progress in understanding the management of ISs. There is no sign that these trends are abating. Nor is there any sign that an organization's needs

TABLE 2 ABC Company Feasibility Analysis: Database Server Project

	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	Totals
Net Benefits	0	75,000	75,000	75,000	75,000	75,000	
Discount Rate	1.0000	0.90909091	0.82644628	0.7513148	0.68301346	0.62092132	
PV of Benefits	0	68,182	61,983	56,349	51,226	46,569	
NPV of all Benefits	0	68,182	130,165	186,514	237,740	284,309	$ 284,309
Sunk Costs	$ (100,000)						
	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	Totals
Net Costs	0	(22,000)	(21,000)	(23,000)	(24,000)	(25,000)	
Discount Rate	1.0000	0.90909091	0.82644628	0.7513148	0.68301346	0.62092132	
PV of Costs	0	(20,000)	(17,355)	(17,280)	(16,392)	(15,523)	
NPV of all Costs	0	(20,000)	(37,355)	(54,636)	(71,028)	(86,551)	$ (186,551)
Overall NPV							$ 97,758
Overall ROI							52%
Break Even Analysis							
Yearly NPV Cashflow	(100,000)	48,182	92,810	131,878	166,712	197,758	
Overall NPV Cashflow	(100,000)	(51,818)	40,992	172,870	339,582	537,340	

Break even is between year 1 and 2.

for ISs will disappear. Thus, even though their platforms, delivery methods, interfaces, and features will change, information systems will persist as indispensable elements of tomorrow's organizations, continuing to satisfy record keeping and reporting needs.

Integration will be a keynote as we look ahead. Integration is a collaborative interplay of components. It is operative at several levels. At the tool level, we should see a trend towards integration of traditionally distinct functions (Watabe et al., 1991). The resultant tools will enable more rapid IS development and the construction of systems having features not readily achieved with conventional means. At the IS level, we should expect to see a trend toward integration across IS classes. For instance, local ISs will increasingly be built to plug into (i.e., operate on the data of) functional and enterprise ISs; the modular ERP tools are designed for cross-functional IS integration; enterprise-wide ISs will increasingly exhibit transorganizational features. Within a very few years, all major ERP tools will be Internet-based and oriented toward collaborative activity.

At the level of business computing systems, the integration trend will manifest itself in efforts to incorporate decision support functionality into conventional management information systems. IS record keeping and reporting will still be essential for operational control, regulatory compliance, and legal purposes, but their records can be further leveraged for decision support. This can already be seen in ISs equipped with ad hoc query facilities. The use of IS records to build data warehouses that are subject to online analytical processing and data mining is another example in this direction. ERP vendors will increasingly offer decision support modules that work with their tools for building enterprise ISs (Holsapple and Sena 1999). Websites that handle transactions (create/update records and produce transaction reports) will increasingly incorporate facilities to support the decisions that underpin those transactions (Holsapple and Singh 2000).

At an organizational level, the rise of transorganizational ISs has been an important factor enabling and facilitating the appearance of network organizations and virtual corporations (Ching et al. 1996). Advances in supply chain, customer relationship, electronic commerce, and collaborative computing tools will add impetus to this trend. Its business driver is the increasingly dynamic and competitive nature of the global marketplace, in which collaboration and agility are vital. Dynamic, open business models are replacing those characterized by static trading partners.

Artificial intelligence technologies have long been recognized in the DSS realm, particularly in the guise of expert systems (Bonczek et al. 1980). It is likely that tools for building ISs will increasingly employ artificial intelligence technologies. This will lead to ISs that interact with users via speech (based on speech recognition/synthesis and natural language technologies), are self-maintaining (based on automated reasoning technologies), and can modify themselves in light of their experiences (based on machine learning and automatic program-generation technologies).

Application service providers will grow in importance as developers and hosts of information systems, particularly for small and mid-sized firms. These are firms that need an IS (e.g., for human resources functions) but cannot afford the infrastructure. For a rental fee, an application service provider (ASP) operates ISs for other companies on its own computing facilities (i.e., servers). ASP customers access their ISs via the Internet or private networks. ASPs specialize in various ways, such as hosting ISs for a particular industry (e.g., health care) or a particular class of ISs (e.g., transorganizational).

Information systems have become the centerpiece of the much-heralded Information Age. However, this 50-year era is rapidly being engulfed by the emerging knowledge economy, populated by knowledge workers and knowledge-based organizations (Holsapple and Whinston 1987). Information is being subsumed within the far richer notion of knowledge. The technology-driven aspect of this knowledge economy is variously known as the digital economy or the network economy, and its fabric is being defined by a multitude of electronic commerce initiatives (Holsapple et al. 2000). To be competitive in the knowledge economy, organizations (both traditional and virtual) strive to ensure that throughout their operations the right knowledge is available to the right processors (human and computer-based) at the right times in the right presentation formats for the right cost. Along with decision support systems, information systems built with tools such as those described here have a major role to play in making this happen.

REFERENCES

Bancroft, N., Seip, H. and Sprengel, A. (1998), *Implementing SAP R/3: How to Introduce a Large System into a Large Organization,* 2nd Ed., Manning, Greenwich, CT.

Barquín, R. C., and Edelstein, H. (1997), *Building, Using, and Managing the Data Warehouse,* Data Warehousing Institute Series, Prentice Hall PTR, Upper Saddle River, NJ.

Barry, M. (1995), "Getting a Grip on Data," *Progressive Grocer,* Vol. 74, No. 11, pp. 75–76.

Bonczek, R., Holsapple, C., and Whinston, A. (1980), "Future Directions for Decision Support Systems," *Decision Sciences,* Vol. 11, pp. 616–631.

Boudreau, M., and Robey, D. (1999), "Organizational Transition to Enterprise Resource Planning Systems: Theoretical Choices for Process Research," in *Proceedings of Twentieth International Conference on Information Systems* (Charlotte, NC), pp. 291–300.

Bowen, T. (1999), "SAP is Banking on the CRM Trend," *Infoworld,* November, pp. 74–76.

Brookshear, J. G. (1999), *Computer Science: An Overview,* Addison-Wesley, Reading, MA.

Brown, C., and Vessey, I. (1999), "ERP Implementation Approaches: Towards a Contingency Framework," in *Proceedings of Twentieth International Conference on Information Systems,* (Charlotte, NC), pp. 411–416.

Brown, S. (1995), "Try, Slice and Dice," *Computing Canada,* Vol. 21, No. 22, p. 44.

Caldwell, B., and Stein, T. (1998), "Beyond ERP: The New IT Agenda," *Information Week,* November 30, pp. 30–38.

Castro, E. (1999), *HTML 4 for the World Wide Web: Visual QuickStart Guide,* Peachpit Press, Berkeley, CA.

Ching, C., Holsapple, C., and Whinston, A. (1996), "Toward IT Support for Coordination in Network Organizations," *Information and Management,* Vol. 30, No. 4, pp. 179–199.

Curran, T., and Ladd, A. (2000), *SAP R/3 Business Blueprint: Understanding Enterprise Supply Chain Management,* 3rd Ed., Prentice Hall, PTR, Upper Saddle River, NJ.

Davenport, T. (1998), "Putting the Enterprise into the Enterprise System," *Harvard Business Review,* Vol. 76, July/August, pp. 121–131.

Davis, B. (1999), "PeopleSoft Fill out Analysis Suite," *Information Week,* Manhasset, RI, July 26.

Deutsch, C. (1998), "Software That Can Make a Grown Company Cry," *The New York Times,* November 8.

Doane, M. (1997), *In the Path of the Whirlwind: An Apprentice Guide to the World of SAP,* The Consulting Alliance, Sioux Falls, SD.

Eliason, A., and Malarkey, R. (1999), *Visual Basic 6: Environment, Programming, and Applications,* Que Education & Training, Indianapolis, IN.

Elmasri, R., and Navathe, S. B. (2000), *Fundamentals of Database Systems,* 3rd Ed., Addison-Wesley, Reading, MA.

Fogarty, K. (1994), "Data Mining Can Help to Extract Jewels of Data," *Network World,* Vol. 11, No. 23.

Francett, B. (1994), "Decisions, Decisions: Users Take Stock of Data warehouse Shelves," *Software Magazine,* Vol. 14, No. 8, pp. 63–70.

Goldbaum, L. (1999), "Another Day, Another Deal in CRM," *Forbes Magazine,* October, pp. 53–57.

Hackathorn, R. D. (1995), *Web Farming for the Data Warehouse,* Morgan Kaufmann Series in Data Management Systems, Vol. 41, Morgan Kaufmann, San Francisco, CA.

Hall, M. (1998), *Core Web Programming,* Prentice Hall PTR, Upper Saddle River, NJ.

Holsapple, C. (1995), "Knowledge Management in Decision Making and Decision Support," *Knowledge and Policy,* Vol. 8, No. 1, pp. 5–22.

Holsapple, C., and Sena, M. (1999), "Enterprise Systems for Organizational Decision Support," in *Proceedings of Americas Conference on Information Systems* (Milwaukee, WI), pp. 216–218.

Holsapple, C., and Singh, M. (2000), "Electronic Commerce: Definitional Taxonomy, Integration, and a Knowledge Management Link," *Journal of Organizational Computing and Electronic Commerce,* Vol. 10, No. 3.

Holsapple, C., and Whinston, A. (1987), "Knowledge-Based Organizations," *The Information Society,* Vol. 5, No. 2, pp. 77–90.

Holsapple, C. W., and Whinston, A. B. (1996), *Decision Support Systems: A Knowledge-Based Approach,* West, Mineapois/St. Paul.

Holsapple, C. W., Joshi, K. D., and Singh, M. (2000), "Decision Support Applications in Electronic Commerce," in *Handbook on Electronic Commerce,* M. Shawn, R. Blanning, T. Strader, and A. Whinston, Eds., Springer, Berlin, pp. 543–566.

Jetly, N. (1999), "ERP's Last Mile," *Intelligent Enterprise,* Vol. 2, No. 17, December, pp. 38–45.

Kirkpatrick, D. (1998), "The E-Ware War: Competition Comes to Enterprise Software," *Fortune,* December 7, pp. 102–112.

Main, M., and Savitch, W. (1997), *Data Structures and Other Objects Using C++,* Addison-Wesley, Chicago.

Marion, L. (1999a), "Big Bang's Return," *Datamation,* October, pp. 43–45.

Marion, L. (1999b), "ERP Transactions from Anywhere," *Datamation,* August, pp. 65–69.

Mattison, R. (1996), "Warehousing Wherewithal," *CIO,* Vol. 9, No. 12, pp. 58–60.

McFadden, F. R., Hoffer, J. A., and Prescott, M. B. (1999), *Modern Database Management,* 5th Ed., Addison-Wesley, Reading, MA.

McLeod, R. J. (1998), *Management Information Systems,* 7th Ed., Prentice Hall, Upper Saddle River, NJ.

Norris, G., Wright, I., Hurley, J. R., Dunleavy, J., and Gibson, A. (1998), *SAP: An Executive's Comprehensive Guide,* John Wiley & Sons, New York.

Rob, P., and Coronel, C. (1997), *Database Systems: Design, Implementation, and Management,* Course Technology, Cambridge, MA.

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E. (2000), *Designing and Managing the Supply Chain,* Irwin/McGraw-Hill, Boston.

Singh, H. (1998), *Data Warehousing: Concepts, Technologies, Implementations, and Management,* Prentice Hall PTR, Upper Saddle River, NJ.

Stedman, C. (1999a), "ERP Guide: Vendor Strategies, Future Plans," *Computerworld,* July 19, pp. 62–64.

Stedman, C. (1999b), "Update: Failed ERP Gamble Haunts Hershey," *Computerworld,* November 18, pp. 34–37.

Stroud, J. (1999), "Oracle Suite Integrates ERP, E-Business Apps," *Internetweek,* October 4, pp. 1–2.

Stroustrup, B. (1994), *The Design and Evolution of C++,* Addison-Wesley, Reading, MA.

Stroustrup, B. (1997), *The C++ Programming Language,* Addison-Wesley, Reading, MA.

Sykes, R. (1999), "Oracle Readies Customer Apps for SAP R/3," IDG News Service.

Watabe, K., Holsapple, C., and Whinston, A. (1991), "Solving Complex Problems via Software Integration," *Journal of Computer Information Systems,* Vol. 31, No. 3, pp. 2–8.

Welti, N. (1999), *Successful SAP R/3 Implementation: Practical Management of ERP Projects,* Addison-Wesley Longman, Reading, MA.

Wu, S., and Wu, M. (1994), *Systems Analysis and Design,* West, Minneapolis/St. Paul.

Zerega, B. (1999), "CRM Rise to the Top," *Red Herring Magazine,* July, pp. 41–44.

CHAPTER 4

Decision Support Systems

ANDREW P. SAGE
George Mason University

1. INTRODUCTION 110
  1.1. Taxonomies for Decision Support 111
  1.2. Frameworks for the Engineering of Decision Support Systems 113
2. DATABASE MANAGEMENT SYSTEMS 114
  2.1. DBMS Design, Selection, and Systems Integration 116
  2.2. Data Models and Database Architectures 117
    2.2.1. Record-Based Models 120
    2.2.2. Structural Models 120
    2.2.3. Expert and Object-Oriented Database Models 122
  2.3. Distributed and Cooperative Databases 124
3. MODEL BASE MANAGEMENT SYSTEMS 125
  3.1. Models and Modeling 126
    3.1.1. Issue, or Problem, Formulation Models 126
    3.1.2. Models for Issue Analysis 127
    3.1.3. Issue Interpretation Models 129
  3.2. Model Base Management 129
4. DIALOG GENERATION AND MANAGEMENT SYSTEMS 131
5. GROUP AND ORGANIZATIONAL DECISION SUPPORT SYSTEMS 134
  5.1. Information Needs for Group and Organizational Decision Making 135
  5.2. The Engineering of Group Decision Support Systems 141
  5.3. Distributed Group Decision Support Systems 145
6. KNOWLEDGE MANAGEMENT FOR DECISION SUPPORT 145
REFERENCES 149

1. INTRODUCTION

In very general terms, a decision support system (DSS) is a system that supports technological and managerial decision making by assisting in the organization of knowledge about ill-structured, semistructured, or unstructured issues. For our purposes, a structured issue is one that has a framework with elements and relations between them that are known and understood within a context brought by the experiential familiarity of a human assessing the issue or situation. The three primary components of a decision support system are generally described as:

- *A database management system* (DBMS)
- *A model-based management system* (MBMS)
- *A dialog generation and management system* (DGMS)

The emphasis in the use of a DSS is upon providing support to decision makers to increase the effectiveness of the decision making effort. This need should be a major factor in the definition of

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

requirements for and subsequent development of a DSS. A DSS is generally used to support humans in the formal steps of problem solving, or systems engineering (Sage 1992, 1995; Sage and Rouse 1999a; Sage and Armstrong 2000), that involve:

- *Formulation* of alternatives
- *Analysis* of their impacts
- *Interpretation* and selection of appropriate options for implementation

Efficiency in terms of time required to evolve the decision, while important, is usually secondary to effectiveness. DSSs are intended more for use in strategic and tactical situations than in operational situations. In operational situations, which are often well structured, an expert system or an accounting system may often be gainfully employed to assist novices or support the myriads of data elements present in these situations. Those very proficient in (experientially familiar with) operational tasks generally do not require support, except perhaps for automation of some routine and repetitive chores. In many application areas the use of a decision support system is potentially promising, including management and planning, command and control, system design, health care, industrial management, and generally any area in which management has to cope with decision situations that have an unfamiliar structure.

1.1. Taxonomies for Decision Support

Numerous disciplinary areas have contributed to the development of decision support systems. One area is computer science, which provides the hardware and software tools necessary to implement decision support system design constructs. In particular, computer science provides the database design and programming support tools that are needed in a decision support system. Management science and operations research provides the theoretical framework in decision analysis that is necessary to design useful and relevant normative approaches to choice making, especially those concerned with systems analysis and model base management. Organizational behavior and behavioral and cognitive science provide rich sources of information concerning how humans and organizations process information and make judgments in a descriptive fashion. Background information from these areas is needed for the design of effective systems for dialog generation and management systems. Systems engineering is concerned with the process and technology management issues associated with the definition, development, and deployment of large systems of hardware and software, including systems for decision support.

Many attempts have been made to classify different types of decisions. Of particular interest here is the decision type taxonomy of Anthony, which describes four types of decisions (Anthony 1965; Anthony et al. 1992):

1. *Strategic planning decisions* are decisions related to choosing highest level policies and objectives, and associated resource allocations.
2. *Management control decisions* are decisions made for the purpose of ensuring effectiveness in the acquisition and use of resources.
3. *Operational control decisions* are decisions made for the purpose of ensuring effectiveness in the performance of operations.
4. *Operational performance decisions* are the day-to-day decisions made while performing operations.

Figure 1 illustrates how these decisions are related and how they normatively influence organizational learning. Generally, low-consequence decisions are made more frequently than high-consequence decisions. Also, strategic decisions are associated with higher consequences and are likely to involve more significant risk and therefore must be made on the basis of considerably less perfect information than are most operational control decisions.

A decision support system should support a number of abilities. It should support the decision maker in the formulation or framing or assessment of the decision situation in the sense of recognizing needs, identifying appropriate objectives by which to measure successful resolution of an issue, and generating alternative courses of action that will resolve the needs and satisfy objectives. It should also provide support in enhancing the decision maker's abilities to assess the possible impacts of these alternatives on the needs to be fulfilled and the objectives to be satisfied. This analysis capability must be associated with provision of capability to enhance the ability of the decision maker to provide an interpretation of these impacts in terms of objectives. This interpretation capability will lead to evaluation of the alternatives and selection of a preferred alternative option. These three steps of formulation, analysis, and interpretation are fundamental for formal analysis of difficult issues. They are the fundamental steps of systems engineering and are discussed at some length in Sage (1991,

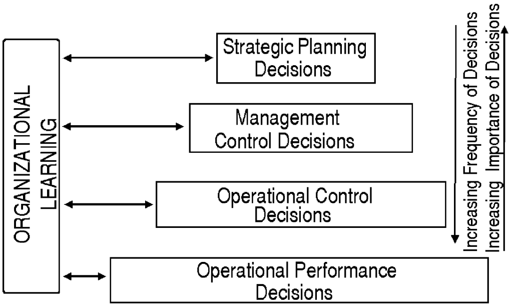

ORGANIZATIONAL LEARNING
Strategic Planning Decisions
Management Control Decisions
Operational Control Decisions
Operational Performance Decisions
Increasing Frequency of Decisions
Increasing Importance of Decisions

Figure 1 Organizational Information and Decision Flow.

1992, 1995), from which much of this chapter is derived. It is very important to note that the purpose of a decision support system is to support humans in the performance of primarily cognitive information processing tasks that involve decisions, judgments, and choices. Thus, the enhancement of information processing in systems and organizations (Sage 1990) is a major feature of a DSS. Even though there may be some human supervisory control of a physical system through use of these decisions (Sheridan 1992), the primary purpose of a DSS is support for cognitive activities that involve human information processing and associated judgment and choice. Associated with these three steps must be the ability to acquire, represent, and utilize information or knowledge and the ability to implement the chosen alternative course of action.

The extent to which a support system possesses the capacity to assist a person or a group to formulate, analyze, and interpret issues will depend upon whether the resulting system should be called a management information system (MIS), a predictive management information system (PMIS), or a decision support system. We can provide support to the decision maker at any of these several levels, as suggested by Figure 2. Whether we have a MIS, a PMIS, or a DSS depends upon the type of automated computer-based support that is provided to the decision maker to assist in reaching the decision. Fundamental to the notion of a decision support system is assistance provided in assessing the situation, identifying alternative courses of action and formulating the decision situation, structuring and analyzing the decision situation, and then interpreting the results of analysis of the alternatives in terms of the value system of the decision maker. In short, a decision support system provides a decision recommendation capability. A MIS or a PMIS does not, although the information provided may well support decisions.

In a classical management information system, the user inputs a request for a report concerning some question, and the MIS supplies that report. When the user is able to pose a “what if?” type question and the system is able to respond with an “if then” type of response, then we have a

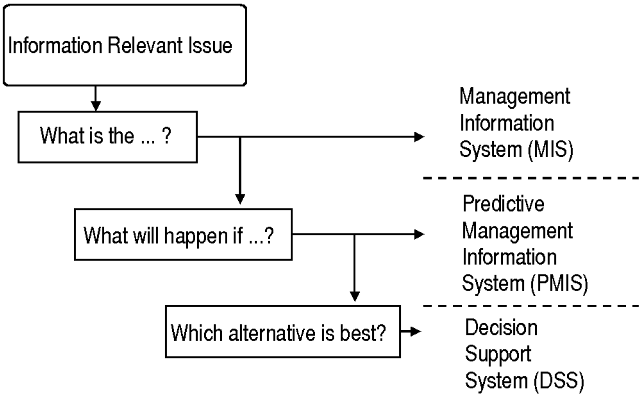

Information Relevant Issue
What is the ... ?
Management Information System (MIS)
What will happen if ...?
Predictive Management Information System (PMIS)
Which alternative is best?
Decision Support System (DSS)

Figure 2 Conceptual Differences Between MIS, PMIS, and DSS.

predictive management information system. In each case there is some sort of formulation of the issue, and this is accompanied by some capacity for analysis. The classic MIS, which needs only to be able to respond to queries with reports, is composed of capabilities for data processing, structured data flows at an operational level, and preparation of summary reports for the system user. The predictive management system would also include an additional amount of analysis capability. This might require an intelligent database query system, or perhaps just the simple use of some sort of spreadsheet or macroeconomic model.

To obtain a decision support system, we would need to add the capability of model-based management to a MIS. But much more is needed, for example, than just the simple addition of a set of decision trees and procedures to elicit examination of decision analysis-based paradigms. We also need a system that is flexible and adaptable to changing user requirements such as to provide support for the decision styles of the decision maker as these change with task, environment, and experiential familiarity of the support system users with task and environment. We need to provide analytical support in a variety of complex situations. Most decision situations are fragmented in that there are multiple decision makers and their staffs, rather than just a single decision maker. Temporal and spatial separation elements are also involved. Further, as Mintzberg (1973) has indicated so very well, managers have many more activities than decision making to occupy themselves with, and it will be necessary for an appropriate DSS to support many of these other information-related functions as well. Thus, the principal goal of a DSS is improvement in the effectiveness of organizational knowledge users through use of information technology. This is not a simple objective to achieve as has been learned in the process of past DSS design efforts.

1.2. Frameworks for the Engineering of Decision Support Systems

An appropriate decision support system design framework will consider each of the three principal components of decision support systems—a DBMS, an MBMS, and a DGMS—and their interrelations and interactions. Figure 3 illustrates the interconnection of these three generic components and illustrates the interaction of the decision maker with the system through the DGMS. We will describe some of the other components in this figure soon.

Sprague and Carlson (1992), authors of an early, seminal book on decision support systems, have indicated that there are three technology levels at which a DSS may be considered. The first of these is the level of DSS tools themselves. This level contains the hardware and software elements that enable use of system analysis and operations research models for the model base of the DSS and the database elements that comprise the database management system. The purpose of these DSS tools is to design a *specific DSS* that is responsive to a particular task or issue. The second level is that of a decision support system generator. The third level is the specific DSS itself. The specific DSS may be designed through the use of the DSS tools only, or it may be developed through use of a generic DSS generator that may call upon elements in the generic MBMS and DBMS tool repository for use in the specific DSS.

Often the best designers of a decision support system are not the specialists primarily familiar with DSS development tools. The principal reason for this is that it is difficult for one person or small group to be very familiar with a great variety of tools as well as to be able to identify the

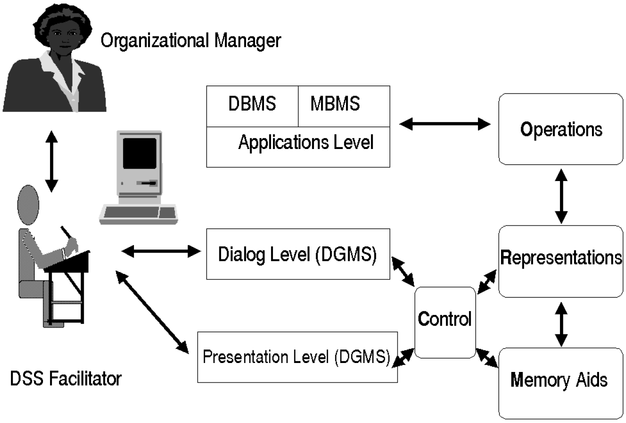


Figure 3 Generic Components in a Decision Support System, Including the ROMC Elements.

requirements needed for a specific DSS and the systems management skills needed to design a support process. This suggests that the decision support generator is a potentially very useful tool, in fact a design level, for DSS system design. The DSS generator is a set of software products, similar to a very advanced generation system development language, which enables construction of a specific DSS without the need to formally use micro-level tools from computer science and operations research and systems analysis in the initial construction of the specific DSS. These have, in effect, already been embedded in the DSS generator. A DSS generator contains an integrated set of features, such as inquiry capabilities, modeling language capabilities, financial and statistical (and perhaps other) analysis capabilities, and graphic display and report preparation capabilities. The major support provided by a DSS generator is that it allows the rapid construction of a prototype of the decision situation and allows the decision maker to experiment with this prototype and, on the basis of this, to refine the specific DSS such that it is more representative of the decision situation and more useful to the decision maker. This generally reduces, often to a considerable degree, the time required to engineer and implement a DSS for a specific application. This notion is not unlike that of software prototyping, one of the principal macro-enhancement software productivity tools (Sage and Palmer 1990) in which the process of constructing the prototype DSS through use of the DSS generator leads to a set of requirements specifications for a DSS that are then realized in efficient form using DSS tools directly.

The primary advantage of the DSS generator is that it is something that the DSS designer can use for direct interaction with the DSS user group. This eliminates, or at least minimizes, the need for DSS user interaction with the content specialists most familiar with micro-level tools of computer science, systems analysis, and operations research. Generally, a potential DSS user will seldom be able to identify or specify the requirements for a DSS initially. In such a situation, it is very advantageous to have a DSS generator that may be used by the DSS engineer, or developer, in order to obtain prototypes of the DSS. The user may then be encouraged to interact with the prototype in order to assist in identifying appropriate requirements specifications for the evolving DSS design.

The third level in this DSS design and development effort results from adding a decision support systems management capability. Often, this will take the form of the dialog generation and management subsystem referred to earlier, except perhaps at a more general level since this is a DGMS for DSS design and engineering rather than a DGMS for a specific DSS. This DSS design approach is not unlike that advocated for the systems engineering of large scale systems in general and DSS in particular.

There are many potential difficulties that affect the engineering of trustworthy systems. Among these are: inconsistent, incomplete, and otherwise imperfect system requirements specifications; system requirements that do not provide for change as user needs evolve over time, and poorly defined management structures. The major difficulties associated with the production of trustworthy systems have more to do with the organization and management of complexity than with direct technological concerns. Thus, while it is necessary to have an appropriate set of quality technological methods and tools, it is also very important that they be used within a well chosen set of lifecycle processes and set of systems management strategies that guide the execution of these processes (Sage 1995; Sage and Rouse 1999a).

Because a decision support system is intended to be used by decision makers with varying experiential familiarity and expertise with respect to a particular task and decision situation, it is especially important that a DSS design consider the variety of issue representations or frames that decision makers may use to describe issues, the operations that may be performed on these representations to enable formulation analysis and interpretation of the decision situation, the automated memory aids that support retention of the various results of operations on the representations, and the control mechanisms that assist decision makers in using these representations, operations, and memory aids. A very useful control mechanism results in the construction of heuristic procedures, perhaps in the form of a set of production rules, to enable development of efficient and effective standard operating policies to be issued as staff directives. Other control mechanisms are intended to encourage the decision maker to personally control and direct use of the DSS and also to acquire new skills and rules based on the formal reasoning-based knowledge that is called forth through use of a decision support system. This process independent approach toward development of the necessary capabilities of a specific DSS is due to Sprague and Carlson (1982) and is known as the ROMC approach (representations, operations, memory aids, and control mechanisms). Figure 3 illustrates the ROMC elements, together with the three principal components (DBMS, MBMS, and DGMS) of a DSS.

2. DATABASE MANAGEMENT SYSTEMS

As we have noted, a database management system is one of the three fundamental technological components in a decision support system. Figure 3 indicates the generic relationship among these components, or subsystems. We can consider a database management system as composed of a

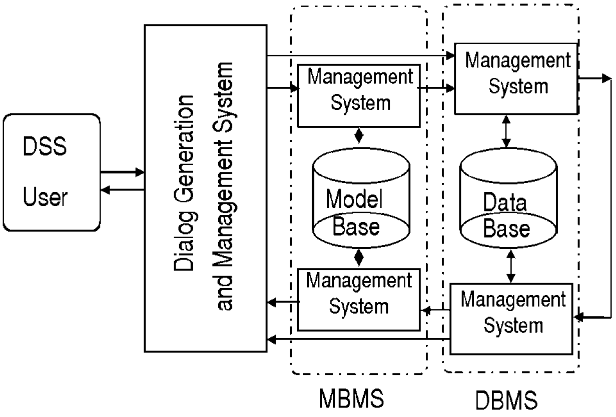

DSS
User
Dialog Generation
and Management System
Management
System
Model
Base
Management
System
Management
System
Data
Base
Management
System
MBMS
DBMS

Figure 4 Expanded View of DSS Indicating Database and Management System for DBMS and Model Base and Management System for MBMS.

database (DB) and a management system (MS). This sort of expansion holds for the model base management system also. Figure 4 indicates the resulting expanded view of a DSS. The DBMS block itself can be expanded considerably, as can the other two subsystems of a DSS. Figure 5 indicates one possible expansion to indicate many of the very useful components of a database management system that we will briefly examine later in this section.

Three major objectives for a DBMS are data independence, data redundancy reduction, and data resource control, such that software to enable applications to use the DBMS and the data processed are independent. If this is not accomplished, then such simple changes to the data structure as adding four digits to the ZIP code number in an address might require rewriting many applications programs. The major advantage to data independence is that DSS developers do not need to be explicitly concerned with the details of data organization in the computer or how to access it explicitly for information processing purposes, such as query processing. Elimination or reduction of data redundancy will assist in lessening the effort required to make changes in the data in a database. It may also assist, generally greatly, in eliminating the inconsistent data problem that often results from updating data items in one part of a database but (unintentionally) not updating this same data that

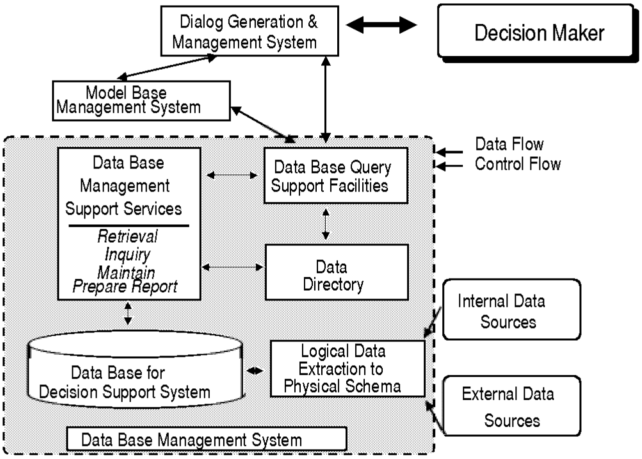

Dialog Generation &
Management System
Decision Maker
Model Base
Management System
Data Flow
Control Flow
Data Base
Management
Support Services
Retrieval
Inquiry
Maintain
Prepare Report
Data Base Query
Support Facilities
Data
Directory
Internal Data
Sources
Data Base for
Decision Support System
Logical Data
Extraction to
Physical Schema
External Data
Sources
Data Base Management System

Figure 5 Generic Components in Decision Support System, with Expansion of Database Management System.

exists elsewhere in the database because of data redundancy. With many people potentially using the same database, resource control is essential. The presence of these three features in a DBMS is the major factor that differentiates it from a file management system.

The management system for a database is composed of the software that is beneficial to the creation, access, maintenance, and updating of a database. A database contains data that are of value to an individual or an organization and that an individual or an organization desires to maintain. Maintenance should be such that the data survive even when the DBMS system hardware and/or software fails. Maintenance of data is one of the important functions provided for in DBMS design.

There are many tasks that we desire to perform using a database. These five are especially important:

1. *Capturing* relevant data for use in the database
2. *Selecting* relevant data from the database
3. *Aggregating* data to form totals, averages, moments, and other items that support decision making
4. *Estimating, forecasting, and predicting* in order to obtain extrapolations of events into the future, and such other activities as
5. *Optimizing* in order to enable selection of a ‘‘best’’ alternative

We note that these database raise issues that are associated with the model base management system. In a similar way, the dialog generation and management system determines how data is viewed and is therefore also important for use of a DBMS.

2.1. DBMS Design, Selection, and Systems Integration

As with any information systems engineering-based activity, DBMS design should start with an identification of the DBMS design situation and the user requirements. From this, identification of the logical or conceptual data requirements follows specification of a logical database structure in accordance with what is known as a data definition language (DDL). After this, the physical database structure is identified. This structure must be very concerned with specific computer hardware characteristics and efficiency considerations. Given these design specifications, an operational DBMS is constructed and DBMS operation and maintenance efforts begin. The logical database design and physical database design efforts can each be disaggregated into a number of related activities. The typical database management system lifecycle will generally follow one of the systems engineering development life cycles discussed in the literature (Sage 1992, 1995; Sage and Rouse 1999a).

The three important DBMS requirements—data independence, redundancy reduction, and increased data resource control—are generally applicable both to logical data and to physical data. It is highly desirable, for example, to be able to change the structure of the physical database without affecting other portions of the DBMS. This is called physical data independence. In a similar way, logical data independence denotes the ability of software to function using a given applications-oriented perspective on the database even though changes in other parts of the logical structure have been made. The requirements specification, conceptual design, logical design, and physical design phases of the DBMS development life cycle are specifically concerned with satisfaction of these requirements.

A number of questions need to be asked and answered successfully in order to design an effective DBMS. Among these are the following (Arden 1980):

1. Are there data models that are appropriate across a variety of applications?
2. What DBMS designs that enable data models to support logical data independence?
3. What DBMS designs enable data models to support logical data independence, and what are the associated physical data transformations and manipulations?
4. What features of a data description language will enable a DBMS designer to control independently both the logical and physical properties of data?
5. What features need to be incorporated into a data description language in order to enable errors to be detected at the earliest possible time such that users will not be affected by errors that occur at a time prior to their personal use of the DBMS?
6. What are the relationships between data models and database security?
7. What are the relationships between data models and errors that may possibly be caused by concurrent use of the database by many users?
8. What are design principles that will enable a DBMS to support a number of users having diverse and changing perspectives?
9. What are the appropriate design questions such that applications programmers, technical users, and DBMS operational users are each able to function effectively and efficiently?

The bottom line question that summarizes all of these is, How does one design a data model and data description language to enable efficient and effective data acquisition, storage, and use? There are many related questions; one concerns the design of what are called standard query languages (SQLs) such that it is possible to design a specific DBMS for a given application.

It is very important that, whenever possible, a specific DBMS be selected prior to design of the rest of the DSS. There are a variety of reasons why this is quite desirable. The collection and maintenance of the data through the DSS are simplified if there is a specified single DBMS structure and architecture. The simplest situation of all occurs when all data collection (and maintenance) is accomplished prior to use of the DSS. The DBMS is not then used in an interactive manner as part of DSS operation. The set of database functions that the DSS needs to support is controlled when we have the freedom to select a single DBMS structure and architecture prior to design of the rest of the DSS. The resulting design of the DSS is therefore simplified. Further, the opportunities for data sharing among potentially distributed databases are increased when the interoperability of databases is guaranteed. Many difficulties result when this is not possible. Data in a DBMS may be classified as *internal data,* stored in an internal database, and *external data,* stored in an external database. Every individual and organization will necessarily have an internal database. While no problem may exist in ensuring DBMS structure and architecture compatibility for internal data, this may be very difficult to do for external data. If both of these kinds of data can be collected and maintained prior to use of a DSS, then there will generally be no data integration needs. Often, however, this will not be possible. Because we will often be unable to control the data structure and architecture of data obtained externally, the difficulties we cite will often be real, especially in what are commonly called real-time, interactive environments. When we have DBMSs that are different in structure and architecture, data sharing across databases is generally difficult, and it is often then necessary to maintain redundant data. This can lead to a number of difficulties in the systems integration that is undertaken to ensure compatibility across different databases.

In this chapter, as well as in much DSS and information system design in practice, we may generally assume that the DBMS is preselected prior to design of the rest of the DSS and that the same DBMS structure and architecture are used for multiple databases that may potentially be used in the DSS. This is often appropriate, for purposes of DSS design, since DBMS design technology (Rob and Coronel 1997; Kroenke 1998; Date 1999; Purba 1999; Atzeni et al. 2000) is now relatively mature in contrast to MBMS and DGMS design. This does not suggest that evolution and improvements to DBMS designs are not continuing. It is usually possible, and generally desirable, to select a DBMS, based on criteria we will soon discuss, and then design the MBMS and DGMS based on the existing DBMS. This will often, but surely not always, allow selection of commercial off-the-shelf (COTS) DBMS software. The alternate approach of designing a MBMS and DGMS first and then specifying the requirements for a DBMS based on these is possible and may in some cases be needed. Given the comparatively developed state of DBMS software development as contrasted with MBMS and DGMS software, this approach will usually be less desirable from the point of view of design economy in engineering the DSS.

2.2. Data Models and Database Architectures

We will now expand very briefly on our introductory comments concerning data model representations and associated architectures for database management systems. Some definitions are appropriate. A data model defines the types of data objects that may be manipulated or referenced within a DBMS. The concept of a logical record is a central one in all DBMS designs. Some DBMS designs are based on mathematical relations, and the associated physical database is a collection of consistent tables in which every row is a record of a given type. This is an informal description of a relational database management system, a DBMS type that is a clear outgrowth of the file management system philosophy. Other DBMS may be based on hierarchical or network structures, which resemble the appearance of the data in the user's world and may contain extensive graphical and interactive support characteristics.

There are several approaches that we might take to describing data models. Date (1983, 1999), for example, discusses a three-level representation in which the three levels of data models are:

1. An external model, which represents a data model at the level of the user's application and is the data model level closest and most familiar to users of a DBMS or DSS
2. A conceptual model, which is an aggregation model that envelopes several external models
3. An internal model, which is a technical-level model that describes how the conceptual model is actually represented in computer storage

Figure 6 is a generic diagram that indicates the mappings and data translations needed to accommodate the different levels of data models and architectures. The relations between the various levels, called mappings. These specify and describe the transformations that are needed in order to obtain one model from another. The user supplies specifications for the source and target data structures in

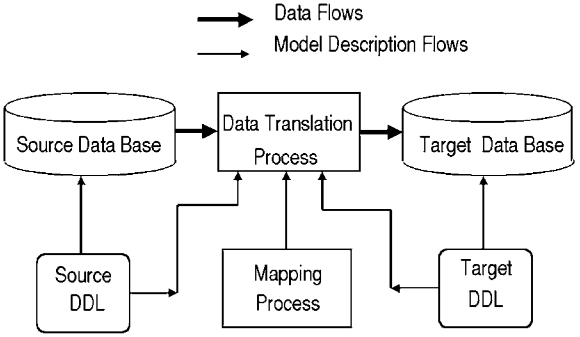


Figure 6 Data Transforms and Model Mappings.

a data description language (DDL) and also describes the mapping that is desired between source and target data. Figure 7 represents this general notion of data transformation. This could, for example, represent target data in the form of a table that is obtained from a source model composed of a set of lists.

Figure 7 is a very simple illustration of the more general problem of mapping between various schemas. Simply stated, a schema is an image used for comparison purposes. A schema can also be described as a data structure for representing generic concepts. Thus, schemas represent knowledge about concepts and are structurally organized around some theme. In terms appropriate here, the user of a database must interpret the real world that is outside of the database in terms of real-world objects, or entities and relationships between entities, and activities that exist and that involve these objects. The database user will interact with the database in order to obtain needed data for use or, alternatively, to store obtained data for possible later use. But a DBMS cannot function in terms of real objects and operations. Instead, a DBMS must use data objects, composed of data elements and relations between them, and operations on data objects. Thus, a DBMS user must perform some sort of mapping or transformation from perceived real objects and actions to those objects' and actions' representations that will be used in the physical database.

The single-level data model, which is conceptually illustrated in Figure 7, represents the nature of the data objects and operations that the user understands when receiving data from the database. In this fashion the DBMS user models the perceived real world. To model some action sequence, or the impact of an action sequence on the world, the user maps these actions to a sequence of operations that are allowed by the specific data model. It is the data manipulation language (DML) that provides the basis for the operations submitted to the DBMS as a sequence of queries or programs. The development of these schemas, which represent logical data, results in a DBMS architecture or DBMS framework, which describes the types of schemas that are allowable and the way in which these schemas are related through various mappings. We could, for example, have a two-schema framework or a three-schema framework, which appears to be the most popular representation at this time. Here

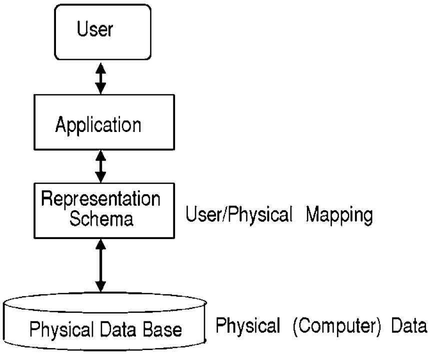


Figure 7 Single-Level Model of Data Schema.

the external schemas define the logical subset of the database that may be presented to a specific DBMS user. The conceptual schemas define conceptual models as represented in the database. The internal schema describes physical data structures in the database. These data structures provide support to the conceptual models. The mappings used establish correspondence between various objects and operations at the three levels. If consistency of the conceptual models is to be ensured, we must be able to map the internal, or physical, and external, or logical, schemas to the conceptual schemas.

Useful documentation about the database structure and architecture is provided through the various schemas, which represent explicit data declarations. These declarations represent data about data. The central repository in which these declarations are kept is called a data dictionary or data directory, and is often represented by the symbol DD. The data directory is a central repository for the definitions of the schemas and the mappings between schemas. Generally, a data dictionary can be queried in the same manner as can the database, thereby enhancing the ability of the DBMS user to pose questions about the availability and structure of data. It is often possible to query a data directory with a high-level, or fourth-generation, query language.

A data dictionary is able to tell us what the records in the data dictionary consist of. It also contains information about the logical relationships that pertain to each of the particular elements in the database. The development of a data dictionary generally begins with formation of lists of desired data items or fields that have been grouped in accordance with the entities that they are to represent. A name is assigned for each of these lists, and a brief statement of the meaning of each is provided for later use. Next, the relationships between data elements should be described and any index keys or pointers should be determined. Finally, the data dictionary is implemented within the DBMS. In many ways, a data dictionary is the central portion of a DBMS. It performs the critical role of retaining high-level information relative to the various applications of the DBMS and thereby enables specification, control, review, and management of the data of value for decision making relative to specific applications.

For very large system designs, it is necessary that the data dictionary development process be automated. A typical data dictionary for a large system may include several thousand entries. It is physically impossible to manually maintain a dictionary of this size or to retain consistent and unambiguous terms for each data element or composite of data elements. Therefore, automated tools are needed for efficient and effective development and maintenance of a data dictionary. These are provided in contemporary DBMSs.

This ability to specify database structures, through use of schemas, enables efficient and effective management of data resources. This is necessary for access control. The use of one or more subschemas generally simplifies access to databases. It may also provide for database security and integrity for authorized users of the DBMS. Since complex physical (computer) structures for data may be specified, independent of the logical structure of the situation that is perceived by DBMS users, it is thereby possible to improve the performance of an existing and operational database without altering the user interface to the database. This provides for simplicity in use of the DBMS.

The data model gives us the constructs that provide a foundation for the development and use of a database management system. It also provides the framework for such tools and techniques as user interface languages. The user of these languages is often called a database administrator (DBA). There are three types of user interface languages:

1. *Data definition languages* (DDLs) provide the basis for definition of schemas and subschemas.
2. *Data manipulation languages* (DMLs) are used to develop database applications.
3. *Data query languages* (DQLs) or simply *query languages* (QLs) are used to write queries and reports. Of course, it is possible to combine a DDL, DML, and DQL into a single database language.

In general, a data model is a paradigm for representation, storing, organizing, and otherwise managing data in a database. There are three component sets in most data models:

1. A set of data structures that define the fields and records that are allowed in the database. Examples of data structures include lists, tables, hierarchies, and networks.
2. A set of operations that define the admissible manipulations that are applied to the fields and records that comprise the data structures. Examples of operations include *retrieve, combine, subtract, add,* and *update*.
3. A set of integrity rules that define or constrain allowable or legal states or changes of state for the data structures that must be protected by the operations.

There are three generically different types of data model representations and several variants within these three representations. Each of these is applicable as an internal, external, or conceptual model. We will be primarily concerned with these three modeling representations as they affect the external, or logical, data model. This model is the one with which the user of a specific decision support system interfaces. For use in a decision support system generator, the conceptual model is of importance because it is the model that influences the specific external model that various users of a DSS will interface with after an operational DSS is realized. This does not mean that the internal data model is unimportant. It is very important for the design of a DBMS. Through its data structures, operations, and integrity constraints, the data model controls the operation of the DBMS portion of the DSS.

The three fundamental data models for use in the external model portion of a DSS are record-based models, structurally based models, and expert system-based models. Each of these will now be briefly described.

2.2.1. Record-Based Models

We are very accustomed to using forms and reports, often prepared in a standard fashion for a particular application. Record-based models are computer implementations of these spreadsheet-like forms. Two types can be identified. The first of these, common in the early days of file processing systems (FPSs) or file management systems (FMSs), is the *individual record model*. This is little more than an electronic file drawer in which records are stored. It is useful for a great many applications. More sophisticated, however, is the relational database data model, in which mathematical relations are used to electronically ‘‘cut and paste’’ reports from a variety of files. Relational database systems have been developed to a considerable degree of sophistication, and many commercial products are available. Oracle and Informix are two leading providers.

The individual record model is surely the oldest data model representation. While the simple single-record tables characteristic of this model may appear quite appealing, the logic operations and integrity constraints that need to be associated with the data structure are often undefined and are perhaps not easily defined. Here, the data structure is simply a set of records, with each record consisting of a set of fields. When there is more than a single type of record, one field contains a value that indicates what the other fields in the record are named.

The relational model is a modification of the individual record model that limits its data structures and thereby provides a mathematical basis for operation on records. Data structures in a relational database may consist only of relations, or field sets that are related. Every relation may be considered as a table. Each row in the table is a record or tuple. Every column in each table or row is a field or attribute. Each field or attribute has a domain that defines the admissible values for that field.

Often there is only a modest difference in structure between this relational model and the individual record model. One difference is that fields in the various records of the individual record model represent relationships, whereas relationships among fields or attributes in a relational model are denoted by the name of the relation.

While the structural differences between the relational model and the individual record model are minimal, there are major differences in the way in which the integrity constraints and operations may affect the database. The operations in a relational database form a set of operations that are defined mathematically. The operations in a relational model must operate on entire relations, or tuples, rather than only on individual records. The operations in a relational database are independent of the data structures, and therefore do not depend on the specific order of the records or the fields. There is often controversy about whether or not a DBMS is truly relational. While there are very formal definitions of a relational database, a rather informal one is sufficient here. A relational database is one described by the following statements.

1. Data are presented in tabular fashion without the need for navigation links, or pointer structures, between various tables.
2. A relational algebra exists and can be used to prepare joins of logical record files automatically.
3. New fields can be added to the database without the necessity of rewriting any programs that used previous versions of the database.

If a DBMS does not satisfy these three criteria, then it is almost surely NOT a relational database.

2.2.2. Structural Models

In many instances, data are intended to be associated with a natural structural representation. A typical hierarchical structure is that of a classic organization. A more general representation than a hierarchy, or tree, is known as a *network*. It is often possible to represent a logical data model with a hierarchical data structure. In a hierarchical model, we have a number of nodes that are connected by links. All links are directed to point from ‘‘child’’ to ‘‘parent,’’ and the basic operation in a hierarchy is that of searching a tree to find items of value. When a query is posed with a hierarchical database, all branches of the hierarchy are searched and those nodes that meet the conditions posed in the query are noted and then returned to the DBMS system user in the form of a report.

Some comparisons of a hierarchical data model with a relational data model are of interest here. The structures in the hierarchical model represent the information that is contained in the fields of the relational model. In a hierarchical model, certain records must exist before other records can exist. The hierarchical model is generally required to have only one key field. In a hierarchical data model, it is necessary to repeat some data in a descendant record that need be stored only once in a relational database regardless of the number of relations. This is so because it is not possible for one record to be a descendant of more than one parent record. There are some unfortunate consequences of the mathematics involved in creating a hierarchical tree, as contrasted with relations among records. Descendants cannot be added without a root leading to them, for example. This leads to a number of undesirable characteristic properties of hierarchical models that may affect our ability to easily add, delete, and update or edit records.

A network model is quite similar to but more general than the hierarchical model. In a hierarchy, data have to be arranged such that one child has only one parent, and in many instances this is unrealistic. If we force the use of a hierarchical representation in such cases, data will have to be repeated at more than one location in the hierarchical model. This redundancy can create a number of problems. A record in a network model can participate in several relationships. This leads to two primary structural differences between hierarchical and network models. Some fields in the hierarchical model will become relationships in a network model. Further, the relationships in a network model are explicit and may be bidirectional. The navigation problem in a network data model can become severe. Because search of the database can start at several places in the network, there is added complexity in searching, as well.

While spreadsheet type relational records are very useful for many purposes, it has been observed (Kent 1979) that not all views of a situation, or human cognitive maps, can be represented by relational data models. This has led to interest in entity and object-oriented data models and to data models based on artificial intelligence techniques. The basic notion in the use of an ER model is to accomplish database design at the conceptual level, rather than at the logical and/or physical levels. ER models (Chen 1976) are relatively simple and easy to understand and use, in large part because of the easy graphical visualization of the database model structure. Also, ER modeling capability is provided by many computer aided software engineering (CASE) tools. While such limitations as lack of a very useful query language are present (Atzeni and Chen 1983), much interest in ER data models, especially at the conceptual level, exists at this time. Figure 8 illustrates the conceptual orientation of the ER modeling approach.

ER models are based on two premises: (a) information concerning entities and relationships exists as a cognitive reality, and (b) this information may be structured using entities and relationships among them as data. An entity relationship data model is a generalization of the hierarchical and network data models. It is based on well-established graph theoretic developments and is a form of structured modeling. The major advantage of the ER approach is that it provides a realistic view of the structure and form for the DBMS design and development effort. This should naturally support the subsequent development of appropriate software. In addition, the approach readily leads to the

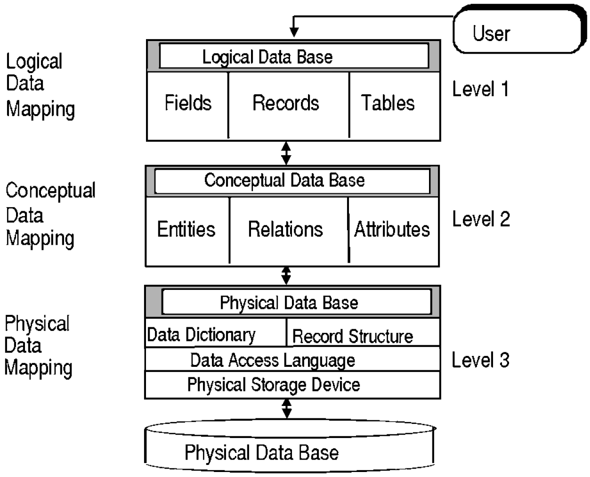


Figure 8 The Potential Role of ER Data Models at the Conceptual Data Level.

development of the data dictionary. The primary difficulties that may impede easy use of the ER method are in the need for selection of appropriate verbs for use as contextual relationships, elimination of redundancy of entities, and assurances that all of the needed ER have been determined and properly used. Often a considerable amount of time may be needed to identify an appropriate set of entities and relations and obtain an appropriate structural model in the form of an ER diagram.

2.2.3. Expert and Object-Oriented Database Models

An expert database model, as a specialized expert database system, involves the use of artificial intelligence technology. The goal in this is to provide for database functionality in more complex environments that require at least some limited form of intelligent capability. To allow this may require adaptive identification of system characteristics, or learning over time as experience with an issue and the environment into which it is embedded increases. The principal approach that has been developed to date is made up of an object-oriented database (OODB) design with any of the several knowledge representation approaches useful in expert system development. The field is relatively new compared to other database efforts and is described much more fully in Kerschberg (1987, 1989), Myloupoulos and Brodie (1989), Parsaye et al. (1989), Debenham (1998), Poe et al. (1998), and Darwen and Date (1998).

The efforts to date in the object oriented area concern what might be more appropriately named object-oriented database management systems design (OODBMS). The objective in OODBMS design is to cope with database complexity, potentially for large distributed databases, through a combination of object-oriented programming and expert systems technology. Object-oriented design approaches generally involve notions of separating internal computer representations of elements from the external realities that lead to the elements. A primary reason for using object-oriented language is that it naturally enables semantic representations of knowledge through use of 10 very useful characteristics of object-oriented approaches (Parsaye et al. 1989): information hiding, data abstraction, dynamic binding and object identity, inheritance, message handling, object oriented graphical interfaces, transaction management, reusability, partitioning or dividing or disaggregating an issue into parts, and projecting or describing a system from multiple perspectives or viewpoints. These could include, for example, social, political, legal, and technoeconomic perspectives (Sage 1992).

There are at least two approaches that we might use in modeling a complex large-scale system: (1) functional decomposition and structuring and (2) purposeful or object decomposition and structuring. Both approaches are appropriate and may potentially result in useful models. Most models of real-world phenomena tend to be purposeful. Most conventional high-level programming languages are functionally, or procedurally, oriented. To use them, we must write statements that correspond to the functions that we wish to provide in order to solve the problem that has been posed. An advantage of object decomposition and structuring is that it enables us to relate more easily the structure of a database model to the structure of the real system. This is the case if we accomplish our decomposition and structuring such that each module in the system or issue model represents an object or a class of objects in the real issue or problem space. Objects in object-oriented methodology are not unlike elements or nodes in graph theory and structural modeling. It is possible to use one or more contextual relations to relate elements together in a structural model. An object may be defined as a collection of information and those operations that can be performed upon it. We request an object to perform one of its allowable operations by instructing it with a message.

Figure 9 illustrates the major conceptual difference between using a conventional programming approach and using an object-oriented approach for construction of a DBMS. In the conventional approach, procedures are at the nexus and procedures update and otherwise manipulate data and return values. In the object-oriented approach, the collection of independent objects are at the nexus and communicate with each other through messages or procedures. Objects investigate requests and behave according to these messages. Object-oriented design often provides a clear and concise interface to the problem domain in that the only way to interact with an object is through the operations or messages to the object. These messages call for operations that may result in a change of state in the object in question. This message will affect the particular object called and only that one, as no other object is affected. This provides a high degree of modularity and increased ability to verify and validate outcomes and thus provides an increased sense of reliability in the resulting DBMS design. Object-oriented languages are, for the most part, very high-level languages used to accomplish precisely the same results as high-level languages. By focusing upon the entities of objects and the relationships between objects, they often provide a simpler way to describe precisely those elements needed in detailed design procedures.

Object-oriented DBMS design is based on appropriate linking of objects, or data items or entities, and the operations on these, such that the information and processing is concentrated on the object classes, attributes, and messages that transfer between them. The features that set object-oriented approaches from other approaches are the capability of object-oriented languages to support abstraction, information hiding, and modularity. The items used in object-oriented DBMS design are objects,

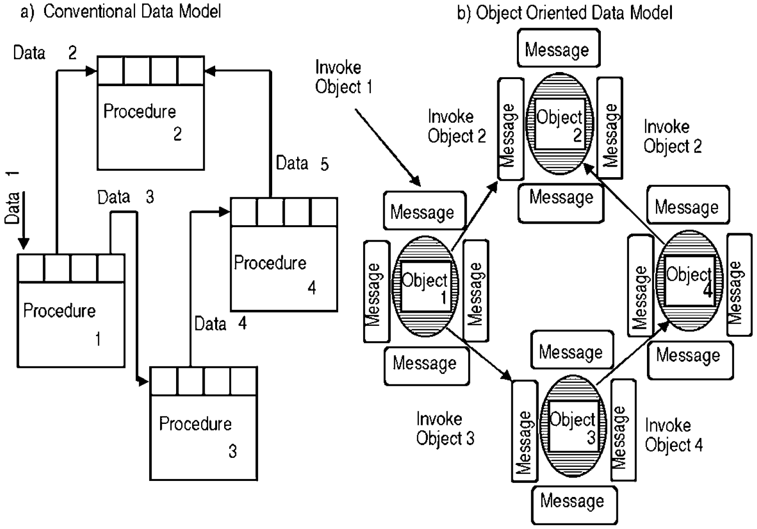


Figure 9 Conventional and Object-Oriented Data Models.

abstracts or physical entities that pertain to the domain of the system; attributes, or characteristics of objects; operations, or dynamic entities that may change with time; and messages, or requests to an object to perform an operation. Objects represent such real-world entities as machines, files, products, signals, persons, things, and places. They may be either physical entities or abstract representations. The attributes are characteristics that describe the properties ascribed to objects. Each attribute may be dynamically associated with a numerical value, and the combination of these values together with the description of the object in the problem domain presents the state of the object. Thus, the attributes are related to the object as subobjects. Attributes include such things as the name of the object, the number of specific objects in a file, or the name for the place, and the like. The basic attribute, one that may not be further decomposed, is the primitive type of object or subobject. Attributes may also be nonnumeric. Operations consist of processes and data structures that apply to the object to which it is directed. These are generally dynamic entities whose value may change over time. Each object may be subjected to a number of operations that provide information relative to control and procedural constructs that are applied to the object. Defining an object such as to have a private part and then assigning a message to address the appropriate processing operation enables us to achieve information hiding. Messages are passed between objects in order to change the state of the object, address the potentially hidden data parts of an object, or otherwise modify an object.

Coad and Yourdon (1990) identify five steps associated with implementing an object-oriented approach:

1. Identify objects, typically by examining the objects in the real world.
2. Identify structures, generally through various abstracting and partitioning approaches that result in a classification structure, an assembly structure, or a combination of these.
3. Identify subjects through examining objects and their structures to obtain this more complex abstract view of the issue. Each classification structure and each assembly structure will comprise one subject.
4. Identify attributes that impart important aspects of the objects that need to be retained for all instances of an object.
5. Identify services such that we are aware of occurrence services, creation or modification of instances of an object, calculation services, or monitoring of other processes for critical events or conditions.

Steps such as these should generally be accomplished in an iterative manner because there is no unique way to accomplish specification of a set of objects, or structural relations between objects.

A major result from using object-oriented approaches is that we obtain the sort of knowledge representation structures that are commonly found in many expert systems. Thus, object-oriented design techniques provide a natural interface to expert system-based techniques for DBMS design. This is so because objects, in object-oriented design, are provided with the ability to store information such that learning over time can occur; process information by initiating actions in response to messages; compute and communicate by sending messages between objects; and generate new information as a result of communications and computations. Object-oriented design also lends itself to parallel processing and distributed environments. Object-oriented design and analysis is a major subject in contemporary computer and information system subject areas and in many application domains as well (Firesmith 1993; Sullo 1994; Jacobson et al. 1995; Zeigler 1997).

2.3. Distributed and Cooperative Databases

We have identified the three major objectives for a DBMS as data independence, data redundancy reduction, and data resource control. These objectives are important for a single database. When there are multiple databases potentially located in a distributed geographic fashion, and potentially many users of one or more databases, additional objectives arise, including:

1. Location independence or transparency to enable DBMS users to access applications across distributed information bases without the need to be explicitly concerned with where specific data is located
2. Advanced data models, to enable DBMS users to access potentially nonconventional forms of data such as multidimensional data, graphic data, spatial data, and imprecise data
3. Extensible data models, which will allow new data types to be added to the DBMS, perhaps in an interactive real-time manner, as required by specific applications.

An important feature of a distributed database management systems (DDBMSs) is that provisions for database management are distributed. Only then can we obtain the needed "fail-safe" reliability and "availability" even when a portion of the system breaks down. There are a number of reasons why distributed databases may be desirable. These include and are primarily related to distributed users and cost savings. Some additional costs are also involved, and these must be justified.

A distributed database management system will generally look much like replicated versions of a more conventional single-location database management system. We can thus imagine replicated versions of Figure 7. For simplicity, we show only the physical database, the database access interface mechanism, and the data dictionary for each database. Figure 10 indicates one possible conceptual

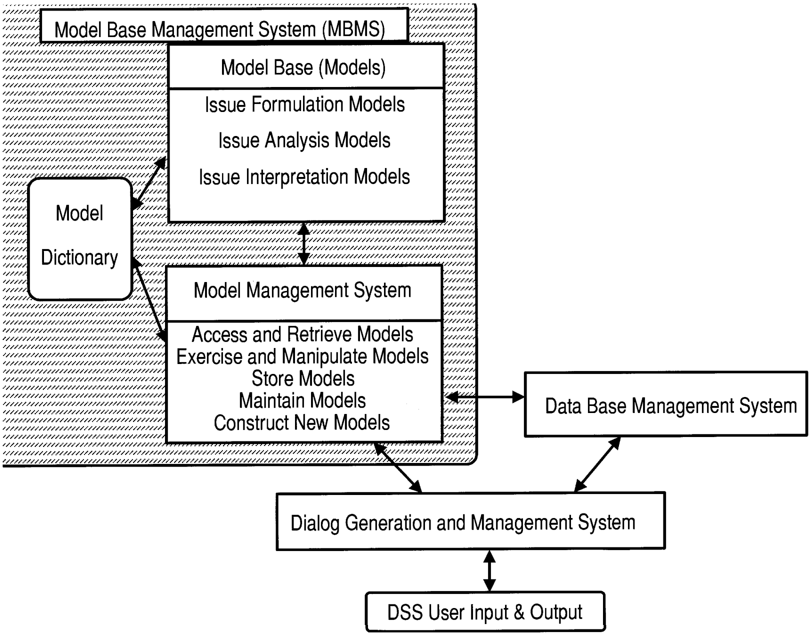

Model Base Management System (MBMS)
Model Base (Models)
Issue Formulation Models
Issue Analysis Models
Issue Interpretation Models
Model Dictionary
Model Management System
Access and Retrieve Models
Exercise and Manipulate Models
Store Models
Maintain Models
Construct New Models
Data Base Management System
Dialog Generation and Management System
DSS User Input & Output

Figure 10 Prototypical Structure of Model Management System.

architecture of a distributed database in which there are two requests for data being simultaneously submitted. Through use of protocols, all requests for data are entered as if those data were stored in the database of the requesting user. The data dictionaries of the distributed system are responsible for finding the requested data elements and delivering them to the requesting user. All of these auxiliary requests are accomplished in a manner that is transparent to the requesting database user. Data transmission rate concerns must be resolved in order for this to be done satisfactorily. Potentially, however, real-time access to all of the data in the entire distributed system can be provided to any local user.

It is important to note that completion of the above steps does not transfer the requested data into the database of the requesting user, except perhaps in some temporary storage fashion. The very good reason why this is not done is that it would result in database redundancy and also increase database security difficulties.

An alternate approach to this distributing of data is to use what is often called cooperative processing or cooperative database management. This involves central database and distributed database concepts, blended to produce better results than can be had with either approach. Central to this are various approaches for the distributing and coordinating of data. We could use function, geography, or organizational level as the basis for distributing and cooperating, and possible system configurations may be based upon background communications, microcomputer-based front-end processing for host applications, and peer-to-peer cooperative processing.

The database component of a DSS provides the knowledge repository that is needed for decision support. Some sort of management system must be associated with a database in order to provide the intelligent access to data that is needed. In this section we have examined a number of existing constructs for database management systems. We now turn to the model-based portion of a DSS.

3. MODEL BASE MANAGEMENT SYSTEMS

There are four primary objectives of a DBMS:

1. To manage a large quantity of data in physical storage
2. To provide logical data structures that interact with humans and are independent of the structure used for physical data storage
3. To reduce data redundancy and maintenance needs and increase flexibility of use of the data, by provision of independence between the data and the applications programs that use it
4. To provide effective and efficient access to data by users who are not necessarily very sophisticated in the microlevel details of computer science.

Many support facilities will typically be provided with a DBMS to enable achievement of these purposes. These include data dictionaries to aid in internal housekeeping and information query, retrieval, and report generation facilities to support external use needs.

The functions of a model base management system (MBMS), a structure for which is illustrated in Figure 10, are quite analogous to those of a DBMS. The primary functions of a DBMS are separation of system users, in terms of independence of the application, from the physical aspects of database structure and processing. In a similar way, a MBMS is intended to provide independence between the specific models that are used in a DSS and the applications that use them. The purpose of a MBMS is to transform data from the DBMS into information that is useful for decision making. An auxiliary purpose might also include representation of information as data such that it can later be recalled and used.

The term *model management system* was apparently first used over 15 years ago (Will 1975). Soon thereafter, the MBMS usage was adopted in Sprague and Carlson (1982) and the purposes of an MBMS were defined to include creation, storage, access, and manipulation of models. Objectives for a MBMS include:

1. To provide for effective and efficient creation of new models for use in specific applications
2. To support maintenance of a wide range of models that support the formulation, analysis, and interpretation stages of issue resolution
3. To provide for model access and integration, within models themselves as well as with the DBMS
4. To centralize model base management in a manner analogous to and compatible with database management
5. To ensure integrity, currency, consistency, and security of models.

Just as we have physical data and logical data processing in a DBMS, so also do we have two types of processing efforts in a MBMS: model processing and decision processing (Applegate et al.

1986). A DSS user would interact directly with a decision processing MBMS, whereas the model processing MBMS would be more concerned with provision of consistency, security, currency, and other technical modeling issues. Each of these supports the notion of appropriate formal use of models that support relevant aspects of human judgment and choice.

Several ingredients are necessary for understanding MBMS concepts and methods. The first of these is a study of formal analytical methods of operations research and systems engineering that support the construction of models that are useful in issue formulation, analysis, and interpretation. Because presenting even a small fraction of the analytical methods and associated models in current use would be a mammoth undertaking, we will discuss models in a somewhat general context. Many discussions of decision relevant models can be found elsewhere in this Handbook, most notably in Chapters 83 through 102. There are also many texts in this area, as well as two recent handbooks (Sage and Rouse 1999a; Gass and Harris 2000).

3.1. Models and Modeling

In this section, we present a brief description of a number of models and methods that can be used as part of a systems engineering-based approach to problem solution or issue resolution. Systems engineering (Sage 1992, 1995; Sage and Rouse 1999a; Sage and Armstrong 2000) involves the application of a general set of guidelines and methods useful to assist clients in the resolution of issues and problems, often through the definition, development, and deployment of trustworthy systems. These may be product systems or service systems, and users may often deploy systems to support process-related efforts. Three fundamental steps may be distinguished in a formal systems-based approach that is associated with each of these three basic phases of system engineering or problem solving:

1. Problem or issue formulation
2. Problem or issue analysis
3. Interpretation of analysis results, including evaluation and selection of alternatives, and implementation of the chosen alternatives

These steps are conducted at a number of phases throughout a systems life cycle. As we indicated earlier, this life cycle begins with definition of requirements for a system through a phase where the system is developed to a final phase where deployment, or installation, and ultimate system maintenance and retrofit occur. Practically useful life cycle processes generally involve many more than three phases, although this larger number can generally be aggregated into the three basic phases of definition, development, and deployment. The actual engineering of a DSS follows these three phases of definition of user needs, development of the DSS, and deployment in an operational setting. Within each of these three phases, we exercise the basic steps of formulation, analysis, and interpretation. Figure 11 illustrates these three steps and phases and much more detail is presented in Sage (1992, 1995), Sage and Rouse (1999a), Sage and Armstrong (2000), and in references contained therein.

3.1.1. Issue, or Problem, Formulation Models

The first part of a systems effort for problem or issue resolution is typically concerned with problem or issue formulation, including identification of problem elements and characteristics. The first step in issue formulation is generally that of definition of the problem or issue to be resolved. Problem

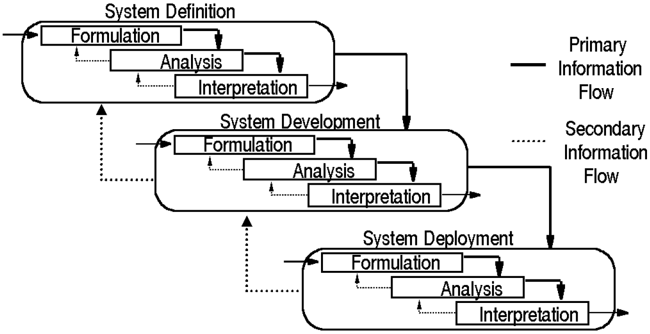


Figure 11 Basic Phases and Steps in Systems Engineering.

definition is generally an outscoping activity because it enlarges the scope of what was originally thought to be the problem. Problem or issue definition will ordinarily be a group activity involving those familiar with or impacted by the issue or the problem. It seeks to determine the needs, constraints, alterables, and social or organizational sectors affecting a particular problem and relationships among these elements.

Of particular importance are the identification and structuring of objectives for the policy or alternative that will ultimately be chosen. This is often referred to as *value system design,* a term apparently first used by Hall (1969) in one of his seminal works in systems engineering. Generation of options (Keller and Ho 1990) or alternative courses of action is a very important and often neglected portion of a problem solving effort. This option generation, or system or alternative synthesis, step of issue formulation is concerned primarily with the answers to three questions:

1. What are the alternative approaches for attaining objectives?
2. How is each alternative approach described?
3. How do we measure attainment of each alternative approach?

The answers to these three questions lead to a series of alternative activities or policies and a set of activities measures.

Several of the methods that are particularly helpful in the identification of issue formulation elements are based on principles of collective inquiry (McGrath 1984). The term *collective inquiry* refers to the fact that a group of interested and motivated people is brought together in the hope that they will stimulate each other's creativity in generating elements. We may distinguish two groups of collective inquiry methods here, depending upon whether or not the group is physically present at the same physical location.

3.1.1.1. Brainstorming, Synectics, and Nominal Group Techniques These approaches typically require a few hours of time, a group of knowledgeable people gathered in one place, and a group leader or facilitator. The nominal group technique is typically better than brainstorming in reducing the influence of dominant individuals. Both methods can be very productive: 50–150 ideas or elements may be generated in less than one hour. Synectics, based on problem analogies, might be very appropriate if there is a need for truly unconventional, innovative ideas. Considerable experience with the method is a requirement, however, particularly for the group leader. The nominal group technique is based on a sequence of idea generation, discussion, and prioritization. It can be very useful when an initial screening of a large number of ideas or elements is needed. Synectics and brainstorming are directly interactive group methods, whereas nominal group efforts are nominally interactive in that the members of the group do not directly communicate.

3.1.1.2. Questionnaires, Survey, and Delphi Techniques These three methods of collective inquiry do not require the group of participants to gather at one place and time, but they typically take more time to achieve results than the methods above. In most questionnaires and surveys a large number of participants is asked, on an individual basis, to generate ideas or opinions that are then processed to achieve an overall result. Generally no interaction is allowed among participants in this effort. Delphi usually provides for written anonymous interaction among participants in several rounds. Results of previous rounds are fed back to participants, and they are asked to comment, revise their views as desired, and so on. The results of using the Delphi technique can be very enlightening, but it usually takes several weeks or months to complete the effort.

Use of some of the many structuring methods, in addition to leading to greater clarity of the problem formulation elements, will typically lead also to identification of new elements and revision of element definitions. Most structuring methods contain an analytical component and may therefore be more properly labeled as analysis methods. The following element structuring aids are among the many modeling aids available that are particularly suitable for the issue formulation step.

There are many approaches to problem formulation (Volkema 1990). In general, these approaches assume that "asking" will be the predominant approach used to obtain issue formulation elements. Asking is often the simplest approach. Valuable information can often be obtained from observation of an existing and evolving system or from study of plans and other prescriptive documents. When these three approaches fail, it may be necessary to construct a "trial" system and determine issue formulation elements through experimentation and iteration with the trial system. These four methods (asking, study of an existing system, study of a normative systems, experimentation with a prototype system) are each very useful for information and system requirements determination (Davis 1982).

3.1.2. Models for Issue Analysis

The analysis portion of a DSS effort typically consists of two steps. First, the options or alternatives defined in issue formulation are analyzed to assess their expected impacts on needs and objectives. This is often called impact assessment or impact analysis. Second, a refinement or optimization effort is often desirable. This is directed towards refinement or fine-tuning a potentially viable alternative through adjustment of the parameters within that alternative so as to obtain maximum performance in terms of needs satisfaction, subject to the given constraints.

Simulation and modeling methods are based on the conceptualization and use of an abstraction, or model, that hopefully behaves in a similar way as the real system. Impacts of alternative courses of action are studied through use of the model, something that often cannot easily be done through experimentation with the real system. Models are, of necessity, dependent on the value system and the purpose behind utilization of a model. We want to be able to determine the correctness of predictions based on usage of a model and thus be able to validate the model. There are three essential steps in constructing a model:

1. Determine those issue formulation elements that are most relevant to a particular problem.
2. Determine the structural relationships among these elements.
3. Determine parametric coefficients within the structure.

We should interpret the word ''model'' here as an abstract generalization of an object or system. Any set of rules and relationships that describes something is a model of that thing. The MBMS of a DSS will typically contain formal models that have been stored into the model base of the support system. Much has been published in the area of simulation realization of models, including a recent handbook (Banks 1998).

Gaming methods are basically modeling methods in which the real system is simulated by people who take on the roles of real-world actors. The approach may be very appropriate for studying situations in which the reactions of people to each others actions are of great importance, such as competition between individuals or groups for limited resources. It is also a very appropriate learning method. Conflict analysis (Fraser and Hipel 1984; Fang et al. 1993) is an interesting and appropriate game theory-based approach that may result in models that are particularly suitable for inclusion into the model base of a MBMS. A wealth of literature concerning formal approaches to mathematical games (Dutta 1999).

Trend extrapolation or time series forecasting models, or methods, are particularly useful when sufficient data about past and present developments are available, but there is little theory about underlying mechanisms causing change. The method is based on the identification of a mathematical description or structure that will be capable of reproducing the data. Then this description is used to extend the data series into the future, typically over the short to medium term. The primary concern is with input–output matching of observed input data and results of model use. Often little attention is devoted to assuring process realism, and this may create difficulties affecting model validity. While such models may be functionally valid, they may not be purposefully or structurally valid.

Continuous-time dynamic-simulation models, or methods, are generally based on postulation and qualification of a causal structure underlying change over time. A computer is used to explore long-range behavior as it follows from the postulated causal structure. The method can be very useful as a learning and qualitative forecasting device. Often it is expensive and time consuming to create realistic dynamic simulation models. Continuous-time dynamic models are quite common in the physical sciences and in much of engineering (Sage and Armstrong 2000).

Input-output analysis models are especially designed for study of equilibrium situations and requirements in economic systems in which many industries are interdependent. Many economic data formats are directly suited for the method. It is relatively simple conceptually, and can cope with many details. Input–output models are often very large.

Econometrics or macroeconomic models are primarily applied to economic description and forecasting problems. They are based on both theory and data. Emphasis is placed an specification of structural relations, based upon economic theory, and the identification of unknown parameters, using available data, in the behavioral equations. The method requires expertise in economics, statistics, and computer use. It can be quite expensive and time consuming. Macroeconomic models have been widely used for short- to medium-term economic analysis and forecasting.

Queuing theory and discrete event simulation models are often used to study, analyze, and forecast the behavior of systems in which probabilistic phenomena, such as waiting lines, are of importance. Queuing theory is a mathematical approach, while discrete-event simulation generally refers to computer simulation of queuing theory type models. The two methods are widely used in the analysis and design of systems such as toll booths, communication networks, service facilities, shipping terminals, and scheduling.

Regression analysis models and estimation theory models are very useful for the identification of mathematical relations and parameter values in these relations from sets of data or measurements. Regression and estimation methods are used frequently in conjunction with mathematical modeling, in particular with trend extrapolation and time series forecasting, and with econometrics. These methods are often also used to validate models. Often these approaches are called system identification approaches when the goal is to identify the parameters of a system, within an assumed structure, such as to minimize a function of the error between observed data and the model response.

Mathematical programming models are used extensively in operations research systems analysis and management science practice for resource allocation under constraints, planning or scheduling, and similar applications. It is particularly useful when the best equilibrium or one-time setting has to be determined for a given policy or system. Many analysis issues can be cast as mathematical programming problems. A very significant number of mathematical programming models have been developed, including linear programming, nonlinear programming, integer programming, and dynamic programming. Many appropriate reference texts, including Hillier and Lieberman (1990, 1994), discuss this important class of modeling and analysis tools.

3.1.3. Issue Interpretation Models

The third step in a decision support systems use effort starts with evaluation and comparison of alternatives, using the information gained by analysis. Subsequently, one or more alternatives are selected and a plan for their implementation is designed. Thus, an MBMS must provide models for interpretation, including evaluation, of alternatives.

It is important to note that there is a clear and distinct difference between the refinement of individual alternatives, or optimization step of analysis, and the evaluation and interpretation of the sets of refined alternatives that result from the analysis step. In some few cases, refinement or optimization of individual alternative decision policies may not be needed in the analysis step. More than one alternative course of action or decision must be available; if there is but a single policy alternative, then there really is no decision to be taken at all. Evaluation of alternatives is always needed. It is especially important to avoid a large number of cognitive biases in evaluation and decision making. Clearly, the efforts involved in the interpretation step of evaluation and decision making interact most strongly with the efforts in the other steps of the systems process. A number of methods for evaluation and choice making are of importance. A few will be described briefly here.

Decision analysis (Raiffa 1968) is a very general approach to option evaluation and selection. It involves identification of action alternatives and possible consequences, identification of the probabilities of these consequences, identification of the valuation placed by the decision maker upon these consequences, computation of the expected value of the consequences, and aggregation or summarization of these values for all consequences of each action. In doing this we obtain an evaluation of each alternative act, and the one with the highest value is the most preferred action or option.

Multiple-attribute utility theory (Keeney and Raiffa 1976) has been designed to facilitate comparison and ranking of alternatives with many attributes or characteristics. The relevant attributes are identified and structured and a weight or relative utility is assigned by the decision maker to each basic attribute. The attribute measurements for each alternative are used to compute an overall worth or utility for each alternative. Multiple attribute utility theory allows for various types of worth structures and for the explicit recognition and incorporation of the decision maker's attitude towards risk in the utility computation.

Policy Capture (or *Social Judgment*) *Theory* (Hammond et al. 1980) has also been designed to assist decision makers in making values explicit and known. It is basically a descriptive approach toward identification of values and attribute weights. Knowing these, one can generally make decisions that are consistent with values. In policy capture, the decision maker is asked to rank order a set of alternatives. Then alternative attributes and their attribute measures or scores are determined by elicitation from the decision maker for each alternative. A mathematical procedure involving regression analysis is used to determine that relative importance, or weight, of each attribute that will lead to a ranking as specified by the decision maker. The result is fed back to the decision maker, who typically will express the view that some of his or her values, in terms of the weights associated with the attributes, are different. In an iterative learning process, preference weights and/or overall rankings are modified until the decision maker is satisfied with both the weights and the overall alternative ranking.

Many efforts have been made to translate the theoretical findings in decision analysis to practice. Klein (1998), Matheson and Matheson (1998), and Hammond et al. (1999) each provide different perspectives relative to these efforts.

3.2. Model Base Management

As we have noted, an effective model base management system (MBMS) will make the structural and algorithmic aspects of model organization and associated data processing transparent to users of the MBMS. Such tasks as specifying explicit relationships between models to indicate formats for models and which model outputs are input to other models are not placed directly on the user of a MBMS but handled directly by the system. Figure 11 presents a generic illustration of a MBMS. It shows a collection of models or model base, a model base manager, a model dictionary, and connections to the DBMS and the DGMS.

A number of capabilities should be provided by an integrated and shared MBMS of a DSS (Barbosa and Herko 1980; Liang 1985) construction, model maintenance, model storage, model manipulation, and model access (Applegate et al. 1986). These involve control, flexibility, feedback, interface, redundancy reduction, and increased consistency:

1. *Control:* The DSS user should be provided with a spectrum of control. The system should support both fully automated and manual selection of models that seem most useful to the user for an intended application. This will enable the user to proceed at the problem-solving pace that is most comfortable given the user's experiential familiarity with the task at hand. It should be possible for the user to introduce subjective information and not have to provide full information. Also, the control mechanism should be such that the DSS user can obtain a recommendation for action with this partial information at essentially any point in the problem-solving process.
2. *Flexibility:* The DSS user should be able to develop part of the solution to the task at hand using one approach and then be able to switch to another modeling approach if this appears preferable. Any change or modification in the model base will be made available to all DSS users.
3. *Feedback:* The MBMS of the DSS should provide sufficient feedback to enable the user to be aware of the state of the problem-solving process at any point in time.
4. *Interface:* The DSS user should feel comfortable with the specific model from the MBMS that is in use at any given time. The user should not have to supply inputs laboriously when he or she does not wish to do this.
5. *Redundancy reduction:* This should occur through use of shared models and associated elimination of redundant storage that would otherwise be needed.
6. *Increased consistency:* This should result from the ability of multiple decision makers to use the same model and the associated reduction of inconsistency that would have resulted from use of different data or different versions of a model.

In order to provide these capabilities, it appears that a MBMS design must allow the DSS user to:

1. Access and retrieve existing models
2. Exercise and manipulate existing models, including model instantiation, model selection, and model synthesis, and the provision of appropriate model outputs
3. Store existing models, including model representation, model abstraction, and physical and logical model storage
4. Maintain existing models as appropriate for changing conditions
5. Construct new models with reasonable effort when they are needed, usually by building new models by using existing models as building blocks

A number of auxiliary requirements must be achieved in order to provide these five capabilities. For example, there must be appropriate communication and data changes among models that have been combined. It must also be possible to locate appropriate data from the DBMS and transmit it to the models that will use it.

It must also be possible to analyze and interpret the results obtained from using a model. This can be accomplished in a number of ways. In this section, we will examine two of them: relational MBMS and expert system control of an MBMS. The objective is to provide an appropriate set of models for the model base and appropriate software to manage the models in the model base; integration of the MBMS with the DBMS; and integration of the MBMS with the DGMS. We can expand further on each of these needs. Many of the technical capabilities needed for a MBMS will be analogous to those needed for a DBMS. These include model generators that will allow rapid building of specific models, model modification tools that will enable a model to be restructured easily on the basis of changes in the task to be accomplished, update capability that will enable changes in data to be input to the model, and report generators that will enable rapid preparation of results from using the system in a form appropriate for human use.

Like a relational view of data, a relational view of models is based on a mathematical theory of relations. Thus, a model is viewed as a virtual file or virtual relation. It is a subset of the Cartesian product of the domain set that corresponds to these input and output attributes. This virtual file is created, ideally, through exercising the model with a wide spectrum of inputs. These values of inputs and the associated outputs become records in the virtual file. The input data become key attributes and the model output data become content attributes.

Model base structuring and organization is very important for appropriate relational model management. Records in the virtual file of a model base are not individually updated, however, as they are in a relational database. When a model change is made, all of the records that comprise the virtual file are changed. Nevertheless, processing anomalies are possible in relational model management. Transitive dependencies in a relation, in the form of functional dependencies that affect only the output attributes, do occur and are eliminated by being projected into an appropriate normal form.

Another issue of considerable importance relates to the contemporary need for usable model base query languages and to needs within such languages for relational completeness. The implementation of joins is of concern in relational model base management just as it is in relational database management. A relational model join is simply the result of using the output of one model as the input to another model. Thus, joins will normally be implemented as part of the normal operation of software, and a MBMS user will often not be aware that they are occurring. However, there can be cycles, since the output from a first model may be the input to a second model, and this may become the input to the first model. Cycles such as this do not occur in relational DBMS.

Expert system applications in MBMS represent another attractive possibility. Four different potential opportunities exist. It might be possible to use expert system technology to considerable advantage in the construction of models (Hwang 1985; Murphy and Stohr 1986), including decisions with respect to whether or not to construct models in terms of the cost and benefits associated with this decision. AI and expert system technology may potentially be used to integrate models. This model integration is needed to join models. AI and expert system technology might be potentially useful in the validation of models. Also, this technology might find potential use in the interpretation of the output of models. This would especially seem to be needed for large-scale models, such as large linear programming models. While MBMS approaches based on a relational theory of models and expert systems technology are new as of this writing, they offer much potential for implementing model management notions in an effective manner. As has been noted, they offer the prospect of data as models (Dolk 1986) that may well prove much more useful than the conventional information systems perspective of “models as data.”

4. DIALOG GENERATION AND MANAGEMENT SYSTEMS

In our efforts in this chapter, we envision a basic DSS structure of the form shown in Figure 12. This figure also shows many of the operational functions of the database management system (DBMS) and the model base management system (MBMS). The primary purpose of the dialog generation and management system (DGMS) is to enhance the propensity and ability of the system user to use and benefit from the DSS. There are doubtless few users of a DSS who use it because of necessity. Most uses of a DSS are optional, and the decision maker must have some motivation and desire to use a DSS or it will likely remain unused. DSS use can occur in a variety of ways. In all uses of a DGMS, it is the DGMS that the user interacts with. In an early seminal text, Bennett (1983) posed three questions to indicate this centrality:

1. What presentations is the user able to *see* at the DSS display terminal?
2. What must the user know about what is seen at the display terminal in order to use the DSS?
3. What can the DSS user *do* with the systems that will aid in accomplishing the intended purpose?

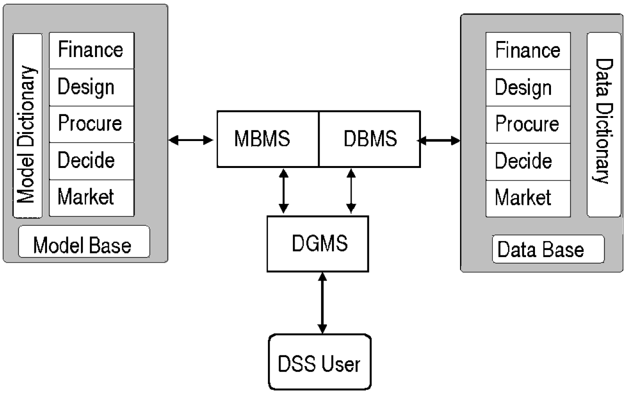


Figure 12 The DGMS as the User-Focused System in a DSS.

Bennett refers to these elements as the *presentation language,* the *knowledge base,* and the *action language*.

It is generally felt that there are three types of languages or modes of human communications: *words, mathematics,* and *graphics*. The presentation and action languages, and the knowledge base in general, many contain any or all of these three language types. The mix of these that is appropriate for any given DSS task will be a function of the task itself, the environment into which the task is embedded, and the nature of the experiential familiarity of the person performing the task with the task and with the environment. This is often called the contingency task structure. The DSS, when one is used, becomes a fourth ingredient, although it is really much more of a vehicle supporting effective use of words, mathematics, and graphics than it is a separate fourth ingredient.

Notions of DGMS design are relatively new, especially as a separately identified portion of the overall design effort. To be sure, user interface design is not at all new. However, the usual practice has been to assign the task of user interface design to the design engineers responsible for the entire system. In the past, user interfaces were not given special attention. They were merely viewed as another hardware and software component in the system. System designers were often not particularly familiar with, and perhaps not even especially interested in, the user-oriented design perspectives necessary to produce a successful interface design. As a result, many user interface designs have provided more what the designer wanted than what the user wanted and needed. Notions of dialog generation and dialog management extend far beyond interface issues, although the interface is a central concern in dialog generation and dialog management. A number of discussions of user interface issues are contained in Part IIIB of this Handbook.

Figure 14 illustrates an attribute tree for interface design based on the work of Smith and Mosier (1986). This attribute tree can be used, in conjunction with the evaluation methods of decision analysis, to evaluate the effectiveness of interface designs. There are other interface descriptions, some of which are less capable of instrumental measurement than these. On the basis of a thorough study of much of the human–computer interface and dialog design literature, Schneiderman (1987) has identified eight primary objectives, often called the "golden rules" for dialog design:

1. Strive for consistency of terminology, menus, prompts, commands, and help screens.
2. Enable frequent users to use shortcuts that take advantage of their experiential familiarity with the computer system.
3. Offer informative feedback for every operator action that is proportional to the significance of the action.
4. Design dialogs to yield closure such that the system user is aware that specific actions have been concluded and that planning for the next set of activities may now take place.
5. Offer simple error handling such that, to the extent possible, the user is unable to make a mistake. Even when mistakes are made, the user should not have to, for example, retype an entire command entry line. Instead, the user should be able to just edit the portion that is incorrect.
6. Permit easy reversal of action such that the user is able to interrupt and then cancel wrong commands rather than having to wait for them to be fully executed.

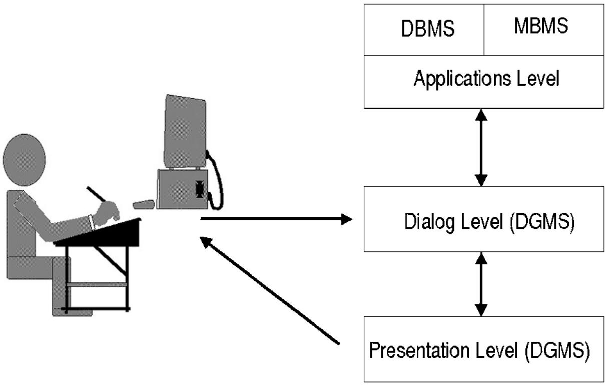


Figure 13 Dialog Generation and Management System Architecture.

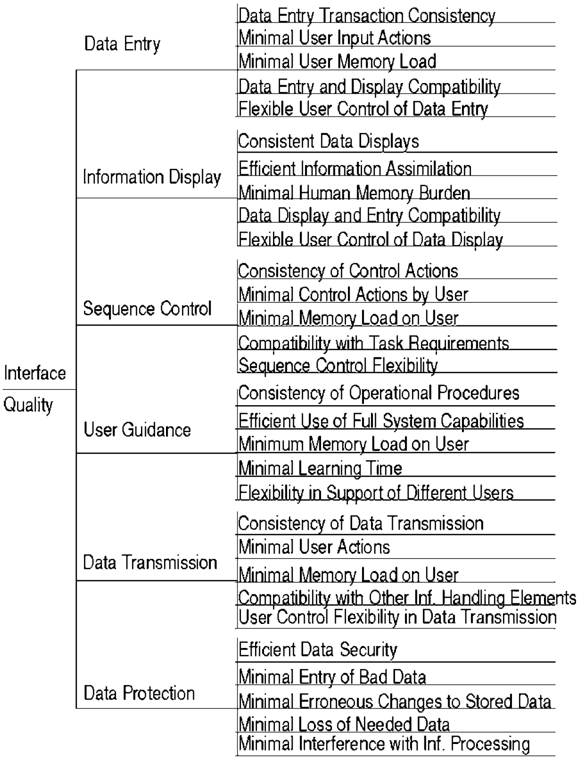

Interface Quality
Data Entry
Data Entry Transaction Consistency
Minimal User Input Actions
Minimal User Memory Load
Data Entry and Display Compatibility
Flexible User Control of Data Entry
Information Display
Consistent Data Displays
Efficient Information Assimilation
Minimal Human Memory Burden
Data Display and Entry Compatibility
Flexible User Control of Data Display
Sequence Control
Consistency of Control Actions
Minimal Control Actions by User
Minimal Memory Load on User
Compatibility with Task Requirements
Sequence Control Flexibility
User Guidance
Consistency of Operational Procedures
Efficient Use of Full System Capabilities
Minimum Memory Load on User
Minimal Learning Time
Flexibility in Support of Different Users
Data Transmission
Consistency of Data Transmission
Minimal User Actions
Minimal Memory Load on User
Compatibility with Other Inf. Handling Elements
User Control Flexibility in Data Transmission
Data Protection
Efficient Data Security
Minimal Entry of Bad Data
Minimal Erroneous Changes to Stored Data
Minimal Loss of Needed Data
Minimal Interference with Inf. Processing

Figure 14 Attribute Tree of Smith and Mosier Elements for Interface Design Evaluation.

7. Support internal locus of control such that users are always the initiators of actions rather than the reactors to computer actions.
8. Reduce short-term memory load such that users are able to master the dialog activities that they perform in a comfortable and convenient manner.

Clearly, all of these will have specific interpretations in different DGMS environments and need to be sustained in and capable of extension for a variety of environments.

Human–computer interaction and associated interface design issues are of major contemporary importance, so it is not surprising that there have been a number of approaches to them. Some of these approaches are almost totally empirical. Some involve almost totally theoretical and formal models (Harrison and Thimbleby 1990). Others attempt approximate predictive models that are potentially useful for design purposes (Card et al. 1983). One word that has appeared often in these discussions is "consistency." This is Schneiderman's first golden rule of dialog design, and many other authors advocate it as well. A notable exception to this advocacy of consistency comes from Grudin (1989), who argues that issues associated with consistency should be placed in a very broad context. He defines three types of consistency:

1. *Internal consistency:* The physical and graphic layout of the computer system, including such characteristics as those associated with command naming and use and dialog forms, are consistent if these internal features of the interface are the same across applications.
2. *External consistency:* If an interface has unchanging use features when compared to another interface with which the user is familiar, it is said to be externally consistent.
3. *Applications consistency:* If the user interface uses metaphors or analogous representations of objects that correspond to those of the real-world application, then the interface may be said to correspond to experientially familiar features of the world and to be applications consistent.

Two of Grudin's observations relevant to interface consistency are that ease of learning can conflict with ease of use, especially as experiential familiarity with the interface grows, and that consistency can work against both ease of use and learning. On the basis of some experiments illustrating these hypotheses, he establishes the three appropriate dimensions for consistency above.

A number of characteristics are desirable for user interfaces. Roberts and Moran (1982) identify the most important attributes of text editors as functionality of the editor, learning time required, time required to perform tasks, and errors committed. To this might be added the cost of evaluation. Harrison and Hix (1989) identify usability, completeness, extensibility, escapability, integration, locality of definition, structured guidance, and direct manipulation as well in their more general study of user interfaces. They also note a number of tools useful for interface development, as does Lee (1990).

Ravden and Johnson (1989) evaluate usability of human computer interfaces. They identify nine top-level attributes: visual clarity, consistency, compatibility, informative feedback, explicitness, appropriate functionality, flexibility and control, error prevention and correction, and user guidance and support. They disaggregate each into a number of more measurable attributes. These attributes can be used as part of a standard multiple-attribute evaluation.

A goal in DGMS design is to define an abstract user interface that can be implemented on specific operating systems in different ways. The purpose of this is to allow for device independence such that, for example, switching from a command line interface to a mouse-driven pull-down-menu interface can be easily accomplished. Separating the application from the user interface should do much towards ensuring portability across operating systems and hardware platforms without modifying the MBMS and the DBMS, which together comprise the applications software portions of the DSS.

5. GROUP AND ORGANIZATIONAL DECISION SUPPORT SYSTEMS

A group decision support system (GDSS) is an information technology-based support system designed to provide decision making support to groups and/or organizations. This could refer to a group meeting at one physical location at which judgments and decisions are made that affect an organization or group. Alternatively, it could refer to a spectrum of meetings of one or more individuals, distributed in location, time, or both. GDSSs are often called *organizational decision support systems,* and other terms are often used, including *executive support systems* (ESSs), which are information technology-based systems designed to support executives and managers, and *command and control systems,* which is a term often used in the military for a decision support system. We will generally use GDSS to describe all of these.

Managers and other knowledge workers spend much time in meetings. Much research into meeting effectiveness suggests that it is low, and proposals have been made to increase this through information technology support (Johansen 1988). Specific components of this information technology-based support might include computer hardware and software, audio and video technology, and communications media. There are three fundamental ingredients in this support concept: technological support facilities, the support processes provided, and the environment in which they are embedded. Kraemer and King (1988) provide a noteworthy commentary on the need for group efforts in their overview of GDSS efforts. They suggest that group activities are economically necessary, efficient as a means of production, and reinforcing of democratic values.

There are a number of predecessors for group decision support technology. Decision rooms, or situation rooms, where managers and boards meet to select from alternative plans or courses of action, are very common. The first computer-based decision support facility for group use is attributed to Douglas C. Engelbart, the inventor of the (computer) mouse, at Stanford in the 1960s. A discussion of this and other early support facilities is contained in Johansen (1988).

Engelbart's electronic boardroom-type design is acknowledged to be the first type of information technology-based GDSS. The electronic format was, however, preceded by a number of nonelectronic formats. The Cabinet war room of Winston Churchill is perhaps the most famous of these. Maps placed on the wall and tables for military decision makers were the primary ingredients of this room. The early 1970s saw the introduction of a number of simple computer-based support aids into situation rooms. The first system that resembles the GDSS in use today is often attributed to Gerald Wagner, the Chief Executive Officer of Execucom, who implemented a planning laboratory made up of a U-shaped table around which people sat, a projection TV system for use as a public viewing screen, individual small terminals and keyboards available to participants, and a minicomputer to which the terminals and keyboards were connected. This enabled participants to vote and to conduct simple spreadsheet-like exercises. Figure 15 illustrates the essential features of this concept. Most present-day GDSS centralized facilities look much like the conceptual illustration of a support room, or situation room, shown in this Figure.

As with a single-user DSS, appropriate questions for a GDSS that have major implications for design concern the perceptions and insights that the group obtains through use of the GDSS and the

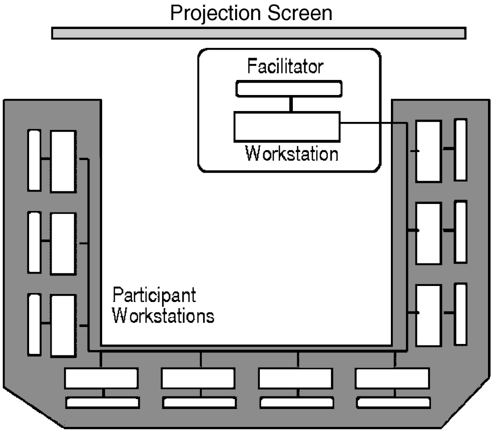


Figure 15 Early Group Decision Support System (GDSS) Situation Room.

activities that can be carried out through its use. Also, additional concerns arise regarding the public screen, interactions between the public screen and individual screens, the characteristics of individual work screens, and contingency task structural variables associated with the individuals in the group using the GDSS (Gray and Olfman 1989).

Huber (1982) has indicated both the needs for GDSS and how an appropriately designed GDSS can meet these needs. He identifies four interacting and complicating concerns:

1. Effective decision making requires not only obtaining an appropriate decision, but also ensuring that participants are happy with the process used to reach the decision and will be willing to meet and work cooperatively in the future.
2. Productivity losses occur because of dominant individuals and group pressures that lead to conformity of thought, or groupthink.
3. Miscommunications are common in group situations.
4. Insufficient time is often spent in situation assessment, problem exploration, and generation of alternative courses of action.

Huber further indicates that a GDSS can help improve the unaided decision situation, which often suffers from imperfect information processing and suboptimal decision selection.

A GDSS is made up of:

1. *Technological components,* in terms of computer hardware and software, and communication equipment
2. *Environmental components,* in terms of the people involved, their locations in time and space, and their experiential familiarity with the task at hand
3. *Process components,* or variables made up of the conventions used to support task performance and enable the other components of decision making to function appropriately.

We have already described the technological design features of DSSs. Thus, our commentary on technological design features of a GDSS will be brief. We do need to provide a perspective on groups and organizations, and we will do this in our next two sections. Then we will turn to some architectural considerations specifically relevant to GDSS.

5.1. Information Needs for Group and Organizational Decision Making

The nature of the decisions and the type of information required differ across each of these four levels identified in Figure 1. Generally, operational activities occur much more frequently than strategic planning activities. Also, there is a difference in the degree to which the knowledge required for each of these levels is structured. In 1960, Herbert Simon (Simon 1960) described decisions as structured or unstructured, depending upon whether the decision making process can be explicitly described prior to the time when it is necessary to make a decision. This taxonomy would seem to lead directly to that in which expert skills (holistic reasoning), rules (heuristics), or formal reasoning (holistic analysis) are normatively used for judgment. Generally, operational performance decisions are much more likely than strategic planning decisions to be prestructured. This gives rise to a number of questions concerning efficiency and effectiveness tradeoffs between training and aiding (Rouse 1991) that occur at these levels.

There are a number of human abilities that a GDSS should augment.

1. It should help the decision maker to formulate, frame, or assess the decision situation. This includes identifying the salient features of the environment, recognizing needs, identifying appropriate objectives by which we are able to measure successful resolution of an issue, and generating alternative courses of action that will resolve the needs and satisfy objectives.
2. It should provide support in enhancing the abilities of the decision maker to obtain and analyze the possible impacts of the alternative courses of action.
3. It should have the capability to enhance the decision maker's ability to interpret these impacts in terms of objectives. This interpretation capability will lead to evaluation of the alternatives and selection of a preferred alternative option.

Associated with each of these three formal steps of formulation, analysis, and interpretation must be the ability to acquire, represent, and utilize information and associated knowledge and the ability to implement the chosen alternative course of action.

Many attributes will affect the quality and usefulness of the information that is obtained, or should be obtained, relative to any given decision situation. These variables are very clearly contingency task dependent. Among these attributes are the following (Keen and Scott Morton 1978).

- *Inherent and required accuracy of available information:* Operational control and performance situations will often deal with information that is relatively accurate. The information in strategic planning and management control situations is often inaccurate.
- *Inherent precision of available information:* Generally, information available for operational control and operational performance decisions is very imprecise.
- *Inherent relevancy of available information:* Operational control and performance situations will often deal with information that is fairly relevant to the task at hand because it has been prepared that way by management. The information in strategic planning and management control situations is often obtained from the external environment and may be irrelevant to the strategic tasks at hand, although it may not initially appear this way.
- *Inherent and required completeness of available information:* Operational control and performance situations will often deal with information that is relatively complete and sufficient for operational performance. The information in strategic planning and management control situations is often very incomplete and insufficient to enable great confidence in strategic planning and management control.
- *Inherent and required verifiability of available information:* Operational control and performance situations will often deal with information that is relatively verifiable to determine correctness for the intended purpose. The information in strategic planning and management control situations is often unverifiable, or relatively so, and this gives rise to a potential lack of confidence in strategic planning and management control.
- *Inherent and required consistency and coherency of available information:* Operational control and performance situations will often deal with information that is relatively consistent and coherent. The information in strategic planning and management control situations is often inconsistent and perhaps even contradictory or incoherent, especially when it comes from multiple external sources.
- *Information scope:* Generally, but not always, operational decisions are made on the basis of relatively narrow scope information related to well-defined events that are internal to the organization. Strategic decisions are generally based upon broad-scope information and a wide range of factors that often cannot be fully anticipated prior to the need for the decision.
- *Information quantifiability:* In strategic planning, information is very likely to be highly qualitative, at least initially. For operational decisions, the available information is often highly quantified.
- *Information currency:* In strategic planning, information is often rather old, and it is often difficult to obtain current information about the external environment. For operational control decisions, very current information is often needed and present.
- *Needed level of detail:* Often very detailed information is needed for operational decisions. Highly aggregated information is often desired for strategic decisions. There are many difficulties associated with information summarization that need attention.

- *Time horizon for information needed:* Operational decisions are typically based on information over a short time horizon, and the nature of the control may change very frequently. Strategic decisions are based on information and predictions based on a long time horizon.
- *Frequency of use:* Strategic decisions are made infrequently, although they are perhaps refined fairly often. Operational decisions are made quite frequently and are relatively easily changed.
- *Internal or external information source:* Operational decisions are often based upon information that is available internal to the organization, whereas strategic decisions are much more likely to be dependent upon information content that can only be obtained external to the organization.

These attributes, and others, could be used to form the basis for an evaluation of information quality in a decision support system.

Information is used in a DSS for a variety of purposes. In general, information is equivalent to, or may be used as, evidence in situations in which it is relevant. Often information is used directly as a basis for testing an hypothesis. Sometimes it is used indirectly for this purpose. There are three different conditions for describing hypotheses (Schum 1987, 1994):

1. Different alternative hypotheses or assessments are possible if evidence is imperfect in any way. A hypothesis may be imperfect if it is based on imperfect information. Imperfect information refers to information that is incomplete, inconclusive, unreliable, inconsistent, or uncertain. Any or all of these alternate hypotheses may or may not be true.
2. Hypotheses may refer to past, present, or future events.
3. Hypotheses may be sharp (specific) or diffuse (unspecified). Sharp hypotheses are usually based on specific evidence rather than earlier diffuse hypotheses. An overly sharp hypothesis may contain irrelevant detail and invite invalidation by disconfirming evidence on a single issue in the hypothesis. An overly diffuse hypothesis may be judged too vague and too uninteresting by those who must make a decision based upon the hypothesis, even though the hypothesis might have been described in a more cogent manner.

The support for any hypothesis can always be improved by either revising a portion of the hypothesis to accommodate new evidence or gathering more evidence that infers the hypothesis. Hypotheses can be potentially significant for four uses:

1. *Explanations:* An explanation usually involves a model, which can be elaborate or simple. The explanation consists of the rationale for why certain events occurred.
2. *Event predictions:* In this case the hypothesis is proposed for a possible future event. It may include the date or period when the possible event will occur.
3. *Forecasting and estimation:* This involves the generation of a hypothesis based on data that does not exist or that is inaccessible.
4. *Categorization:* Sometimes it is useful to place persons, objects, or events into certain categories based upon inconclusive evidence linking the persons, objects, or events to these categories. In this case the categories represent hypotheses about category membership.

Assessment of the validity of a given hypothesis is inductive in nature. The generation of hypotheses and determination of evidence relevant to these hypotheses involve deductive and abductive reasoning. Hypotheses may be generated on the basis of the experience and prior knowledge that leads to analogous representations and recognitional decision making, as noted by Klein (1990, 1998).

Although no theory has been widely accepted on how to quantify the value of evidence, it is important to be able to support a hypothesis in some logical manner. Usually there is a major hypothesis that is inferred by supporting hypotheses, and each of these supporting hypotheses is inferred by its supporting hypothesis, and so on. Evidence is relevant to the extent that it causes one to increase or decrease the likeliness of an existing hypothesis, or modify an existing hypothesis, or create a new hypothesis. Evidence is direct if it has a straightforward bearing on the validity of the main hypothesis. Evidence is indirect if its effect on the main hypothesis is inferred through at least one other level of supporting hypothesis.

In many cases, it is necessary to acquire, represent, use, and/or communicate knowledge that is imperfect. This is especially important in group decision situations. In describing the structure of the beliefs and the statements that people make about issues that are of importance to them, the nature of the environment that surrounds them, as well as the ways in which people reason and draw conclusions about the environment and issues that are embedded into the environment, especially when there are conflicting pieces of information and opinions concerning these, people often attempt to use one or more of the forms of logical reasoning. Many important works deal with this subject. Of particular interest here is the work of Toulmin and his colleagues (Toulmin et al. 1979), who have described an explicit model of logical reasoning that is subject to analytical inquiry and computer implementation. The model is sufficiently general that it can be used to represent logical reasoning in a number of application areas.

Toulmin assumes that whenever we make a claim, there must be some ground on which to base our conclusion. He states that our thoughts are generally directed, in an inductive manner, from the grounds to the claim, each of which are statements that may be used to express both facts and values. As a means of explaining observed patterns of stating a claim, there must be some reason that can be identified with which to connect the grounds and the claim. This connection, called the warrant, gives the grounds–claim connection its logical validity.

We say that the grounds support the claim on the basis of the existence of a warrant that explains the connection between the grounds and the claim. It is easy to relate the structure of these basic elements with the process of inference, whether inductive or deductive, in classical logic. The warrants are the set of rules of inference, and the grounds and claim are the set of well-defined propositions or statements. It will be only the sequence and procedures, as used to formulate the three basic elements and their structure in a logical fashion, that will determine the type of inference that is used.

Sometimes, in the course of reasoning about an issue, it is not enough that the warrant will be the absolute reason to believe the claim on the basis of the grounds. For that, there is a need for further backing to support the warrant. It is this backing that provides for reliability, in terms of truth, associated with the use of a warrant. The relationship here is analogous to the way in which the grounds support the claim. An argument will be valid and will give the claim solid support only if the warrant is relied upon and is relevant to the particular case under examination. The concept of logical validity of an argument seems to imply that we can make a claim only when both the warrant and the grounds are certain. However, imprecision and uncertainty in the form of exceptions to the rules or low degree of certainty in both the grounds and the warrant do not prevent us on occasion from making a *hedge* or, in other words, a vague claim. Often we must arrive at conclusions on the basis of something less than perfect evidence, and we put those claims forward not with absolute and irrefutable truth but with some doubt or degree of speculation.

To allow for these cases, modal qualifiers and possible rebuttals may be added to this framework for logical reasoning. Modal qualifiers refer to the strength or weakness with which a claim is made. In essence, every argument has a certain modality. Its place in the structure presented so far must reflect the generality of the warrants in connecting the grounds to the claim. Possible rebuttals, on the other hand, are exceptions to the rules. Although modal qualifiers serve the purpose of weakening or strengthening the validity of a claim, there may still be conditions that invalidate either the grounds or the warrants, and this will result in deactivating the link between the claim and the grounds. These cases are represented by the possible rebuttals.

The resulting structure of logical reasoning provides a very useful framework for the study of human information processing activities. The order in which the six elements of logical reasoning have been presented serves only the purpose of illustrating their function and interdependence in the structure of an argument about a specific issue. It does not represent any normative pattern of argument formation. In fact, due to the dynamic nature of human reasoning, the concept formation and framing that result in a particular structure may occur in different ways. The six-element model of logical reasoning is shown in Figure 16.

Computer-based implementations of Figure 16 may assume a Bayesian inferential framework for processing information. Frameworks for Bayesian inference require probability values as primary inputs. Because most events of interest are unique or little is known about their relative frequencies of occurrence, the assessment of probability values usually requires human judgment. Substantial

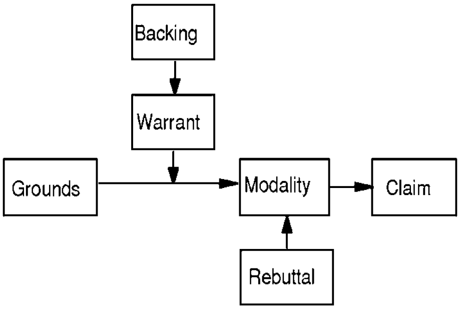


Figure 16 Representation of Toulmin Logic Structure.

psychological research has shown that people are unable to elicit probability values consistent with the rules of probabilities or to process information as they should, according to the laws of probability, in revising probability assessment when new information is obtained. For example, when people have both causal and diagnostic implications, they should weigh the causal and diagnostic impacts of the evidence. More often, however, unaided humans will reach judgments that suggest they apparently assess conditional probabilities primarily in terms of the direct causal effect of the impacts. If A is perceived to be the cause of B, for example, people will usually associate higher probabilities with $P(B|A)$ than they will with $P(A|B)$. Studies concerning the elicitation of probability values often report that individuals found it easier and showed more confidence in assessing $P(A|B)$ if B was causal to A. This strongly suggests that the choice of which form of inference to invoke depends more on the level of familiarity of the observer with the task at hand and the frame adopted to initially represent the knowledge. Again, this supports the wisdom of appropriate structuring of decision situations.

In general, grounds can be categorized by the several means in which are warranted:

- Empirical observation
- Expert judgment
- Enumerative induction (statistics)
- Experiment (hypothesis test)
- Direct fact

A decision assessment system based on this concept is described in Janssen and Sage (2000) together with application to the issues in public policy decision making associated with agricultural pest control. This system can support several user groups, who can use the support system to state arguments for or against an important policy issue and to assist in identifying and evaluating alternatives for implementation as policy decisions. The resulting decision support system assists in improving the clarity of the lines of reasoning used in specific situations; the warrants, grounds, and backings that are used to support claims and specific lines of reasoning; and the contradictions, rebuttals, and arguments surrounding each step in the reasoning process associated with evaluating a claim or counterclaim. Thus, experts and decision makers with differing views and backgrounds can better understand each other's thought processes in complex situations. The net effect is enhanced communications and understanding of the whole picture and, in many cases, consensus on decisions to be taken.

A number of human information processing capabilities and limitations interact with organizational arrangements and task requirements to strongly influence resource allocations for organizational problem solving and decision making. These needs have provided much motivation for the development of group decision support systems (GDSSs). The purpose of these GDSSs as computerized aids to planning, problem solving, and decision making include:

1. Removing a number of common communication barriers in groups and organizations
2. Providing techniques for the formulation, analysis, and interpretation of decisions
3. Systematically directing the discussion process and associated problem solving and decision making in terms of the patterns, timing, and content of the information that influences the actions that follow from decisions that have been taken

A number of variations and permutations are possible in the provision of group and organizational decision support. These are associated with specific realization or architectural format for a GDSS to support a set of GDSS performance objectives for a particular task in a particular environment.

The same maladies that affect individual decision making and problem solving behavior, as well as many others, can result from group and organizational limitations. A considerable body of knowledge, generally qualitative, exists relative to organizational structure, effectiveness, and decision making in organizations. The majority of these studies suggest that a bounded rationality or satisficing perspective, often heavily influenced by bureaucratic political considerations, will generally be the decision perspective adopted in actual decision making practice in organizations. To cope with this effectively requires the ability to deal concurrently with technological, environmental, and process concerns as they each, separately and collectively, motivate group and organizational problem solving issues.

The influencers of decision and decision process quality are particularly important in this. We should sound a note of caution regarding some possibly overly simplistic notions relative to this. Welch (1989) identifies a number of potential imperfections in organizational decision making and discusses their relationship to decision process quality. In part, these are based on an application of seven symptoms identified in Herek et al. (1987) to the Cuban Missile Crisis of 1962. These potential imperfections include:

1. Omissions in surveying alternative courses of action
2. Omissions in surveying objectives
3. Failure to examine major costs and risks of the selected course of action (COA)
4. Poor information search, resulting in imperfect information
5. Selective bias in processing available information
6. Failure to reconsider alternatives initially rejected, potentially by discounting favorable information and overweighing unfavorable information
7. Failure to work out detailed implementation, monitoring, and contingency plans

The central thrust of this study is that the relationship between the quality of the decision making process and the quality of the outcome is difficult to establish. This strongly suggests the usefulness of the contingency task structural model construct and the need for approaches that evaluate the quality of processes, as well as decisions and outcomes, and that consider the inherent embedding of outcomes and decisions within processes that lead to these.

Organizational ambiguity is a major reason why much of the observed ‘‘bounded rationality’’ behavior is so pervasive. March (1983) and March and Wessinger-Baylon (1986) show that this is very often the case, even in situations when formal rational thought or ‘‘vigilant information processing’’ (Janis and Mann 1977) might be thought to be a preferred decision style. March (1983) indicates that there are at least four kinds of opaqueness or equivocality in organizations: *ambiguity of intention, ambiguity of understanding, ambiguity of history,* and *ambiguity of human participation.* These four ambiguities relate to an organization’s structure, function, and purpose, as well as to the perception of these decision making agents in an organization. They influence the information that is communicated in an organization and generally introduce one or more forms of information imperfection. The notions of organizational management and organizational information processing are indeed inseparable. In the context of human information processing, it would not be incorrect to define the central purpose of management as development of a consensual grammar to ameliorate the effects of equivocality or ambiguity. This is the perspective taken by Karl Weick (1979, 1985) in his noteworthy efforts concerning organizations.

Starbuck (1985) notes that much direct action is a form of deliberation. He indicates that action should often be introduced earlier in the process of deliberation than it usually is and that action and thought should be integrated and interspersed with one another. The basis of support for this argument is that probative actions generate information and tangible results that modify potential thoughts. Of course, any approach that involves ‘‘act now, think later’’ behavior should be applied with considerable caution.

Much of the discussion to be found in the judgment, choice, and decision literature concentrates on what may be called formal reasoning and decision selection efforts that involve the issue resolution efforts that follow as part of the problem solving efforts of issue formulation, analysis, and interpretation that we have discussed here. There are other decision making activities, or decision-associated activities, as well. Very important among these are activities that allow perception, framing, editing and interpretation of the effects of actions upon the internal and external environments of a decision situation. These might be called information selection activities. There will also exist information retention activities that allow admission, rejection, and modification of the set of selected information or knowledge such as to result in short-term learning and long-term learning. Short-term learning results from reduction of incongruities, and long-term learning results from acquisition of new information that reflects enhanced understanding of an issue. Although the basic GDSS design effort may well be concerned with the short-term effects of various problem solving, decision making, and information presentation formats, the actual knowledge that a person brings to bear on a given problem is a function of the accumulated experience that the person possesses, and thus long-term effects need to be considered, at least as a matter of secondary importance.

It was remarked above that a major purpose of a GDSS is to enhance the value of information and, through this, to enhance group and organizational decision making. Three attributes of information appear dominant in the discussion thus far relative to value for problem solving purposes and in the literature in general:

1. *Task relevance:* Information must be relevant to the task at hand. It must allow the decision maker to know what needs to be known in order to make an effective and efficient decision. This is not as trivial a statement as might initially be suspected. Relevance varies considerably across individuals, as a function of the contingency task structure, and in time as well.
2. *Representational appropriateness:* In addition to the need that information be relevant to the task at hand, the person who needs the information must receive it in a form that is appropriate for use.

3. *Equivocality reduction:* It is generally accepted that high-quality information may reduce the imperfection or equivocality that might otherwise be present. This equivocality generally takes the form of uncertainty, imprecision, inconsistency, or incompleteness. It is very important to note that it is neither necessary nor desirable to obtain decision information that is unequivocal or totally ''perfect.'' Information need only be sufficiently unequivocal or unambiguous for the task at hand. To make it better may well be a waste of resources!

Each of these top-level attributes may be decomposed into attributes at a lower level. Each is needed as fundamental metrics for valuation of information quality. We have indicated that some of the components of equivocality or imperfection are uncertainty, imprecision, inconsistency, and incompleteness. A few of the attributes of representational appropriateness include naturalness, transformability to naturalness, and concision. These attributes of information presentation system effectiveness relate strongly to overall value of information concerns and should be measured as a part of the DSS and GDSS evaluation effort even though any one of them may appear to be a secondary theme.

We can characterize information in many ways. Among attributes that we noted earlier and might use are accuracy, precision, completeness, sufficiency, understandability, relevancy, reliability, redundancy, verifiability, consistency, freedom from bias, frequency of use, age, timeliness, and uncertainty. Our concerns with information involve at least five desiderata (Sage 1987):

1. Information should be presented in very clear and very familiar ways, such as to enable rapid comprehension.
2. Information should be such as to improve the precision of understanding of the task situation.
3. Information that contains an advice or decision recommendation component should contain an explication facility that enables the user to determine how and why results and advice are obtained.
4. Information needs should be based upon identification of the information requirements for the particular situation.
5. Information presentations and all other associated management control aspects of the support process should be such that the decision maker, rather than a computerized support system, guides the process of judgment and choice.

It will generally be necessary to evaluate a GDSS to determine the extent to which these information quality relevant characteristics are present.

5.2. The Engineering of Group Decision Support Systems

There are two fundamental types of decision making in an organization: individual decisions, made by a single person, and group or organizational decisions, made by a collection of two or more people. It is, of course, possible to disaggregate this still further. An individual decision may, for example, be based on the value system of one or more people and the individual making the decision may or may not have his or her values included. In a multistage decision process, different people may make the various decisions. Some authors differentiate between group and organizational decisions (King and Star 1992), but we see no need for this here, even though it may be warranted in some contexts. There can be no doubt at all, however, that a GDSS needs to be carefully matched to an organization that may use it.

Often groups make decisions differently from the way an individual does. Groups need protocols that allow effective inputs by individuals in the group or organization, a method for mediating a discussion of issues and inputs, and algorithms for resolving disagreements and reaching a group consensus. Acquisition and elicitation of inputs and the mediation of issues are usually local to the specific group, informed of personalities, status, and contingencies of the members of the group. Members of the group are usually desirous of cooperating in reaching a consensus on conflicting issues or preferences. The support for individual vs. group decisions is different and hence DSSs and GDSSs may require different designs. Because members of a group have different personalities, motivations, and experiential familiarities with the situation at hand, a GDSS must assist in supporting a wide range of judgment and choice perspectives.

It is important to note that the group of people may be centralized at one spot or decentralized in space and/or time. Also, the decision considered by each individual in a decision making group may or may not be the *ultimate* decision. The decision being considered may be sequential over time and may involve many component decisions. Alternatively, or in addition, many members in a decision making group may be formulating and/or analyzing options and preparing a short list of these for review by a person with greater authority or responsibility over a different portion of the decision making effort.

Thus, the number of possible types of GDSS may be relatively extensive. Johansen (1988) has identified no less than 17 approaches for computer support in groups in his discussion of groupware.

1. *Face-to-face meeting facilitation services:* This is little more than office automation support in the preparation of reports, overheads, videos, and the like that will be used in a group meeting. The person making the presentation is called a ''facilitator'' or ''chauffeur.''
2. *Group decision support systems:* By this, Johansen essentially infers the GDSS structure shown in Figure 15 with the exception that there is but a single video monitor under the control of a facilitator or chauffeur.
3. *Computer-based extensions of telephony for use by work groups:* This involves use of either commercial telephone services or private branch exchanges (PBXs). These services exist now, and there are several present examples of conference calling services.
4. *Presentation support software:* This approach is not unlike that of approach 1, except that computer software is used to enable the presentation to be contained within a computer. Often those who will present it prepare the presentation material, and this may be done in an interactive manner to the group receiving the presentation.
5. *Project management software:* This is software that is receptive to presentation team input over time and that has capabilities to organize and structure the tasks associated with the group, often in the form of a Gantt chart. This is very specialized software and would be potentially useful for a team interested primarily in obtaining typical project management results in terms of PERT charts and the like.
6. *Calendar management for groups:* Often individuals in a group need to coordinate times with one another. They indicate times that are available, potentially with weights to indicate schedule adjustment flexibility in the event that it is not possible to determine an acceptable meeting time.
7. *Group authoring software:* This allows members of a group to suggest changes in a document stored in the system without changing the original. A lead person can then make document revisions. It is also possible for the group to view alternative revisions to drafts. The overall objective is to encourage, and improve the quality and efficiency of, group writing. It seems very clear that there needs to be overall structuring and format guidance, which, while possibly group determined, must be agreed upon prior to filling out the structure with report details.
8. *Computer-supported face-to-face meetings:* Here, individual members of the group work directly with a workstation and monitor, rather than having just a single computer system and monitor. A large screen video may, however, be included. This is the sort of DSS envisioned in Figure 14. Although there are a number of such systems in existence, the Colab system at Xerox Palo Alto Research Center (Stefik et al. 1987) was one of the earliest and most sophisticated. A simple sketch of a generic facility for this purpose might appear somewhat as shown in Figure 17. Generally, both public and private information are contained in these systems. The public information is shared, and the private information, or a portion of it, may be converted to public programs. The private screens normally start with a menu screen from which participants can select activities in which they engage, potentially under the direction of a facilitator.
9. *Screen sharing software:* This software enables one member of a group to selectively share screens with other group members. There are clearly advantages and pitfalls in this. The primary advantage to this approach is that information can be shared with those who have a reason to know specific information without having to bother others who do not need it. The disadvantage is just this also, and it may lead to a feeling of ganging up by one subgroup on another subgroup.
10. *Computer conferencing systems:* This is the group version of electronic mail. Basically, what we have is a collection of DSSs with some means of communication among the individuals that comprise the group. This form of communication might be regarded as a product hierarchy in which people communicate.
11. *Text filtering software:* This allows system users to search normal or semistructured text through the specification of search criteria that are used by the filtering software to select relevant portions of text. The original system to accomplish this was Electronic Mail Filter (Malone et al. 1987). A variety of alternative approaches are also being emphasized now.
12. *Computer-supported audio or video conferences:* This is simply the standard telephone or video conferencing, as augmented by each participant having access to a computer and appropriate software.
13. *Conversational structuring:* This involves identification and use of a structure for conversations that is presumably in close relationship to the task, environment, and experiential fa-

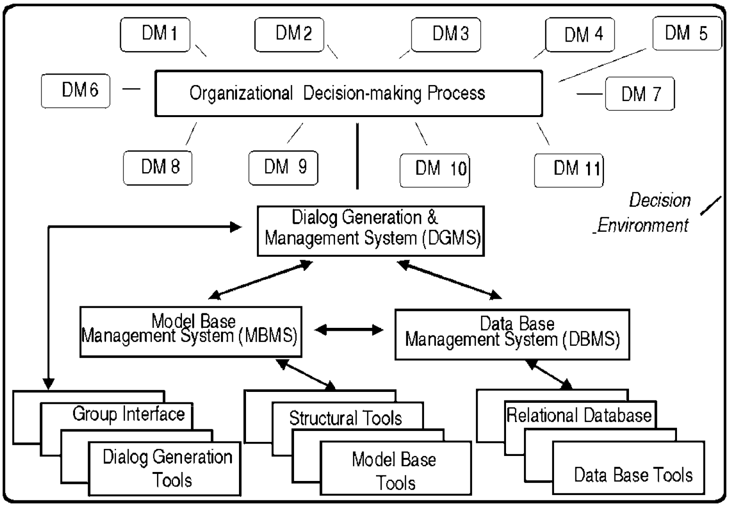


Figure 17 Level II (No Facilitator) and Level III (with Facilitator) GDSS.

miliarity of the group with the issues under consideration (Winograd and Flores 1986). For these group participants, structured conversations should often provide for enhanced efficiency and effectiveness or there may be a perception of unwarranted intrusions that may defeat the possible advantages of conversational structuring.

14. *Group memory management:* This refers to the provision of support between group meetings such that individual members of a group can search a computer memory in personally preferred ways through the use of very flexible indexing structures. The term *hypertext* (Nielson 1989) is generally given to this flexible information storage and retrieval. One potential difficulty with hypertext is the need for a good theory of how to prepare the text and associated index such that it can be indexed and used as we now use a thesaurus. An extension of hypertext to include other than textual material is known as hypermedia.
15. *Computer-supported spontaneous interaction:* The purpose of these systems is to encourage the sort of impromptu and extemporaneous interaction that often occurs at unscheduled meetings between colleagues in informal setting, such as a hallway. The need for this could occur, for example, when it is necessary for two physically separated groups to communicate relative to some detailed design issue that needs to be resolved in order to continue product development.
16. *Comprehensive work team support:* This refers to integrated and comprehensive support, such as perhaps might be achieved through use of the comprehensive DSS design philosophy described above.
17. *Nonhuman participants in team meetings:* This refers to the use of unfacilitated DSS and expert systems that automate some aspects of the process of decision making.

According to Johansen (1988), the order in which these are described above also represents the order of increasing difficulty of implementation and successful use. These scenarios of support for decision making are also characterized in terms of support for face-to-face meetings (1, 2, 4, 8); support for electronic meetings (3, 9, 10, 11, 12); and support between meetings (5, 6, 7, 13, 14, 15, 16). There is much interest in groupware as a focused subject within modern information technology developments (Shapiro et al. 1996; Chaffey 1998; Smith 1999). A number of groupware products are available, Lotus Notes arguably being the best known.

A GDSS may and doubtless will influence the process of group decision making, perhaps strongly. A GDSS has the potential for changing the information-processing characteristics of individuals in the group. This is one reason why organizational structure and authority concerns are important ingredients in GDSS designs. In one study of the use of a GDSS to facilitate group consensus (Watson et al. 1988), it was found that:

1. GDSS use tended to reduce face-to-face interpersonal communication in the decision making group.

2. GDSS use posed an intellectual challenge to the group and made accomplishment of the purpose of their decision making activity more difficult than for groups without the GDSS.
3. The groups using the GDSS became more process oriented and less specific-issue oriented than the groups not using the GDSS.

Support may be accomplished at any or all of the three levels for group decision support identified by DeSanctis and Gallupe (1987) in their definitive study of GDSS. A GDSS provides a mechanism for group interaction. It may impose any of various structured processes on individuals in the group, such as a particular voting scheme. A GDSS may impose any of several management control processes on the individuals on the group, such as that of imposing or removing the effects of a dominant personality. The design of the GDSS and the way in which it is used are the primary determinants of these.

DeSanctis and Gallupe have developed a taxonomy of GDSS. A Level I GDSS would simply be a medium for enhanced information interchange that might lead ultimately to a decision. Electronic mail, large video screen displays that can be viewed by a group, or a decision room that contains these features could represent a Level I GDSS. A Level I GDSS provides only a mechanism for group interaction. It might contain such facilities as a group scratchpad, support for meeting agenda development, and idea generation and voting software.

A Level II GDSS would provide various decision structuring and other analytic tools that could act to reduce information imperfection. A decision room that contained software that could be used for problem solution would represent a Level II GDSS. Thus, spreadsheets would primarily represent a Level II DSS. A Level II GDSS would also have to have some means of enabling group communication. Figure 17 represents a Level II GDSS. It is simply a communications medium that has been augmented with some tools for problem structuring and solution with no prescribed management control of the use of these tools.

A Level III GDSS also includes the notion of management control of the decision process. Thus, a notion of facilitation of the process is present, either through the direct intervention of a human in the process or through some rule-based specifications of the management control process that is inherent in Level III GDSS. Clearly, there is no sharp transition line between one level and the next, and it may not always be easy to identify at what level a GDSS is operating. The DSS generator, such as discussed in our preceding section, would generally appear to produce a form of Level III DSS. In fact, most of the DSSs that we have been discussing in this book are either Level II or Level III DSSs or GDSSs. The GDSS of Figure 17, for example, becomes a Level III GDSS if is supported by a facilitator.

DeSanctis and Gallupe (1987) identify four recommended approaches:

1. Decision room for small group face-to-face meetings
2. Legislative sessions for large group face-to-face meetings
3. Local area decision networks for small dispersed groups
4. Computer-mediated conferencing for large groups that are dispersed

DeSanctis and Gallupe discuss the design of facilities to enable this, as well as techniques whereby the quality of efforts such as generation of ideas and actions, choosing from among alternative courses of action, and negotiating conflicts may be enhanced. On the basis of this, they recommend six areas as very promising for additional study: GDSS design methodologies; patterns of information exchange; mediation of the effects of participation; effects of (the presence or absence of) physical proximity, interpersonal attraction, and group cohesion; effects on power and influence; and performance/satisfaction trade-offs. Each of these supports the purpose of computerized aids to planning, problem solving, and decision making: removing a number of common communication barriers; providing techniques for structuring decisions; and systematically directing group discussion and associated problem-solving and decision making in terms of the patterns, timing, and content of the information that influences these actions.

We have mentioned the need for a GDSS model base management system (MBMS). There are a great variety of MBMS tools. Some of the least understood are group tools that aid in the issue formulation effort. Because these may represent an integral part of a GDSS effort, it is of interest to describe GDSS issue formation here as one component of a MBMS.

Other relevant efforts and interest areas involving GDSS include the group processes in computer mediated communications study; the computer support for collaboration and problem solving in meetings study of Stefik et al. (1987); the organizational planning study of Applegate et al. (1987), and the knowledge management and intelligent information sharing systems study of Malone et al. (1987). Particularly interesting current issues surround the extent to which cognitive science and engineering studies that involve potential human information processing flaws can be effectively dealt with, in the sense of design of debiasing aids, in GDSS design. In a very major way, the purpose of a GDSS is to support enhanced organizational communications. This is a very important contemporary subject, as evidenced by recent efforts concerning virtual organizations (DeSanctis and Monge 1999; Jablin and Putnam 2000).

5.3. Distributed Group Decision Support Systems

A single-user DSS must provide single-user-single-model and single-user-multiple-model support, whereas a GDSS model base management system (MBMS) must support multiple-user-single-model and multiple-user-multiple-model support. A centralized GDSS induces three basic components:

1. A group model base management subsystem (GMBMS)
2. A group database management subsystem (GDBMS)
3. A group dialog generation and management subsystem (GDGMS)

These are precisely the components needed in a single-user DSS, except for the incorporation of group concerns. In the GDSS, all data are stored in the group database and models are stored in the group model base. The GMBMS controls all access to the individually owned and group-owned models in the model base.

A distributed GDSS allows each user to have an individual DSS and a GDSS. Each individual DSS consists of models and data for a particular user. The GDSS maintains the GMBMS and GDBMS, as well as controlling access to the GMBMS and GDBMS and coordination of the MBMSs of various individual DSSs (Liang 1988). During actual GDSS system use, the difference between the centralized and distributed GDSSs should be transparent to the users.

We generally use models to help define, understand, organize, study, and solve problems. These range from simple mental models to complex mathematical simulation models. An important mission of a GDSS is to assist in the use of formal and informal models by providing appropriate model management. An appropriate model base management system for a GDSS can provide the following four advantages (Hwang 1985):

1. *Reduction of redundancy,* since models can be shared
2. *Increased consistency,* since more than one decision maker will share the same model
3. *Increased flexibility,* since models can be upgraded and made available to all members of the group
4. *Improved control over the decision process,* by controlling the quality of the models adopted

A model base management system provides for at least the following five basic functions (Blanning and King 1993): construction of new models, storage of existing and new models, access to and retrieval of existing models, execution of existing models, and maintenance of existing models. MBMSs should also provide for model integration and selection. Model integration by using the existing model base as building blocks in the construction of new or integrated models is very useful when ad hoc or prototype models are desired. Model integration is needed in the production of operational MBMSs.

In this section, we have provided a very broad overview of group decision support systems that potentially support group and organizational decision making functions. Rather than concentrate on one or two specific systems, we have painted a picture of the many requirements that must be satisfied in order to produce an acceptable architecture and design for these systems. This provides much fertile ground for research in many GDSS-relevant cognitive systems engineering areas (Rasmussen et al. 1995; Andriole and Adelman 1995).

6. KNOWLEDGE MANAGEMENT FOR DECISION SUPPORT

There can be no doubt that contemporary developments in information technology have changed engineering and business practices in many ways. The information revolution has, for example, created entirely new ways of marketing and pricing such that we now see very changed relationships among producers, distributors, and customers. It has also led to changes in the ways in which organizations are managed and structured, and deal with their products and services. In particular, it creates a number of opportunities and challenges that affect the way in which data is converted into information and then into knowledge. It poses many opportunities for management of the environment for these transfers, such as enhancing the productivity of individuals and organizations. Decision support is much needed in these endeavors, and so is knowledge management. It is fitting that we conclude our discussions of decision support systems with a discussion of knowledge management and the emergence in the 21st century of integrated systems to enhance knowledge management and decision support.

Major growth in the power of computing and communicating and associated networking is fundamental to the emergence of these integrated systems and has changed relationships among people, organizations, and technology. These capabilities allow us to study much more complex issues than was formerly possible. They provide a foundation for dramatic increases in learning and both individual and organizational effectiveness. This is due in large part to the networking capability that enables enhanced coordination and communications among humans in organizations. It is also due to the vastly increased potential availability of knowledge to support individuals and organizations in their efforts, including decision support efforts. However, information technologies need to be appropriately integrated within organizational frameworks if they are to be broadly useful. This poses a transdisciplinary challenge (Somerville and Rapport 2000) of unprecedented magnitude if we are to move from high-performance information technologies and high-performance decision support systems to high-performance organizations.

In years past, broadly available capabilities never seemed to match the visions proffered, especially in terms of the time frame of their availability. Consequently, despite these compelling predictions, traditional methods of information access and utilization continued their dominance. As a result of this, comments something like ‘‘computers are appearing everywhere except in productivity statistics’’ have often been made (Brynjolfsson and Yang 1996). In just the past few years, the pace has quickened quite substantially, and the need for integration of information technology issues with organizational issues has led to the creation of related fields of study that have as their objectives:

- Capturing human information and knowledge needs in the form of system requirements and specifications
- Developing and deploying systems that satisfy these requirements
- Supporting the role of cross-functional teams in work
- Overcoming behavioral and social impediments to the introduction of information technology systems in organizations
- Enhancing human communication and coordination for effective and efficient workflow through knowledge management

Because of the importance of information and knowledge to an organization, two related areas of study have arisen. The first is concerned with technologies associated with the effective and efficient acquisition, transmission, and use of information, or *information technology*. When associated with organizational use, this is sometimes called *organizational intelligence* or *organizational informatics*. The second area, known as *knowledge management*, refers to an organization’s capacity to gather information, generate knowledge, and act effectively and in an innovative manner on the basis of that knowledge. This provides the capacity for success in the rapidly changing or highly competitive environments of knowledge organizations. Developing and leveraging organizational knowledge is a key competency and, as noted, it requires information technology as well as many other supporting capabilities. Information technology is necessary for enabling this, but it is not sufficient in itself. Organizational productivity is not necessarily enhanced unless attention is paid to the human side of developing and managing technological innovation (Katz 1997) to ensure that systems are designed for human interaction.

The human side of knowledge management is very important. Knowledge capital is sometimes used to describe the intellectual wealth of employees and is a real, demonstrable asset. Sage (1998) has used the term *systems ecology* to suggest managing organizational change to create a knowledge organization and enhance and support the resulting intellectual property for the production of sustainable products and services. Managing information and knowledge effectively to facilitate a smooth transition into the Information Age calls for this systems ecology, a body of methods for systems engineering and management (Sage 1995; Sage and Rouse, 1999a,b) that is based on analogous models of natural ecologies. Such a systems ecology would enable the modeling, simulation, and management of truly large systems of information and knowledge, technology, humans, organizations, and the environments that surround them.

The information revolution is driven by technology and market considerations and by market demand and pull for tools to support transaction processing, information warehousing, and knowledge formation. Market pull has been shown to exert a much stronger effect on the success of an emerging technology than technology push. Hardly any conclusion can be drawn other than that society shapes technology (Pool 1997) or, perhaps more accurately stated, that technology and the modern world shape each other in that only those technologies that are appropriate for society will ultimately survive.

The potential result of this mutual shaping of information technology and society is knowledge capital, and this creates needs for knowledge management. Current industrial and management efforts are strongly dependent on access to information. The world economy is in a process of globalization, and it is possible to detect several important changes. The contemporary and evolving world is much more service oriented, especially in the more developed nations. The service economy is much more information and knowledge dependent and much more competitive. Further, the necessary mix of job skills for high-level employment is changing. The geographic distance between manufacturers and consumers and between buyers and sellers is often of little concern today. Consequently, organizations from diverse locations compete in efforts to provide products and services. Consumers potentially benefit as economies become more transnational.

Information technology-based systems may be used to support taking effective decisions. Ideally, this is accomplished through both critical attention to the information needs of humans in problem solving and decision making task and provision of technological aids, including computer-based systems of hardware and software and associated processes, to assist in these tasks. There is little question but that successful information systems strategies seek to meaningfully evolve the overall architecture of systems, the systems' interfaces with humans and organizations, and their relations with external environments. In short, they seek to enhance systems integration effectiveness (Sage and Lynch 1998).

Although information technology and information systems do indeed potentially support improvement of the designs of existing organizations and systems, they also enable fundamentally new ones, such as virtual corporations (DeSanctis and Monge 1999), and they also enable major expansions of organizational intelligence and knowledge. They do this not only by allowing for interactivity in working with clients to satisfy present needs, but also through proactivity in planning and plan execution. An ideal organizational knowledge strategy accounts for future technological, organizational, and human concerns to support the graceful evolution of products and services that aid clients. Today, we realize that human and organizational considerations are vital to success in using information technology to better support decisions. A major challenge is to ensure that people gain maximal benefit from these capabilities. This is why information technology must be strongly associated with information ecology, knowledge management, and other efforts that we discuss here and that will ultimately lead to an effective systems ecology (Sage 1998).

There are three keys to organizations prospering in this type of environment: speed, flexibility, and discretion (Rouse 1999). Speed means rapid movement in understanding a situation, such as a new product development or market opportunity; formulating a plan for pursuing this opportunity, such as an intended joint venture for the new product; and deploying this plan so as to proceed through to product availability in stores. Flexibility is crucial for reconfiguring and redesigning organizations, and consequently reallocating resources. Functional walls must be quite portable, and generally the few of them the better.

Discretion transforms flexibility into speed. Distributed organizations must be free to act. While they may have to play by the contemporary and evolutionary rules of the game, they need to be able to adapt rapidly when things are not working well. Resources can thus be deployed effectively and speedily and results monitored quickly. Resource investments that are not paying off in the anticipated time frame can be quickly redeployed elsewhere, thereby ensuring adaptive and emergent evolution of the organization.

A major determinant of these organizational abilities is the extent to which an organization possesses intellectual capital, or knowledge capital, such that it can create and use innovative ideas to produce productive results. The concept of intellectual capital has been defined in various ways (Brooking 1996; Edvisson and Malone 1997; Stewart 1997; Klein 1998; Roos et al. 1998). We would add communications to the formulation of Ulrick (1998), representing intellectual capital to yield:

$$\text{Intellectual capital} = \text{Competence} \times \text{Commitment} \times \text{Communications}$$

Other important terms, such as *collaboration* and *courage,* could be added to this generic equation.

Loosely structured organizations and the speed, flexibility, and discretion they engender in managing intellectual capital fundamentally affect knowledge management (Myers 1997; Ruggles 1997; Prusak 1997; Albert and Bradley 1997; Liebowitz and Wilcox, 1997). Knowledge workers are no longer captive and hence know-how is not "owned" by the organization. What matters most is the ability to make sense of market and technology trends, quickly decide how to take advantage of these trends, and act faster than other players. Sustaining competitive advantage requires redefining market-driven value propositions and quickly leading in providing value in appropriate new ways. Accomplishing this in an increasingly information-rich environment is a major challenge, both for organizations experiencing contemporary business environments (Allee 1997; Stacey 1996) and for those who devise and provide decision support systems for supporting these new ways. The major interactions involving knowledge work and intellectual capital and the communications-driven information and knowledge revolution suggest many and profound, complex, adaptive system-like changes in the economy of the 21st century (Hagel and Armstrong 1997; Shapiro and Varian 1999; Kelly 1998; Hagel and Singer 1999). In particular, this has led to the notion of virtual enterprises and virtual communities and markets where customers make the rules that enhance net gain and net worth.

All of this creates major challenges for the evolution of knowledge organizations and appropriate knowledge management. One of the major challenges is that of dealing in an appropriate manner with the interaction among humans, organizations, and technologies and the environment surrounding these. Davenport and Prusak (1998) note that when organizations interact with environments, they absorb information and turn it into knowledge. Then they make decisions and take actions. They suggest five modes of knowledge generation:

1. Acquisition of knowledge that is new to the organization and perhaps represents newly created knowledge. Knowledge-centric organizations need to have appropriate knowledge available when it is needed. They may buy this knowledge, potentially through acquisition of another company, or generate it themselves. Knowledge can be leased or rented from a knowledge source, such as by hiring a consultant. Generally, knowledge leases or rentals are associated with knowledge transfer.
2. Dedicated knowledge resource groups may be established. Because time is required for the financial returns on research to be realized, the focus of many organizations on short-term profit may create pressures to reduce costs by reducing such expenditures. Matheson and Matheson (1998) describe a number of approaches that knowledge organizations use to create value through strategic research and development.
3. Knowledge fusion is an alternative approach to knowledge generation that brings together people with different perspectives to resolve an issue and determine a joint response. Nonaka and Takeuchi (1995) describe efforts of this sort. The result of knowledge fusion efforts may be creative chaos and a rethinking of old presumptions and methods of working. Significant time and effort are often required to enable group members to acquire sufficient shared knowledge, work effectively together, and avoid confrontational behavior.
4. Adaptation through providing internal resources and capabilities that can be utilized in new ways and being open to change in the established ways of doing business. Knowledge workers who can acquire new knowledge and skills easily are the most suitable to this approach. Knowledge workers with broad knowledge are often the most appropriate for adaptation assignments.
5. Knowledge networks may act as critical conduits for innovative reasoning. Informal networks can generate knowledge provided by a diversity of participants. This requires appropriate allocation of time and space for knowledge acquisition and creation.

In each of these efforts, as well as in much more general situations, it is critical to regard technology as a potential enabler of human effort, not as a substitute for it. There are, of course, major feedbacks here because the enabled human efforts create incentives and capabilities that lead to further enhanced technology evolution.

Knowledge management (Nonaka and Takeuchi 1995; Cordata and Woods 1999; Bukowitz and Williams, 1999) refers to management of the environment for knowledge creation, transfer, and sharing throughout the organization. It is vital in fulfilling contemporary needs in decision support. Appropriate knowledge management considers knowledge as the major organizational resource and growth through enhanced knowledge as a major organizational objective. While knowledge management is dependent to some extent upon the presence of information technology as an enabler, information technology alone cannot deliver knowledge management. This point is made by McDermott (1999), who also suggests that the major ingredient in knowledge management, leveraging knowledge, is dramatically more dependent upon the communities of people who own and use it than upon the knowledge itself.

Knowledge management is one of the major organizational efforts brought about by the realization that knowledge-enabled organizations are best posed to continue in a high state of competitive advantage. Such terms as *new organizational wealth* (Sveiby 1997), *intellectual capital* (Brooking 1996; Edvinsson and Malone 1997; Klein 1998; Roos et al. 1998), the *infinite resource* (Halal 1998), and *knowledge assets* (Boisot 1998) are used to describe the knowledge networking (Skyrme, 1999) and working knowledge (Davenport and Prusak 1998) in knowledge-enabled organizations (Liebowitz and Beckman 1998; Tobin 1998). Major objectives of knowledge management include supporting organizations in turning information into knowledge (Devin 1999) and, subsequent to this through a strategy formation and implementation process (Zack 1999), turning knowledge into action (Pfeffer and Sutton 2000). This latter accomplishment is vital because a knowledge advantage is brought to fruition only through an action advantage. Appropriate information technology and knowledge management tools (Ruggles 1997) are necessary but not at all sufficient to enable knowledge creating organizations or knowing organizations. The result of appropriate creation of knowledge in organizations (Prusak 1997) leads to a very important result, a knowing organization in which organizations use information to construct meaning, to create knowledge, and to use this knowledge in taking decisions. The term *fourth generation R&D* has been given to the efforts necessary for management of knowledge, technology, and innovation. A framework to enable this is described in Miller and Morris (1999).

The major increase in interest in knowledge management in recent years has been brought about by the reality that contemporary engineering and business success are progressively more linked to abilities associated with the management of data, information, and knowledge. Knowledge management is concerned with knowledge creation, knowledge acquisition, knowledge packaging, knowledge transfer, and knowledge use and reuse by humans and organizations. Knowledge management and decision support each support a common purpose: making decisions and actions taken on the basis of information and knowledge more effective. Thus, it is reasonable to suggest that an objective of knowledge management is to make appropriate knowledge available in a timely and cost-effective manner to decision makers. It seems very safe to predict that computer-based decision support systems will increasingly employ or be associated with various knowledge management techniques to enhance the representation and processing of information such that it can be best associated with contingency task structures to become the beneficial knowledge that is absolutely needed for effective decision support.

The development of intellectual capital such as to optimize organizational value is of particular importance in effective decision support. This is accomplished by using information to construct meaning and create knowledge and thereby enable appropriate decision and action. Such an organization has been called a "knowing organization" (Choo 1998). The major challenge in all of this is very often that of making sense of a wealth of opportunities concerning alternative schools and strategies. These include dramatic increases in available data, as well as an ever-growing set of potentially useful methods and tools—and the major need to convert data into information and thence into knowledge—and to manage knowledge successfully in the networked economy. Sage and Rouse (1999a,b) identifies 10 important challenges and paradigms: systems modeling, emergent and complex phenomena, uncertainties and control, access to and utilization of information and knowledge, information and knowledge requirements, information and knowledge support systems, inductive reasoning, learning organizations, planning and design, and measurement and evaluation. Ongoing trends in information technology and knowledge management pose substantial challenges for information systems frontiers. Addressing the 10 key challenges elaborated here requires a new, broader perspective on the nature of information access and utilization, as well as knowledge management. Satisfactorily addressing these 10 key challenges, will require that decision support systems engineering efforts move beyond structure-bound views of the world and the natural tendency to nail down requirements and constraints before proceeding. The current dynamics of information technology and knowledge management make such "givens" obsolete almost as quickly as they are envisioned. These appears to be the major decision support systems engineering challenges today, and in a very real sense they are challenges for all of engineering and engineering management. In large part, they provide a major motivation for this Handbook.

REFERENCES

Albert, S., and Bradley, K. (1997), *Managing Knowledge: Experts, Agencies, and Organizations,* Cambridge University Press, Cambridge.

Alberts, D. S., and Papp, D. S., Eds. (1997), *The Information Age: An Anthology of Its Impacts and Consequences,* National Defense University Press, Washington, DC.

Allee, V. (1997), *The Knowledge Evolution: Expanding Organizational Intelligence,* Butterworth-Heinemann, Boston.

Andriole, S., and Adelman, L. (1995), *Cognitive Systems Engineering for User–Computer Interface Design, Prototyping, and Evaluation,* Lawrence Erlbaum Associates, Hillsdale, NJ.

Anthony, R. N. (1965), *Planning and Control Systems: A Framework for Analysis*. Harvard University Press, Cambridge, MA.

Anthony, R. N., Dearden, N. J., and Govindarajan, V. (1992), *Management Control Systems,* Richard D. Irwin, Homewood, IL.

Applegate, L. M, Konsynski, B. R., and Nunamaker, J. F. (1986), "Model Management Systems: Design for Decision Support," *Decision Support Systems,* Vol. 2, No. 1, pp. 81–91.

Applegate, L. M., Chen, T. T., Konsynski, B. R., and Nunamaker, J. F. (1987), "Knowledge Management in Organizational Planning," *Journal of Management Information Systems,* Vol. 3, No. 4, pp. 20–38.

Arden, B. W., Ed. (1980), "What Can Be Automated?," Chapter 10 in *The Computer Science and Engineering Research Study,* MIT Press, Cambridge, MA.

Atzeni, P., and Chen, P. P. (1983), "Completeness of Query Languages for the Entity-Relationship Model," in *Entity-Relationship Approach to Information Modeling and Analysis,* Ed. P. P. Chen, North Holland, Amsterdam.

Atzeni, P., Ceri, S., Paraboschi, S., and Torione, R. (2000), *Database Systems: Concepts, Languages and Architectures,* McGraw-Hill, New York.

Banks, J., Ed. (1998), *Handbook of Simulation,* John Wiley & Sons, New York.

Barbosa, L. C., and Herko, R. G. (1980), ''Integration of Algorithmic Aids into Decision Support Systems,'' *MIS Quarterly,* Vol. 4, No. 3, pp. 1–12.

Bennet, J. L., Ed. (1983), *Building Decision Support Systems*, Addison-Wesley, Reading, MA.

Blanning, R. W., and King, D. R. (1993), *Current Research in Decision Support Technology,* IEEE Computer Society Press, Los Altos, CA.

Boisot, M. H. (1998), *Knowledge Assets: Securing Competitive Advantage in the Information Economy,* Oxford University Press, New York.

Brooking, A. (1996), *Intellectual Capital: Core Asset for the Third Millennium Enterprise,* Thompson Business Press, London.

Brynjolfsson, E., and Yang, S. (1996), ''Information Technology and Productivity: A Review of the Literature,'' in *Advances in Computers,* Vol. 43, pp. 179–214.

Bukowitz, W. R., and Williams, R. L. (1999), *The Knowledge Management Fieldbook,* Financial Times Prentice Hall, London.

Card, S.K., Moran, T. P., and Newell, A. (1993), *The Psychology of Human Computer Interaction.* Lawrence Erlbaum, Hillsdale, NJ.

Chaffey, D. (1998), *Groupware, Workflow and Intranets: Reengineering the Enterprise with Collaborative Software,* Digital Press, Boston.

Chen, P. P. S. (1976), ''The Entity-Relationship Model: Towards a Unified View of Data,'' *ACM Transactions on Database Systems,* Vol. 1, pp. 9–36.

Choo, C. W. (1998), *The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge, and Make Decisions,* Oxford University Press, New York.

Coad, P., and Yourdon, E. (1990), *Object-Oriented Analysis,* Prentice Hall, Englewood Cliffs, NJ.

Cordata, J. W., and Woods, J. A., Eds. (1999), *The Knowledge Management Yearbook 1999–2000,* Butterworth-Heinemann, Woburn, MA.

Darwen, H., and Date, C. J. (1998), *Foundation for Object/Relational Databases: The Third Manifesto,* Addison-Wesley, Reading MA.

Date, C. J. (1983), *Database: A Primer,* Addison-Wesley, Reading, MA.

Date, C. J. (1999), *An Introduction to Database Systems,* 7th Ed., Addison-Wesley, Reading, MA.

Davenport, T. H., and Prusak, L. (1994), *Working Knowledge: How Organizations Manage What They Know,* Harvard Business School Press, Boston.

Davis, G. B. (1982), ''Strategies for Information Requirements Determination,'' *IBM Systems Journal*, Vol. 21, No. 1, pp. 4–30.

Debenham, J. (1998), *Knowledge Engineering: Unifying Knowledge Base and Database Design,* Springer-Verlag, Berlin.

DeSanctis, G., and Gallupe, R. B. (1987), ''A Foundation for the Study of Group Decision Support Systems,'' *Management Science,* Vol. 33, pp. 547–588.

DeSanctis, G., and Monge, P., Eds. (1999), Special Issue: Communications Processes for Virtual Organizations, *Organizational Science,* Vol. 10, No. 6, November–December.

Devin, K. (1999), *Infosense: Turning Information into Knowledge,* W. H. Freeman & Co., New York.

Dolk, D. R. (1986), ''Data as Models: An Approach to Implementing Model Management,'' *Decision Support Systems,* Vol. 2, No. 1, pp, 73–80.

Dutta, P. K. (1999), *Strategies and Games: Theory and Practice,* MIT Press, Cambridge, MA.

Edvinsson, L., and Malone, M. S. (1997), *Intellectual Capital: Realizing your Company's True Value by Finding Its Hidden Brainpower,* HarperCollins, New York.

Fang, L., Hipel, K. W., and Kilgour, D. M. (1993), *Interactive Decision Making: The Graph Model for Conflict Resolution,* John Wiley & Sons, New York.

Firesmith, D. G. (1993), *Object Oriented Requirements and Logical Design,* John Wiley & Sons, New York.

Fraser, N. M., and Hipel, K. W. (1984), *Conflict Analysis: Models and Resolution.* North Holland, New York.

Gass, S. I., and Harris, C. M., Eds. (2000), *Encyclopedia of Operations Research and Management Science,* Kluwer Academic Publishers, Boston.

Gray, P., and Olfman, L. (1989), ''The User Interface in Group Decision Support Systems,'' *Decision Support Systems,* Vol. 5, No. 2, pp. 119–137.

Grudin, J. (1989), "The Case Against User Interface Consistency," *Communications of the ACM,* Vol. 32, pp. 1164–1173.

Hagel, J., and Armstrong, A. G. (1997), *Net Gain: Expanding Markets Through Virtual Communities,* Harvard Business School Press, Boston.

Hagel, J., and Singer, M. (1999), *Net Worth,* Harvard Business School Press, Boston.

Halal, W. E., Ed. (1998), *The Infinite Resource: Creating and Leading the Knowledge Enterprise* Jossey-Bass, San Francisco.

Hall, A. D. (1969), "Three Dimensional Morphology of Systems Engineering," *IEEE Transactions on Systems Science and Cybernetics,* Vol. 5, pp. 156–160.

Hammond, J. S., Keeney, R. L., and Raiffa, H. (1999), *Smart Choices: A Practical Guide to Making Better Decisions,* Harvard Business School Press, Boston.

Hammond, K. R., McClelland, G. H., and Mumpower, J. (1980), *Human Judgment and Decision Making: Theories, Methods, and Procedures,* Praeger, New York.

Harrison, H. R., and Hix, D. (1989), "Human Computer Interface Development." *ACM Computing Surveys,* Vol. 21, No. 1, pp. 5–92.

Harrison, M., and Thimbleby, H., Eds. (1990), *Formal Methods in Human–Computer Interaction,* Cambridge University Press, Cambridge.

Herek, G. M., Janis, I. L., and Hurth, P. (1987), "Decision Making During International Crises: Is Quality of Process Related to Outcome?" *Journal of Conflict Resolution,* Vol. 31, No. 2, pp. 203–226.

Hillier, F. S., and Lieberman, G. J. (1990), *Operations Research,* 5th Ed., Holden Day, San Francisco.

Hillier, F. S., and Lieberman, G. J. (1994), *Introduction to Mathematical Programming,* 2nd Ed., McGraw-Hill, New York.

Huber, G. P. (1982), "Group Decision Support Systems as Aids in the Use of Structured Group Management Techniques," in *Proceedings of the 2nd International Conference on Decision Support Systems* (San Francisco), pp. 96–108.

Hwang, S. (1985), "Automatic Model Building Systems: A Survey," in *DSS-85 Transactions,* J. J. Elam, Ed., pp. 22–32.

Jablin, F., and Putnam, L., Eds. (2000), *New Handbook of Organizational Communication,* Sage, Thousand Oaks, CA.

Jacobson, I., Ericsson, M., and Jacobson, A. (1995), *The Object Advantage: Business Process Reengineering with Object Technology.* Addison-Wesley, Reading MA.

Janis, I. J., and Mann, L. (1977), *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment,* Free Press, New York.

Janssen, T. J., and Sage, A. P. (2000), A Group Decision Support System for Science Policy Conflict Resolution," in Special Issue on Public Policy Engineering Management, E. G. Beroggi, Ed., *International Journal of Technology Management,* Vol. 19, No. 3.

Johansen, R. (1988), *Groupware: Computer Support for Business Teams,* Free Press, New York.

Katz, R., Ed. (1997), *The Human Side of Managing Technological Innovation,* Oxford University Press, New York.

Keen, P. G. W., and Scott Morton, M. S. (1978), *Decision Support Systems: An Organizational Perspective,* Addison-Wesley, Reading, MA.

Keeney, R. L., and Raiffa, H. (1976), *Decisions with Multiple Objectives,* John Wiley & Sons, New York.

Keller, L. R., and Ho, J. L. (1990), "Decision Problem Structuring, in *Concise Encyclopedia of Information Processing in Systems and Organizations,* A. P. Sage, Ed., Pergamon Press, Oxford, pp. 103–110.

Kelly, K. (1998), *New Rules for the New Economy,* Viking Press, New York.

Kent, W. (1979), "Limitation of Record Based Information Models," *ACM Transactions on Database Systems,* Vol. 4, pp. 107–131.

Kerschberg, L., Ed. (1987, 1989), *Proceedings from the First International Conference on Expert Database Systems* (Charleston, SC), Vol. 1 (1987), Vol. 2 (1989), Benjamin-Cummings, Menlo Park, CA.

King, J. L., and Star, S. L. (1992), "Organizational Decision Support Processes as an Open Systems Problem," in *Information Systems and Decision Processes,* T. Stohr and B. Konsynski, Eds., IEEE Press, Los Altos, CA, pp. 150–154.

Klein, D. A., Ed. (1998), *The Strategic Management of Intellectual Capital,* Butterworth-Heineman, Boston.

Klein, G. A. (1990), "Information Requirements for Recognitional Decision Making," in *Concise Encyclopedia of Information Processing in Systems and Organizations,* A. P. Sage, Ed., Pergamon Press, Oxford, pp. 414–418.

Klein, G. A. (1998), *Sources of Power: How People Make Decisions,* MIT Press, Cambridge.

Kraemer, K. L., and King, J. L. (1988), "Computer Based Systems for Cooperative Work and Group Decision Making," *ACM Computing Surveys,* Vol. 20, No. 2, pp. 115–146.

Kroenke, D. M. (1998), *Database Processing: Fundamentals, Design, and Implementation,* Prentice Hall, Englewood Cliffs, NJ.

Lee, E. (1990), "User-Interface Development Tools," *IEEE Software,* Vol. 7, No. 3, pp. 31–36.

Liang, B. T. (1988), "Model Management for Group Decision Support," *MIS Quarterly,* Vol. 12, pp. 667–680.

Liang, T. P. (1985), "Integrating Model Management with Data Management in Decision Support Systems," *Decision Support Systems.* Vol. 1, No. 3, pp. 221–232.

Liebowitz, J., and Beckman, T. (1998), *Knowledge Organizations: What Every Manager Should Know,* CRC Press, Boca Raton, FL.

Liebowitz, J., and Wilcox, L. C., Eds. (1997), *Knowledge Management and Its Integrative Elements,* CRC Press, Boca Raton, FL.

Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A., and Cohen, M. D. (1987), "Intelligent Information Sharing Systems," *Communications of the ACM,* Vol. 30, pp. 390–402.

Marakas, G. M., *Decision Support Systems in the 21st Century,* Prentice Hall, Englewood Cliffs, NJ.

March, J. G. (1983), "Bounded Rationality, Ambiguity, and the Engineering of Choice," *Bell Journal of Economics,* Vol. 9, pp. 587–608.

March, J., and Wessinger-Baylon, T., Eds. (1986), *Ambiguity and Command: Organizational Perspectives on Military Decisionmaking,* Pitman, Boston.

Matheson, D., and Matheson, J. (1998), *The Smart Organization: Creating Value Through Strategic R&D,* Harvard Business School Press, Boston.

McDermott, R. (1999), "Why Information Technology Inspired but Cannot Deliver Knowledge Management," *California Management Review,* Vol. 41, No. 1, pp. 103–117.

McGrath, J. E. (1984), *Groups: Interaction and Performance,* Prentice Hall, Englewood Cliffs, NJ.

Miller, W. L., and Morris, L. (1999), *Fourth Generation R&D: Managing Knowledge, Technology and Innovation,* John Wiley & Sons.

Mintzberg, H. (1973), *The Nature of Managerial Work,* Harper & Row, New York.

Murphy, F. H., and Stohr, E. A. (1986), "An Intelligent System for Formulating Linear Programs," *Decision Support Systems,* Vol. 2, No. 1, pp. 39–47.

Myers, P. S. (1997), *Knowledge Management and Organizational Design,* Butterworth-Heinemann, Boston.

Mylopoulos, J., and Brodie, M. L. Eds. (1998), *Artificial Intelligence and Databases,* Morgan Kaufmann, San Mateo, CA.

Nielson, J. (1989), *Hypertext and Hypermedia,* Academic Press, San Diego.

Nonaka, I., and Takeuchi, H. (1995), *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation,* Oxford University Press, New York.

Parsaei, H. R., Kolli, S., and Handley, T. R., Eds. (1997), *Manufacturing Decision Support Systems,* Chapman & Hall, New York.

Parsaye, K., Chignell, M., Khoshafian, S., and Wong, H. (1989), *Intelligent Databases: Object-Oriented, Deductive, Hypermedia Technologies,* John Wiley & Sons, New York.

Pfeffer, J., and Sutton, R. I. (2000), *The Knowing–Doing Gap: How Smart Companies Turn Knowledge into Action,* Harvard Business School Press, Boston.

Poe, V., Klauer, P., and Brobst, S. (1998), *Building a Data Warehouse for Decision Support,* 2nd Ed., Prentice Hall, Englewood Cliffs, NJ.

Pool, R. (1997), *Beyond Engineering: How Society Shapes Technology,* Oxford University Press, New York.

Prusak, L., Ed. (1997), *Knowledge in Organizations,* Butterworth-Heinemann, Woburn, MA.

Purba, S., Ed. (1999), *Data Management Handbook,* 3rd Ed., CRC Press, Boca Raton, FL.

Raiffa, H. (1968), *Decision Analysis,* Addison-Wesley, Reading, MA.

Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1995), *Cognitive Systems Engineering,* John Wiley & Sons, New York.

Ravden, S., and Johnson, G. (1989), *Evaluating Usability of Human–Computer Interfaces: A Practical Method,* John Wiley & Sons, Chichester.

Rob, P., and Coronel, C. (1997), *Database Systems,* 3rd Ed., International Thomson Publishing, London.

Roberts, T. L., and Moran, T. P. (1982), "A Methodology for Evaluating Text Editors," in *Proceedings of the IEEE Conference on Human Factors in Software Development,* B. Curtis, Ed., Gaithersburg, MD.

Roos, J., Roos, G., Edvinsson, L., and Dragonetti, N. C. (1998), *Intellectual Capital: Navigating in the New Business Landscape,* New York University Press, New York.

Rouse, W. B. (1991), "Conceptual Design of a Computational Environment for Analyzing Tradeoffs Between Training and Aiding," *Information and Decision Technologies,* Vol. 17, pp. 143–152.

Rouse, W. B. (1999), "Connectivity, Creativity, and Chaos: Challenges of Loosely-Structured Organizations," *Information, Knowledge, and Systems Management,* Vol. 1, No. 2, pp. 117–131.

Ruggles, R. L., III, Ed. (1997), *Knowledge Management Tools,* Butterworth-Heinemann, Woburn, MA.

Sage, A. P. (1987), "Information Systems Engineering for Distributed Decision Making," *IEEE Transactions on Systems, Man and Cybernetics,* Vol. 17, No. 6, pp. 920–936.

Sage, A. P. Ed. (1990), *Concise Encyclopedia of Information Processing in Systems and Organizations,* Pergamon Press, Oxford.

Sage, A. P. (1991), *Decision Support Systems Engineering,* John Wiley & Sons, New York.

Sage, A. P. (1992), *Systems Engineering,* John Wiley & Sons, New York.

Sage, A. P. (1995), *Systems Management for Information Technology and Software Engineering,* John Wiley & Sons, New York.

Sage, A. P. (1998), "Towards Systems Ecology," *IEEE Computer,* Vol. 31, No. 2, pp. 107–110.

Sage, A. P., and Armstrong, J. E. (2000), *Introduction to Systems Engineering,* John Wiley & Sons, New York.

Sage, A. P., and Lynch, C. L. (1998), "Systems Integration and Architecting: An Overview of Principles, Practices, and Perspectives," *Systems Engineering,* Vol. 1, No. 3, pp. 176–227.

Sage, A. P., and Palmer, J. D. (1990), *Software Systems Engineering,* Wiley Interscience, New York.

Sage, A. P., and Rouse, W. B., Eds. (1999a), *Handbook of Systems Engineering and Management,* John Wiley & Sons, New York.

Sage, A. P., and Rouse, W. B. (1999b), "Information System Frontiers in Knowledge Management," *Information System Frontiers,* Vol. 1, No. 3, pp. 195–204.

Schneiderman, B. (1987), *Designing the User Interface: Strategies for Effective Human Computer Interaction,* Addison-Wesley, Reading, MA.

Schum, D. A. (1987), *Evidence and Inference for the Intelligent Analyst,* 2 Vols., University Press of America, Lanham, MD.

Schum, D. A. (1994), *Evidential Foundations of Probabilistic Reasoning,* John Wiley & Sons, New York.

Shapiro, C., and Varian, H. R. (1999), *Information Rules: A Strategic Guide to the Network Economy,* Harvard Business School Press, Boston.

Shapiro, D. et al., Eds. (1996), *The Design of Computer Supported Cooperative Work and Groupware Systems,* North Holland, Amsterdam.

Sheridan, T. B. (1992), *Telerobotics, Automation, and Human Supervisory Control,* MIT Press, Cambridge, MA.

Simon, H. A. (1960), *The New Science of Management Decisions,* Harper, New York.

Skyrme, D. J. (1999), *Knowledge Networking: Creating the Collaborative Enterprise,* Butterworth-Heinemann, Boston.

Smith, D. E., Ed. (1999), *Knowledge, Groupware, and the Internet,* Butterworth-Heinemann, Boston.

Smith, S. L., and Mosier, J. N. (1986), "Guidelines for Designing User Interface Software," MITRE Corporation Technical Report MTR-10090, ESD-TR-86-278, Bedford, MA.

Somerville, M. A., and Rapport, D., Eds. (1999), *Transdisciplinarity: Re-creating Integrated Knowledge,* EOLSS, Oxford.

Sprague, R. H., Jr., and Carlson, E. D. (1982), *Building Effective Decision Support Systems,* Prentice Hall, Englewood Cliffs, NJ.

Sprague, R. H., Jr., and Watson, H. J. (1995), *Decision Support for Management,* Prentice Hall, Englewood Cliffs, NJ.

Stacey, R. D. (1996), *Complexity and Creativity in Organizations,* Berrett-Koehler, San Francisco.

Starbuck, W. E. (1985), ''Acting First and Thinking Later,'' in *Organizational Strategy and Change,* Jossey-Bass, San Francisco.

Stefik, M., Foster, G., Bobrow, D. G., Kahn, K., Lanning, S., and Suchman, L. (1987), ''Beyond the Chalkboard: Computer Support for Collaboration and Problem Solving in Meetings,'' *Communications of the ACM,* Vol. 30, No. 1, pp. 32–47.

Stewart, T. A. (1997). *Intellectual Capital: The New Wealth of Organizations,* Currency Doubleday, New York.

Sullo, G. C. (1994), *Object Oriented Engineering: Designing Large-Scale Object-Oriented Systems,* John Wiley & Sons, New York.

Sveiby, K. E. (1997), *The New Organizational Wealth: Managing and Measuring Knowledge Based Assets,* Berrett-Koehler, San Francisco.

Tobin, D. R. (1998), *The Knowledge Enabled Organization: Moving from Training to Learning to Meet Business Goals,* AMACOM, New York.

Toulmin, S., Rieke, R., and Janik, A. (1979), *An Introduction to Reasoning,* Macmillan, New York.

Ulrich, D. (1998), ''Intellectual Capital = Competence × Commitment,'' *Sloan Management Review,* Vol. 39, No. 2, pp. 15–26.

Utterback, J. M. (1994), *Mastering the Dynamics of Innovation: How Companies Can Seize Opportunities in the Face of Technological Change,* Harvard Business School Press, 1994.

Volkema, R. J. (1990), ''Problem Formulation, in *Concise Encyclopedia of Information Processing in Systems and Organizations,* A. P. Sage, Ed., Pergamon Press, Oxford.

Watson, R. T. (1998), *Database Management: An Organizational Perspective,* John Wiley & Sons, New York.

Watson, R. T., DeSanctis, G., and Poole, M. S. (1988), ''Using a GDSS to Facilitate Group Consensus: Some Intended and Unintended Consequences,'' *MIS Quarterly,* Vol. 12, pp. 463–477.

Weick, K. E. (1979), *The Social Psychology of Organizing,* Addison-Wesley, Reading, MA.

Weick, K. E. (1985), ''Cosmos vs. Chaos: Sense and Nonsense in Electronic Context,'' *Organizational Dynamics,* Vol. 14, pp. 50–64.

Welch, D. A. (1989), ''Group Decision Making Reconsidered,'' *Journal of Conflict Resolution,* Vol. 33, No. 3, pp. 430–445.

Will, H. J. (1975), ''Model Management System,'' in *Information Systems and Organizational Structure,* E. Grochia and N. Szyperski, Eds., de Gruyter, Berlin, pp. 468–482.

Winograd, T., and Flores, F. (1986), *Understanding Computers and Cognition,* Ablex Press, Los Altos, CA.

Zack, M. H., Ed. (1999), *Knowledge and Strategy,* Butterworth-Heinemann, Boston.

Zeigler, B. P. (1997), *Objects and Systems,* Springer, New York.

CHAPTER 5

Automation Technology

Chin-Yin Huang
Tunghai University, Taiwan

Shimon Y. Nof
Purdue University

1. INTRODUCTION 155
2. PHYSICAL AUTOMATION TECHNOLOGY 156
3. AUTOMATIC CONTROL SYSTEMS 156
    3.1. Fundamentals of Control 157
    3.2. Instrumentation of an Automatic Control System 158
    3.3. Basic Control Models 159
        3.3.1. Control Modeling 159
        3.3.2. Control Models 160
    3.4. Advanced Control Models 160
4. TECHNOLOGIES OF ARTIFICIAL INTELLIGENCE 160
    4.1. Knowledge-Based Systems 160
    4.2. Artificial Neural Networks 162
        4.2.1. Training Stage 163
        4.2.2 Control Stage 163
    4.3. Fuzzy Logic 163
    4.4. Genetic Algorithms 164
    4.5. Hybrid Intelligent Control Models 164
5. INTEGRATION TECHNOLOGY 164
    5.1. Networking Technologies 165
    5.2. Object Orientation and Petri Net Techniques 166
    5.3 Distributed Control vs. Central Control 166
    5.4. Robot Simulator/Emulator 167
6. EMERGING TRENDS 167
    6.1. Virtual Machines 168
    6.2. Tool Perspective Environment 168
        6.2.1. Facility Description Language (FDL) 171
        6.2.2 Concurrent Flexible Specifications (CFS) 172
    6.3. Agent-Based Control Systems 174
7. CONCLUSION 175
ACKNOWLEDGMENTS 175
REFERENCES 175

1. INTRODUCTION

Accompanying the evolution of human society, tools have been developed to assist humans to perform all kinds of activities in humans' daily life. Tools not only reduce the effort that men have to put into those activities, they also help men perform activities that would otherwise be impossible due to the limitations of the human body—for example, using a telescope to see objects on the moon. However, humans are not satisfied with developing tools to enhance productivity or conquer their limitations. People have always dreamed of building machines to do their work for them, allowing them more leisure time. This dream can be realized through the help of modern computer and communication technologies. Activities in the manufacturing enterprises that are automated by computer and communication technologies can be summarized into (but not limited by) the following seven categories (Huang and Nof 1999):

1. *Design:* Powerful computation speeds up activities in enterprises. For example, design activities are improved because of powerful CAD workstations.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2. *Decisions:* Powerful computation allows many simulation trials to find a better solution in decision making. For example, an optimal material handling equipment selection can be obtained through repeated simulation runs.
3. *Sensing:* Input devices (e.g., sensors, bar code readers) can gather and communicate environmental information to computers or humans. A decision may be made by computer systems based on the input information. The decision may also trigger output devices (e.g., robot arms, monitors) to realize the decisions.
4. *Recovery:* Computer systems may apply techniques of artificial intelligence (AI) (e.g., fuzzy rules, knowledge-based logic, neural networks) to improve the quality of activities. For example, a robot system may be recovered automatically from error conditions through decisions made by AI programs.
5. *Collaboration:* Distributed designers can work together on a common design project through a computer supported collaborative work (CSCW) software system.
6. *Partners:* A computer system in an organization may automatically find cooperative partners (e.g., vendors, suppliers, and subcontractors) from the Internet to fulfill a special customer order without any increase in the organization's capacity.
7. *Logistics:* Logistics flows of products and packages are monitored and maintained by networked computers.

Although these seven categories reflect the impact of computer and communication technologies, they are driven by four automation technologies: physical automation systems, automatic control systems, artificial intelligence systems, and integration technology. Physical automation systems and automatic control systems represent two early and ongoing achievements in automation technology. Through automatic control theories, most systems can be controlled by the set points defined by users. With the support of both automatic control theories and modern digital control equipment, such as the programmable logic controller (PLC), physical automation systems that consist of processing machines (e.g., CNCs), transportation equipment (e.g., robots and AGVs), sensing equipment (e.g., bar code readers) can be synchronized and integrated.

Artificial intelligence systems and integration technology are two relatively recent technologies. Many AI techniques, such as artificial neural networks, knowledge-based systems, and genetic algorithms, have been applied to automate the complex decision making processes in design, planning, and managerial activities of enterprises. Additionally, integration techniques, such as electronic data interchange (EDI), client-server systems, and Internet-based transactions, have automated business processes even when the participants are in remote sites.

In this chapter, we will discuss the above four technologies to give readers comprehensive knowledge of automation technology. Section 2 addresses physical automation technologies that are applied in the processing, transportation, and inspection activities. Section 3 introduces classical automatic control theory. The purpose of addressing automatic control theory is to review the traditional methods and explain how a system can be automatically adjusted to the set point given by users. Section 4 addresses artificial intelligence techniques and introduces basic application approaches. Section 5 introduces integration technology, which is based mostly on today's information technology. Section 6 introduces the emerging trends of automation technologies, which include virtual machines, tool perspective environment, and autonomous agents. Section 7 makes some concluding remarks.

2. PHYSICAL AUTOMATION TECHNOLOGY

Since MIT demonstrated the first numerically controlled machine tool in 1952, information technologies have revolutionized and automated the manufacturing processes. See Chapter 12 for physical automation techniques such as robots. In general, physical automation technology can be applied in three areas: processing, transportation/storage, and inspection. Representative examples of automated equipment are:

1. *Automated processing equipment:* CNC machine tools, computer-controlled plastic-injection machines, etc.
2. *Automated transportation/storage equipment:* Industrial robots, automatic guided vehicles (AGV), an automatic storage/retrieval systems (AS/RS), etc.
3. *Automated inspection equipment:* Coordination measuring machines (CMM), machine vision systems, etc.

3. AUTOMATIC CONTROL SYSTEMS

Control is the fundamental engineering and managerial function whose major purpose is to measure, evaluate, and adjust the operation of a process, a machine, or a system under dynamic conditions so that it achieves desired objectives within its planned specifications and subject to cost and safety considerations. A well-planned system can perform effectively without any control only as long as no variations are encountered in its own operation and its environment. In reality, however, many changes occur over time. Machine breakdown, human error, variable material properties, and faulty information are a few examples of why a system must be controlled.

When a system is more complex and there are more potential sources of dynamic variations, a more complicated control is required. Particularly in automatic systems where human operators are replaced by machines and computers, a thorough design of control responsibilities and procedures is necessary. Control activities include automatic control of individual machines, material handling, equipment, manufacturing processes, and production systems, as well as control of operations, inventory, quality, labor performance, and cost. Careful design of correct and adequate controls that continually identify and trace variations and disturbances, evaluate alternative responses, and result in timely and appropriate actions is therefore vital to the successful operation of a system.

3.1. Fundamentals of Control

Automatic control, as the term is commonly used, is "self-correcting," or feedback, control; that is, some control instrument is continuously monitoring certain output variables of a controlled process and is comparing this output with some preestablished desired value. The instrument then compares the actual and desired values of the output variable. Any resulting error obtained from this comparison is used to compute the required correction to the control setting of the equipment being controlled. As a result, the value of the output variable will be adjusted to its desired level and maintained there. This type of control is known as a servomechanism.

The design and use of a servomechanism control system requires a knowledge of every element of the control loop. For example, in Figure 1 the engineer must know the dynamic response, or complete operating characteristics, of each pictured device:

1. The indicator or sampler, which senses and measures the actual output
2. The controller, including both the error detector and the correction computer, which contain the decision making logic
3. The control value and the transmission characteristics of the connecting lines, which communicate and activate the necessary adjustment
4. The operating characteristics of the plant, which is the process or system being controlled

Dynamic response, or *operating characteristics,* refer to a mathematical expression, for example, differential equations, for the transient behavior of the process or its actions during periods of change in operating conditions. From it one can develop the transfer function of the process or prepare an experimental or empirical representation of the same effects.

Because of time lags due to the long communication line (typically pneumatic or hydraulic) from sensor to controller and other delays in the process, some time will elapse before knowledge of changes in an output process variable reaches the controller. When the controller notes a change, it must compare it with the variable value it desires, compute how much and in what direction the control valve must be repositioned, and then activate this correction in the valve opening. Some time is required, of course, to make these decisions and correct the valve position.

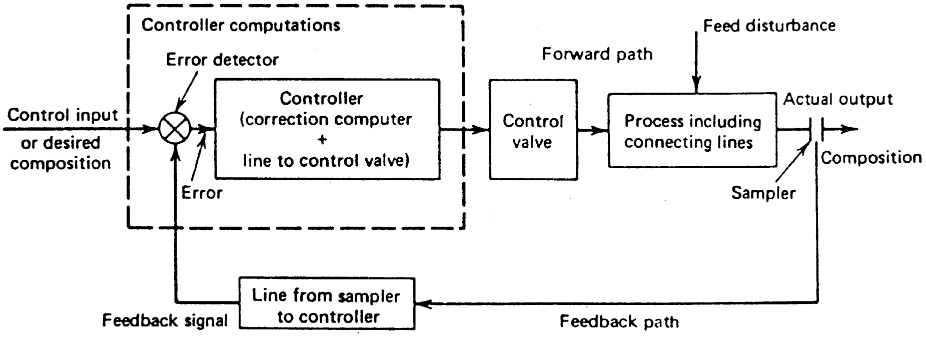


Figure 1 Block Diagram of a Typical Simple, Single Control Loop of a Process Control System. (From Williams and Nof 1992)

Some time will also elapse before the effect of the valve correction on the output variable value can reach the output itself and thus be sensed. Only then will the controller be able to know whether its first correction was too small or too large. At that time it makes a further correction, which will, after a time, cause another output change. The results of this second correction will be observed, a third correction will be made, and so on.

This series of measuring, comparing, computing, and correcting actions will go around and around through the controller and through the process in a closed chain of actions until the actual process valve is finally balanced again at the desired level by the operator. Because from time to time there are disturbances and modifications in the desired level of the output, the series of control actions never ceases. This type of control is aptly termed *feedback control.* Figure 1 shows the direction and path of this closed series of control actions. The closed-loop concept is fundamental to a full understanding of automatic control.

Although the preceding example illustrates the basic principles involved, the actual attainment of automatic control of almost any industrial process or other complicated device will usually be much more difficult because of the speed of response, multivariable interaction, nonlinearity, response limitations, or other difficulties that may be present, as well as the much higher accuracy or degree of control that is usually desired beyond that required for the simple process just mentioned.

As defined here, automatic process control always implies the use of a feedback. This means that the control instrument is continuously monitoring certain output variables of the controlled process, such as a temperature, a pressure, or a composition, and is also comparing this output with some preestablished desired value, which is considered a reference, or a set point, of the controlled variable. An error that is indicated by the comparison is used by the instrument to compute a correction to the setting of the process control valve or other final control element in order to adjust the value of the output variable to its desired level and maintain it there.

If the set point is altered, the response of the control system to bring the process to the new operating level is termed that of a servomechanism or self-correcting device. The action of holding the process at a previously established level of operation in the face of external disturbances operating on the process is termed that of a regulator.

3.2. Instrumentation of an Automatic Control System

The large number of variables of a typical industrial plant constitute a wide variety of flows, levels, temperatures, compositions, positions, and other parameters to be measured by the sensor elements of the control system. Such devices sense some physical, electrical, or other informational property of the variable under consideration and use it to develop an electrical, mechanical, or pneumatic signal representative of the magnitude of the variable in question. The signal is then acted upon by a transducer to convert it to one of the standard signal levels used in industrial plants (3–15 psi for pneumatic systems and 1–4, 4–20, or 10–50 mA or 0–5 V for electrical systems). Signals may also be digitized at this point if the control system is to be digital.

The signals that are developed by many types of sensors are continuous representations of the sensed variables and as such are called analog signals. When analog signals have been operated upon by an analog-to-digital converter, they become a series of bits, or on–off signals, and are then called digital signals. Several bits must always be considered together in order to represent properly the converted analog signal (typically, 10–12 bits).

As stated previously, the resulting sensed variable signal is compared at the controller to a desired level, or set point, for that variable. The set point is established by the plant operator or by an upper-level control system. Any error (difference) between these values is used by the controller to compute the correction to the controller output, which is transmitted to the valve or other actuator of the system's parameters.

A typical algorithm by which the controller computes its correction is as follows (Morriss 1995). Suppose a system includes components that convert inputs to output according to relationships, called gains, of three types: proportional, derivative, and integral gains. Then the controller output is

$$\text{Output} = K_P e(t) + K_I \int e(t)\,dt + K_D \frac{d(e(t))}{dt}$$

where K_P, K_D, and K_I = proportional, derivative, and integral gains, respectively, of the controller.

The error at time t, $e(t)$, is calculated as

$$e(t) = \text{set point} - \text{feedback signal}$$

3.3. Basic Control Models

3.3.1. Control Modeling

Five types of modeling methodologies have been employed to represent physical components and relationships in the study of control systems:

1. Mathematical equations, in particular, differential equations, which are the basis of classical control theory (transfer functions are a common form of these equations)
2. Mathematical equations that are used on state variables of multivariable systems and associated with modern control theory
3. Block diagrams
4. Signal flow graphs
5. Functional analysis representations (data flow diagram and entity relationships)

Mathematical models are employed when detailed relationships are necessary. To simplify the analysis of mathematical equations, we usually approximate them by linear, ordinary differential equations. For instance, a characteristic differential equation of a control loop model may have the form

$$\frac{d^2x}{dt^2} + 2\alpha\frac{dx}{dt} + \beta^2 x = f(t)$$

with initial conditions of the system given as

$$x(0) = X_0$$
$$x'(0) = V_0$$

where $x(t)$ is a time function of the controlled output variable, its first and second derivatives over time specify the temporal nature of the system, α and β are parameters of the system properties, $f(t)$ specifies the input function, and X_0 and V_0 are specified constants.

Mathematical equations such as this example are developed to describe the performance of a given system. Usually an equation or a transfer function is determined for each system component. Then a model is formulated by appropriately combining the individual components. This process is often simplified by applying Laplace and Fourier transforms. A graph representation by block diagrams (see Figure 2) is usually applied to define the connections between components.

Once a mathematical model is formulated, the control system characteristics can be analytically or empirically determined. The basic characteristics that are the object of the control system design are:

1. Response time
2. Relative stability
3. Control accuracy

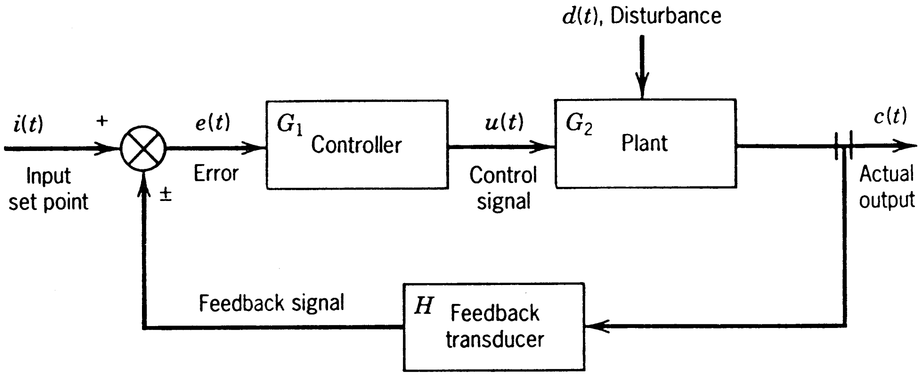


Figure 2 Block Diagram of a Feedback Loop. (From Williams and Nof 1992)

They can be expressed either as functions of frequency, called frequency domain specifications, or as functions of time, called time domain specifications. To develop the specifications the mathematical equations have to be solved. Modern computer software, such as MATLAB (e.g., Kuo 1995), has provided convenient tools for solving the equations.

3.3.2. Control Models

Unlike the open-loop control, which basically provides a transfer function for the input signals to actuators, the feedback control systems receive feedback signals from sensors then compare the signals with the set point. The controller can then control the plant to the desired set point according to the feedback signal. There are five basic feedback control models (Morriss 1995):

1. *On/off control:* In on/off control, if the $e(t)$ is smaller than 0, the controller may activate the plant; otherwise the controller stays still. Most household temperature thermostats follow this model.
2. *Proportional (PE) control:* In PE control, the output is proportional to the $e(t)$ value, i.e., $e(t) = K_P e(t)$. In PE, the plant responds as soon as the error signal is non-zero. The output will not stop exactly at the set point. When it approaches the set point, the $e(t)$ becomes smaller. Eventually, the output is too small to overcome opposing force (e.g., friction). Attempts to reduce this small $e(t)$, also called steady state error, by increasing K_P can only cause more overshoot error.
3. *Proportional-integral (PI) control:* PI control tries to solve the problem of steady state error. In PI, output $= K_P e(t) + K_I \int e(t)\, dt$. The integral of the error signal will have grown to a certain value and will continue to grow as soon as the steady state error exists. The plant can thus be drawn to close the steady state error.
4. *Proportional-derivative (PD) control:* PD control modifies the rate of response of the feedback control system in order to prevent overshoot. In PD,

$$\text{Output} = K_P e(t) + K_D \frac{d(e(t))}{dt}$$

   When the $e(t)$ gets smaller, a negative derivative results. Therefore, overshoot is prevented.
5. *Proportional-integral-derivative (PID) control:* PID control takes advantage of PE, PI, and PD controls by finding the gains (K_P, K_I, and K_D) to balance the proportional response, steady state reset ability, and rate of response control, so the plant can be well controlled.

3.4. Advanced Control Models

Based on the control models introduced in Section 3.3, researchers have developed various advanced control models for special needs. Table 1 shows the application domains and examples of the models, rather than the complicated theoretical control diagram. Interested readers can refer to Morriss (1995).

4. TECHNOLOGIES OF ARTIFICIAL INTELLIGENCE

Automation technologies have been bestowed intelligence by the invention of computers and the evolution of artificial intelligence theories. Because of the introduction of the technologies of artificial intelligence, automated systems, from the perspective of control, can intelligently plan, actuate, and control their operations in a reasonable time limit by handling/sensing much more environmental input information (see the horizontal axis in Figure 3). Meanwhile, artificial intelligence increases the decision making complexity of automation technology, but the cost of the system that is automated by the technologies of artificial intelligence is relatively low compared with the system automated by the automatic control theory (see the vertical axis in Figure 3). Figure 3 shows a roadmap of how various automation technologies influence the development of automation systems in two axes. In the following subsections, those technologies of artificial intelligence, including neural networks (NN), genetic algorithms (GA), knowledge-based systems (KBS), fuzzy control, and the hybrid of the above-mentioned technologies will be introduced in general.

4.1. Knowledge-Based Systems

The structure of a knowledge-based system (KBS) in control is presented in Figure 4. The control decisions are achieved by reasoning techniques (inference engine) rather than quantitative computation and can deal with uncertainties and unstructured situations. The knowledge base is updated by continuously acquiring the knowledge from experts, the decisions made by the operator and the feedback from the process. Nowadays, KBSs have been applied in many fields to diagnose causes of problems, such as in medical diagnosis, vehicle troubleshooting, and loan strategy planning in banks.

TABLE 1 System Characteristics and Examples of Advanced Control Models

Control Models	When to Apply (System Characteristics)	Examples
(1) Nested Control Loops	More than one characteristic of a system's output to be controlled at the same time	Robot controllers
(2) Directed Synthesis Control	Time lag between output and feedback reception is long	Satellite guidance systems
(3) Adaptive Directed Synthesis Control	Allow a direct synthesis control system to be adjusted when actual feedback is received	Modern machine tools
(4) Feedforward Control; Cascade Control	Set points are given by the sensed conditions upstream of the process.	Chemical process control
(5) Ratio Control	Use a sensor in one stream to control the processes in one or more parallel streams	Chemical process control
(6) Multiple Output Control	A single control provides the output signal to a group of parallel actuators. Then the result of the multiple actuation is fed back to the controller.	Complex hydraulic cylinders
(7) Constraint Control	More than one controller in the system. Under normal conditions, one of the controllers does the controlling, but if a preset limit is exceeded at a sensor, the second controller overrides the first and takes over control.	Chemical process control

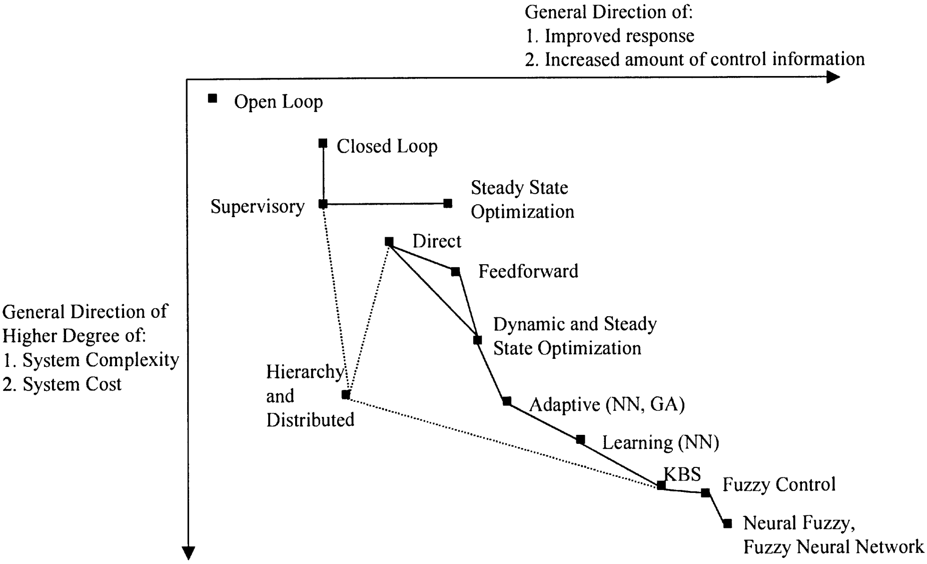

General Direction of:
1. Improved response
2. Increased amount of control information
Open Loop
Closed Loop
Supervisory
Steady State Optimization
Direct
Feedforward
Dynamic and Steady State Optimization
Hierarchy and Distributed
Adaptive (NN, GA)
Learning (NN)
KBS
Fuzzy Control
Neural Fuzzy, Fuzzy Neural Network
General Direction of Higher Degree of:
1. System Complexity
2. System Cost

Figure 3 Roadmap of Control Models. (Revised from Williams and Nof 1992)

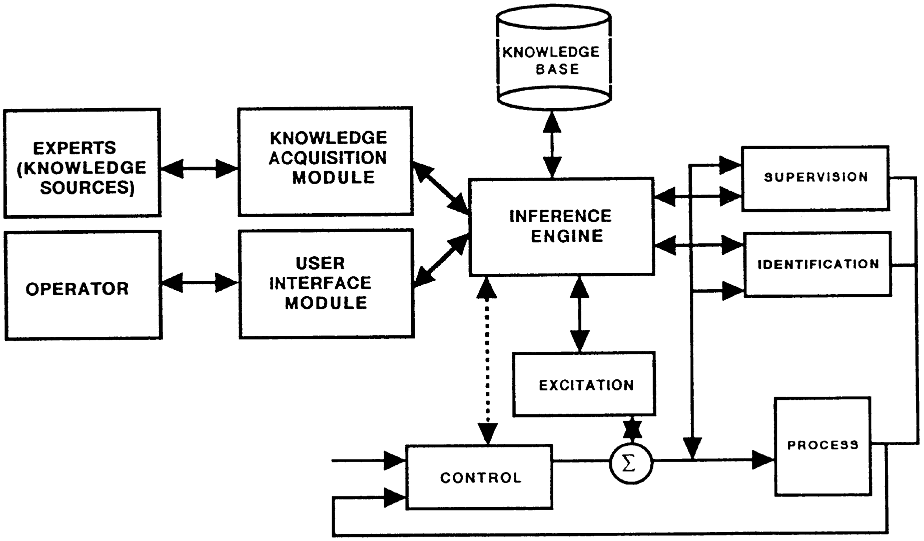


Figure 4 Block Diagram of a Knowledge-Based Control System. (From Williams and Nof 1992)

The numerous industrial engineering applications of control models in computer information systems can be classified into two types: (1) development of decision support systems, information systems that provide the information, and decisions to control operations, and (2) maintenance of internal control over the quality and security of the information itself. Because information systems are usually complex, graphic models are typically used.

Any of the control models can essentially incorporate an information system, as indicated in some of the examples given. The purpose of an information system is to provide useful, high-quality information; therefore, it can be used for sound planning of operations and preparation of realistic standards of performance. Gathering, classifying, sorting, and analyzing large amounts of data can provide timely and accurate measurement of actual performance. This can be compared to reference information and standards that are also stored in the information system in order to immediately establish discrepancies and initiate corrective actions. Thus, an information system can improve the control operation in all its major functions by measuring and collecting actual performance measures, analyzing and comparing the actual to the desired set points, and directing or actuating corrective adjustments.

An increasing number of knowledge-based decision support and control systems have been applied since the mid-1980s. Typical control functions that have been implemented are:

- Scheduling
- Diagnosis
- Alarm interpretation
- Process control
- Planning
- Monitoring

4.2. Artificial Neural Networks

The powerful reasoning and inference capabilities of artificial neural networks (ANN) in control are demonstrated in the areas of

- Adaptive control and learning
- Pattern recognition/classification
- Prediction

To apply ANN in control, the user should first answer the following questions:

1. If there are training data, ANN paradigm with supervised learning may be applied; otherwise, ANN paradigm with unsupervised learning is applied.
2. Select a suitable paradigm, number of network layers, and number of neurons in each layer.
3. Determine the initial weights and parameters for the ANN paradigm.

A widely applied inference paradigm, back propagation (BP), is useful with ANN in control. There are two stages in applying BP to control: the training stage and the control stage (Figure 5).

4.2.1. Training Stage

1. Prepare a set of training data. The training data consist of many pairs of data in the format of input–output.
2. Determine the number of layers, number of neurons in each layer, the initial weights between the neurons, and parameters.
3. Input the training data to the untrained ANN.
4. After the training, the trained ANN provides the associative memory for linking inputs and outputs.

4.2.2. Control Stage

Input new data to the trained ANN to obtain the control decision. For instance, given a currently observed set of physical parameter values, such as noise level and vibration measured on a machine tool, automatically adapt to a new calculated motor speed. The recommended adjustment is based on the previous training of the ANN-based control. When the controller continues to update its training ANN over time, we have what is called learning control.

Other application examples of neural networks in automated systems are as follows:

- Object recognition based on robot vision
- Manufacturing scheduling
- Chemical process control

The ANN could be trained while it is transforming the inputs on line to adapt itself to the environment. Detailed knowledge regarding the architectures of ANN, initial weights, and parameters can be found in Dayhoff (1990), Freeman and Skapura (1991), Fuller (1995), Lin and Lee (1996).

4.3. Fuzzy Logic

Figure 6 shows a structure of applying fuzzy logic in control. First, two types of inputs must be obtained: numerical inputs and human knowledge or rule extraction from data (i.e., fuzzy rules). Then the numerical inputs must be fuzzified into fuzzy numbers. The fuzzy rules consist of the fuzzy membership functions (knowledge model) or so-called fuzzy associative memories (FAMs). Then the

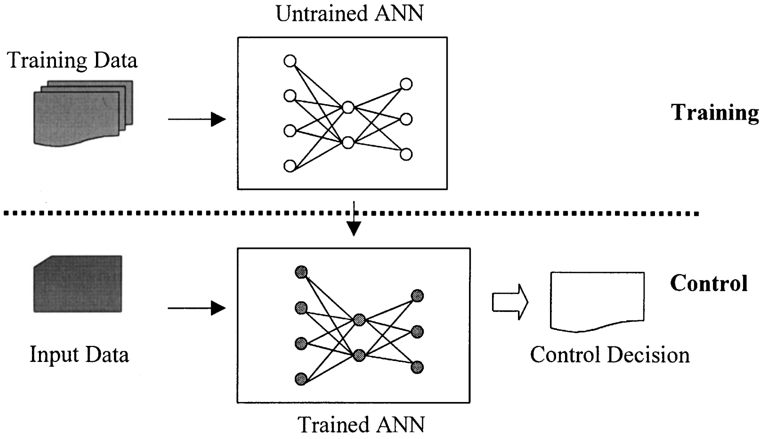


Figure 5 Structure of Applying Back-Propagation ANN in Control.

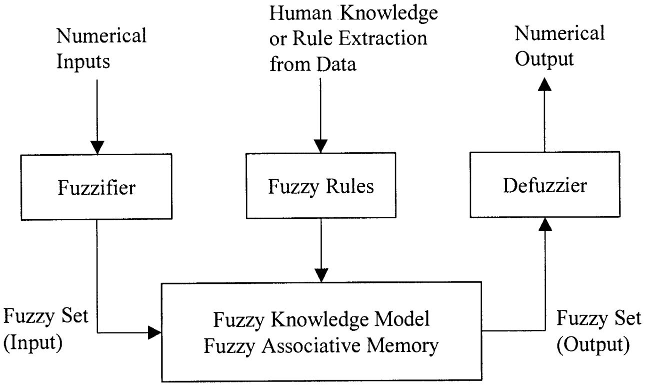


Figure 6 The Structure of Applying Fuzzy Logic in Control.

FAMs map fuzzy sets (inputs) to fuzzy sets (outputs). The output fuzzy sets should be defuzzified into numerical values to control the plant. Due to the powerful ability of fuzzy sets in describing system linguistic and qualitative behavior and imprecise and/or uncertain information, many industrial process behavior and control laws can be modeled by fuzzy logic-based approaches. Fuzzy logic has been applied in a wide range of automated systems, including:

- Chemical process control
- Autofocusing mechanism on camera and camcorder lens
- Temperature and humidity control for buildings, processes, and machines

4.4. Genetic Algorithms

Genetic algorithms (GAs), also referred to as evolutionary computation, are highly suitable for certain types of problems in the areas of optimization, product design, and monitoring of industrial systems. A GA is an automatically improving (evolution) algorithm. First the user must encode solutions of a problem into the form of chromosomes and an evaluation function that would return a measurement of the cost value of any chromosome in the context of the problem. A GA consists of the following steps:

1. Establish a base population of chromosomes.
2. Determine the fitness value of each chromosome.
3. Create new chromosomes by mating current chromosomes; apply mutation and recombination as the parent chromosomes mate.
4. Delete undesirable members of the population.
5. Insert the new chromosomes into the population to form a new population pool.

GA are useful for solving large-scale planning and control problems. Several cases indicate that GA can effectively find an acceptable solution for complex product design, production scheduling, and plant layout planning.

4.5. Hybrid Intelligent Control Models

Intelligent control may be designed in a format combining the techniques introduced above. For example, fuzzy neural networks use computed learning and the adaptive capability of neural networks to improve the computed learning's associative memory. Genetic algorithms can also be applied to find the optimal structure and parameters for neural networks and the membership functions for fuzzy logic systems. In addition, some techniques may be applicable in more than one area. For example, the techniques of knowledge acquisition in KBSs and fuzzy logic systems are similar.

5. INTEGRATION TECHNOLOGY

Recent communication technologies have enabled another revolution in automation technologies. Stand-alone automated systems are integrated via communication technologies. Integration can be identified into three categories:

1. *Networking:* Ability to communicate.
2. *Coordinating:* Ability to synchronize the processes of distributed automated systems. The coordination is realized by either a controller or an arbitrary rule (protocol).
3. *Integration:* Distributed automated systems are able to cooperate or collaborate with other automated systems to fulfill a global goal while satisfying their individual goals. The cooperation/collaboration is normally realized via a protocol that is agreed upon by distributed automated systems. However, the agreement is formed through the intelligence of the automated systems in protocol selection and adjustment.

In this section, automated technologies are introduced based on the above categories. However, truly integrated systems (technologies) are still under development and are mostly designed as agent-based systems, described in Section 6. Only technologies from the first two categories are introduced in this section.

5.1. Networking Technologies

To connect automated systems, the simplest and the most economic method is to wire the switches on the automated systems to the I/O modules of the programmable logic controller (PLC). Ladder diagrams that represent the control logic on the automated systems are popularly applied in PLCs. However, as the automated systems are remotely distributed, automated systems become more intelligent and diversified in the communication standards that they are built in, and the coordination decisions become more complex, the simple messages handled by the PLC-based network are obviously not enough for the control of the automated systems. Hence, some fundamental technologies, such as field bus (Mahalik and Moore 1997) and local area networks (LANs), and some advanced communication standards, such as LonWorks, Profibus, Manufacturing Automation Protocol (MAP), Communications Network for Manufacturing Applications (CNMA), and SEMI* Equipment Communication Standard (SECS) are developed.

In a field bus, automated devices are interconnected. Usually the amount of data transmitted in the bus is not large. In order to deliver message among equipment's timely in a field bus, the seven layers of the open system interconnection (OSI) are simplified into three layers: *physical layer, data link layer,* and *application layer.* Unlike a field bus, which usually handles the connections among devices, office activities are automated and connected by a LAN. Usually more than one file server is connected in a LAN for data storage, retrieval, and sharing among the connected personal computers. Three technological issues have to be designed/defined in a LAN topology, media, and access methods (Cohen and Apte 1997).

For the advanced communication standard, MAP and CNMA are two technologies that are well known and widely applied. Both technologies are based on the Manufacturing Message Specification (MMS) (SISCO 1995), which was developed by the ISO Industrial Automation Technical Committee Number 184. MMS provides a definition to specify automated equipment's external behaviors (Shanmugham et al. 1995). Through the specification, users can control the automated equipment with little knowledge about the internal message conversion within MMS and the equipment.

SECS is currently a popular communication standard in semiconductor industries (SEMI 1997). The standard was developed by SEMI based on two standards, SECS-I (SEMI Equipment Communications Standard Part 1 Message Transfer) and SECS-II (SEMI Equipment Communication Standard 2 Message Content). The relationship between SECS-I and SECS-II, as shown in Figure 7, shows that SECS-I transmits the message that is defined by SECS-II to RS-232. SECS-II's message

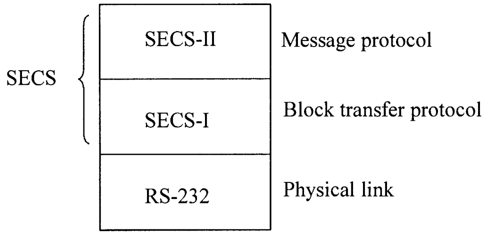


Figure 7 Basic Structure of SECS.

---

* SEMI stands for Semiconductor Equipment and Materials International.

is added as control information by the SECS-I so the transmission message can conform the format of RS-232 for message delivery. For SECS-II, it provides a set of interequipment communication standards under various situations. Hence, engineers only need to know and follow the SECS-II to control the connected automated equipment, rather than taking time to define the detailed message conversion.

5.2. Object Orientation and Petri Net Techniques

Object orientation and Petri net are automation techniques in modeling and analysis levels. Automated systems and their associated information and resources can be modeled by object models. The coordination and communication among the automated systems can then be unified with the message passing among the objects. However, the complex message passing that is used to coordinate behaviors of the automated systems relies on the technique of Petri net. The graphical and mathematically analyzable characteristics make Petri net a very suitable tool for synchronizing the behaviors and preventing deadlocks among automated systems. Combinations of both techniques have been developed and applied in the controllers of flexible manufacturing systems (Wang 1996; Wang and Wu 1998).

5.3. Distributed Control vs. Central Control

The rapid development of microprocessor technology has made distributed control possible and attractive. The use of reliable communication between a large number of individual controllers, each responsible for its own tasks rather than for the complete operation, improves the response of the total system. We can take PLC-based control as a typical example of central control system and LonWorks, developed by Echelon, as an example of distributed control system. In LonWorks, each automated device is controlled by a control module—LTM-10 (Figure 8). The control modules are connected on a LonTalk network that provides an ISO/OSI compatible protocol for communication.

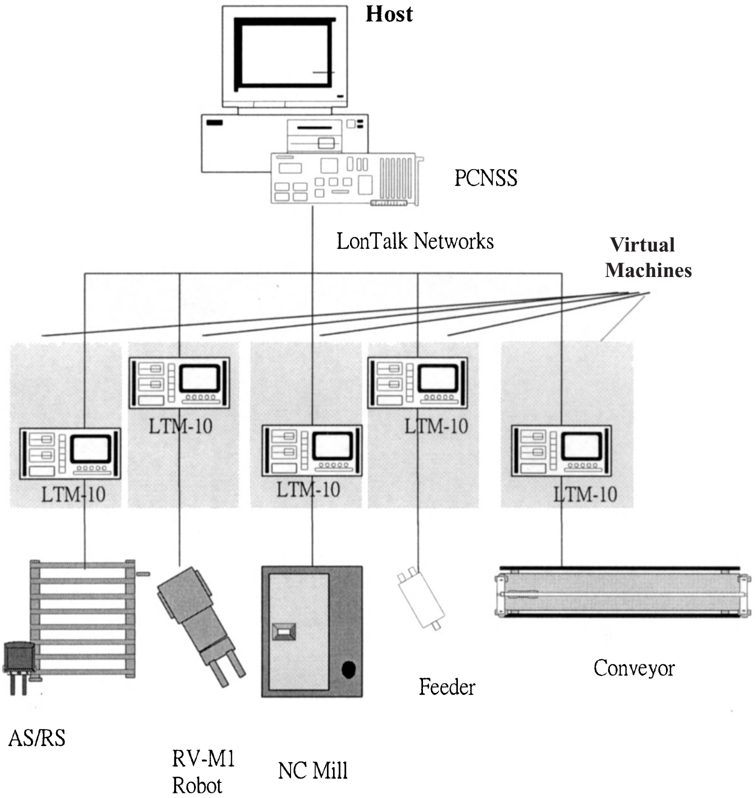


Figure 8 Applying LonWorks to Develop Virtual Machines.

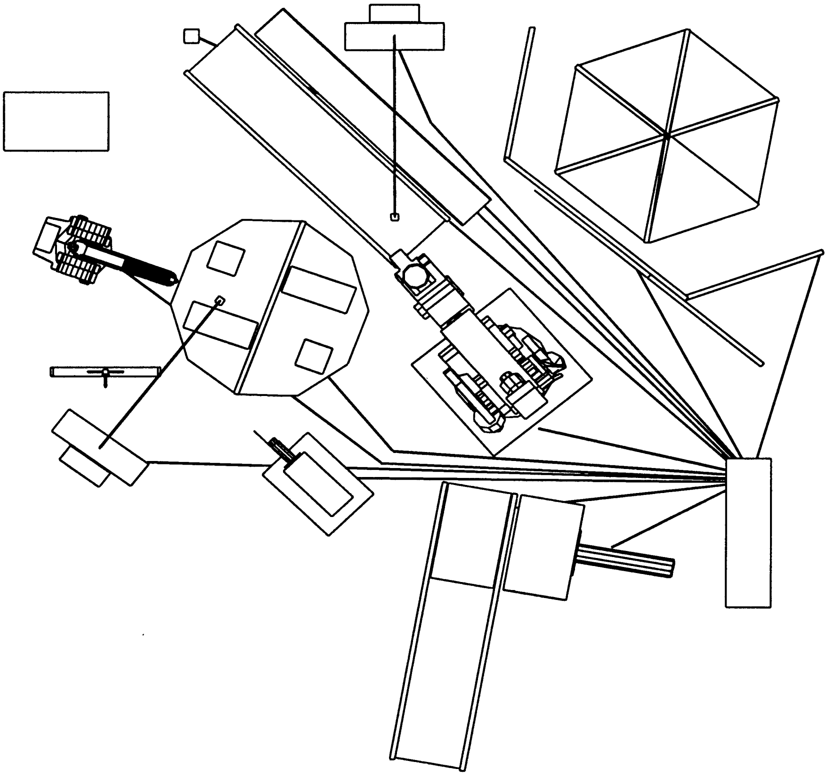


Figure 9 An Example of a Robot Emulator. (From Witzerman and Nof 1995)

Usually the control modules are loaded with Neuron C programs from a host computer that has a PCNSS network card for network management and Neuron C programming. Hence, under running mode the host computer is not necessary and the control modules can work under a purely distributed environment.

Distributed automated systems have the following advantages over centralized automated systems:

- Higher system reliability
- Better response to local demands
- Lower cost in revising the system control programs when automated equipment is added or deleted from the system

5.4. Robot Simulator/Emulator

In recent years, powerful robot simulators/emulators have been developed by several companies (Nof 1999). Examples include ROBCAD by Tecnomatix and RAPID by Adept Technologies. With these highly interactive, graphic software, one can program, model, and analyze both the robots and their integration into a production facility. Furthermore, with the geometric robot models, the emulation can also check for physical reachability and identify potential collisions. Another important feature of the robot simulators/emulators is off-line programming of robotic equipment, which allows designers to compare alternative virtual design before the program is transferred to actual robots (see Figure 9).

6. EMERGING TRENDS

Automation technology has reached a new era. An automation system is automated not only to reach the setting point but to situate itself intelligently in its complex environment. The individual automated systems are also networked to accomplish collaborative tasks. However, networking automated systems is not an easy task. It involves the technologies from the physical levels, which handle the interface and message conversion between automation systems, to the application level, which handles mostly the social interactions between the automation systems. Three typical ongoing research directions are described next to present the trends of automation technology.

6.1. Virtual Machines

Traditionally, a hierarchy structure of the computer control system for a fully automated manufacturing system can be divided into four levels (Figure 10). The supervisory control is responsible for managing the direct digit control by sending control commands, downloading process data, monitoring the process, and handling exception events. From the perspective of message transmission, the hierarchy in Figure 10 can be classified into three standard levels, as shown in Figure 11:

1. *Production message standard:* a standard for obtaining/sending production information from /to low-level computers. The production information can be a routing, a real-time statistics, etc. Usually the information is not directly related to the equipment control.
2. *Control message standard:* a standard for controlling the equipment logically. The control message includes commands, the parameters associated with the command, and data. SECS-II is a standard that can fulfill such a need (Elnakhal and Rzehak 1993).

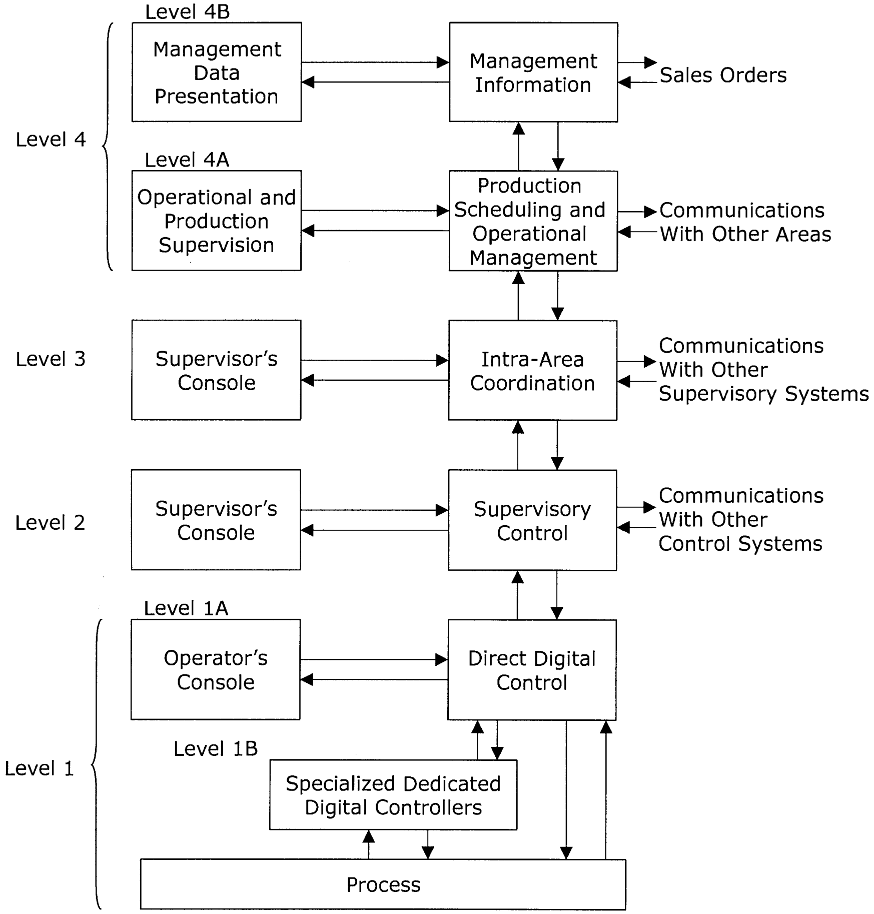

Level 4B
Management Data Presentation
Management Information
Sales Orders
Level 4
Level 4A
Operational and Production Supervision
Production Scheduling and Operational Management
Communications With Other Areas
Level 3
Supervisor's Console
Intra-Area Coordination
Communications With Other Supervisory Systems
Level 2
Supervisor's Console
Supervisory Control
Communications With Other Control Systems
Level 1A
Operator's Console
Direct Digital Control
Level 1
Level 1B
Specialized Dedicated Digital Controllers
Process

Figure 10 The Hierarchy Structure of the Computer Control System for a Fully Automated Industrial Plant. (From Williams and Nof 1992)

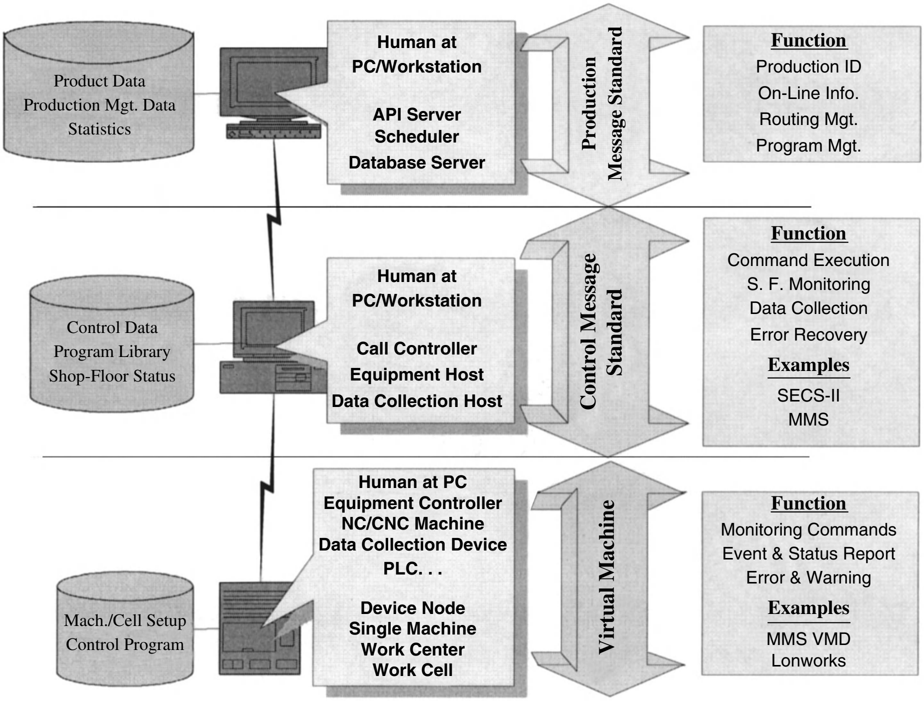

Product Data
Production Mgt. Data
Statistics
Human at PC/Workstation
API Server
Scheduler
Database Server
Production Message Standard
Function
Production ID
On-Line Info.
Routing Mgt.
Program Mgt.
Control Data
Program Library
Shop-Floor Status
Human at PC/Workstation
Call Controller
Equipment Host
Data Collection Host
Control Message Standard
Function
Command Execution
S. F. Monitoring
Data Collection
Error Recovery
Examples
SECS-II
MMS
Mach./Cell Setup
Control Program
Human at PC
Equipment Controller
NC/CNC Machine
Data Collection Device
PLC. . .
Device Node
Single Machine
Work Center
Work Cell
Virtual Machine
Function
Monitoring Commands
Event & Status Report
Error & Warning
Examples
MMS VMD
Lonworks

Figure 11 Three Layers of Message Transmission.

3. *Virtual machine:* a mechanism that converts the logical commands based on the control message standard, e.g., SECS-II, to commands format that can be accepted by a physical machine. Figure 12 shows that a virtual machine is laid between a host that delivers SECS-II commands and a real device that receives and executes the commands in its format. A virtual machine is therefore responsible for bridging the message format gap.

The development of virtual machines reduces the problems of incompatibility in an automated system, especially when the automated equipment follows different communication standards and protocols. It also enables the high-level controller to focus on his or her control activities, such as sequencing and scheduling, rather than detailed command incompatibility problems. Further information on virtual machines and automation can be found in Burdea and Coiffet (1999).

6.2. Tool Perspective Environment

Modern computer and communication technologies not only improve the quality, accuracy, and timeliness of design and decisions but also provide tools for automating the design and decision making processes. Under such a tool perspective, human experts focus on developing tools, then users can apply the tools to model and resolve their automation problems (see the right-side diagram in Figure 13). In contrast with traditional approaches, difficulties occur only during the tool development. For traditional approaches, “experts” are the bottlenecks in design projects. They must understand not only the problems but also the applications and limitations in practice. In practice, the costs of looking for and working with the experts are the main expense. For tool perspective, the cost of training the users becomes the major expense. However, frequent environment changes usually result in the design of a flexible system to respond to repeated needs to modify and evaluate the existing system. Modification and evaluation needs can be fulfilled by rapidly applying computerized tools, which provide relatively high flexibility in design. Some researchers have noted the importance of modeling manufacturing systems with the tool perspective. A sample survey is presented in Table 2. The three modeling concerns ((1) conflicts among designers, (2) constraints in physical environment, (3) the information flows in manufacturing systems) are part of the criteria in the survey table. It is found

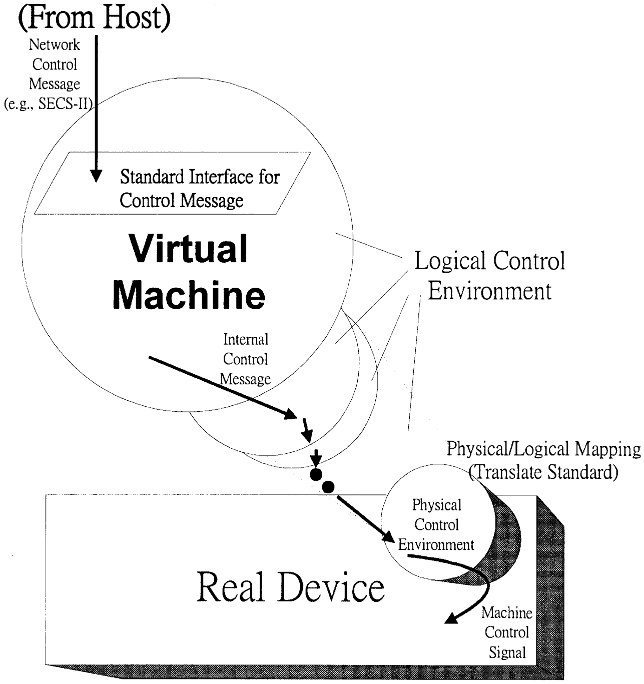

(From Host)
Network Control Message (e.g., SECS-II)
Standard Interface for Control Message
Virtual Machine
Logical Control Environment
Internal Control Message
Physical/Logical Mapping (Translate Standard)
Physical Control Environment
Real Device
Machine Control Signal

Figure 12 Role of Virtual Machine in Automation Applications.

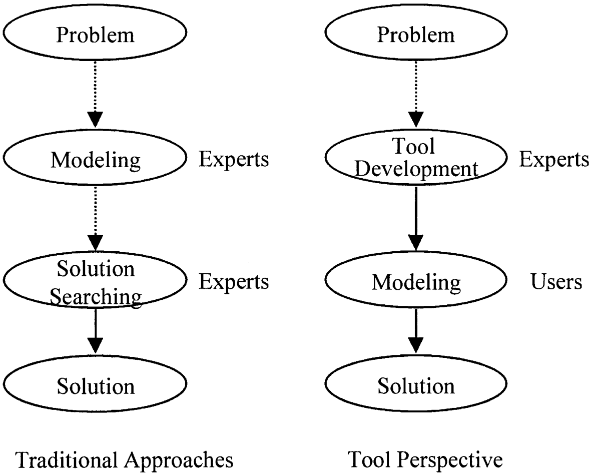

Problem
Modeling
Experts
Solution Searching
Experts
Solution
Traditional Approaches
Problem
Tool Development
Experts
Modeling
Users
Solution
Tool Perspective

Figure 13 A Comparison of Traditional Approaches and Tool Perspective. (From Huang and Nof 1998)

TABLE 2 A Sample Survey of Models of Material Handling and Robotics with the Tool Perspective

No.	Modeling Approach/Tool, Reference	Methodologies Applied	Measured Criteria or Goal of the Model	Three Modeling Concerns Addressed?
1.	Nadoli and Rangaswami (1993)	Expert system, simulator	Waiting time, congestion level, etc.	No
2.	Chadha et al. (1994)	IDEF0, Data flow diagram, extended entity relationship	Develop an integrated information system	(3) only
3.	Prakash and Chen (1995)	SIMAN IV simulator	Speed of AGV and dispatching rules	No
4.	Sly (1995)	AutoCAD with FactoryFLOW	Minimum distance of material flows	No
5.	*FDL* (Witzerman and Nof 1995, 1996); FDL/CR (Lara et al. 2000)	ROBCAD and simulator	Integrated tool for distributed engineers; computer support for conflict resolution	(1), (2), and (3)
6.	*CFS* (Furtado and Nof 1995)	Data/control flow diagram, Petri net	Integrated tool to incorporate material flows, information flows, and control signals	(1) and (3) only

Adapted from Huang and Nof (1998).

that the concerns are not addressed by the first four models. The approach of the first four models still follows the traditional approaches, except for the second model. In the following paragraphs, two automated, integrated tools for manufacturing system design are introduced to present the above concept: facility description language (FDL) and the other is concurrent flexible specification (CFS).

6.2.1. Facility Description Language (FDL)

FDL provides an integrated modeling tool for distributed engineers working on various aspects of manufacturing system design (Witzerman and Nof 1995, 1996; Lara et al. 2000). FDL is implemented in a 3D emulation environment, ROBCAD (Tecnomatix 1989), which provides a dynamic and physically visible (comparing with ‘‘iconically visible’’ in simulation animation) CAD environment. Therefore, all the materials, operations of robots and material handling facilities, and machines are shown by their true physical relationships.

In the framework of FDL, the manufacturing systems are modeled by computer information and graphics. Various information teams are included in the model that are not seen in traditional models (Witzerman and Nof 1995):

1. Organizational relationship among facility components
2. Specification of working locations for robots
3. Flow of parts and materials through the facility
4. Location of personnel and equipment aisles
5. Control relationships between devices
6. Designation of sensors and their targets

These information items are supported by the modeling functions of FDL (Table 3). An example of modeling an aisle by an FDL function is:

TABLE 3 FDL Modeling Functions

Function Class	Function
FDL Definition Input Functions	Aisle, Capabiity, Control, Define (aisle, path, perimeter), Device, Facility, Part, Process, ProcessPart, Sensor, Transfer, Workpoint
FDL Manipulation Input Functions	Attach, Delete, Detach, Display (top, lower, bottom, name), Moveback, ShiftBy, ShiftTo
FDL Utility Input Functions	Comment, Print, Save
FDL Evaluation Function	Reconcile Database, Evaluate Aisles, Evaluate Device Reach, Display Material Flow Paths, Evaluate Fields of View

Aisle	*Redefine aspects pertaining to an aisle.*	
	syntax	***aisle*** *action path parent size*
	action	char (A, C, D) add, change, delete record (mandatory)
	path	char[16] path name (ROBCAD name)
	parent	char[16] parent name from device table
	size	char “HUMAN”, “AGV”, “SMALL_FL”, or “LARGE_FL”

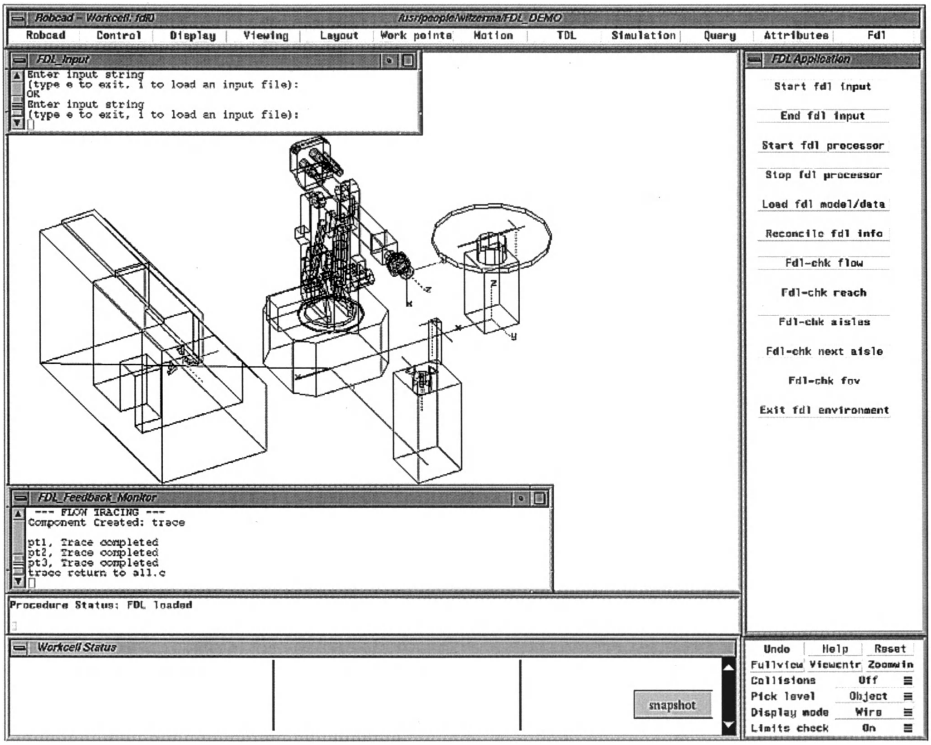
The model in FDL then becomes a list of syntax. The list of syntax triggers the ROBCAD program to construct a 3D emulation model (see the model inside the window of Figure 14). Figure 14 is an example of FDL in a ROBCAD system. The upper left window is used to input the information of the system, including the geometric information of facilities, material flow information, and material flow control information. The lower left window is used to show the output information (e.g., a collision occurring to the robot during the material handling). In addition, FDL provides a reconciliation function (see the right function menu in Figure 14). Therefore, all the control and physical conflicts on the manufacturing systems can be resolved according to the built in algorithm. The reconciliation function may change the positions of robots or machines to avoid the collision or unreachability of material handling. Recently, FDL/CR has been developed to provide knowledge-based computer support for conflict resolution among distributed designers.

Because FDL provides such direct syntax specifications, the designers can use the syntax to model and develop their subsystems. When the designers are in different locations, their subsystems can submit input to the host ROBCAD system to construct the entire system, then use the reconciliation function to adjust the subsystems if conflicts occur. Therefore, the cooperation of designers in different locations for different subsystems can be achieved in FDL. In the FDL working environment, two types of information are exchanged among the designers: (1) the design based on FDL syntax and (2) the operations of the facilities described by a task description language (TDL). TDL represents the control functions of the facilities. Thus, not only the models of the material flow but also the control information are presented and shared among the designers.

6.2.2. Concurrent Flexible Specifications (CFS)

By specification, engineers describe the way that a system should be constructed. Concurrent, flexible specifications for manufacturing systems, in other words, are provided by several engineers to model manufacturing systems with flexibility to design changes. In a manufacturing system, there is a physical flow of raw materials, parts, and subassemblies, together with an information and control flow consisting of status (system state) and control signals. The control and status signals govern the behavior of the physical flows. In order to simultaneously achieve optimal capacity loading with maintained or increased flexibility, an exact definition of material and information flow becomes necessary. This approach is followed by CFS modeling (Furtado and Nof 1995). The specification should represent not only the logical structures of different functions but also their logical connections, such as the structures of material and information flow in a manufacturing cell (Csurgai et al. 1986).

Another important requirement of specification is the ability to define precisely the real-time behavior of the system. In a real-time system, many of the inputs to the system are signals that indicate the occurrence of events. These inputs do not pass data to the system to be processed. Generally they occur in streams over time and their purpose is to trigger some process in the system


Robcad – Workcell: fdl0
/usr/people/witzerma/FDL_DEMO
Robcad
Control
Display
Viewing
Layout
Work points
Motion
TDL
Simulation
Query
Attributes
Fdl
FDL_Input
FDL Application
Start fdl input
End fdl input
Start fdl processor
Stop fdl processor
Load fdl model/data
Reconcile fdl info
Fdl-chk flow
Fdl-chk reach
Fdl-chk aisles
Fdl-chk next aisle
Fdl-chk fov
Exit fdl environment
FDL_Feedback_Monitor
Workcell Status
snapshot
Undo
Help
Reset
Fullview
Viewcntr
Zoomwin
z
x
y
z

```
Enter input string
(type e to exit, i to load an input file):
OK
Enter input string
(type e to exit, i to load an input file):
```

```
 --- FLOW TRACING ---
Component Created: trace

pt1, Trace completed
pt2, Trace completed
pt3, Trace completed
trace return to all.c
```

```
Procedure Status: FDL loaded
```

```
Collisions      Off
Pick level      Object
Display mode    Wire
Limits check    On
```

Figure 14 An Example of Facility Description Language in ROBCAD System.

repeatedly. Furthermore, many systems are made up of subsystems, any of which may be active or nonactive at a particular time during system operation. For this reason, the treatment of timing is an essential element of the specification.

To integrate the above requirements, tools that can incorporate material flows, information flows, and control signals are required. Therefore, different representations of specifications through different tools should not be independent or mutually exclusive but should support each other by forming a concurrent, comprehensive specification of the system. For manufacturing systems, the specification of *functional* and *real-time logic* is important because these attributes describe what and how the system is executing. This information is necessary to determine how the processes are to be implemented with physical equipment. By utilizing two complementary representations, both these aspects of system behavior can be specified concurrently.

Data/control flow diagrams (DFD/CFDs), which are enhanced with real-time extensions when used in conjunction with Petri nets, provide a suitable framework for concurrent specification of functional and real-time state logic. The main reason for this is the ability to maintain identical structural decompositions in both representations at all levels of detail in the specification. This model is accomplished by maintaining identical partitioning of processes in both specifications.

With DFD/CFDs, partitioning is accomplished by hierarchical decomposition of bubbles that represent processes or tasks. An identical hierarchical partitioning can be created with Petri nets by representing processes with subnets at higher levels and then showing the detailed, internal net at lower levels of detail. The DFD/CFDs provide a process definition model of the system, while the Petri nets provide a process analysis model for the study of real-time state behavior. Even though object-oriented modeling is becoming a more popular technique of system design, data/control flow diagrams are still an acceptable technique in our case study. Researchers have proved the possibility of transforming data flow diagrams to object models (Alabiso 1988).

Both these techniques are realized by two software packages: Teamwork and P-NUT. Teamwork is a computer aided software engineering (CASE) tool family that automates standard structured methodologies using interactive computer graphics and multiuser workstation power. P-NUT is a set of tools developed by the Distributed Systems Project in the Information and Computer Science

Department of the University of California at Irvine (Razouk 1987) to assist engineers in applying various Petri net-based analysis methods.

6.3. Agent-Based Control Systems

Early agents were defined for distributed artificial intelligence, which includes two main areas: distributed problem solving (DPS) and multi-agent systems (MASs). DPS focuses on centrally designed systems solving global problems and applying build-in cooperation strategies. In contrast, MAS deals with heterogeneous agents whose goal is to plan their utility-maximizing coexistence. Examples of DPS are mobile robots exploring uncertain terrain, and task scheduling in manufacturing facilities. Both can be operated with centralized programs, but in relatively more distributed environments they are usually more effective with autonomous programs, or agents. Examples of MAS are collaborative product design and group behavior of several mobile robots.

Recent research has explored the concept of autonomous agents in control. An agent is a computing system that can autonomously react and reflex to the impacts from the environment in accordance with its given goal(s). An agent reacts to the environment by executing some preloaded program. Meanwhile, there is an autonomous adjustment mechanism to provide a threshold. When the environmental impacts are higher than the threshold, the agent reflexes; otherwise it is intact. An agent may seek collaboration through communicating with other agents. The communication among agents is regulated by protocols, structure of dialogue, to enhance the effectiveness and efficiency of communication.

An important difference between autonomous agents and other techniques is that an autonomous agent evaluates the rules that it will perform. It may even automatically change its goal to keep itself alive in a harsh environment. Autonomous agents have been applied in many control systems, including air traffic control, manufacturing process control, and patient monitoring.

Usually an agent functions not alone, but as a member of a group of agents or an agent network. The interaction and communication among agents can be explained by the analogy of organizational communication. An organization is an identifiable social pursuing multiple objectives through the coordinated activities and relations among members and objects. Such a social system is open ended and depends for its effectiveness and survival on other individuals and subsystems in the society of all related organizations and individuals. (It is actually similar for both human societies and agent societies.) Following this analogy, three characteristics of an organization and of an agent network can be observed (Weick 1990):

1. Entities and organization
2. Goals and coordinated activities
3. Adaptability and survivability of the organization

Five motivations have been observed for organizational and agent network communication (Jablin 1990):

1. Generate and obtain information
2. Process and integrate information
3. Share information needed for the coordination of interdependent organizational tasks
4. Disseminate decisions
5. Reinforce a group's perspective or consensus

These five motivations can serve as a checklist for developing protocols. One of the most influential factors affecting interpersonal or interagent communication patterns among group members is the characteristic of the task on which they are working. As task certainty increases, the group coordinates itself more through formal rules and plans than through individualized communication modes. Therefore, the interacting behaviors and information exchanges among agents have to follow interaction and communication protocols.

Although different agent applications will require different agent design, five general areas have to be addressed:

1. Goal identification and task assignment
2. Distribution of knowledge
3. Organization of the agents
4. Coordination mechanism and protocols
5. Learning and adaptive schemes

Research into intelligent, collaborative agents is extremely active and in its preliminary stages (Nof 1999; Huang and Nof 2000). While the best-known applications have been in Internet search and remote mobile robot navigation, emerging examples combine agents through computer networks with remote monitoring for security, diagnostics, maintenance and repair, and remote manipulation of robotic equipment. Emerging agent applications will soon revolutionize computer and communication usefulness. Interaction and communication with and among intelligent tools, home appliances, entertainment systems, and highly reliable, safe mobile service robots will change the nature of manufacturing, services, health care, food delivery, transportation, and virtually all equipment-related activities.

7. CONCLUSION

This chapter discusses automation technology in various levels:

1. *Single process:* automatic control theory and technology
2. *Single machine:* artificial intelligence and knowledge-based technology
3. *Distributed machines and distributed systems:* integration theory and technology

Additionally, the trends in future automation technology are classified into another three levels:

1. *Machine integration:* virtual machines
2. *Human–machine integration:* tool-oriented technologies
3. *Machine autonomy:* agent-based technologies

In our postulation, how to automate machines in terms of operations will soon be a routine problem because of the technological maturity of actuators, sensors, and controllers. The application of automation technology will then focus on the issues of intelligence, integration, and autonomy. Another emerging trend involves the incorporation of micro-electromechanical systems (MEMSs) as sensors and controllers of automation and the microscale. However, education and training of workers who interact with intelligent and autonomous machines may be another important research issue in the future.

Acknowledgments

The authors would like to thank Professor Ted J. Williams of Purdue University and Professor Li-Chih Wang, Director of the Automation Laboratory, Tunghai University, Taiwan, for their valuable contributions to this chapter.

REFERENCES

Alabiso, B. (1988), “Transformation of Data Flow Diagram Analysis Models to Object-Oriented Design,” in *OOPSLA ’88: Object-Oriented Programming Systems, Languages and Applications Conference Proceedings* (San Diego, CA, September 25–30, 1988), N. Meyrowitz, Ed., Association for Computing Machinery, New York, pp. 335–353.

Burdea, G. C., and Coiffet, P. (1999), “Virtual Reality and Robotics,” in *Handbook of Industrial Robotics, 2nd Ed.,* S. Y. Nof, Ed., John Wiley & Sons, New York, pp. 325–333.

Chadha, B., Fulton, R. E., and Calhoun, J. C. (1994). “Design and Implementation of a CIM Information System,” *Engineering with Computers,* Vol. 10, No. 1, pp. 1–10.

Cohen, A., and Apte, U. (1997), *Manufacturing Automation,* McGraw-Hill, New York.

Csurgai, G., Kovacs, V., and Laufer, J. (1986), “A Generalized Model for Control and Supervision of Unmanned Manufacturing Cells,” in *Software for Discrete Manufacturing: Proceedings of the 6th International IFIP/IFAC Conference on Software for Discrete Manufacturing, PROLAMAT 85* (Paris, France, June 11–13, 1985), J. P. Crestin and J. F. McWaters, Eds., North-Holland, Amsterdam, pp. 103–112.

Dayhoff, J. E. (1990), *Neural Network Architecture,* Van Nostrand Reinhold, New York.

Elnakhal, A. E., and Rzehak, H. (1993), “Design and Performance Evaluation of Real Time Communication Architectures,” *IEEE Transactions on Industrial Electronics,* Vol. 40, No. 4, pp. 404–411.

Freeman, J. A., and Skapura, D. M. (1991), *Neural Networks: Algorithms, Applications, and Programming Techniques,* Addison-Wesley, Reading, MA.

Fuller, R. (1995), *Neural Fuzzy Systems,* http://www.abo.fi/~rfuller/nfs.html.

Furtado, G. P., and Nof, S. Y. (1995), ‘‘Concurrent Flexible Specifications of Functional and Real-Time Logic in Manufacturing,’’ Research Memorandum No. 95-1, School of Industrial Engineering, Purdue University.

Huang, C. Y., and Nof, S. Y. (1998), ‘‘Development of Integrated Models for Material Flow Design and Control—A Tool Perspective,’’ *Robots and CIM,* Vol. 14, 441–454.

Huang, C. Y., and Nof, S. Y. (1999), ‘‘Enterprise Agility: A View from the PRISM Lab,’’ *International Journal of Agile Management Systems,* Vol. 1, No. 1, pp. 51–59.

Huang, C. Y., and Nof, S. Y. (2000), ‘‘Formation of Autonomous Agent Networks for Manufacturing Systems,’’ *International Journal of Production Research,* Vol. 38, No. 3, pp. 607–624.

Jablin, F. M. (1990), ‘‘Task/Work Relationships: A Life-Span Perspective,’’ in *Foundations of Organizational Communication: A Reader,* S. R. Corman, S. P. Banks, C. R. Bantz, M. Mayer, and M. E. Mayer, Eds., Longman, New York, pp. 171–196.

Kuo, B. C. (1995), *Automatic Control Systems,* 7th Ed., John Wiley & Sons, New York.

Lara, M. A., Witzerman, J. P., and Nof, S. Y. (2000), ‘‘Facility Description Language for Integrating Distributed Designs,’’ *International Journal of Production Research,* Vol. 38, No. 11, pp. 2471–2488.

Lin, C.-T., and Lee, G. (1996), *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems,* Prentice Hall, Upper Saddle River, NJ.

Mahalik, N. P., and Moore, P. R. (1997), ‘‘Fieldbus Technology Based, Distributed Control in Process Industries: A Case Study with LonWorks Technology,’’ *Integrated Manufacturing Systems,* Vol. 8, Nos. 3–4, pp. 231–243.

Morriss, S. B. (1995), *Automated Manufacturing Systems: Actuators, Controls, Sensors, and Robotics,* McGraw-Hill, New York.

Nadoli, G., and Rangaswami, M. (1993), ‘‘An Integrated Modeling Methodology for Material Handling Systems Design,’’ in *Proceedings of the 1993 Winter Simulation Conference* (Los Angeles, December 12–15, 1993), G. W. Evans, Ed., Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 785–789.

Nof, S. Y., Ed. (1999), *Handbook of Industrial Robotics,* 2nd Ed., John Wiley & Sons, New York.

Prakash, A., and Chen, M. (1995), ‘‘A Simulation Study of Flexible Manufacturing Systems,’’ *Computers and Industrial Engineering,* Vol. 28, No. 1, pp. 191–199.

Razouk, R. R. (1987), ‘‘A Guided Tour of P-NUT (Release 2.2),’’ ICS-TR-86-25, Department of Information and Computer Science at the University of California, Irvine.

Semiconductor Equipment and Materials International (SEMI) (1997), *SEMI International Standards,* SEMI, Mt View, CA.

Shanmugham, S. G., Beaumariage, T. G., Roberts, C. A., and Rollier, D. A. (1995), ‘‘Manufacturing Communication: A Review of the MMS Approach,’’ *Computers and Industrial Engineering,* Vol. 28, No. 1, pp. 1–21.

SISCO Inc. (1995), *Overview and Introduction to the Manufacturing Message Specification* (*MMS*) *Revision 2.*

Sly, D. P. (1995), ‘‘Plant Design for Efficiency Using AutoCAD and FactoryFLOW,’’ in *Proceedings of the 1995 Winter Simulation Conference* (Arlington, VA, December 3–6, 1995), C. Alexopoulos, K. Kong, W. R. Lilegdon, and D. Goldsman, Eds., Society for Computer Simulation, International, San Diego, CA, pp. 437–444.

Tecnomatix Technologies (1989), *ROBCAD TDL Reference Manual,* version 2.2.

Wang, L.-C. (1996), ‘‘Object-Oriented Petri Nets for Modelling and Analysis of Automated Manufacturing Systems,’’ *Computer Integrated Manufacturing Systems,* Vol. 9, No. 2, pp. 111–125.

Wang, L. C., and Wu, S.-Y. (1998), ‘‘Modeling with Colored Timed Object-Oriented Petri Nets for Automated Manufacturing Systems,’’ *Computers and Industrial Engineering,* Vol. 34, No. 2, pp. 463–480.

Weick, K. (1990), ‘‘An Introduction to Organization,’’ In *Foundations of Organizational Communication: A Reader,* S. R. Corman, S. P. Banks, C. R. Bantz, M. Mayer, and M. E. Mayer, Eds., Longman, New York, pp. 124–133.

Williams, T. J., and Nof, S. Y. (1992), ‘‘Control Models,’’ in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, pp. 211–238.

Witzerman, J. P., and Nof, S. Y. (1995), ‘‘Facility Description Language,’’ in *Proceedings of the 4th Industrial Engineering Research Conference* (Nashville, TN, May 24–25, 1995), B. Schmeiser and R. Vzsoy, Eds., Institute of Industrial Engineers, Norcross, GA, pp. 449–455.

Witzerman, J. P. and Nof, S. Y. (1996), ‘‘Integration of Simulation and Emulation with Graphical Design for the Development of Cell Control Programs,’’ *International Journal of Production Research,* Vol. 33, No. 11, pp. 3193–3206.

CHAPTER 6

Computer Integrated Technologies and Knowledge Management

Frank-Lothar Krause
Technical University Berlin

Kai Mertins
Andreas Edler
Peter Heisig
Ingo Hoffmann
Fraunhofer Society

Markus Helmke
Technical University Berlin

1. INTRODUCTION 178

2. DESIGN TECHNIQUES, TOOLS, AND COMPONENTS 178

- 2.1. Computer-Aided Design 178
  - 2.1.1. 2D Design 178
  - 2.1.2. 3D Design 180
  - 2.1.3. Parametrical Design 183
  - 2.1.4. Feature Technology 185
  - 2.1.5. Assembly Technology 185
  - 2.1.6. Further Technologies 187
- 2.2. CAD Interfaces 191
  - 2.2.1. General Explanations 191
  - 2.2.2. Standarization of CAD Interfaces 191
- 2.3. Engineering Data Management 195
- 2.4. Architecture and Components 196
- 2.5. Calculation Methods 199
  - 2.5.1. General Explanations 199
  - 2.5.2. Finite Element Methods 199

3. CONCEPTS 203

- 3.1. Process Chains 203
  - 3.1.1. Tasks of Process Control and Monitoring 205
- 3.2. Integrated Modeling 205
- 3.3. Methodical Orientation 206
- 3.4. Rapid Prototyping 207
  - 3.4.1. Systemizaion of Rapid Prototyping Processes 208
- 3.5. Digital Mock-up 209

4. ARCHITECTURES 210

- 4.1. General Explanations 210
- 4.2. Integration of Application Modules and Core Modeler 211
- 4.3. Shared System Architectures 212

5. KNOWLEDGE MANAGEMENT 213

- 5.1. Origin and Background 213
- 5.2. Knowledge 213
- 5.3. Knowledge Management Is Business and Process Oriented 214
  - 5.3.1. The Core Process of Knowledge Management 215
  - 5.3.2. Design Fields of Knowledge Management 215
- 5.4. Approaches to the Design of Business Process and Knowledge Management 217
- 5.5. A Method for Business Process-Oriented Knowledge Management 218
- 5.6. Knowledge-Management Tools 220
  - 5.6.1. Technologies for Knowledge Management 220

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

5.6.2. Architecture for Knowledge Management 222
5.7. Future Trends 222
REFERENCES 223

1. INTRODUCTION

The rapidly growing globalization of industrial processes is defining the economic structure worldwide. The innovative power of a company is the deciding factor for success in gaining market acceptance. Strengthening innovation, however, relies on the ability to structure the design phase technically so that innovative products are generated with consideration of certain success factors such as time, expense, quality, and environment. Furthermore, these products should not only be competitive but also be able to maintain a leading position on the world market. The opportunity to increase innovation lies in deciding on novel approaches for generating products based on future-oriented, computer-integrated technologies that enable advances in technology and expertise. In the end, this is what guarantees market leadership.

In recent years, progress in product development has been made principally through the increasing use of information technology and the benefits inherent in the use of these systems. In order to fulfill the various requirements on the product development process in respect to time, costs, and quality, a large number of different approaches have been developed focusing on the optimization of many aspects of the entire development process.

This chapter provides an overview of the computer integrated technologies supporting product development processes and of the current situation of knowledge management. Section 2 discusses the design techniques affected by computer-aided technologies and the corresponding tools. Section 3 gives a general introduction of concepts of computer integrated technologies and describes basic technologies and organizational approaches. Section 4 explains possible architectures for software tools that support product development processes. Section 5 gives a wide overview of the application and development of knowledge management.

2. DESIGN TECHNIQUES, TOOLS, AND COMPONENTS

2.1. Computer-Aided Design

2.1.1. 2D Design

Like traditional design, 2D geometrical processing takes place through the creation of object contours in several 2D views. The views are then further detailed to show sections, dimensions, and other important drawing information. As opposed to 3D CAD systems, 2D CAD design occurs through conventional methods using several views. A complete and discrepancy-free geometrical description, comparable to that provided by 3D CAD systems, is not possible. Therefore, the virtual product can only be displayed to a limited extent. Two-dimensional pictures still have necessary applications within 3D CAD systems for hand-sketched items and the creation of drawing derivations.

Despite their principal limitation compared to 3D CAD systems, 2D systems are widely used. A 3D object representation is not always necessary, for example, when only conventional production drawings are required and no further computer aided processing is necessary. In addition, a decisive factor for their widespread use is ease of handling and availability using relatively inexpensive PCs.

The main areas of 2D design are the single-component design with primitives geometry, sheet metal, and electrical CAD design for printed circuit boards.

Depending on the internal data structure of the computer, classical drawing-oriented and parametrical design-oriented 2D CAD systems can be differentiated. This distinction has an essential influence on the fundamental processing principles and methodical procedures followed during the design phase. The drawing-oriented systems are then implemented efficiently in the detailing phases while new-generation parametrical systems support the development in early concept phases. With efficient equation algorithms running in the background, extensive calculations and parametrical dependencies for simulation purposes can be applied.

The essential goal in using design-oriented, 2D CAD systems is creation of manufacturing plans and documentation. From the exact idea of the topology and dimensions of the technical entity to the input of the geometry and ending with the detailing of the technical drawing and preparation of the parts list, this computer-supported process corresponds to a large degree with conventional strategies.

The technical drawings are created with the aid of geometrical elements available within the system. The following geometrical elements can be used: points, straight lines, parallel lines, tangents, circles, arcs, fillets, ellipses, and elliptical arcs such as splines. Furthermore, equidistant lines may be created. Editing commands such as *paste, delete, join, extend, trim,* and orientation operations such as translation and rotation, exist for element manipulation purposes. Further design clarification and simplification is provided by the help functions of scaling, duplicating, locating, and mirroring on symmetrical axes. Detail work may be simplified with the use of zoom functions, and snap points aid in dimensioning and positioning of geometrical elements. The technical drawings are then completed with the use of hatching, dimensioning, and labeling.

The following strategies are recommended for the completion of drawings using 2D CAD systems:

- Do default settings for drawing parameters such as scale ratios, line characteristics and hatching, font, and measurement standards.
- Call up and blend in predefined frames with text and form fields.
- Use construction lines for added support.
- Create contour and surface in different views.
- Apply eventual cutouts and hatching
- Set measurements, tolerances, etc.
- Transfer generated objects to predefined printing or storage mediums.

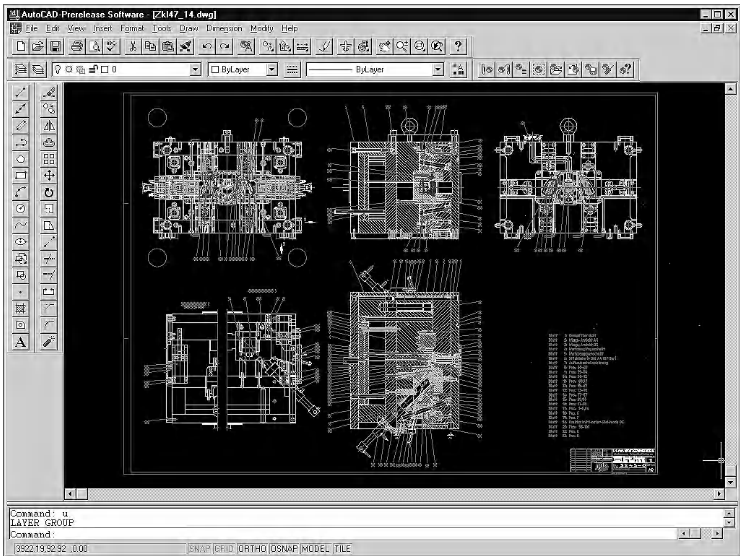
An example of a complex assembly drawing is presented in Figure 1.

With the use of layers, drawing information of all types may be distributed over any desired number of named and editable layers. The layers simplify the drawing process because momentarily irrelevant information can be easily blended out. Additional text in different languages or measurement systems or norms can also be called up and implemented.

Grouping techniques allow for the summarizing of geometrical elements through various selection processes and in turn ease further working processes.

The majority of systems provide function libraries for the simplification of drawing preparation and the efficient use of results. The following areas may be simplified with the use of function libraries:

- Creation and manipulation of geometrical elements
- Geometrical calculations
- Access to saved elements using search criterion
- Creation of groups and layers
- Utilization of drawing elements such as dimensioning, hatching, and text
- Use of standard dialogs for value, point, or font input



```
Command:  u
LAYER GROUP
Command:
```

3922.19,92.92 ,0.00

Figure 1 An Example of a 2D Drawing. (AutoCAD System. Reproduced with permission of Autodesk®)

Geometry macros are internal computer representations of dimension and form-variable component geometries of any complexity. The macros are easily called up from a database and copied by the user. The user may then easily insert the copied macro into the current object representations. The advantages of the geometry macros are seen especially when working with standard and repeat parts.

A spatial representation is desirable for a variety of tasks where primarily 2D representations are utilized. Therefore, in many systems an existing function is available for isometric figures that provide an impression of the object's spatial arrangement without the need for a 3D model. This approach is sufficient for the preparation of spare parts catalogs, operating manuals, and other similar tasks. Prismatic bodies and simplified symmetric bodies may also be processed. For the use of this function it is necessary to present the body in front and side views as well as plan and rear views where required. After the generation of the isometric view, the body must be completed by blending out hidden lines. This happens interactively because the system actually does not have any information available to do this automatically. So-called 2½ systems generate surface-oriented models from 2D planes through translation or rotation of surface copies. The main application for these models lies in NC programming and documentation activities.

To reduce redundant working steps of geometrical modifications, the dimensioning of the part must be finished before the computer-supported work phases. This approach, which underlies many CAD systems today, is connected to the software's technical structure. In geometry-oriented dimensioning, the measurement alone is not inferred but rather the totality of geometrical design of the entity. The supporting role played by the system for interactive functioning is limited to automatic or semiautomatic dimension generation. The dimensional values are derived from the geometry specified by the designer. The designer is forced to delete and redo the geometry partially or completely when he or she wishes to make changes. This creates a substantial redundancy in the work cycle for the project.

The principal procedures of dimension-oriented modeling in parametric 2D CAD systems often correspond to conventional 2D systems. For the generation of geometrical contours, the previously mentioned functions, such as lines, circles, tangents, trimming, move, and halving, are used. The essential difference is the possibility for free concentration on the constructive aspects of the design. In parametrical systems, on the other hand, the measurements are not fixed but rather are displayed as variable parameters. Associative relationships are built between the geometry and measurements, creating a system of equations. Should the geometry require modification, the parameters may be varied simply by writing over the old dimensions directly on the drawing. This in turn changes the geometrical form of the object to the newly desired values. The user can directly influence the object using a mouse, for example, by clicking on a point and dragging it to another location, where the system then calculates the new measurement. The changes cause a verification of the geometrical relationships as well as the consistency of the existing dimension in the system.

Another advantageous element of the parametrical system is the ability to define the above-mentioned constraints, such as parallel, tangential, horizontal, vertical, and symmetrical constraints. Constraints are geometrical relationships fixed by formula. They specify the geometrical shape of the object and the relevant system of equations necessary for geometry description. Thus, a tangent remains a tangent even when the respective arc's dimension or position is changed. With the constraints, further relationships can be established, such as that two views of an object may be linked to one another so that changes in one view occur the related view. This makes the realization of form variations fairly simple.

Modern CAD systems include effective sketching mechanisms for the support of the conceptual phase. For the concept design, a mouse- or pen-controlled sketching mechanism allows the user to portray hand-drawn geometries relatively quickly on the computer. Features aiding orientation, such as snap-point grids, and simplifying performing the alignment and correction of straight, curved, orthogonal, parallel, and tangential elements within predetermined tolerances. This relieves the user of time-consuming and mistake-prone calculations of coordinates, angles, or points of intersection for the input of geometrical elements. When specifying shape, the parametrical system characteristics come into play, serving to keep the design intent. The design intent containing the relative positioning is embodied by constraints that define position and measurement relationships. Component variations are easily realizable through changes of parameter values and partially with the integration of calculating table programs in which table variations may be implemented.

The existing equation system may be used for the simulation of kinematic relations. Using a predetermined, iterative variation of a particular parameter and the calculation of the dependent parameter, such as height, simple kinematic analysis of a geometrical model is possible. This simplifies packaging analysis, for example.

2.1.2. 3D Design

2.1.2.1. Surface Modeling The geometrical form of a physical object can be internally represented completely and clearly using 3D CAD systems, is contrast to 2D CAD systems. Therefore, the user has the ability to view and present reality-like, complex entities. Thus, 3D CAD systems provide the basis for the representation of virtual products. Through 3D modeling and uniform internal computer representation, a broader application range results so that the complete product creation process can be digitally supported. This starts with the design phase and goes through the detailing, calculation, and drawing preparation phases and on to the production. Because of the complete and discrepancy-free internal computer representation, details such as sectional drawings may be derived automatically. Also, further operations such as finite element analysis and NC programs are better supported. The creation of physical prototypes can be avoided by using improved simulation characteristics. This in turn reduces the prototype creation phase.

Due to the required complete description of the geometry, a greater design expenditure results. Because of the greater expenditure, special requirements for user-friendliness of component geometry generation of components and their links to assemblies and entire products are necessary.

3D CAD systems are classified according to their internal computer representations as follows (Grätz 1989):

- Edge-oriented models (wire frames)
- Surface-oriented models
- Volume-oriented models

The functionality and modeling strategy applied depend on the various internal computer representations.

Edge-oriented models are the simplest form of 3D models. Objects are described using end points and the connections made between these points. As in 2D CAD systems, various basic elements, such as points, straight lines, circles and circular elements, ellipses, and free forming, are available to the user and may be defined as desired. Transformation, rotation, mirroring, and scaling possibilities resemble those of 2D processing and are related to the 3D space.

Neither surfaces nor volumes are recognizable in wire frame models. Therefore, it is not possible to make a distinction between the inside and outside of the object, and section cuts of the object cannot be made. Furthermore, quantifiable links from simple geometrical objects to complex components are not realizable. This is a substantial disadvantage during the practical design work. An arbitrary perspective is possible, but in general a clear view is not provided. Among other things, the blending out of hidden lines is required for a graphical representation. Because of the lack of information describing the object's surfaces, this cannot take place automatically but must be carried out expensively and interactively. Models of medium complexity are already difficult to prepare in this manner. The modeling and drawing preparation are inadequately supported. Another problem is the consistency of the geometry created. The requirements for the application of a 3D model are indeed given, but because of the low information content the model does not guarantee a mistake-free wire frame presentation (Grätz 1989).

Surface-oriented systems are able to generate objects, known as free-form surfaces, whose surfaces are made up of numerous curved and analytically indescribable surfaces. One feature visible in an internal computer representation of free-form surfaces is their interpolated or approximated nature. Therefore, various processes have been developed, such as the Bézier approximation, Coon's surfaces, the NURBS representations, and the B-spline interpolation (see Section 4.1).

As with all other models, the computer internal representation of an object implies characteristic methods and techniques used in generating free-form surfaces. In general, basis points and boundary conditions are required for the surface description. Free-form surfaces, such as the one portrayed in Figure 2, can be generated using a variety of techniques. A few examples for the generation of free-form surfaces will be described later.

In addition to the simple color-shaded representations, photorealistic model representations may be used for the review and examination of designs. The models are portrayed and used in a reality-like environment and are supported by key hardware components. This allows for the consideration of various environmental conditions such as type and position of the light source, the observer's position and perspective, weather conditions, and surface characteristics such as texture and reflectivity. Also, animation of the object during design is possible for the review of movement characteristics. Through these visualization methods it is possible to assess a product's design before prototype production begins and aid the preparation of sales brochures and other documents.

A characteristic feature and major disadvantage of surface models is the fact that the surfaces are not automatically correlated to form a body. This has important consequences for the designer. It results in a distinct modeling technique for the user that offers only limited functionality with regard to realization of the design tasks. For example, from the outset it is not defined whether a surface should face the inside or the outside of the object. Therefore, the designer must be careful to ensure a closed and complete model during design phases. Intersections are not recognizable by the system

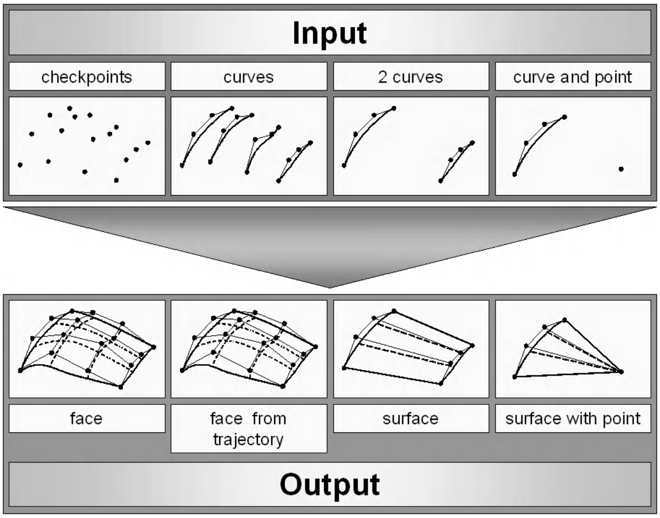


Figure 2 Modeling Processes for Free-Form Surfaces.

because of missing information describing the solid during cutting operations. The intersections are represented only by contour lines that are interactively hatched.

Surface models are applied when it is necessary to generate a portrait of free-formed objects. Such models are used in the automobile, shipbuilding, and consumer goods industries, for example. The surface formations of the model, which are distinguished into functional and esthetic surfaces, can be used as ground-level data for various applications employed later on. This includes later applications such as the generation of NC data from multiaxis processing. Another area of application is in FEM analysis, where the surface structure is segmented into a finite element structure that is subsequent to extensive mechanical or thermal examinations. With surface models, movement simulations and collision analyses can be carried out. For example, mechanical components dependent on each other can be reviewed in extreme kinematic situations for geometrical overlapping and failure detection.

2.1.2.2. Solid Modeling A very important class of 3D modelers is generated by the volume systems. The appropriate internal computer representation is broken down into either boundary representation models (B-reps) or Constructive Solid Models (CSGs). The representation depends on the basic volume modeling principles and the deciding differences between the 3D CAD system classes.

Volume modelers, also known as solid modelers, can present geometrical objects in a clear, consistent, concise manner. This provides an important advantage over previously presented modeling techniques. Because each valid operation quasi-implicitly leads to a valid model, the creativity of the user is supported and the costly verification of geometric consistency is avoided.

The main goal of volume modeling is the provision of fundamental data for virtual products. This data includes not only the geometry but also all information gathered throughout the development process, which is then collected and stored in the form of an integrated product model. The information allows for a broad application base and is available for later product-related processes. This includes the preparation of drawings, which consumes a large portion of the product development time. The generation of various views takes place directly, using the volume model, whereby the formation of sections and the blending out of hidden lines occurs semiautomatically.

Normally, volume models are created using a combination of techniques. The strengths of the 3D volume modeling systems lie in the fact that they include not only the complete spectrum of well-known modeling techniques for 2D and 3D wire frames and 3D surface modeling but also the new, volume-oriented modeling techniques. These techniques contribute significantly to the reduction of costly input efforts for geometrically complex objects.

The volume model can be built up by linking basic solid bodies. Therefore, another way of thinking is required in order to describe a body. The body must be segmented into basic bodies or provided by the system as basic elements (Spur and Krause 1984). Each volume system comes with a number of predefined simple geometrical objects that are automatically generated. The objects are generated by a few descriptive system parameters. Spheres, cubes, cones, truncated cones, cylinders, rings, and tetrahedrons are examples of the basic objects included in the system.

Many complex components can be formed from a combination of positioning and dimensioning of the basic elements.

Complex components are created in a step-by-step manner using the following connecting operations:

- Additive connection (unification)
- Substractive connection (differential)
- Section connection (average)
- Complementing

These connections are known as Boolean or set-theoretical operations. For object generation using Boolean operations, the user positions two solids in space that touch or intersect one another. After the appropriate functions are called up (unification, average, complementing, or differential), all further steps are carried out by the system automatically and then present the geometrical object.

The sequence of the operations when the set-theoretical operations are applied is of decisive importance. Some CAD systems work with terms used for manufacturing techniques in order to provide the designer with intuitive meanings. This allows for the simple modeling of drilling, milling, pockets, or rounding and chamfering. This results because each process can be represented by a geometrical object and further defined as a tool. The tool is then defined in Boolean form and the operation to perform the removal of material from the workpiece is carried out. To correct the resulting geometry, it is necessary that the system back up the complete product creation history. This is required to make a step by step pattern that can be followed to return the set-theoretical operations to their previous forms.

Set-theoretical operations provide an implementation advantage where complex model definitions must be created with relatively few input commands. These definitions are, if even realizable, extremely costly and time consuming by conventional methods. The disadvantage is that a relatively high conceptual capability is required of the user.

Solid models with a number of free-form surfaces are realized using surface-oriented modeling techniques that correspond to the surface model presented. With one of these techniques, *sweeping,* one attains a 3D body through an intermediate step when creating a 2D contour. The 2D contour is expanded to three dimensions along a set curve. The desired contours are generated with known 2D CAD system functions and are called up as bases for sweep operations.

In rotational sweeping, objects are generated through the rotation of surfaces, as well as closed or open, but restricted, contour lines around a predefined axis. The axis may not cut the surface or contour. A prismatic body originates from the expansion of a closed contour line. Shells or full bodies can be generated with sweep operations. With sweep-generated shells, movement analysis and assembly inspection can occur. The analysis takes place by the bodies' movement along a curve in space.

The modeling strategy implemented depends heavily on the actual task and the basic functionality offered by the CAD system. This means that the strategy implemented cannot be arbitrarily specified, as most CAD systems have a variety of alternative approaches within the system. For the user, it is important that a CAD system offer as many modeling methods as possible. Modeling, such as surface and solid modeling, is important, and the ability to combine and implement these methods in geometrical models is advantageous to the user. Because of the numerous modeling possibilities offered, the CAD system should have as few restrictions as possible in order to allow the user to develop a broad modeling strategy. Figure 3 shows a surface-oriented model of a bumper that was created using a simple sweep operation of a spline along a control curve. Using a variable design history, the air vent may be positioned anywhere on the bumper's surface without time-consuming design changes during modeling. The cylindrical solid model that cuts and blends into the surface of the bumper demonstrates that no difference exists between a solid and surface model representation. Both models may be hybridized within components. A CAD system supporting hybridized modeling creates a major advantage because the designer realizes a greater freedom of design.

2.1.3. Parametrical Design

Another development in CAD technology is parametrical modeling (Frei 1993). The function of parametrical systems displays basic procedure differences when compared to conventional CAD systems. The principal modeling differences were explained in the section on parametrical 2D CAD systems.

In parametrical model descriptions, the model can be varied in other areas through the selection and characterization of additional constraints.

Boundary conditions can be set at the beginning of the design phase without fixing the final form of the component. Therefore, generation of the geometry takes place through exact numerical input and sketching. The designer can enter the initial designs, such as contour, using a 2D or 3D sketcher. In this manner, geometrical characteristics, such as parallel or perpendicular properties, that lie within

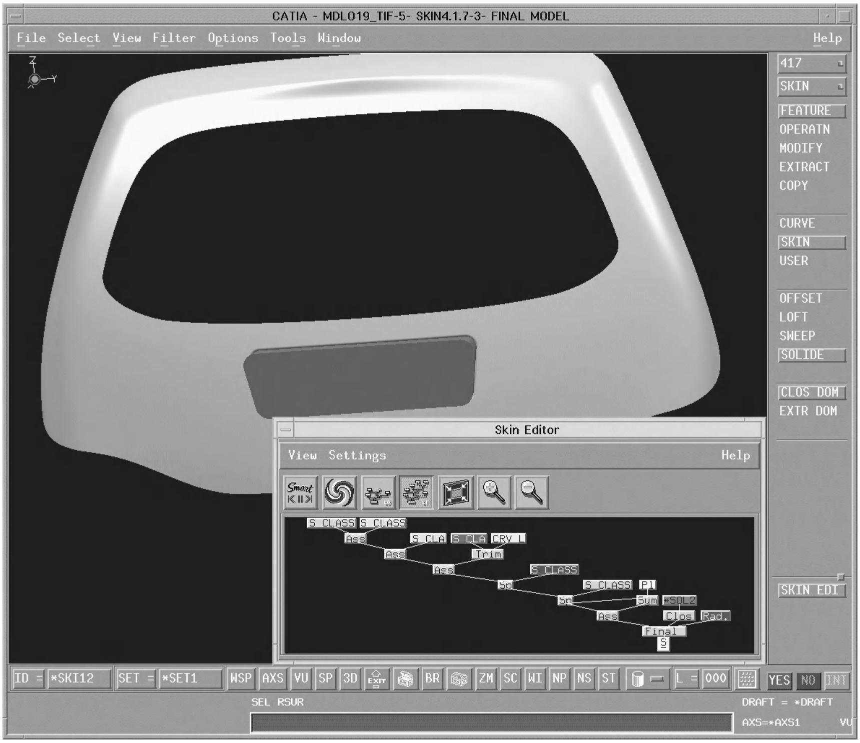

CATIA - MDL019_TIF-5- SKIN4.1.7-3- FINAL MODEL
File Select View Filter Options Tools Window Help
417
SKIN
FEATURE
OPERATN
MODIFY
EXTRACT
COPY
CURVE
SKIN
USER
OFFSET
LOFT
SWEEP
SOLIDE
CLOS DOM
EXTR DOM
SKIN EDI
Skin Editor
View Settings Help
Smart
S CLASS
S CLASS
Ass
S CLA
S CLA
CRV L
Ass
Trim
Ass
S CLASS
Sp
S CLASS
P1
Sp
Sum
*SOL2
Ass
Clos
Rad.
Final
S
ID = *SKI12
SET = *SET1
WSP AXS VU SP 3D EXIT BR ZM SC WI NP NS ST
L = 000
YES NO INT
SEL RSUR
DRAFT = *DRAFT
AXS=*AXS1
VU

Figure 3 Hybrid Surface and Solid Model. (CATIA System, Dassault Systèmes)

set tolerances are automatically and correctly recognized. The characteristics may also be set through predetermined constraints. Subsequently, the user can develop the initial dimensions using operations that essentially resemble those of conventional measurement determination. Through associativity, the adaptation of the model presented takes places automatically.

Beginning with the first contour sketches, the designer attains the desired solid object through further modeling using operations such as sweeping. The design history is of great importance at this point of the development process. The history can subsequently be modified, influencing the geometry of the model. Because of the associative relationship between the dimensioning and the solid model, the geometry of a component is easily varied in every design phase. Based on the parametrical method implemented, the designer can define conditions for parallel, tangential, or perpendicularity characteristics. Also, complex parametrical dependencies from the system can be solved simultaneously using nonlinear equations.

Because of the parametrical relationships created in the background of the process, the user is in a position to modify the dimensions and topology of the geometry later in the project. Costly reworking of the model due to changes of the boundary conditions is avoided.

Complete process chains may be built up associatively. Downstream activities, including drawing creation and detailing, make up a large percentage of the design time. Here the generation of various views directly from the existing solid model may take place because the formation of sections and the blending out of hidden lines occurs automatically. The main advantage of this approach over previous CAD systems is due to the above-mentioned preservation of the bidirectional relationship between the solid model and the drawing derived from the model. This associative connection allows the designer to take geometrical modifications directly from a drawing, even if the designer has progressed substantiality into the detailing phase. Conversely, changes to the solid model are automatically reflected in the drawing.

The effect and the functionality of the parametric concepts do not end with the completion of the geometry. Rather, they are encompassed in all areas of product development. Solid modeling, calculations, drawing preparation, NC programming, simulation, and documentation are connected together in an associative chain so that modifications made in one area automatically take place in the other areas.

For the designer, it is important that both parametric and nonparametric geometry can be combined and implemented during the development of a component. Therefore, complex components must not be completely described parametrically. The designer decides in which areas it is advantageous to present the geometry parametrically.

2.1.4. Feature Technology

Most features in commercial CAD systems are built upon a parametrical modeler. The ability to define one's own standard design elements also exists. These elements can be tailored to the designer's way of thinking and style of expression. The elements are then stored in a library and easily selected for implementation in the design environment. Additional attributes such as quantity, size, and geometrical status can be predetermined to create logical relationships. These constraints are then considered and maintained with every change made to an element. An example of a design feature is a rounding fillet. The fillet is selected from the features library and its parameters modified to suit the required design constraints.

The library, containing standard elements, can be tailored to the user's requirements for production and form elements. This ensures consistency between product development and the ability to manufacture the product. Therefore, only standard elements that match those of available production capabilities are made available. Feature definition, in the sense of concurrent engineering, is an activity in which all aspects of product creation in design, production, and further product life phases are considered.

In most of today's commercially available CAD systems, the term *feature* is used mainly in terms of form. Simple parametric and positional geometries are made available to the user. Usually features are used for complicated changes to basic design elements such as holes, recesses, and notches. Thus, features are seen as elements that simplify the modeling process, not as elements that increase information content. With the aid of feature modeling systems, user-defined or company-specific features can be stored in their respective libraries. With features, product design capabilities should be expanded from a pure geometric modeling standpoint to a complete product modeling standpoint. The designer can implement fully described product parts in his or her product model. Through the implementation of production features, the ability to manufacture the product is implicitly considered. A shorter development process with higher quality is then achieved.

By providing generated data for subsequent applications, features represent an integration potential.

2.1.5. Assembly Technology

Complex products are made up of numerous components and assemblies. The frequently required design-in-context belongs to the area of assembly modeling. Thereby, individual parts that are relevant to the design of the component are displayed on the screen. For instance, with the support of EDM functionality, the parts can be selected from a product structure tree and subsequently displayed. The part to be designed is then created in the context already defined. If the assembly is then worked on by various coworkers, the methods for shared design are employed. For example, changes to a component can be passed on to coworkers working on related and adjacent pieces through system-supported information exchange.

An example of a complex assembly in which shared design processes are carried out is the engine in Figure 4. The goal is to be able to build and then computer analyze complex products. The assembly modeling will be supported by extensive analysis and representation methods such as interference checks as well as exploded views and section views.

Besides simple geometrical positioning of the component, assembly modeling in the future will require more freedom of modeling independent of the various components. Exploiting the association between components permits changes to be made in an assembly even late in the design phase, where normally, because of cost and time, these modifications must be left out. Because the amount of data input for assemblies and the processing time required to view an assembly on the screen are often problematic, many systems allow a simplified representation of components in an assembly, where suppression of displaying details is possible. But often the characteristics of simplified representation are to be defined by the user. It is also possible for the CAD system to determine the precision with which the component is displayed, depending on the model section considered. Zooming in on the desired location then provides a representation of the details.

A representation of the assembly structure is helpful for the modeling of the assembly. Besides the improved overview, changes to the assembly can be carried out within the structure tree and parts lists are also easily prepared.

The combination of assemblies as a digital mock-up enables control over the entire structure during the design process. Thereby, interference checks and other assembly inspections are carried out using features that make up the so-called packaging analyses of the software. For the check of collision, voxel-based representations are suitable as well. Importing and portraying assemblies from other CAD systems is also advantageous. This is possible with the creation of adequate interfaces.

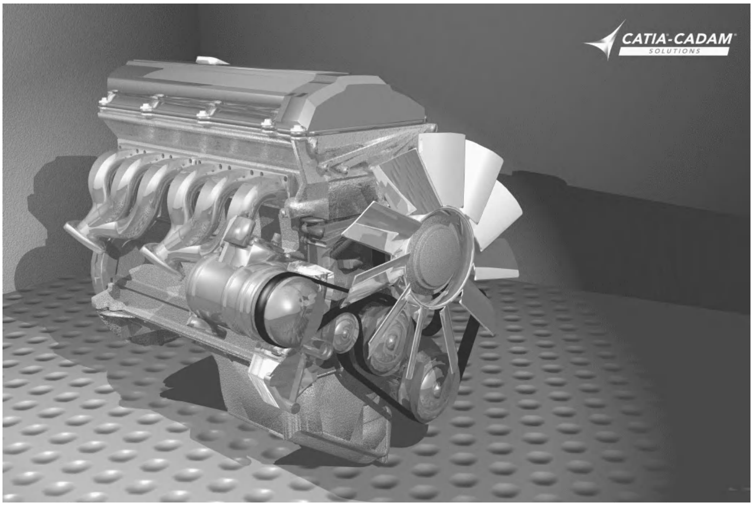
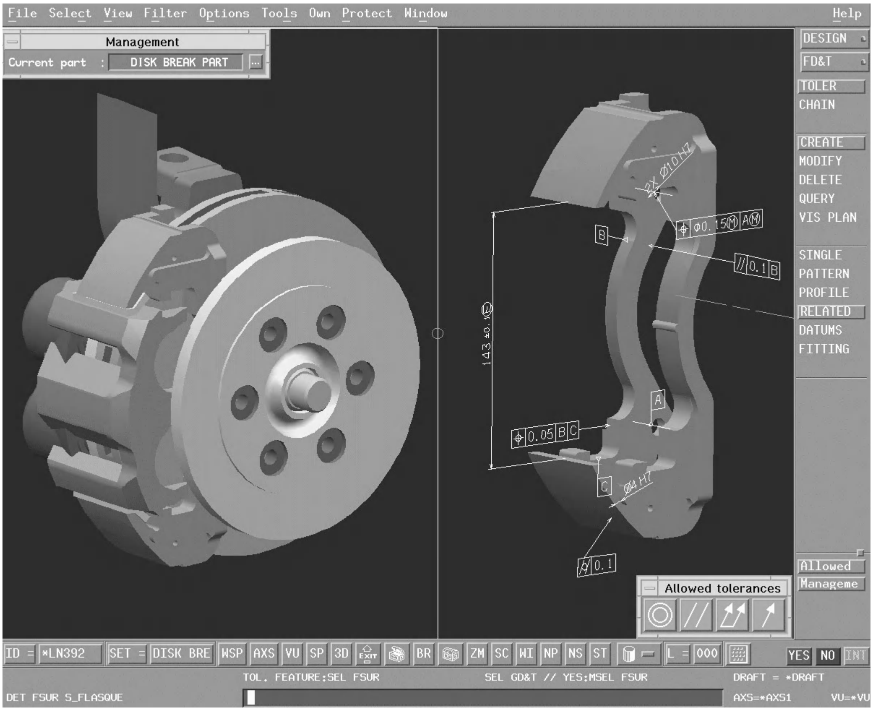

CATIA-CADAM
SOLUTIONS

Figure 4 An Engine as a Complex Assembly. (CATIA System, Dassault Systèmes)

An important functionality in assembly modeling is the definition of geometric and functional tolerances that are supported by the respective CAD module (Figure 5). Thereby, the geometrical and functional tolerance definitions are carried out based on assembly specifications and international standards. Special analysis functions are utilized for the control of complex assemblies and products.


File Select View Filter Options Tools Own Protect Window
Help
Management
DESIGN
FD&T
TOLER
CHAIN
CREATE
MODIFY
DELETE
QUERY
VIS PLAN
SINGLE
PATTERN
PROFILE
RELATED
DATUMS
FITTING
2X Ø10 H7
Φ0.15Ⓜ AⓂ
B
//0.1 B
143 ±0.1
A
Φ0.05 B C
C
Ø4 H7
0.1
Allowed
Manageme
Allowed tolerances
WSP AXS VU SP 3D EXIT BR ZM SC WI NP NS ST
YES NO INT

Current part : DISK BREAK PART

ID = *LN392 SET = DISK BRE L = 000

TOL. FEATURE:SEL FSUR SEL GD&T // YES:MSEL FSUR DRAFT = *DRAFT

DET FSUR S_FLASQUE AXS=*AXS1 VU=*VU

Figure 5 Definition of Tolerances in an Assembly. (CATIA System, Dassault Systèmes)

For example, the fly-through function allows the user to navigate in real time through complex assemblies. This enables visual inspections, which among other things, help in the recognition of component interference.

The design of complex assemblies requires sharing of the design processes through teamwork. This shared group work is supported by software that enables conferences over the Internet/intranet in which text, audio, and video information may be exchanged.

2.1.6. Further Technologies

Further applications in the process chain of product development can be carried out based on the 3D models created in early phases. For example, most commercial CAD systems are composed of modules and are implemented for the realization of a geometric model and other development-oriented solutions. Because the CAD systems are compatible with one another, an unhindered exchange of data is possible right from the beginning of the design through production stages. The systems must meet the requirements for continuity and openness, which are indispensable for the development of complex products. Continuity means that the product data are entered only once. They are saved in a central location and are available for use in carrying out other tasks without extensive conversion being required between the various forms of data. Openness means that, besides the ability to process data on various hardware platforms, it is possible to connect with other program systems within the same company and to external sites such as suppliers and other service providers. Further differentiation results from homogeneous and heterogeneous system environments. Using a common database for the application modules, for example in STEP format, means that conversion or interface problems between the individual applications are avoided. External application programs and data exchange between the purchaser and the supplier are simplified by the use of data exchange formats such as IGES, SET, and VDAFS. A uniform user interface improves the usability and comfort of the user.

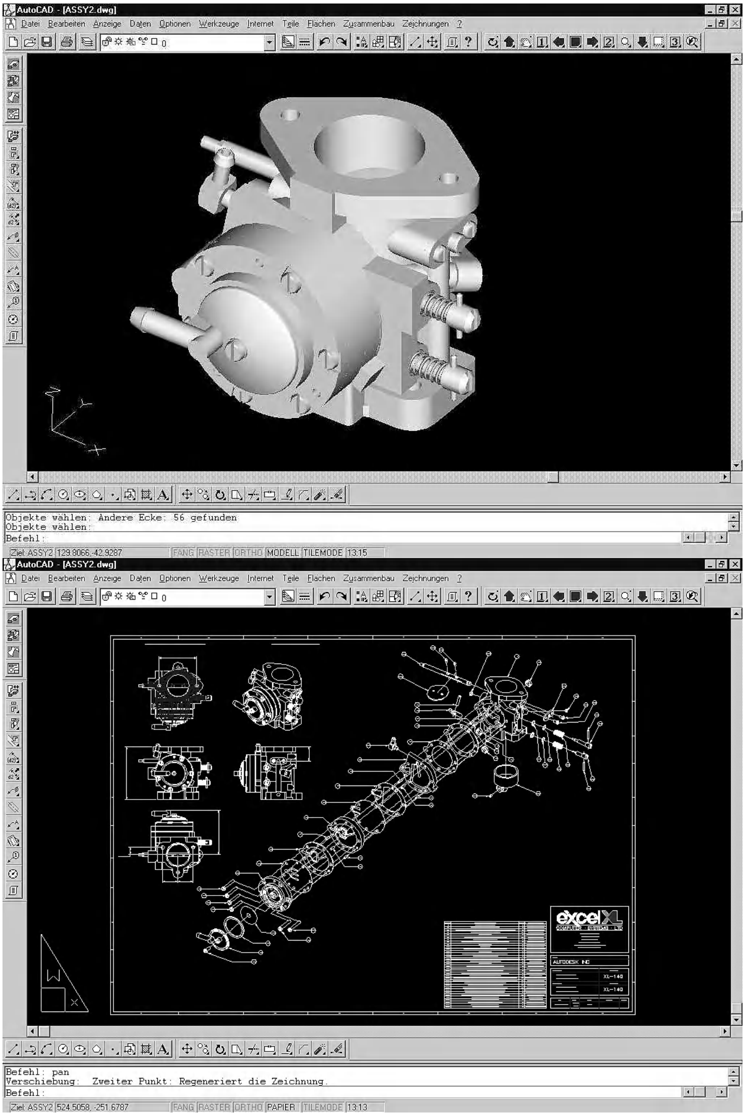
The application modules can be related to the various tasks within the development process. In the various phases of the virtual product development, different modules come into play depending on the task involved. The following principal tasks are presented:

- *Drawing preparation:* Drawings can be derived from the solid model and displayed in various views and measurements (Figure 6). This includes the representation of details as well as component arrangements. For higher work efficiency, it should be possible to switch between the 2D drawing mode and the 3D modeling mode at any time without suffering from conversion time and problems. A further simplification of the work results if the drawing and measurement routines provide and support the drawings in standardized formats, for example.
- *Parts list generator:* The preparation of a parts list is important for product development, procurement, and manufacturing. This list is automatically generated from the CAD component layout drawing. The parts list can then be exchanged through interfaces with PPS systems.
- *Analysis and simulation applications:* Analysis and simulation components enable the characteristics of a product to be attained and optimized earlier in the development phase. The costly and time-consuming prototype production and product development are thereby reduced to a minimum.

The strength characteristics can be calculated with finite element methods (FEMs). Besides stress–strain analyses, thermodynamic and fluid dynamic tests can be carried out. Preprocessors enable automated net generation based on the initial geometric model. Most CAD systems have interfaces for common FEM programs (e.g., NASTRAN, ANSYS) or have their own equation solver for FEM tests. Based on the substantial data resulting from an FEM test, it is necessary for postprocessing of results. The results are then portrayed in deformation plots, color-coded presentations of the stress–strain process, or animations of the deformation.

Material constants, such as density, can be defined for the components of the solid model. Then other characteristics, such as volume, mass, coordinates of the center of gravity, and moment of inertia can be determined. These data are essential for a dynamic simulation of the product. For kinematic and dynamic simulations, inelastic solid elements with corresponding dynamic characteristics are derived from the solid model (Figure 7). Various inelastic solid elements can be coupled together from a library. The complete system is built up of other elements such as a spring and damper and then undergoes a loading of external forces. The calculation of the system dynamics is usually carried out using a commercial dynamics package. A postprocessor prepares a presentation of the results for display on a screen. An animation of the system can then be viewed.

- *Piping:* Piping can be laid within the 3D design using conduit application modules, as shown in Figure 8. Design parameters such as length, angle, and radius can be controlled. Intersections or overlapping of the conduit with adjacent assemblies can also be determined and repaired. The design process is supported through the use of libraries containing standard piping and accessories. The parts lists are automatically derived from the piping design.


AutoCAD - [ASSY2.dwg]
Datei Bearbeiten Anzeige Daten Optionen Werkzeuge Internet Teile Flächen Zusammenbau Zeichnungen ?
AutoCAD - [ASSY2.dwg]
Datei Bearbeiten Anzeige Daten Optionen Werkzeuge Internet Teile Flächen Zusammenbau Zeichnungen ?
excelXL
AUTODESK INC
XL-140
XL-140

```
Objekte wählen: Andere Ecke: 56 gefunden
Objekte wählen:
Befehl:
```

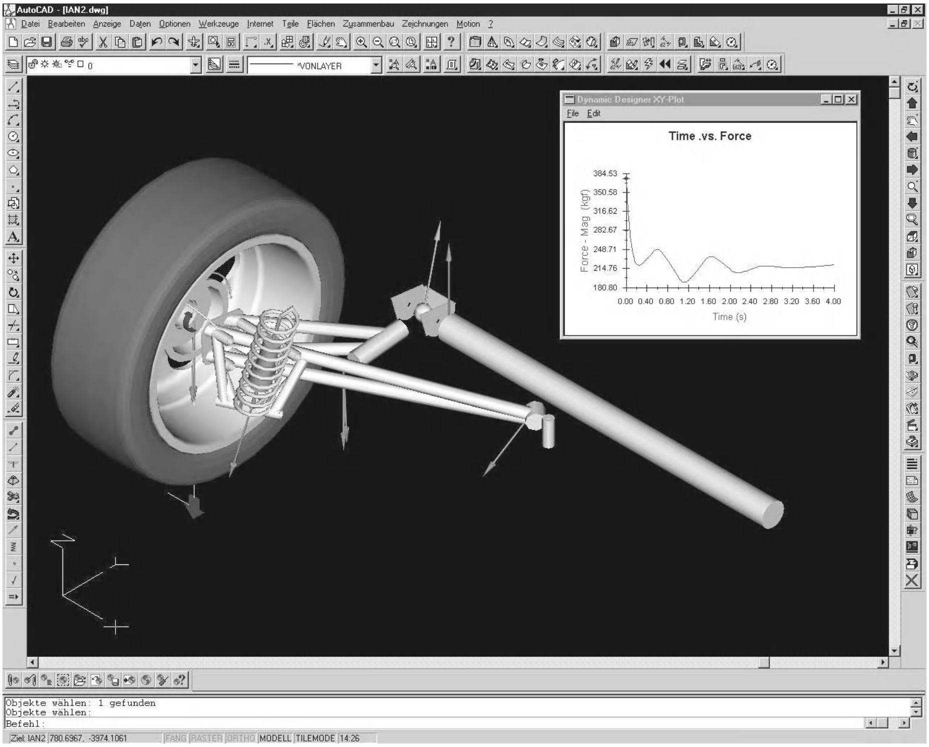
```
Ziel: ASSY2 129.8066,-42.9287  FANG RASTER ORTHO MODELL TILEMODE 13:15
```

```
Befehl: pan
Verschiebung:  Zweiter Punkt: Regeneriert die Zeichnung.
Befehl:
```

```
Ziel: ASSY2 524.5058, -251.6787  FANG RASTER ORTHO PAPIER TILEMODE 13:13
```

Figure 6 3D Model and the Drawing Derived from the Model. (AutoCAD System. Reproduced with permission of Autodesk®)

- *Weld design:* Weld design can also be supported by using appropriate application modules (Figure 9). Together, designers and production engineers can determine the required joining techniques and conditions for the assemblies. Thereby, the weld type and other weld parameters such as weld spacing and electrodes are chosen for the material characteristics of the components to be joined. Process information such as weld length and cost and time requirements can be derived from the weld model.
- *Integration of ECAD design:* The design of printed circuit boards and cable layouts are typical application examples of 2D designs that are related to the surrounding 3D components. With the design of printed circuit boards, layout information, such as board size and usable and unusable area, can be exchanged and efficiently stored between and within the MCAD and ECAD systems. For the design of cable layouts (Figure 10), methods similar to those used for conduit design are implemented. This simplifies the placement of electrical conductors within assemblies.


AutoCAD - [IAN2.dwg]
Datei Bearbeiten Anzeige Daten Optionen Werkzeuge Internet Teile Flächen Zusammenbau Zeichnungen Motion ?
0
VONLAYER
Dynamic Designer XY-Plot
File Edit
Time .vs. Force
384.53
350.58
316.62
282.67
248.71
214.76
180.80
Force - Mag (kgf)
0.00 0.40 0.80 1.20 1.60 2.00 2.40 2.80 3.20 3.60 4.00
Time (s)
Ziel: IAN2
780.6967, -3974.1061
FANG RASTER ORTHO MODELL TILEMODE 14:26

```
Objekte wählen: 1 gefunden
Objekte wählen:
Befehl:
```

Figure 7 Kinematic Simulation of a 3D Model. (System AutoCAD. Reproduced with permission of Autodesk®)

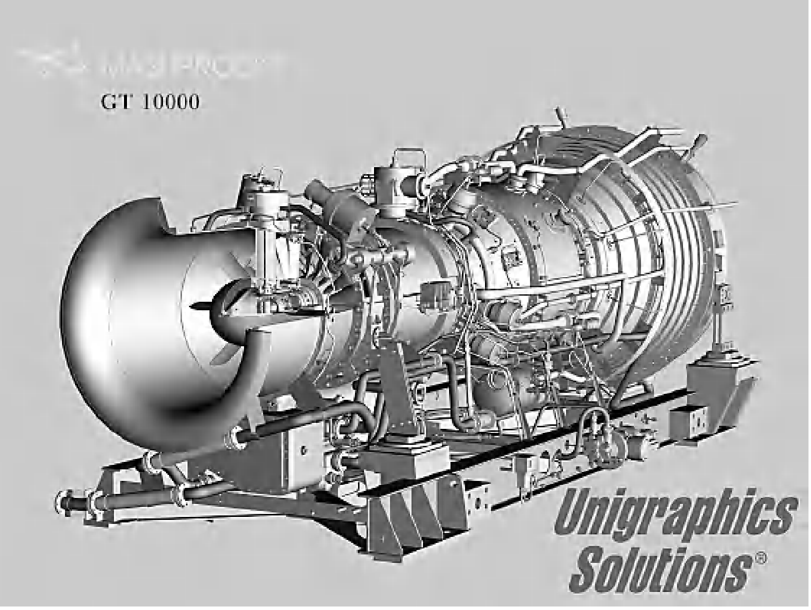

GT 10000
Unigraphics Solutions®

Figure 8 Pipe Laying in a 3D Assembly. (Unigraphics. Reproduced with permission of Unigraphics Solutions)

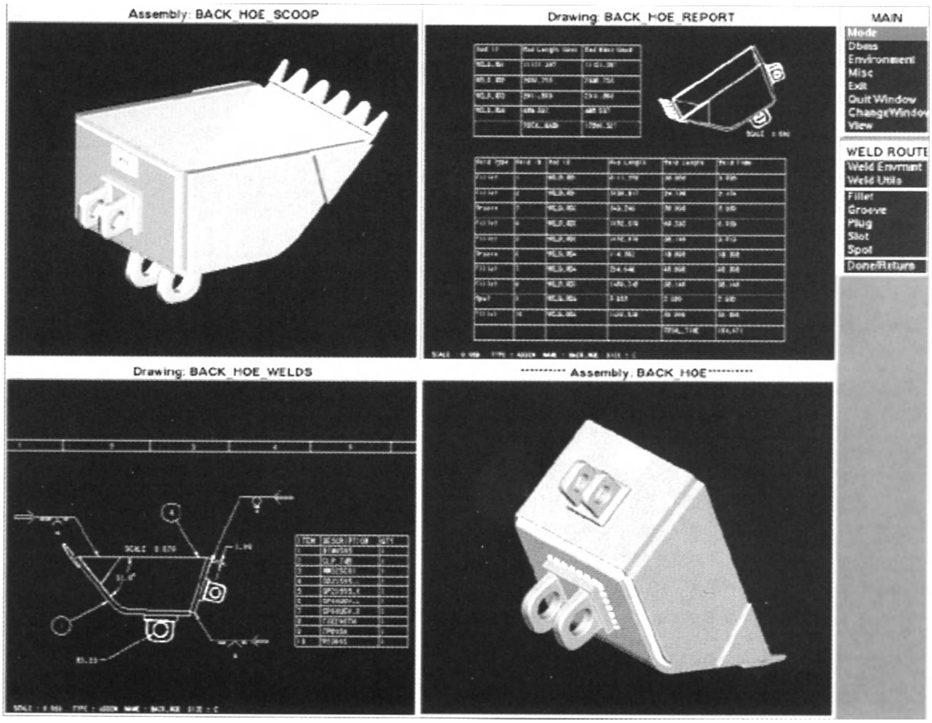

Assembly: BACK_HOE_SCOOP
Drawing: BACK_HOE_REPORT
MAIN
Mode
Dbms
Environment
Misc
Exit
Quit Window
ChangeWindo
View
WELD ROUTE
Weld Envrmnt
Weld Utils
Fillet
Groove
Plug
Slot
Spot
Done/Return
Drawing: BACK_HOE_WELDS
********** Assembly: BACK_HOE**********

Figure 9 Example of a Weld Design. (System Pro/WELDING, PTC) Reprinted by permission of PTC.

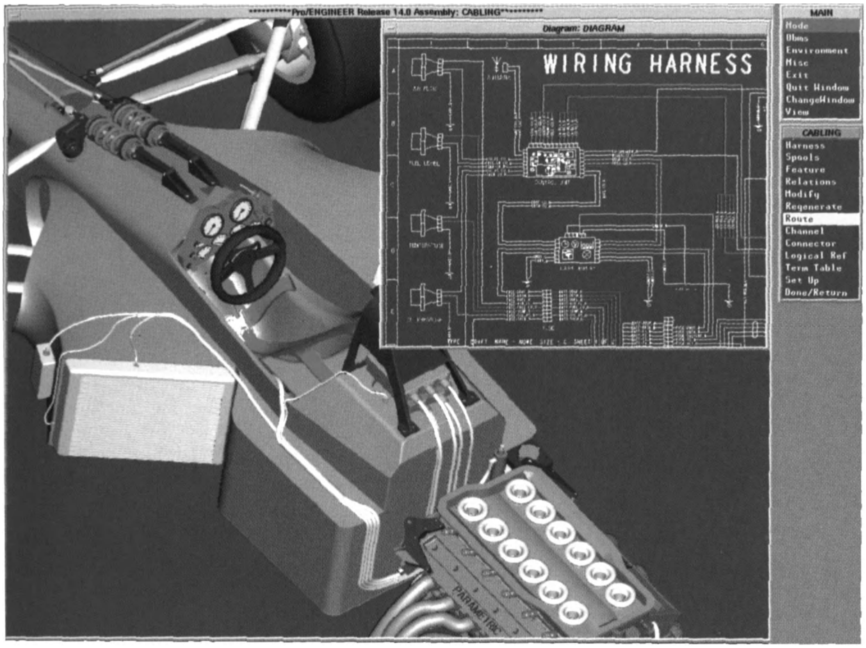

**********Pro/ENGINEER Release 14.0 Assembly: CABLING**********
Diagram: DIAGRAM
WIRING HARNESS
MAIN
Mode
Dbms
Environment
Misc
Exit
Quit Window
ChangeWindow
View
CABLING
Harness
Spools
Feature
Relations
Modify
Regenerate
Route
Channel
Connector
Logical Ref
Term Table
Set Up
Done/Return
PARAMETRIC

Figure 10 2D Design of a Wiring Harness. (Pro/DIAGRAM System, PTC) and Parametric Cable Layout in the 3D Assembly. (Pro/CABLING, PTC) Reprinted by permission of PTC.

- *Integration of process planning:* Product data can be taken directly from production planning. The machine tool design and planning can take place parallel to the design of the product. NC programs and control data can be derived from the finished 3D description of the geometry, although the model requires supplemental information for mountings and other devices. The tool paths are determined under consideration of the additional technical information of the different processes, such as drilling, milling, turning, eroding, and cutting. The generation of NC data in diverse formats, such as example COMPACT II, CLDATA, and APT, is carried out using a postprocessor. The NC programs derived from an optimized design can be checked by simulation.

  In addition to the three- and five-axis milling machines used for conventional production tasks, rapid prototyping processes are applied for quick design verification. For example, the process of stereo lithography can be applied. Rapid prototyping reduces the period of development and allows for the quick generation of physical prototypes for product review and evaluation purposes. A variety of materials, from plastics to metals, are used in rapid prototyping, enabling the prototypes to be presented in the most frequently used materials so that material characteristics can be easily reviewed (Gebhardt and Pflug 1995). The control data for the laser stereo lithography are generated from the geometry data.

  When molds for injection molding are made, die geometry is taken into consideration because of the possibility of material shrinkage. Rheological calculations are carried out using preprocessors (e.g., Moldflow, Cadmould 3D). The set-up of the tools takes place using catalogs provided by the machine manufacturer. Subsequently, NC data are generated for the part shape and machining plates.

  With complete product models, simulation and robot programming can take place for the various manufacturing processes. The programming and simulation take place offline. The programs generated are then sent on to the robots for execution of the process.

Only through comprehensive integration of the complete virtual product development process, with the many different working steps, can the existing potential for development time, costs, and overall productivity be optimized (VDI-EKV 1992, Krause 1992).

2.2. CAD Interfaces

2.2.1. General Explanations

Despite the lack of a clear definition for the term *interface,* its use has become quite frequent. In general, an interface can be defined as a link forming a common boundary between integrating systems. It is a system of conditions, rules, and agreements that defines the terms of information exchange between communicating systems or system components (Anderl 1993).

CAD interfaces interconnect internal CAD system components and provide a link to other software and hardware components. They connect internal CAD system components such as geometrical modeling, object representation, and realization of calculation operations to a CAD database and standardized CAD files.

CAD interfaces can be classed as device interfaces (hardware interfaces) and software interfaces (Grabowski and Anderl 1990) or, alternatively, as internal and external interfaces. Internal interfaces provide links within the application environment, enabling the exchange of information between other system programs or ensuring undisturbed communication with the user. External interfaces require a common or uniform representation of the information to be exchanged. They transmit product data, by means of graphical peripheries, to the user. The standardization of external interfaces is of special importance for the transmission of product data. Because of the exchange of data between various CAD systems, uniform data formats are required.

In Figure 11, CAD components and interfaces are displayed.

From an application in use, access to the product data and methods of a CAD system is provided through application interfaces. In the course of using a CAD system through a standardized interface, the integration of such an external application interface becomes increasingly important. Standardized application interfaces allow access to the methods employed by the various CAD systems.

Product data represent all data such as geometry, topology, technological information and organizational data to define the product. This data is generated during the design process, enhanced by the application and then internally mapped.

2.2.2. Standardization of CAD Interfaces

The primary objectives for the standardization of interfaces are:

- The unification of interfaces
- The prevention of dependencies between system producer and user
- The prevention of repeated working of redundant data

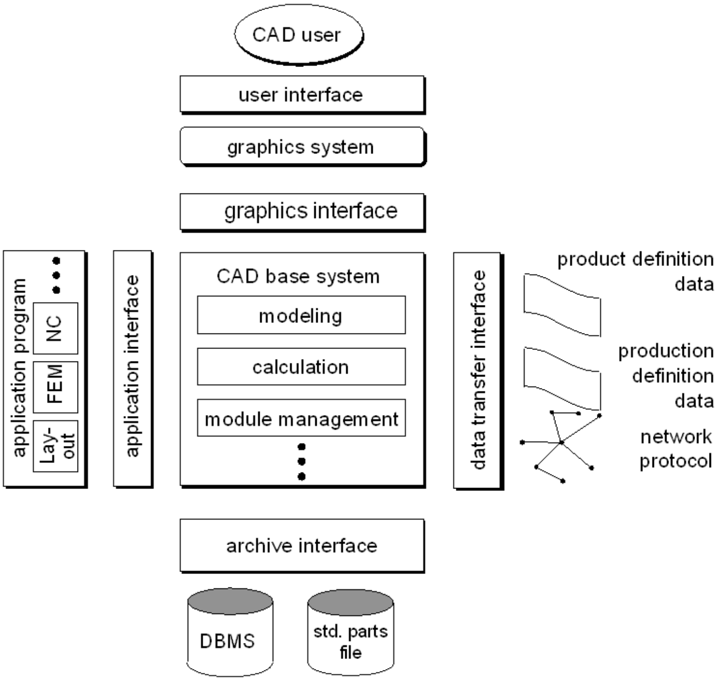

CAD user
user interface
graphics system
graphics interface
application program
Lay-out
FEM
NC
application interface
CAD base system
modeling
calculation
module management
data transfer interface
product definition data
production definition data
network protocol
archive interface
DBMS
std. parts file

Figure 11 CAD Components and Interfaces.

The standardization of interfaces leads to the following advantages (Stanek 1989):

- The possibility of an appropriate combination of system components
- Selection between various systems
- Exchange of individual components
- Expandability to include new components
- Coupling of various systems and applications
- Greater freedom in the combining of hard and software products

2.2.2.1. Interfaces for Product Data Exchange The description of product data over the entire product life cycle can be effectively improved through the use of computer-supported systems. For this reason, uniform, system-neutral data models for the exchange and archiving of product data are desired. These enable the exchange of data within CAD systems and other computer-aided applications.

The internal computer models must be converted into one another through so-called pre- and postprocessor conversion programs. To minimize the costs of implementing such processors, a standardized conversion model is applied. The conversion models standardized to date are still of limited use because they can only display extracts of an integrated product model. Interfaces such as

- initial graphics exchange specification (IGES)
- standard d'echange et de transfert (SET)
- Verband der Automobilindustrie-FlächenSchnittstelle (VDA-FS)

have been designed and to some extent nationally standardized for geometry data exchange, mainly in the area of mechanical design (Anderl 1989).

IGES is recognized as the first standardized format for product-defined data to be applied industrially (Anderl 1993; Grabowski and Anderl 1990; Rainer 1992). The main focus is the transfer of design data. IGES is for the mapping of:

- 2D line models
- 3D wire models
- 3D dimensional surface models
- 3D solid models
- Presentation models for technical drawings

Considering the further demands on CAD/CAM systems, the integration of additional data in IGES format has been realized. Examples are the data for FEM, factory planning, and electronic/electrical applications.

The interface standard VDA-FS was developed by the German Automotive Industry Association for the transfer of CAD free-form surface data. VDA-FS is standardized and described in Standard DIN 66301 and has proven efficient for use in many areas, but its efficiency has been demonstrated principally in the German automotive industry (Grabowski and Glatz 1986; Nowacki 1987; Scheder 1991).

The STEP product model (Standard for the Exchange of Product Model Data) offers the only standardized possibility for efficiently forming the product data exchange and system integration in the CAx world (Wagner and Bahe 1994).

With the development of ISO 10303 (Product Data Representation and Exchange), also called STEP, the objective is to standardize a worldwide accepted reference model for the transfer, storage and archiving, and processing of all data needed for the entire product life cycle (Anderl 1993; Düring and Dupont 1993; Krause et al. 1994; McKay et al. 1994).

STEP can be seen as a toolbox for describing application-oriented product information models using basic elements, so-called integrated resources, while considering previously defined rules and standardized methods (Holland and Machner 1995).

ISO 10303 can be subdivided into:

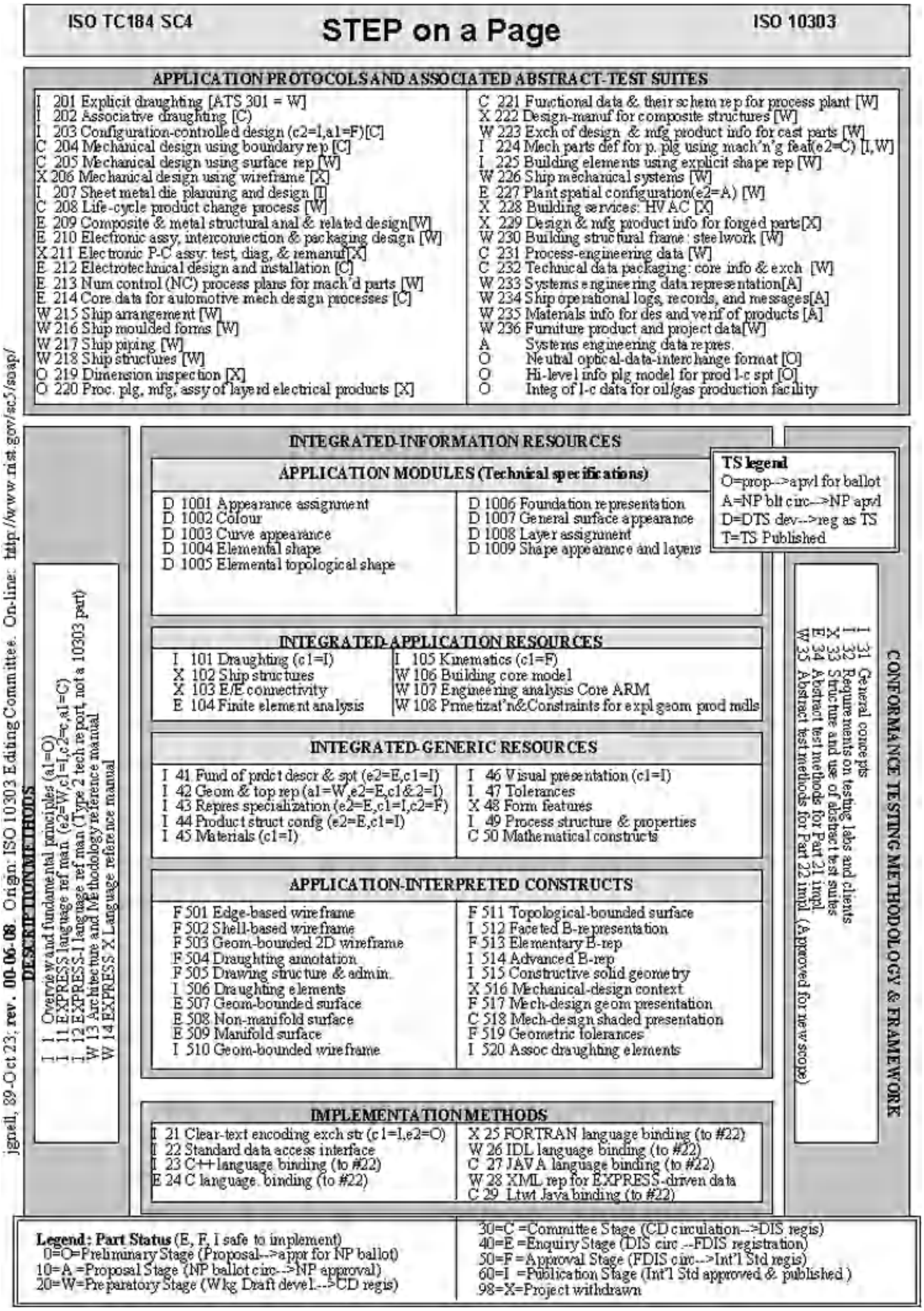
- Generic and application-oriented information models
- Application protocols
- Methods for specification and implementation
- Concepts for examination and testing

Independent application specifications are referred to as generic models. An overview of the series is as follows (Figure 12):

- Series 0: fundamentals, principles
- Series 10: description or specification methods
- Series 20: implementation methods
- Series 30: test methods and criterion
- Series 40: application-independent base models
- Series 100: application-dependent base models
- Series 200: application protocols
- Series 300: abstract test methods
- Series 500: application-specific, interpreted design

The core of consists of information models in which the representational form of the product data is defined. Information models can be broken down into three categories: generic resources, application-related resources, and application protocols. Similar to a toolbox system, the generic resource models define application-independent base elements that may be used in application-specific resources or directly in application protocols. An example is geometrical information, which is required in most application models. Application-related resources are especially tailored to the demands of specific branches but also serve for the definition of application protocols. Application protocols form implementable parts for STEP. They define a section of the information structure that is needed to support a specified application.

The description methods used for defining the information structure, validation, and test methods as well as implementation methods are standardized. The object-oriented modeling language EXPRESS as well as its graphical form EXPRESS-G are implemented for the presentation of the information structures (Scholz-Reiter 1991). Validation and test methods define the specifications with which the STEP processors are examined. The following elements are included in partial models of the generic resources:



ISO TC184 SC4 **STEP on a Page** ISO 10303

APPLICATION PROTOCOLS AND ASSOCIATED ABSTRACT-TEST SUITES

I 201 Explicit draughting [ATS 301 = W]
I 202 Associative draughting [C)
I 203 Configuration-controlled design (c2=I,a1=F)[C]
C 204 Mechanical design using boundary rep [C]
C 205 Mechanical design using surface rep [W]
X 206 Mechanical design using wireframe [X]
I 207 Sheet metal die planning and design [I]
C 208 Life-cycle product change process [W]
E 209 Composite & metal structural anal & related design[W]
E 210 Electronic assy, interconnection & packaging design [W]
X 211 Electronic P-C assy: test, diag, & remanuf[X]
E 212 Electrotechnical design and installation [C]
E 213 Num control (NC) process plans for mach'd parts [W]
E 214 Core data for automotive mech design processes [C]
W 215 Ship arrangement [W]
W 216 Ship moulded forms [W]
W 217 Ship piping [W]
W 218 Ship structures [W]
O 219 Dimension inspection [X]
O 220 Proc. plg, mfg, assy of layerd electrical products [X]
C 221 Functional data & their schem rep for process plant [W]
X 222 Design-manuf for composite structures [W]
W 223 Exch of design & mfg product info for cast parts [W]
I 224 Mech parts def for p. plg using mach'n'g feat(e2=C) [I,W]
I 225 Building elements using explicit shape rep [W]
W 226 Ship mechanical systems [W]
E 227 Plant spatial configuration(e2=A) [W]
X 228 Building services: HVAC [X]
X 229 Design & mfg product info for forged parts[X]
W 230 Building structural frame: steelwork [W]
C 231 Process-engineering data [W]
C 232 Technical data packaging: core info & exch [W]
W 233 Systems engineering data representation[A]
W 234 Ship operational logs, records, and messages[A]
W 235 Materials info for des and verif of products [A]
W 236 Furniture product and project data[W]
A Systems engineering data repres.
O Neutral optical-data-interchange format [O]
O Hi-level info plg model for prod l-c spt [O]
O Integ of l-c data for oil/gas production facility

jgnell, 89-Oct 23; **rev. 00-06-08**. Origin: ISO 10303 Editing Committee. On-line: http://www.nist.gov/sc5/soap/

DESCRIPTION METHODS

I 1 Overview and fundamental principles (a1=O)
I 11 EXPRESS language ref man (e2=W,c1=I,c2=c,a1=C)
I 12 EXPRESS-I language ref man (Type 2 tech report, not a 10303 part)
W 13 Architecture and Methodology reference manual
W 14 EXPRESS X Language reference manual

INTEGRATED-INFORMATION RESOURCES

APPLICATION MODULES (Technical specifications)

D 1001 Appearance assignment
D 1002 Colour
D 1003 Curve appearance
D 1004 Elemental shape
D 1005 Elemental topological shape
D 1006 Foundation representation
D 1007 General surface appearance
D 1008 Layer assignment
D 1009 Shape appearance and layers

TS legend
O=prop-->apvl for ballot
A=NP blt circ-->NP apvl
D=DTS dev-->reg as TS
T=TS Published

INTEGRATED-APPLICATION RESOURCES

I 101 Draughting (c1=I)
X 102 Ship structures
X 103 E/E connectivity
E 104 Finite element analysis
I 105 Kinematics (c1=F)
W 106 Building core model
W 107 Engineering analysis Core ARM
W 108 Prmetizat'n&Constraints for expl geom prod mdls

INTEGRATED-GENERIC RESOURCES

I 41 Fund of prdct descr & spt (e2=E,c1=I)
I 42 Geom & top rep (a1=W,e2=E,c1&2=I)
I 43 Repres specialization (e2=E,c1=I,c2=F)
I 44 Product struct confg (e2=E,c1=I)
I 45 Materials (c1=I)
I 46 Visual presentation (c1=I)
I 47 Tolerances
X 48 Form features
I 49 Process structure & properties
C 50 Mathematical constructs

APPLICATION-INTERPRETED CONSTRUCTS

F 501 Edge-based wireframe
F 502 Shell-based wireframe
F 503 Geom-bounded 2D wireframe
F 504 Draughting annotation
F 505 Drawing structure & admin.
I 506 Draughting elements
E 507 Geom-bounded surface
E 508 Non-manifold surface
E 509 Manifold surface
I 510 Geom-bounded wireframe
F 511 Topological-bounded surface
I 512 Faceted B-representation
F 513 Elementary B-rep
I 514 Advanced B-rep
I 515 Constructive solid geometry
X 516 Mechanical-design context
F 517 Mech-design geom presentation
C 518 Mech-design shaded presentation
F 519 Geometric tolerances
I 520 Assoc draughting elements

IMPLEMENTATION METHODS

I 21 Clear-text encoding exch str (c1=I,e2=O)
I 22 Standard data access interface
I 23 C++ language binding (to #22)
E 24 C language. binding (to #22)
X 25 FORTRAN language binding (to #22)
W 26 IDL language binding (to #22)
C 27 JAVA language binding (to #22)
W 28 XML rep for EXPRESS-driven data
C 29 Ltwt Java binding (to #22)

CONFORMANCE TESTING METHODOLOGY & FRAMEWORK

I 31 General concepts
I 32 Requirements on testing labs and clients
X 33 Structure and use of abstract test suites
E 34 Abstract test methods for Part 21 impl.
W 35 Abstract test methods for Part 22 impl. (Approved for new scope)

Legend: Part Status (E, F, I safe to implement)
0=O=Preliminary Stage (Proposal-->appr for NP ballot)
10=A =Proposal Stage (NP ballot circ-->NP approval)
20=W=Preparatory Stage (Wkg Draft devel.-->CD regis)
30=C =Committee Stage (CD circulation-->DIS regis)
40=E =Enquiry Stage (DIS circ --FDIS registration)
50=F =Approval Stage (FDIS circ-->Int'l Std regis)
60=I =Publication Stage (Int'l Std approved & published)
98=X=Project withdrawn

Figure 12 Overview of the ISO 10303 Structure.

- Base model
- Geometric and topological model
- Tolerance model
- Material model
- Product structure model
- Representation model and
- Process model

The significance of STEP goes much further than the exchange and archiving of product models. The innovative character of STEP development sets new standards for CAD systems. The functional requirements for new systems stem from partial models such as tolerance, form/feature, or PSCM models. In certain circumstances, STEP may be considered a reference model for the internal computer representation of CAD systems.

2.3. Engineering Data Management

Engineering data management (EDM, also an abbreviation for electronic document management or enterprise data management) is data modeling of a complete enterprise. In other words, the data modeling goes beyond the intrinsic product data. A common synonym for EDM is ''product data management'' (PDM), in which emphasis is put on the handling of product-related engineering data. *Engineering data management* was coined as a broader term. All definitions combined provide the perception that EDM is the management and support of all information within an enterprise at any point in time (Ploenzke 1997).

For the support of data modeling within a company, software systems, so-called EDM systems, exist. With these systems, various tasks are carried out depending on the product, company, and level of integration, such as the management of (Kiesewetter 1997):

- Drawing data
- CAD model data
- Parts lists
- Standard libraries
- Project data
- NC data
- Software
- Manufacturing plans
- Tools and equipment production facilities

In addition, EDM systems serve for classification and object parameter management. Generally, EDM systems provide a set of specific functions for the modeling of product and process data. Because of their specific origins, the various EDM systems focus on a variety of main functions and strengths, but all provide a basic functionality for document management as well as functions for the support of change, version, and release management. The complete functionality of an EDM system as a central information system for the product development process is made up of a broad spectrum of functions. These functions can be broken down into application-related and system-overlapping functions (Ploenzke 1997).

- *Application-related functions:* The main priority for an EDM system is the management of all product data and documentation. Application-related functions provide task-specific support of the data management. Classical data management functions such as additions, modifications, and deletions are extended to integrate additional capabilities. For example, in the case of drawing management, automatic extraction of metadata from the CAD drawing header is implemented in the EDM database or used for classification of the components.
- *System-overlapping functions:* Data and document management requires an application-neutral infrastructure that provides functions for an organized handling of the management processes. These overlapping system and application functions support the data and document management through functions created for variant and version management and ensuring the editing status of the document. Also, these functions support the provision of central services such as user management, privacy, and data protection.

Support provided by the use of an EDM system is aimed at the integration of the information flow and processes into the business processes. Integration and transparency are the essential aspects for the EDM system to process every demand and provide the right information at the right time. The storage, management, and provision of all product-related data create the basis for:

- The integration of application systems for technical and commercial processes as well as for office systems in a common database
- The task-oriented supply of all operations with actual and consistent data and documentation
- The control and optimization of business processes

Metadata are required for the administration of product data and documentation, enabling the identification and localization of product data. Metadata represent the information about the creator, the data of generation, the release status, and the repository. The link between all relevant data is made by the workflow object, which is a document structure in which documents of various formats are filled out along a process chain. This structure is equivalent to part of the product structure and, after release, is integrated into the product structure. In general, this is the last step of a workflow.

The modeling and control of a process chain are the task of the workflow management, whereas access control is the task of document management. The users concerned in the workflow are managed by groups, roles and rights. The systems applied are integrated in the EDM system. Linking a special document with the editing system is the responsibility of the document management program.

2.4. Architecture and Components

The reference architecture, from Ploenzke (1997) (Figure 13) describes the individual components of an EDM system from a functional point of view and illustrates the connections between the respective system components.

The core of the reference architecture is the engineering data. The application-oriented and system-overlapping functions are applied to the core data. The engineering data are found within the model data and broken down into product-defined data and metadata. Model data, such as drawing data, parts lists, text files, raster data from document, and data in other formats, are not interpreted by the EDM system but are still managed by the EDM System. Metadata is interpreted by the EDM system and contains information regarding the management and organization of the product data and documents.

The application-oriented functions support the management of the model data. Although the model data may be made up of various kinds of data, the management of these data, with the help of application oriented functions, must be able to summarize the data logically in maps and folders. Also, relationships between the documents and product data can then be determined. Important application-oriented functions are (Ploenzke 1997):

- General data management
- Drawing data management

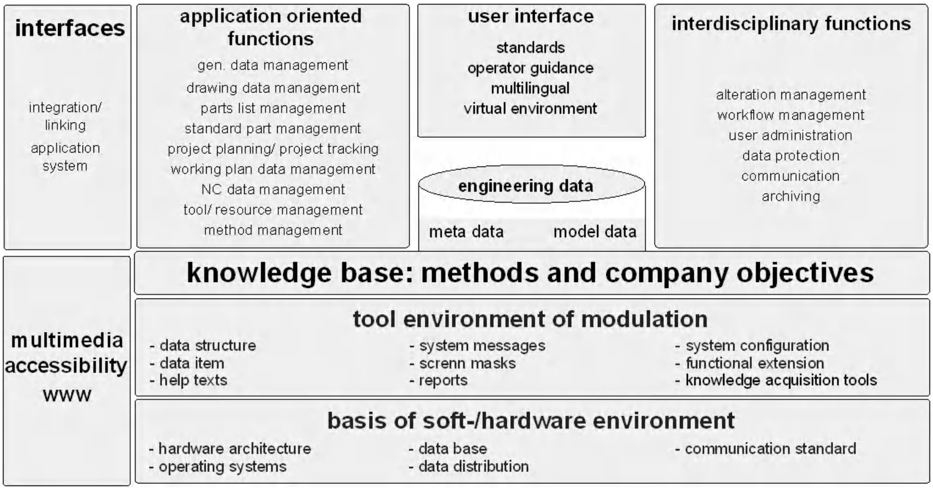

interfaces
integration/ linking
application system
application oriented functions
gen. data management
drawing data management
parts list management
standard part management
project planning/ project tracking
working plan data management
NC data management
tool/ resource management
method management
user interface
standards
operator guidance
multilingual
virtual environment
engineering data
meta data
model data
interdisciplinary functions
alteration management
workflow management
user administration
data protection
communication
archiving
multimedia accessibility
www
knowledge base: methods and company objectives
tool environment of modulation
- data structure
- data item
- help texts
- system messages
- screnn masks
- reports
- system configuration
- functional extension
- knowledge acquisition tools
basis of soft-/hardware environment
- hardware architecture
- operating systems
- data base
- data distribution
- communication standard

Figure 13 Reference Architecture of an EDM System. (From Ploenzke 1994. Reprinted by permission of CSC Ploenzke.)

- Classification
- Parts list management
- Standard parts management
- Project planning and management
- Production plan management
- NC data management
- Machine tool and equipment management
- Method management

System-overlapping functions support the business processes over numerous steps, such as change and workflow management. Furthermore, central system services such as user management and data security are provided. Important system-overlapping functions are (Ploenzke 1997):

- Change management
- Workflow management
- User administration
- Data security
- Data protection
- Communication
- Archiving

Besides the core components of the reference architecture, the following system components belong to the system environment (Ploenzke 1997):

- User interface
- Machine environment
- Interfaces
- Basic software and hardware environment

The user interface forms the direct interface to the user and thus must provide the look and functions desired by the user. The application environment provides methods for customizing the interface. This allows the adaptation of the EDM system to company-specific conditions and for maintenance of the system. Programming interfaces known as application procedural interfaces (APIs) enable the launching and integration of application modules for the functional extension of an EDM system. The basic software and hardware environment forms the respective operation platform of an EDM system. With the status of the technology today, client–server-based EDM systems come into operation with connecting databases. These provide the various users with the necessary client programs, increasing the economical utilization of today's workstations (Krause et al. 1996).

EDM can also be termed an enabling technology. The reference architecture is an integration platform for systems, functions, and data. Company-specific conditions and a dynamic control process, however, cause the transformation of each EDM system installation into an almost nontransferable case. Nevertheless, the EDM reference architecture forms the basis of system integration for CAD integration, PPS coupling, or archiving, regardless of the company-specific focus.

The necessity for integration and transparency of the data leads to the broad scope of functionality enabled by an EDM system. The primary functions are combined in numerous modules. The level of performance of the individual modules varies from system to system and is determined by the company philosophy and strengths of the supplier (Figure 14).

Many functions that serve for the basic control of data in the EDM already stem from other systems such as CAD or PPS systems. These functions are combined in the EDM system and make feasible the creation and management of an information pool. This supports company-wide availability and enables faster information searches (Figure 15). The information pool, known as the vault, is a protected storage area that enables the availability of all documents and ensures that no unauthorized access occurs. Access to documents held in other systems is possible through logins to the various systems. After logging off, the document is available only in a read-only form. Therefore, logging off copies the document to local memory. Logging in and opening a changed document is associated with the saving of a new version of the document. All customary EDM systems on the market are able to manage a centralized or decentralized vault.

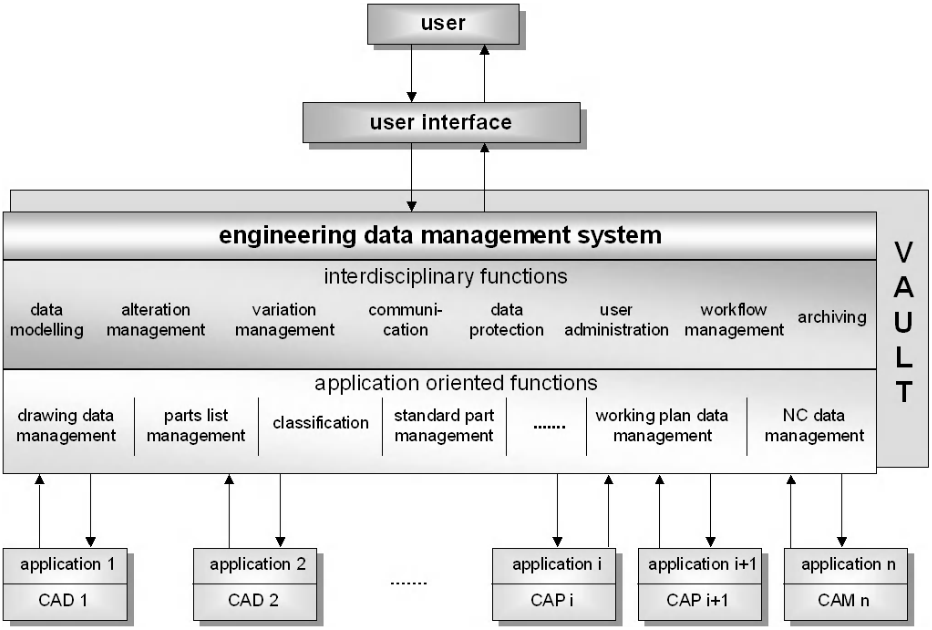

user
user interface
engineering data management system
interdisciplinary functions
data modelling
alteration management
variation management
communi-cation
data protection
user administration
workflow management
archiving
application oriented functions
drawing data management
parts list management
classification
standard part management
....
working plan data management
NC data management
VAULT
application 1
CAD 1
application 2
CAD 2
....
application i
CAP i
application i+1
CAP i+1
application n
CAM n

Figure 14 Architecture of an EDM System with Integrated Applications. (From Stephan 1997)

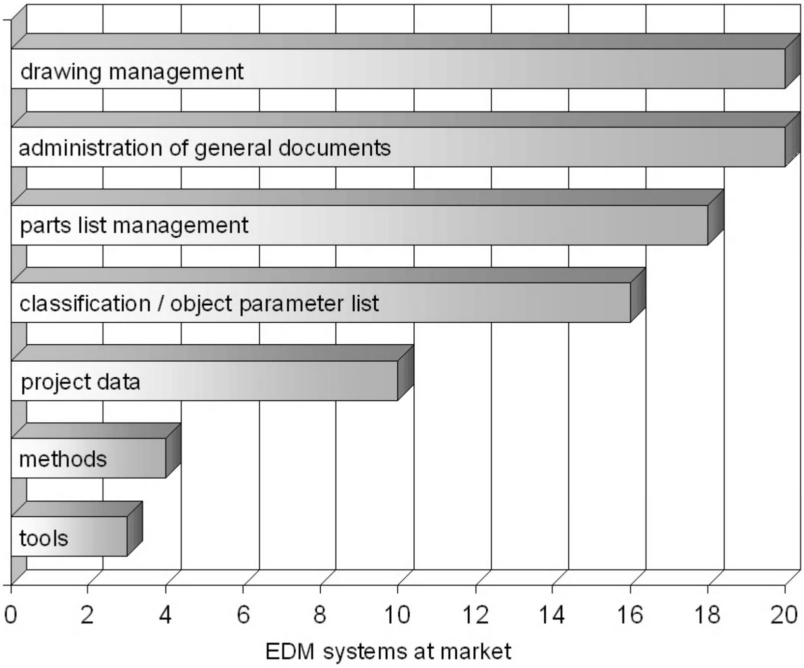


Figure 15 Availability of Function Modules in EDM Systems. (From Ploenzke 1997. Reprinted by permission of CSC Ploenzke.)

2.5. Calculation Methods

2.5.1. General Explanation

Increasing demands for product performance necessitates a secure component configuration. The factor of safety is determined by calculations performed on the system. The following goals should be achieved:

- Assurance against failure
- Testing of product functionality
- Assessment of external effects
- High strength-to-weight ratio
- Optimal material utilization
- Achievement of an economical production process

The most important computer-supported calculation processes are:

- Finite element methods (FEM)
- Boundary element methods (BEM)
- Finite different methods (FDM)

These methods can be applied for problems in which differential equations describe any continua. The methods have different approaches for performing calculations. FEM assumes a variational formulation, BEM a formulation by means of integrals, and FDM uses differential equations. Due to the mathematical formulation, it is irrelevant whether the computational problem comes from mechanics, acoustics, or fluid mechanics. In all three methods, the discretization of the structure is common and is required in order to derive the mathematical formulations for the desired tasks.

The processes commercially available have integrated interfaces that enable them to work with geometries already developed in CAD systems. To prepare the geometry model, various support mechanisms for the particular calculation systems are offered.

Beyond FEM, BEM, and FDM, there are other calculation systems that are based on problem-specific model creation. These systems, however, are usually applied only in conjunction with an associated model. Two examples are the calculation of suspensions and the determination and layout of weld contacts.

The general sequence of a calculation is equivalent to that of an information handling process. Input in the form of the geometry, material, and forces is transformed using mathematical and physical rules to calculated results. In the majority of cases, the object geometry and material properties are simplified. Also, the stress–strain or loading characteristics are often idealized, which eases the calculation task and reduces time. In many cases assumptions and simplifications are made because otherwise the calculation of the problem might be too extensive or impossible.

2.5.2. Finite Element Methods

FEM is the most commonly used calculation process today. Their implementation spectrum covers many different areas. For example, they are applied in:

- Statics in civil engineering
- Crash research in the automotive industry
- Electromagnetic field research for generator design
- Material strength and life cycle determination in mechanical engineering
- Bone deformation in biomechanics

FEM has found its place in the field of structural mechanics. It is used for calculations in:

- Stress and deformation
- Natural shape and eigenfrequency
- Stability problems

Because complex structures, in respect to their mechanical or thermal behavior, are no longer solvable analytically, the structure must be broken down into smaller elements. The FEM process enables the breakdown of a larger structure into elements, thus enabling the description of component behavior. Therefore, very complex entities are solvable. Calculation problems from real-world applications are usually quite complex. For example, in crash testing it is necessary to reproduce or simulate the complete vehicle structure even when it consists of many different components and materials (Figure 16).

The finite elements are described geometrically using a series of nodes and edges, which in turn form a mesh. The formation of the mesh is calculable. The complete behavior of a structure can then be described through the composition of the finite elements.

An FEM computation can be divided into the following steps:

- Breakdown of the structure into finite elements
- Formulation of the physical and mathematical description of the elements
- Composition of a physical and mathematical description for the entire system
- Computation of the description according to requirements
- Interpretation of the computational results

The FEM programs commercially available today process these steps in essentially three program phases:

1. *Preprocessing:* Preparation of the operations, mesh generation
2. *Solving:* Actual finite element computation
3. *Postprocessing:* Interpretation and presentation of the results

Preprocessing entails mainly geometric and physical description operations. The determination of the differential equations is implemented in the FEM system, whereby the task of the user is limited to selecting a suitable element type.

In the following sections, the three phases of an FEM computation will be presented from a user's point of view.

2.5.2.1. FEM Preprocessing The objective of the preprocessing is the automation of the meshing operation. The following data are generated in this process:

- Nodes
- Element types
- Material properties
- Boundary conditions
- Loads

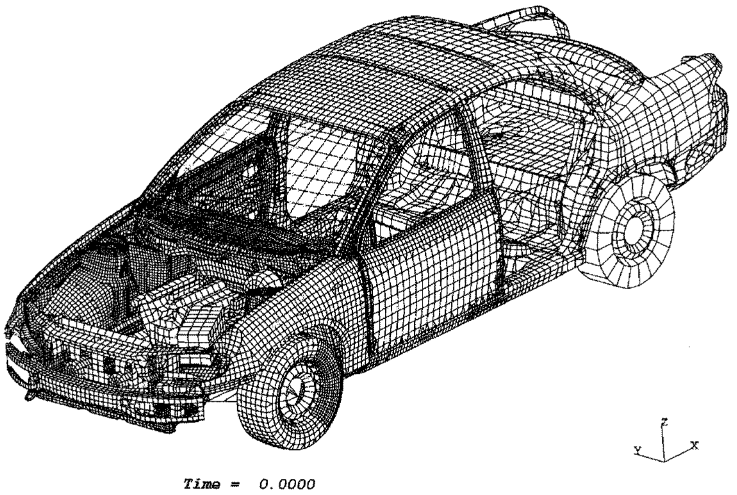


Figure 16 Opel Omega B Crash Model, 76,000 elements. (From Kohlhoff et al. 1994. Reprinted by permission of VDI Verlag GmbH.)

The data sets may be generated in various ways. For the generation of the mesh, there are basically two possibilities:

- Manual, interactive modeling of the FEM mesh with a preprocessor
- Modeling of the structure in a CAD system with a subsequent automated or semiautomated mesh generation

The generation of the mesh should be oriented towards the expected or desired calculation results. This is meaningful in order to simplify the geometry and consider a finer meshed area sooner in the FEM process. In return, the user must have some experience with FEM systems in order to work efficiently with the technology available today. Therefore, completely automatic meshing for any complex structure is not yet possible (Weck and Heckmann 1993).

Besides interactive meshing, commercially available mesh generators exist. The requirements of an FE mesh generator also depend on the application environment. The following requirements for automatic mesh generators, for both 2D and 3D models, should be met (Boender 1992):

- The user of an FE mesh generator should have adequate control over the mesh density for the various parts of the component. This control is necessary because the user, from experience, should know which areas of the part require a higher mesh density.
- The user must specify the boundary conditions. For example, it must be possible to determine the location of forces and fixed points on a model.
- The mesh generator should require a minimum of user input.
- The mesh generator must be able to process objects that are made up of various materials.
- The generation of the mesh must occur in a minimal amount of time.

The mesh created from the FE mesh generator must meet the following requirements:

- The mesh must be topologically and geometrically correct. The elements may not overlap one another.
- The quality of the mesh should be as high as possible. The mesh can be compared to analytical or experimental examinations.
- Corners and outside edges of the model should be mapped exactly using suitable node positioning.
- The elements should not cut any surfaces or edges. Further, no unmeshed areas should exist. At the end of the mesh refinement, the mesh should match the geometrical model as closely as possible. A slight simplification of the geometry can lead to large errors (Szabo 1994).
- The level of refinement should be greater in the areas where the gradient of the function to be calculated is high. This is determined with an automatic error estimation during the analysis and provides a new point for further refinement.

Various processes have been developed for automatic mesh generation. The motivation to automate the mesh generation process results from the fact that manual generation is very time consuming and quite prone to mistakes. Many processes based on 2D mesh generation, however, are increasingly being suggested for 3D processes.

The most frequently used finite element forms for 2D are three- and four-sided elements and for 3D are tetrahedrons and hexahedrons. For automatic mesh generation, triangles and tetrahedrons are suitable element shapes, whereas hexahedrons provide better results in the analysis phase (Knothe and Wessels 1992).

2.5.2.2. FEM Solution Process In the solution process, equation systems are solved that are dependent on the type of examination being carried out. As a result, various algorithms must be provided that allow for efficient solution of the problem. The main requirements for such algorithms are high speed and high accuracy.

Linear statistical examinations require only the solution of a system of linear equations. Dynamically nonlinear problems, on the other hand, require the application of highly developed integration methods, most of which are based on further developments of the Runge–Kutta method.

To reduce the computing time, matrices are often converted. This permits a more effective handling of the calculations. The objective is to arrange the coefficients with a resulting diagonal matrix. An example of an FEM calculation sequence is shown in Figure 17.

2.5.2.3. FEM Postprocessing Because the calculation results of an FEM computation only deliver nodes and their displacement and elements with the stresses or eigenforms in numerical rep-

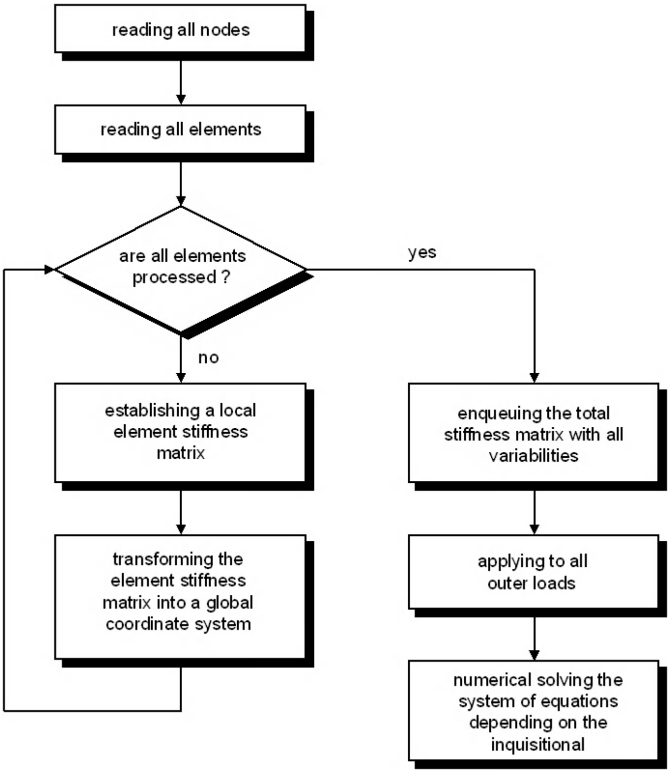

reading all nodes
reading all elements
are all elements processed ?
yes
no
establishing a local element stiffness matrix
transforming the element stiffness matrix into a global coordinate system
enqueuing the total stiffness matrix with all variabilities
applying to all outer loads
numerical solving the system of equations depending on the inquisitional

Figure 17 Sequence of a Finite Element Calculation.

resentation, postprocessors are required in order to present the results in a graphical form (Figure 18).

The postprocessing manages the following tasks:

- Visualization of the calculated results
- Plausibility control of the calculation

The performance and capability of postprocessors are constantly increasing and offer substantial presentation possibilities for:

- Stress areas and main stresses
- Vector fields for forces, deformation and stress characteristics
- Presentation of deformations
- Natural forms
- Temporary deformation analysis
- Temperature fields and temperature differences
- Velocities

It is possible to generate representations along any curve of the FEM model or, for example, to present all forces on a node. Stress fields of a component can be shown on the surface or within the component. The component can thus be reviewed in various views.

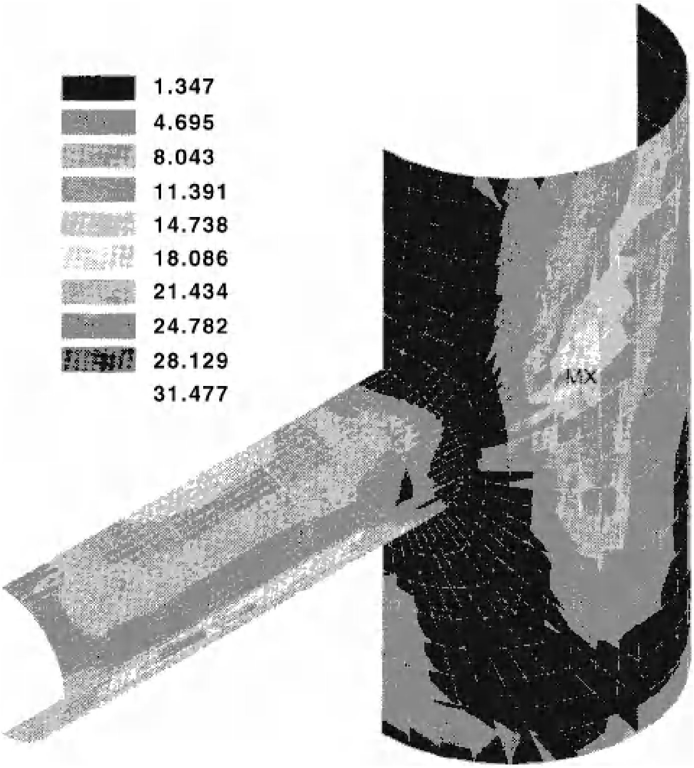


Figure 18 FEM Postprocessing Resulting from a Standard Pipe Formation. (From CAD-FEM 1995)

3. CONCEPTS

3.1. Process Chains

A process chain is a set of rules or steps carried out in a specific order in order to carry out the completion of a defined process. Processes can be executed either in sequence or in parallel. A typical example of a product development process chain is portrayed in Figure 19.

A process chain is characterized as:

- A process, divided into subtasks, in which contributions to the creation of virtual products take place
- A series of systems in which computer support for product-related subtasks is guaranteed
- A series of systems for organized support of the complete proces
- Mechanisms for the adequate exchange of data between systems

Computer support of process chains is related, on one hand, to the processing and management of product data and, on the other hand, to the organization of the supported process and the handling of associated data.

The processing and management of product data entail all tasks that concern the generation, storage, transfer, and function-related editing and provision of data (Reinwald 1995).

General classification of process chains suggests a determination of the depth of integration of the product data handling. Classification can be divided into:

- Coupled/linked process chains
- Integrated process chains
- Process chains with common data management

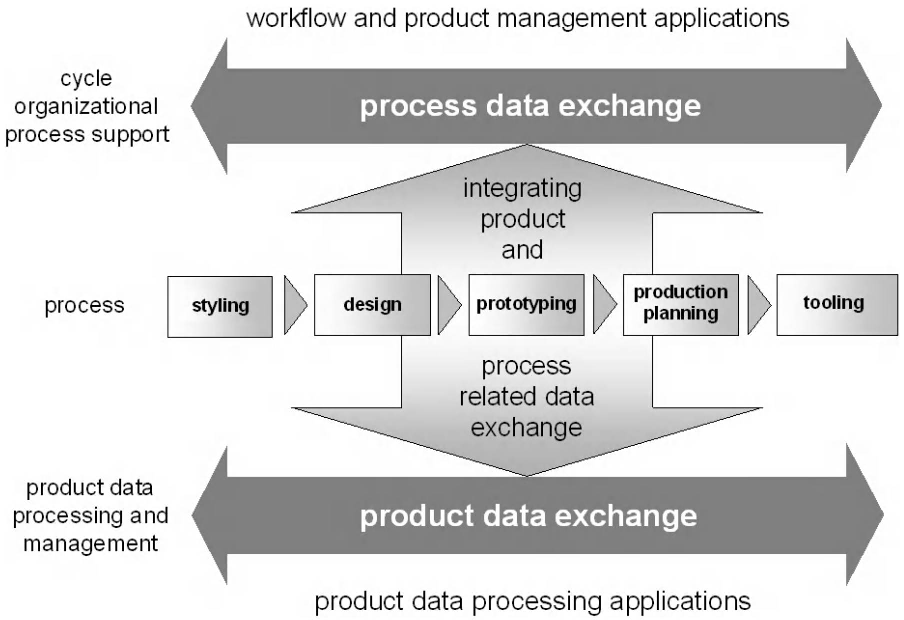

workflow and product management applications
cycle organizational process support
process data exchange
integrating product and
process
styling
design
prototyping
production planning
tooling
process related data exchange
product data processing and management
product data exchange
product data processing applications

Figure 19 Layout of a Computer-Supported Process Chain.

The exchange of data within a coupled process chain takes place directly between the supporting systems. Thereby, information transferred between two respective systems is aided with the help of system-specific interfaces. This realization of information transfer quickly reaches useful limits as the complexity of the process chains increases. This is because the fact that the necessary links grow disproportionately and the information from the various systems is not always transferable (Anderl 1993).

Characteristic of integrated process chains is a common database in the form of an integrated product model that contains a uniform information structure. The information structure must be able to present in a model all relevant product data in the process chain and to portray the model with the use of a homogenous mechanism. A corresponding basis for the realization of such information models is offered by the Standard for the Exchange of Product Model Data (ISO 10303—STEP) (ISO 1994).

Basic to a process chain with common product data management based on an EDM system is the compromise between coupled and integrated process chains. With the use of common data management systems, organizational deficiencies are avoided. Thus, the EDM system offers the possibility of arranging system-specific data in comprehensive product structures. The relationship between documents such as CAD models and the respective working drawing can be represented with corresponding references. A classification system for parts and assemblies aids in locating the required product information (Jablonski 1995).

On the other hand, with joint product data management, conversion problems for system-specific data occur in process chains based on an EDM system. These problems exist in a similar form in coupled process chains. This makes it impossible to combine individual documents into an integrated data model. References within various systems, related to objects within a document rather than the document itself, cannot be made.

Besides requiring support for the processing of partial tasks and the necessary data transfer, process chains also require organized, sequential support. This includes planning tasks as well as process control and monitoring. Process planning, scheduling, and resource planning are fundamental planning requirements.

The subject of process planning involves segmenting the complete process into individual activities and defining the sequence structure by determining the order of correlation.

In the use of process plans, the following points are differentiated:

- One-time, detailed process planning (all processes are executed similarly)
- Case-by-case planning of individual processes

- One-time, rough planning of a generic process, followed by deeper detailing and specification of individual processes.

The first basic approach is typical for the majority of commercial work flow management systems. Because these systems are intended for routine tasks, their application is unproblematic only when they concern the support of completely determinable processes. Product development processes are characterized by the fact that their achievements are not precisely predictable in advance. For that reason they are treated as nondeterministic processes.

The second approach requires an independent sequence for each process path. This corresponds to the generation of a network plan using methods such as CPM (critical path method), MPM (metra potential method), and PERT (program evaluation and review technique). These techniques, as part of the project management, are employable for activities such as proper scheduling (Burghardt 1988).

The third approach involves the strategic use of rough supplementary outlines of generic processes with those of detailed, individual processes. A requirement for successful implementation of the strategy is the ability of the supporting system to map process structures hierarchically.

For scheduling, a closed sequence plan is assumed to exist. The activities and correlations must be determined beforehand. Therefore, the minimum task of the scheduling part of the process is the calculation and determination of time limits, critical paths, and buffer times. The CPM, MPM, and PERT methods are implemented for this purpose, and the incorporation of these operations leads to the calculation of complete process and buffer times using both forward and backward calculations.

The duration of a procedure depends significantly on the use of resources required for fulfilment of the process. In this respect, interaction exists between the scheduling and capacity planning (Burghardt 1988).

Capacity planning for the product-development process involves proper quantity and time allocation of coworkers, application system capacities, and other partial tasks of an individual process (Golm 1996). Therefore, the goal of capacity planning is to ensure the on-schedule processing of the entire process and the uniform utilization of resources.

3.1.1. Tasks of Process Control and Monitoring

The task of process control and monitoring is to guarantee the fulfilment of the process based on the sequence, schedule, and capacity planning determinants. The following individual tasks must then be resolved:

- Activation of processable actions
- Process monitoring of contents
- Process monitoring of time

3.2. Integrated Modeling

Modeling of virtual products includes all phases of the product life cycle, from product planning to product disposal. The aim is complete integration of all the development processes for efficient product development.

The following system characteristics are strived for through the integration of design and production planning systems:

- Increased productivity during product development
- Acceleration of the product development process
- Improvement of product and technical documentation quality

An increase in the productivity of product development phases through the integration of design and manufacturing plans is based on preventing the loss of information. A loss of information often occurs between coupled and unrelated systems because of integration difficulties. With a lack of proper integration, detection and editing of data take place numerous times for the same information content. Therefore, using product models derived from information within both the design and manufacturing data, it is possible to integrate specific work functions. The models then serve as a basis for all system functions and support in a feature approach-type form.

The feature approach is based on the idea that product formation boundaries or limits are explicitly adhered to during design and production planning stages. This ensures that the end result of the product envisioned by the designer is not lost or overlooked during the modeling phase. To realize the form and definition incorporated by the designer in a product, it is essential that an integrated design and production planning feature exist.

Accelerating the product development process is achievable through the minimization of unnecessary data retrieval and editing. Efforts to parallel the features of concurrent engineering tasks require integrated systems of design and production planning.

Higher product and documentation quality is achievable with the use of integrated systems. Through integration, all functions of a product model are made available to the user. Any errors that may occur during data transmission between separate systems are avoided.

Because of the high responsibility for cost control during product development, estimates for design decisions are extremely necessary. Cost estimates reveal the implications of design decisions and help in avoiding costly design mistakes. For effective estimation of costs to support the design, it is necessary to be able to access production planning data and functions efficiently. Integrated systems provide the prerequisites for the feedback of information from the production planners to the designers, thus promoting qualitatively better products. The following forms of feedback are possible:

- Abstract production planning experience can be made available to the designer in the form of rules and guidelines. A constant adaptation or adjustment of production planning know-how has to be guaranteed.
- Design problems or necessary modifications discovered during the production planning phases can be directly represented in an integrated model.
- Necessary modifications can be carried out in both production planning and design environments.

3.3. Methodical Orientation

The virtualization of product development is a process for the acquisition of information in which, at the end of the product development, all necessary information generated is made available. Assuming that humans are at the heart of information gathering and that the human decision making process drives product development, the virtual product-creation methods must support the decision maker throughout the product creation process.

Product-creation processes are influenced by a variety of variables. The form the developed product takes is determined by company atmosphere, market conditions, and the designer's own decisions. Also influential are the type of product, materials, technology, complexity, number of model variations, material costs, and the expected product quantity and batch sizes.

Product development is not oriented toward the creation of just any product, but rather a product that meets the demands and desires of the consumer while fulfilling the market goals of the company. The necessary mechanisms must provide a correlation between the abstract company goals and the goals of the decisions made within the product development process. For example, to ensure the achievement of cost-minimization objectives, product developers can use mechanisms for early cost estimation and selection, providing a basis for the support of the most cost-effective solution alternatives (Hartung and Elpet 1986). To act upon markets characterized by individual consumer demands and constant preference changes, it is necessary to be able to supply a variety of products within relatively short development times (Rathnow 1993; Eversheim 1989). This means that product structuring must take into account the prevention of unnecessary product variation and that parallel execution of concurrent engineering is of great importance.

Methods that are intended to support the product developer must be oriented toward not only the contents of the problem but also the designer's way of thinking. Therefore, a compromise must be made between the methodical problem solving strategy and the designer's creative thought process.

The product-development methods can be separated into process-oriented and product-oriented categories. The main concern in process-oriented product development is the design. The objective is the indirect improvement of the design or, more precisely, a more efficient design process. Product-oriented methods concern the product itself and the direct improvement of the product.

Various process-oriented methods are discussed below. Fundamental to VDI 2221 guidelines is the structuring of the product-development process into partial processes or phases. This structuring takes place independently of the developed product, the company, market conditions, and the decision maker. The demand for a general strategy results in individual steps being described at very abstract levels. Consideration of product, company, and market-related influences is incorporated into the design throughout all processes. The impact of individual influences on methods can only be clarified with examples. The worth of a new method lies in the general systemizing of the product development process. This method is not suitable for immediate derivation of concrete product development processes. Rather, it describes an ideal, flexible product-development process. The structure and content of the development process are then molded by the product, company, or market-related influences (VDI-Gesellschaft Entwicklung Konstruktion 1993).

Another example of process-oriented methods is the concurrent engineering method, which covers all topics from product development to equipment and production planning. The primary objective of simultaneous engineering is the reconfiguration of the development process, with the intention of reducing development time while improving product quality. As opposed to the traditional approach, in which a series of steps is followed and feedback occurs through long loops, the development tasks within concurrent engineering allow for the parallel execution of many tasks. Therefore, development times are substantially reduced and the old system of following predetermined steps is avoided. This concept relies heavily on the exchange of information between the various departments. This should extend beyond company boundaries to include the equipment manufacturer in order to tie production equipment planning into the product development process. The main focus in the implementation of these methods, besides the use of computer aided technology, is the reorganization of the company structure. In the foreground are measures to stimulate the exchange of information, such as the formation of interdisciplinary teams (Eversheim et al. 1995; Krause et al. 1993; Bullinger and Warschat 1996).

The design-to-cost method is a product-development approach that bases design decisions on cost-sensitive criteria. Here, evaluation standards are not only production costs but also the costs incurred throughout the life of the product. This method is particularly applicable for complex products with relatively long life cycles. The method is broken down into target costing, cost-oriented design, and cost control. Within target costing, a goal for final costs is broken down for individual product components. The final cost is determined from the results of a market analysis or from a clear depiction of customer demands. Relative cost and manufacturing cost data can be used to aid the development of design alternatives that fall within the early cost-oriented design process. With the aid of cost estimate models, the life cycle costs of each alternative are assessed and the most cost effective variant selected. In the area of cost control, the design costs are compared to and contrasted with the original end cost objective. If the cost objective is not met, deviation analyses are necessary to determine required alternatives.

Product-oriented techniques are characterized by product formation, task formulation, and the objectives pursued. Product-oriented tasks vary considerably, according to the multitude of products and tasks. A few examples of these techniques and technique groupings are presented later in this chapter.

Design rules and guidelines specify how a product is designed according to previously determined objectives. Well-known design rules and guidelines exist particularly for production, assembly, ergonomic, and logistics-oriented processes as well as resource-efficient and recycling-oriented product design (Pawellek and Schulte 1987; Krause 1996; Pahl and Beitz 1993; Kriwet 1995).

Simulation technology has become important in the field of product development. The primary objective in the implementation of simulation technology is the early acquisition of information on product characteristics before the product even exists. The knowledge gained aids in the assessment of the respective development results. Costly and time-consuming mistakes during development phases can be recognized early and avoided. Through an iterative strategy, it is also possible to optimize systematically particular product characteristics or qualities (Frepoli and Botta 1996; Schönbach 1996).

One requirement for the application of simulation techniques is the creation of a model, which allows for the investigation and breakdown of real product tasks. With the use of computers, it is possible to design complex models that allow answers to relevant design questions to be found. These models were previously realizable only through the tedious production of prototypes. The application of simulation technology is especially helpful when numerous product variations must be considered, a situation not economically feasible with prototypes.

In the future, systems must be able to provide all relevant information in the user-desired form and make additional mechanisms available that reveal the consequences of a decision made during product development. The provision of information must include not only product specific data but also comprehensive information. Data covering all topics from design guidelines and rules to design, solution, and measurement catalogs to company individual knowhow is necessary. Solutions developed for similar products need to be accessible not only in general form but in a form that takes product-related tasks and context into account.

3.4. Rapid Prototyping

Due to the ever-increasing demand for shorter product development cycles, a new technology known as rapid prototyping (RP) has emerged. RP is the organizational and technical connection of all processes in order to construct a physical prototype. The time required from the date the order is placed to when the prototype is completed can be substantially reduced with the application of RP technology (Figure 20).

RP is made up of generative manufacturing processes, known as RP processes, and conventional NC processes as well as follow-up technologies. In contrast to conventional working processes, RP processes such as stereo lithography, selective laser sintering, fused deposition modeling, and laminated object manufacturing enable the production of models and examples without the use of forming tools or molds. RP processes are also known as free-form manufacturing or layer manufacturing.

Prototypes can be divided into design, function, and technical categories. To support the development process, design prototypes are prepared in which proportion and ergonomic models are

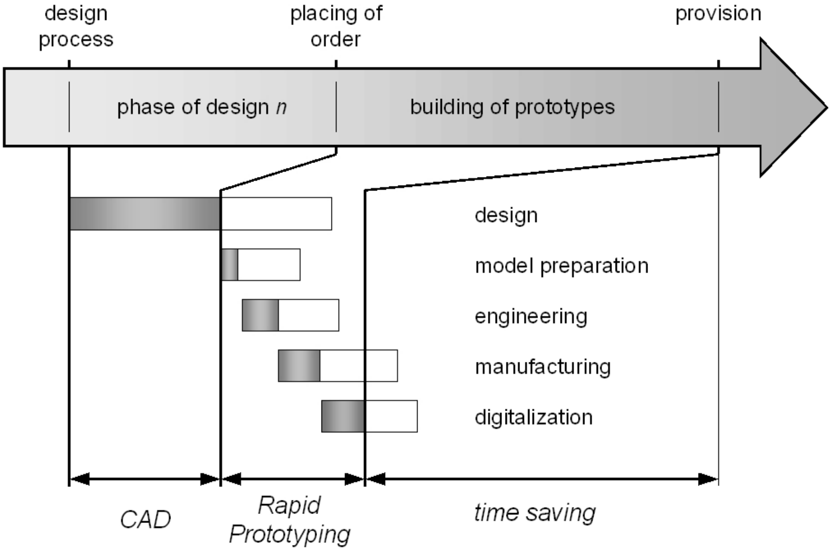


Figure 20 Acceleration Potential through CAD and RP Application.

incorporated. Design prototypes serve for the verification of haptic, esthetic, and dimensional requirements as well as the constructive conception and layout of the product. RP-produced examples are made from polycarbonates, polyamides, or wood-like materials and are especially useful for visualization of the desired form and surface qualities.

For function verification and optimization, a functional product example is required. The application of production series materials is not always necessary for this purpose. Functional prototypes should, however, display similar material strength characteristics. On the other hand, technical prototypes should be produced using the respective production series materials and, whenever possible, intended production line equipment. The latter serve for purposes such as customer acceptance checks and the verification of manufacturing processes.

Prototypes can be utilized for individual parts, product prototypes, and tool prototypes. Normally, product prototypes consist of various individual prototypes and therefore require some assembly. This results in higher levels of dimension and shape precision being required.

The geometric complexity represents an essential criterion for the selection of the suitable prototype production process. If, for example, the prototype is rotationally symmetric, the conventional NC turning process is sufficient. In this case, the presently available RP processes provide no savings potential. However, the majority of industry-applied prototypes contain complex geometrical elements such as free-form surfaces and cut-outs. The production of these elements belongs to some of the most demanding tasks in the area of prototype production and is, because of the high amount of manual work involved, one of the most time-consuming and cost-intensive procedures. RP processes, however, are in no way subjected to geometrical restrictions, so the build time and costs of producing complex geometrical figures are greatly reduced (König et al. 1994).

Another criterion for process selection is the required component measurements and quality characteristics such as shape, dimensional accuracy, and surface quality.

3.4.1. Systemization of Rapid Prototyping Processes

With the implementation of CAD/CAM technology, it is possible to produce prototypes directly based on a virtual model. The generation of the geometry using RP processes takes place quickly without the requirement of molds and machine tools. The main feature of the process is the formation of the workpiece. Rather than the conventional manufacturing process of a clamped workpiece and material removal techniques, RP processes entail the layering of a fluid or powder in phases to form a solid shape.

Using CAD models, the surfaces of the components are fragmented into small triangles through a triangulation process. The fragments are then transformed into the de facto RP standard format known as STL (stereo lithography format). The STL format describes the component geometry as a closed surface composed of triangles with the specification of a directional vector. Meanwhile, most

CAD systems now provide formatting interfaces as part of the standard software. Another feature of this process is that CAD-generated NC code describing the component geometry allows for a slice process in which layers of the object may be cut away in desired intervals or depths.

Starting with the basic contour derived from the slicing process, the workpiece is subsequently built up in layers during the actual forming process. Differences exist between the RP processes in process principles and execution.

RP processes can be classified by either the state of the raw material or the method of prototype formation. The raw materials for RP processes are in either fluid, powder, or solid states (Figure 21). As far as the method of prototype creation, the component forming can either be processed into direct, 3D objects or undergo a continuous process of layers built upon one another (Figure 22).

3.5. Digital Mock-up

Today the verification and validation of new products and assemblies relies mainly on physical mock-ups. The increasing number of variants and the need for higher product and design quality require a concurrent product validation of several design variants that are based on digital mock-ups. A digital mock-up (DMU) can be defined as "a computer-internal model for spatial and functional analysis of the structure of a product model, its assembly and parts respectively" (Krause et al. 1999).

The primary goal of DMU is to ensure the ability to assemble a product at each state of its development and simultaneously to achieve a reduction in the number of physical prototypes. Oriented to the product development cycle, tools for DMU provide methods and functionality for design and analysis of product components and their function. Modeling methods are divided into several categories: space management, multiple views, configuration management of product variants and versions, management of relations between components, and also incomplete models. Simulations are targeted to analyze the assembly and disassembly of a product as well as the investigation and verification of ergonomic and functional requirements. Constituent parts of a DMU tool are distinguished into either components or applications. Components are the foundation of applications and consist of modules for data organization, visualization, simulation models, and DMU/PMU correlations. Main applications are collision checks, assembly, disassembly, and simulation of ergonomics, functionality, and usage aspects. Because a complex product like an airplane can consist of more than 1 million parts, handling a large number of parts and their attributes efficiently is necessary. This requires that DMU methods are optimized data structures and algorithms.

Besides basic functions such as the generation of new assemblies, the modification of existing assembles and the storage of assemblies as product structures with related parts in heterogeneous databases, advanced modeling methods are required in order to ensure process-oriented modeling on the basis of DMU. The main methods in this context are (BRITE-EURAM 1997):

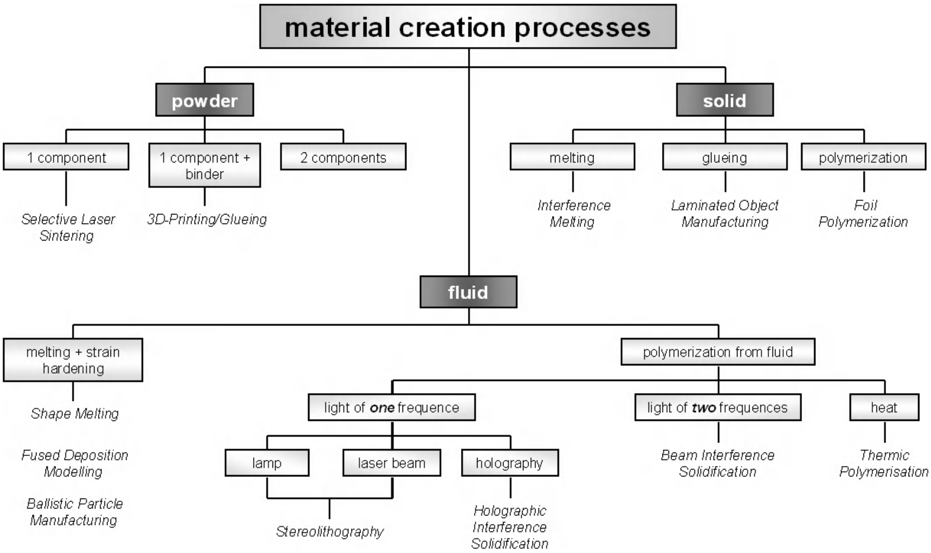

material creation processes
powder
solid
1 component
1 component + binder
2 components
Selective Laser Sintering
3D-Printing/Glueing
melting
glueing
polymerization
Interference Melting
Laminated Object Manufacturing
Foil Polymerization
fluid
melting + strain hardening
Shape Melting
Fused Deposition Modelling
Ballistic Particle Manufacturing
polymerization from fluid
light of one frequence
lamp
laser beam
holography
Stereolithography
Holographic Interference Solidification
light of two frequences
Beam Interference Solidification
heat
Thermic Polymerisation

Figure 21 Classification of RP Processes with Respect to the Material-Generation Process. (From Kruth 1991)

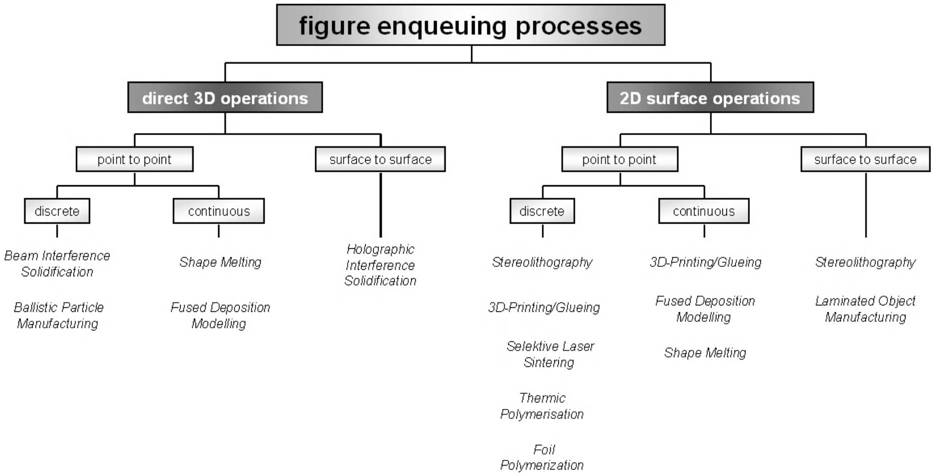


figure enqueuing processes
direct 3D operations
2D surface operations
point to point
surface to surface
point to point
surface to surface
discrete
continuous
discrete
continuous
Beam Interference Solidification
Ballistic Particle Manufacturing
Shape Melting
Fused Deposition Modelling
Holographic Interference Solidification
Stereolithography
3D-Printing/Glueing
Selektive Laser Sintering
Thermic Polymerisation
Foil Polymerization
3D-Printing/Glueing
Fused Deposition Modelling
Shape Melting
Stereolithography
Laminated Object Manufacturing

Figure 22 Classification of RP Processes with Regard to the Form-Generation Process. (From Kruth 1991)

- *Organization of spaces:* Allocating and keeping open functional and process-oriented spaces in DMU applications. The major problem is the consideration of concurrent requirements concerning the spaces.
- *Organization of views:* A view is an extract of the entire structure in a suitable presentation dedicated to particular phases and applications.
- *Handling of incomplete models:* Incomplete and inconsistent configurations during development process are allowable in order to fulfill the user's requirements regarding flexibility. Examples are symbolically and geometrically combined models with variable validity.
- *Configuration management of product structure:* Management of versions, variants, and multiuse.

Besides these methods in a distributed, cooperative environment, consequent safety management has to be taken into account. Therefore, a role- and process-related access mechanism must be implemented that allows the administrator to define restrictions of modeling related to roles. The application of such technologies enables a company to manage outsourced development services.

4. ARCHITECTURES

4.1. General Explanations

Product modeling creates product model data and is seen as a decisive constituent of the computer-supported product development activities. Product creation includes all the tasks or steps of product development, from the initial concept to the tested prototypes. During product modeling, a product model database is created and must support all relevant data throughout the product's life cycle.

Product modeling is made up of interconnected parts: the product model and the process chain. The product model is related to the product database and the management and access algorithms. The process chain, besides usually being related to the operational sequence of the product development, is in this context all the necessary product modeling processes required to turn the initial idea into a finished product. The product modeling processes consist of technical and management-related functions. The product model data are the most important factor determined from the development and planning activities.

The term *product model* can be logically interpreted to mean the accumulation of all product-related information within the product life cycle. This information is stored in the form of digitalized product model data and is provided with access and management functions. Modeling systems serve for the processing and handling of the product model data.

As with many other systems, the underlying architecture for CAD systems is important for extensibility and adaptability to special tasks and other systems. Compatibility with other systems, not just other CAD systems, is extremely important for profitable application. Product data should be

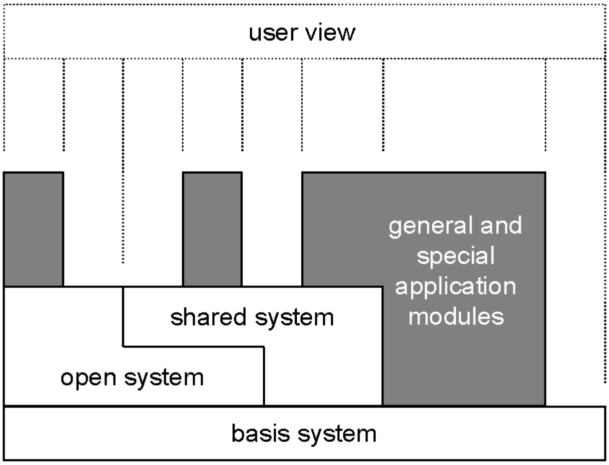


Figure 23 Classes of CAD Systems.

made universally employable throughout the product development process because bridges between media and format changes are error-prone and not often automated.

Enhancements of existing CAD systems, in the form of general and specialty application modules, are important in order to allow customer-specific adaptations to already-provided functionality.

The various classes of CAD systems, which may appear in combination with one another (Figure 23), include the basic system, general and specialty application software, and open and shared systems.

A basic CAD system is made up of a computer internal representation (CIR) of the product model, a core modeler with functionality for management and processing of the CIR, and a user interface for visualization of the CIR and interaction with the user (Figure 24).

4.2. Integration of Application Modules and Core Modeler

Many companies require not only a modeler for geometrical elements but also application modules to integrate into their product development process and for computer internal imaging of the processes. Calculation modules and simulation software belong in this category. Such application modules must integrate with the modeler to form a functional entity (Figure 25).

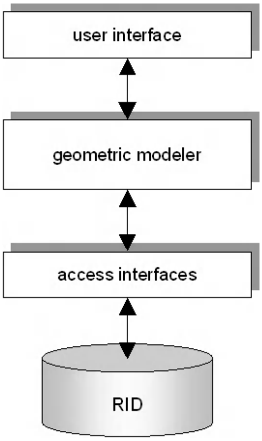


Figure 24 Basic CAD System.

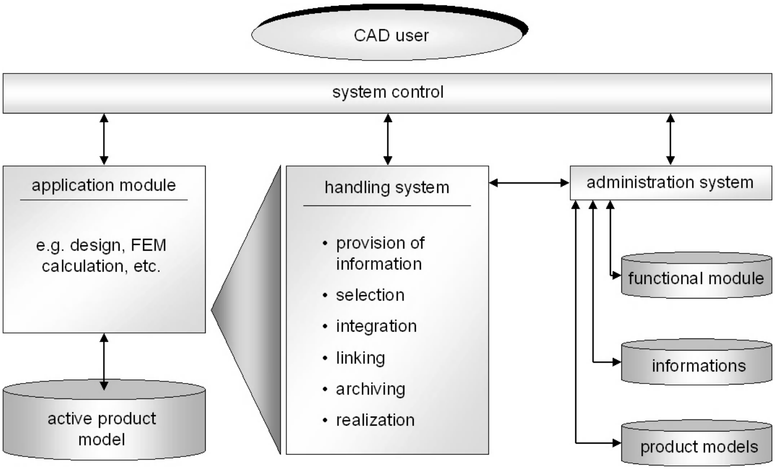


Figure 25 Modeler with Application Modules. (From Krause et al. 1990)

The application modules can be subdivided, according to their area of employment, into general and specialty modules. General application modules for widespread tasks, such as FEM modules or interfaces to widely used program systems, are commercially marketed by the supplier of the CAD system or by a system supplier of complementary software. Specially adapted extensions, on the other hand, are usually either created by the user or outsourced.

A thoroughly structured and functionally complete application programming interface (API) is a particular requirement due to the inability of the user to look deeper into the system.

Honda implemented a CAD system with around 200 application software packages (Krause and Pätzold 1992), but the user was presented with a Honda interface as a uniform medium for accessing the new plethora of programs. This enabled relative autonomy from the underlying software in that as the user, switching between systems, always dealt with the same user interface.

A further step towards autonomy, stemming from the basic system in use, is met only when application modules use an interface already provided. If this does not occur, the conversion to another base system is costly and time consuming. Sometimes it is even cheaper to install a new application module.

Often special application modules are required that cannot be integrated seamlessly into the CAD system. Therefore, various automobile manufacturers, such as Ford, General Motors, Nissan, VW, Audi, and Skoda, have implemented special surface modelers besides their standard CAD systems to enable free-form surface designs to be carried out.

4.3. Shared System Architectures

The scope and time restrictions of most design projects demand the collaboration of many designers. The implementation of shared CAD systems significantly simplifies this teamwork. Here it is possible for several designers to work on the same CIR (computer internal representation). It is not necessary to copy the CIR to various computers in order to work on it and then manually integrate the changes afterwards. The management of this process is taken over by the CAD system. These tasks, however, remain widely concealed from the user. Before this work is begun, only the design area of the respective designer must be defined in order to avoid overlapping.

Such shared CAD systems use a common database. This is usually accessible from the individual CAD stations through a client–server architecture (Figure 26). Usually an Engineering Data Management System (EDMS), implemented in conjunction with a CAD system, is used.

This technique is often found applied in local area networks (LANs), but new problems emerge when a changeover is made to wide area networks (WANs):

- The bandwidth available in a WAN is significantly lower than in a LAN and is often too small for existing shared systems.
- The data security must be absolutely ensured. Therefore, six aspects must be considered:
  - *Access control:* exclusive retrieval of data by authorized personnel
  - *Confidentiality:* prevention of data interception during transmission

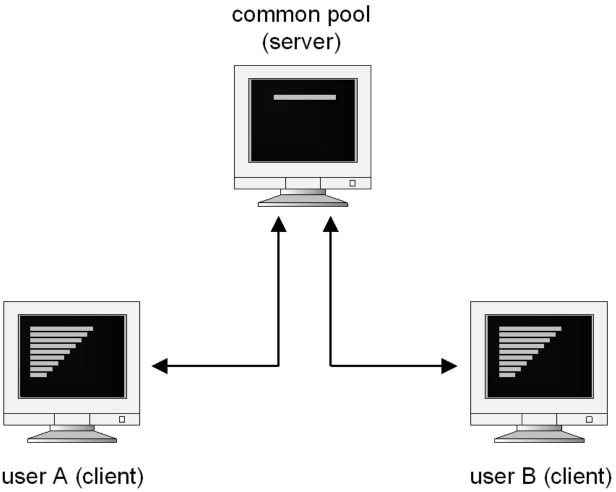


Figure 26 Shared CAD System with a Common Database and a Client-server Architecture.

- *Authentication:* the origin of the data transferred can be reliably identified
- *Integrity:* data cannot be modified during transmission
- *Nonrepudiation:* the sending of data cannot be refused or denied
- *Availability:* the data must always be available to authorized users

Newer implementations of shared systems rely on software architectures for the exchange of objects. Here, not only data is exchanged but, according to the object, also the data methods.

5. KNOWLEDGE MANAGEMENT

5.1. Origin and Background

The concept of knowledge management has been used in different disciplines, mostly in knowledge engineering (De Hoog 1997; Schreiber et al. 2000) and artificial intelligence (Göbler 1992; Forkel 1994). AI research often reduces the concept of knowledge management to the description of the development and use of expert systems (Gödicke 1992) and decision support systems. However, Davenport et al. (1996) found only one expert system application within 30 knowledge work improvement projects. Analysis of approximately 100 case studies published before February 1998 shows that IT-based approaches towards knowledge management are dominant. IT-based knowledge management approaches focus mainly on the storage (databases, DMS) and distribution (intranet and Internet applications, push and/or pull) of explicit, electronically documented knowledge, thus ignoring the tacit dimension of knowledge (Mertins 2001).

The improvements in data processing and network technologies have enabled access to data and information via the Internet at every time and every place in the world. Increasing market demands for reduction in time-to-market, more flexibility, and higher quality at lowest costs have contributed to a new discussion of the concept of knowledge management. These approaches differ from the above-mentioned ones in giving more emphasis to critical success factors such as culture and motivation of the employees and aiming to combine human-based and IT-based methods and tools for knowledge management (Davenport and Prusak 1998; Probst et al. 1998; Skyrme and Amidon 1997; Wiig 1995, 1997; Willke 1998).

5.2. Knowledge

What is knowledge? That is the question most frequently asked by people interested in knowledge management. The discussion about knowledge has a very long tradition. More than 2 thousand years ago, Socrates asked his students, “Why do we have to know what knowledge is? How can we know what knowledge is? What do we know about knowledge?” (see Plato, *Theaetetus*). Today there are numerous descriptions and definitions of knowledge. Romhardt (1998) finds 40 dichotomies of knowledge, such as explicit vs. implicit or tacit and individual vs. collective. Von Krogh and Venzin (1995) create seven categories of knowledge to be used in management and organization theory: tacit, embodied, encoded, embrained, embedded, event and procedural. Holsapple and Whinston (1992) discuss six types of knowledge that are important for knowledge management and decision support systems: descriptive, procedural, reasoning, linguistic, assimilative, and presentation. Moreover, Schreiber et al. (2000) ask the question, "Why bother?" because even physicists will often have difficulty giving an exact definition of energy. This does not prevent them, however, producing energy and other products.

The concepts most often mentioned by authors in the context of knowledge management are data, information, and knowledge. Some even add wisdom. This classification, if not properly understood and used, could lead to a philosophical discussion of the "right" distinction between the categories. The transition from one to the other is not always clear-cut. Instead of a hierarchy, a continuum ranging from data via information to knowledge has proved to be the most practical scheme for knowledge management (Probst et al. 1998; Heisig 2000).

Data means the individual facts that are found everywhere in a company. These facts can be easily processed electronically, and gathering of large amounts of data is not problematic today. However, this process alone does not lead to appropriate, precise, and objective decisions. Data alone are meaningless. Data become information when they are relevant and fulfill a goal. Relevant information is extracted as a response to a flood of data.

However, deciding which knowledge is sensible and useful is a subjective matter. The receiver of information decides whether it is really information or just noise. In order to give data meaning and thus change it into information, it is necessary to condense, contextualize, calculate, categorize, and correct (Tiwana 2000). When data are shared in a company, their value is increased by different people contributing to their meaning.

As opposed to data, knowledge has a value that can be anywhere between true and false. Knowledge can be based on assumption, preconception, or belief. Knowledge-management tools must be able to deal with such imprecision (e.g., documentation of experiences).

> Knowledge is simply actionable information. Actionable refers to the notion of *relevant, and nothing but the relevant* information being available in the right place at the right time, in the right context, and in the right way so anyone (not just the producer) can bring it to bear on decisions being made every minute. Knowledge is the key resource in intelligent decision making, forecasting, design, planning, diagnosis, analysis, evaluation, and intuitive judgment making. It is formed in and shared between individual and collective minds. It does *not* grow out of databases but evolves with experience, successes, failures, and learning over time." (Tiwana 2000, p. 57)

Taking all these aspects into consideration, knowledge is the result of the interaction between information and personal experience. Typical questions for data and information are Who? What? Where? and When? Typical questions for knowledge are How? and Why? (Eck 1997).

One important differentiation is often made between tacit and explicit knowledge. Tacit knowledge is stored in the minds of employees and is difficult to formalize (Polanyi 1962; Nonaka and Takeuchi 1995). Explicit knowledge is the kind that can be codified and transferred. Tacit knowledge becomes explicit by means of externalization. With the introduction of CNC machines in mechanical workshops, experienced and highly skilled workers often felt insecure about their ability to control the process. They missed the "right sound" of the metal and the "good vibrations" of the machine. These signals were absorbed by the new CNC machines and hence workers were not able to activate their tacit knowledge in order to produce high-quality products (Martin 1995; Carbon and Heisig 1993). Similar problems have been observed with the introduction of other CIM technologies, such as CAD/CAM in the design and process-planning department and MRP systems for order management. The information supply chain could not fully substitute the informal knowledge transfer chain between the different departments (Mertins et al. 1993; Fleig and Schneider 1995). A similar observation is quoted from a worker at a paper manufacturing plant: "We know the paper is right when it smells right" (Victor and Boynton 1998, p. 43) However, this kind of knowledge is not found only in craftwork or industrial settings. It can be found in high-tech chip production environments (Luhn 1999) as well as social settings. From the noise of the pupils, experienced teachers can distinguish what they have to do in order to progress (Bromme 1999).

5.3. Knowledge Management Is Business and Process Oriented

Nearly all approaches to knowledge management emphasize the process character of interlinked tasks or activities. The wording and number of knowledge-management tasks given by each approach differ markedly. Probst (1998) proposes eight building blocks: the identification, acquisition, development, sharing, utilization, retention, and assessment of knowledge and the definition of knowledge goals. Another difference is the emphasis given by authors to the steps of the process- or knowledge-management tasks. Nonaka and Takeuchi (1995) describe processes for the creation of knowledge, while Bach et al. (1999) focus on the identification and distribution of the explicit, electronically documented objects of knowledge.

5.3.1. The Core Process of Knowledge Management

The analysis of different knowledge-management approaches (Probst et al. 1998; Davenport and Prusak 1998; Nonaka and Takeuchi 1995; Bach et al. 1999; Bukowitz 1999; Weggemann 1998) and the empirical results (Heisig and Vorbeck 2001) lead to the design of an integrated core process in which all activities are supported by organizational, motivational, and technical aspects. The core process can be further broken down into the core activities "define the goals of knowledge," "identify knowledge," "create (new) knowledge," "store knowledge," "distribute knowledge," and "apply knowledge." The quality of these stages is guaranteed by certain design fields for knowledge management. These fields include a company's process organization, available information technology, management systems, corporate culture, management of human resources, and control.

- *Create (new) knowledge:* Measures and instruments that promote the creation of knowledge include the acquisition of external knowledge (mergers, consultants, recruiting, patent acquisition), the setting up of interdisciplinary project teams that include the customers, and the application of lessons learned and methods to elicit tacit knowledge.
- *Store knowledge:* The stored knowledge in manuals, databases, case studies, reports, and even corporate processes and rules of thumb makes up one column of the other core activities. The other column consists of the knowledge stored in the brains of thousands of employees who leave their respective organizations at the end of each working day.
- *Distribute knowledge:* Provision of the right knowledge to the right person at the right time is the aim of the core task of distribution of knowledge. The methods and tools are dominated by IT applications such as the Internet or intranet. However, these tools provide added value only if trust and mutual understanding pervade the atmosphere of the entire company as well as project teams. The development of a common language is an important task. Other aspects of the distribution of knowledge are the transfer of experiences to new employees by training on the job, mentoring, or coaching techniques.
- *Apply knowledge:* According to our survey, the application of knowledge is the most essential task of knowledge management. Knowledge management mainly provides methods to overcome the barriers of the "not invented here" syndrome: the one-sided thinking and the development of preferred solution by existing information pathologies.

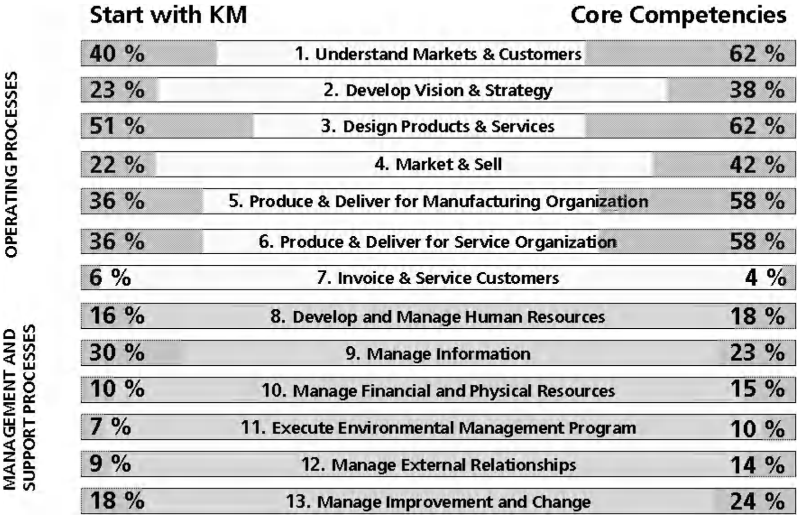
The close relationship between process and knowledge management is underscored by the critical success factors named by companies in the Europe-wide survey. Nearly one out of four companies (24%) mentioned aspects of the design of structures and processes as a critical factor in the success of knowledge management. Knowledge management is understood by practitioners from manufacturing and the service industry mainly as part of corporate culture and a business-oriented method: "The sum of procedures to generate, store, distribute and apply knowledge to achieve organizational goals" (Heisig and Vorbeck 2001).

Furthermore, the survey results indicate that companies focus on specific business processes to implement knowledge management. One out of every two companies starts its KM initiatives in the R&D area, two out of five focus on the process "Understanding Markets and Customers," and more than one out of every three of the companies begins in the area "Production and Delivery of Products and/or Services." The process "Manage Information" is ranked fourth in our overall sample and second in the service industry sample (Heisig and Vorbeck 2001). The companies locate their core competencies in these business processes too (Figure 27). Knowledge-management activities are started mainly within the areas identified as core competencies.

5.3.2. Design Fields of Knowledge Management

The second important step is to set up the link between knowledge management and the general organizational design areas, such as business processes, information systems, leadership, corporate culture, human resource management, and control (Figure 28).

- The business processes are the application areas for the core process of knowledge management. Existing knowledge has to be applied and new knowledge has to be generated to fulfill the needs of internal and external customers. The core activities have to be permanently aligned with the operating and value-creating business processes. Furthermore, knowledge-management activities could be linked with existing process-documentation programs (e.g., ISO certification) and integrated into business process reengineering approaches.
- Information technology is currently the main driving factor in knowledge management. This is due to considerable technological improvements in the field of worldwide data networking through Internet/intranet technologies. IT builds the infrastructure to support the core activities of storing and distributing knowledge. Data warehouses and data mining approaches will enable



Start with KM		Core Competencies
40 %	1. Understand Markets & Customers	62 %
23 %	2. Develop Vision & Strategy	38 %
51 %	3. Design Products & Services	62 %
22 %	4. Market & Sell	42 %
36 %	5. Produce & Deliver for Manufacturing Organization	58 %
36 %	6. Produce & Deliver for Service Organization	58 %
6 %	7. Invoice & Service Customers	4 %
16 %	8. Develop and Manage Human Resources	18 %
30 %	9. Manage Information	23 %
10 %	10. Manage Financial and Physical Resources	15 %
7 %	11. Execute Environmental Management Program	10 %
9 %	12. Manage External Relationships	14 %
18 %	13. Manage Improvement and Change	24 %

OPERATING PROCESSES
MANAGEMENT AND SUPPORT PROCESSES

Figure 27 Where Companies Start with Knowledge Management and Where They Locate Their Core Competencies.

companies to analyze massive databases and therefore contribute to the generation of new knowledge.

- The success of knowledge-management strategies is to a large degree determined by the support through top and mid-level managers. Therefore, leadership is a critical success factor. Each manager has to promote and personify the exchange of knowledge. He has to act as a multiplier and catalyst within day-to-day business activities. Special leadership training and change programs have to be applied to achieve the required leadership style.
- If the knowledge-management diagnosis indicates that the current corporate culture will not sustain knowledge management, wider change-management measures have to be implemented.

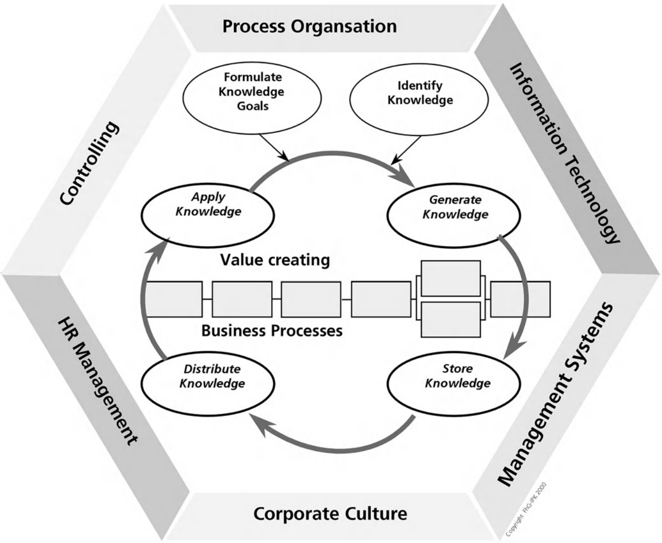

Process Organsation
Information Technology
Management Systems
Corporate Culture
HR Management
Controlling
Formulate Knowledge Goals
Identify Knowledge
Apply Knowledge
Generate Knowledge
Value creating
Business Processes
Distribute Knowledge
Store Knowledge
Copyright FhG-IPK 2000

Figure 28 Core Process and Design Fields of Knowledge Management.

The required company culture could be characterized by openness, mutual trust, and tolerance of mistakes, which would then be considered necessary costs of learning.

- Personnel-management measures have to be undertaken to develop specific knowledge-management skills such as the ability to develop and apply research and retrieval strategies as well as adequately structure and present knowledge and information. Furthermore, incentives for employees to document and share their knowledge have to be developed. Career plans have to be redesigned incorporating aspects of knowledge acquisition of employees. Performance-evaluation schemes have to be expanded towards the employees' contribution to knowledge generation, sharing, and transfer.
- Each management program has to demonstrate its effectiveness. Therefore, knowledge-controlling techniques have to be developed to support the goal-oriented control of knowledge creation and application with suitable control indicators. While strategic knowledge control supports the determination of knowledge goals, operative knowledge control contributes to the control of short-term knowledge activities.

Empirical results confirmed the great potential for savings and improvements that knowledge management offers (Figure 29). Over 70% of the companies questioned had already attained noticeable improvements through the use of knowledge management. Almost half of these companies had thus saved time and money or improved productivity. About 20% of these companies had either improved their processes, significantly clarified their structures and processes, increased the level of customer satisfaction, or facilitated decisions and forecasts through the use of knowledge management (Heisig and Vorbeck 2001).

However, some differences were apparent between the answers provided by service and by manufacturing companies. Twenty-eight percent of the service firms indicated an improvement in customer satisfaction due to knowledge management, as compared with only 16% of the manufacturing companies. Twenty-three percent of manufacturing companies stressed improvements in quality, as compared to only 15% of the service companies. Answers to questions about the clarity of structures and processes showed yet another difference. Twenty-six percent of the service companies indicated improvement with the use of knowledge management, as opposed to only 14% of manufacturing companies.

5.4. Approaches to the Design of Business Process and Knowledge Management

One primary design object in private and public organizations are the business processes that structure work for internal and external clients. Known as business process reengineering (BPR) (Hammer 1993), the design of business processes became the focus of management attention in the 1990s. Various methods and tools for BPR have been developed by research institutes, universities, and consulting companies. Despite these developments, a comparative study of methods for business process redesign conducted by the University of St. Gallen (Switzerland) concludes: "To sum up,

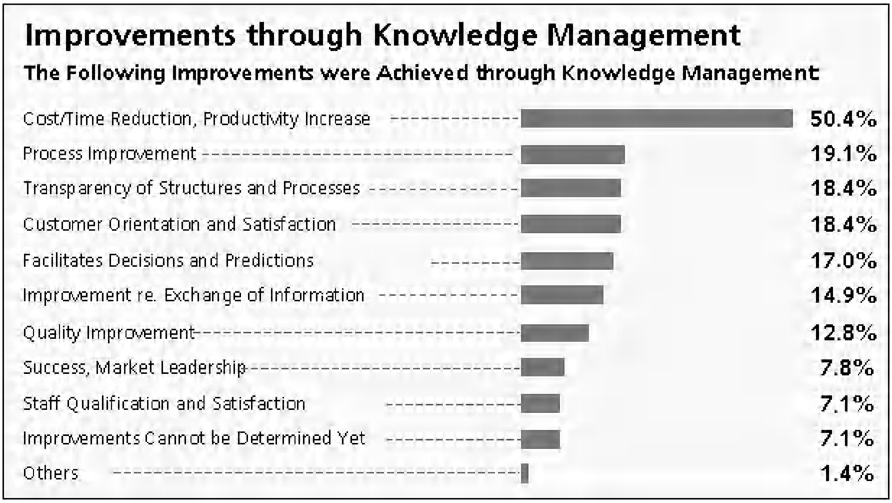


Figure 29 Improvements through Knowledge Management.

we have to state: hidden behind a more or less standard concept, there is a multitude of the most diverse methods. A standardized design theory for processes has still not emerged" (Hess and Brecht 1995, p. 114).

> BPR's focus is typically on studying and changing a variety of factors, including work flows and processes, information flows and uses, management and business practices, and staffing and other resources. However, most BPR efforts have not focused much on knowledge, if at all. This is indeed amazing considering that knowledge is a principal success factor—or in many judgment, the major driving force behind success. Knowledge-related perspectives need to be part of BPR. (Wiig 1995, p. 257)

Nearly all approaches to knowledge management aim at improving the results of the organization. These results are achieved by delivering a product and/or service to a client. This again is done by fulfilling certain tasks, which are linked to each other, thereby forming processes. These processes have been described as business processes. Often knowledge is understood as a resource used in these processes. Nevertheless, very few approaches to knowledge management have explicitly acknowledged this relation. And even fewer approaches have tried to develop a systematic method to integrate knowledge-management activities into the business processes. The following approaches aim to support the analysis and design of knowledge within business processes:

- CommonKADS methodology (Schreiber et al. 2000)
- The business knowledge management approach (Bach et al. 1999)
- The knowledge value chain approach (Weggemann 1998)
- The building block approach (Probst et al. 1998)
- The model-based knowledge-management approach (Allweyer 1998)
- The reference model for knowledge management (Warnecke et al. 1998).

None of the approaches presented to knowledge management has been developed from scratch. Their origins range from KBS development and information systems design to intranet development and business process reengineering. Depending on their original focus, the approaches still show their current strengths within these particular areas. However, detailed criteria for the analysis and design of knowledge management are generally missing.

Due to their strong link to information system design, all approaches focus almost exclusively on explicit and documented knowledge as unstructured information. Their design scope is mainly limited to technology-driven solutions. This is surprising because the analysis of 30 knowledge work-improvement projects suggests a modified use of traditional business process design approaches and methods including nontechnical design strategies (Davenport et al. 1996). Only the business knowledge management approach (Bach et al. 1999) covers aspects such as roles and measurements.

5.5. A Method for Business Process-Oriented Knowledge Management

Since the late 1980s, the division of Corporate Management at the Fraunhofer Institute for Production Systems and Design Technology (Fraunhofer IPK) has developed the method of integrated enterprise modeling (IEM) to describe, analyze, and design processes in organizations (Figure 30) (Spur et al. 1993). Besides traditional business process design projects, this method has been used and customized for other planning tasks such as quality management (Mertins and Jochem 1999) (Web and process-based quality manuals for ISO certification) for the design and introduction of process-based controlling in hospitals and benchmarking. The IEM method is supported by the software tool MO^{2}GO (Methode zur objektorientierten Geschäftsprozessoptimierung—method for object-oriented business process optimization).

The method of integrated enterprise modeling (IEM) distinguishes between the three object classes "product," "order," and "resource." These object classes are combined by the construct "Action" within a generic activity model. Five elements are provided to link the objects to the actions (Figure 31). The IEM approach offers the possibility of describing knowledge as an object within the process model. According to the overall modeling task, knowledge can be modeled as a subclass of the superordinated class "resource" and broken down into different sub-subclasses in the form of knowledge domains. The subclass "knowledge" can be linked to other "resource" subclasses such as "staff," "EDP-Systems," "Databases," "Documents," and so on that are relevant for the analysis and improvement of the business process. The final objective of every business process consists of the fulfillment of the internal and/or external customer demand with a product and/or service. Knowledge is required to produce and/or deliver this service or/and product and thus becomes implemented in the object "product." This implemented knowledge could be divided into subclasses as well. The object "order" that triggers the actions could be translated into knowledge goals if appropriate.

The business-oriented knowledge-management approach starts with the selection of the business process to be improved. Davenport (1996) characterizes knowledge work processes as possessing a

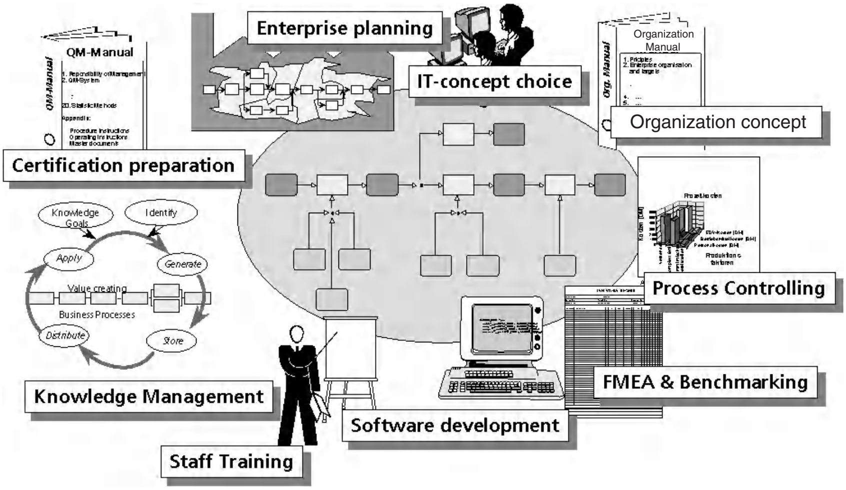

Enterprise planning
IT-concept choice
QM-Manual
Organization Manual
Organization concept
Certification preparation
Knowledge Goals
Identify
Apply
Generate
Value creating Business Processes
Distribute
Store
Process Controlling
FMEA & Benchmarking
Knowledge Management
Software development
Staff Training

Figure 30 Application Fields of the IEM Business Process Models.

high degree of variety and exception rather than routine and requiring a high level of skills and expertise. The description of the real-world business process is carried out with the modeling constructs of the IEM method. After the description of the real-world business process, the analysis starts with the evaluation of each business task. The result is a knowledge activity profile that shows the level and quality of support provided by the current operational task towards the individual core

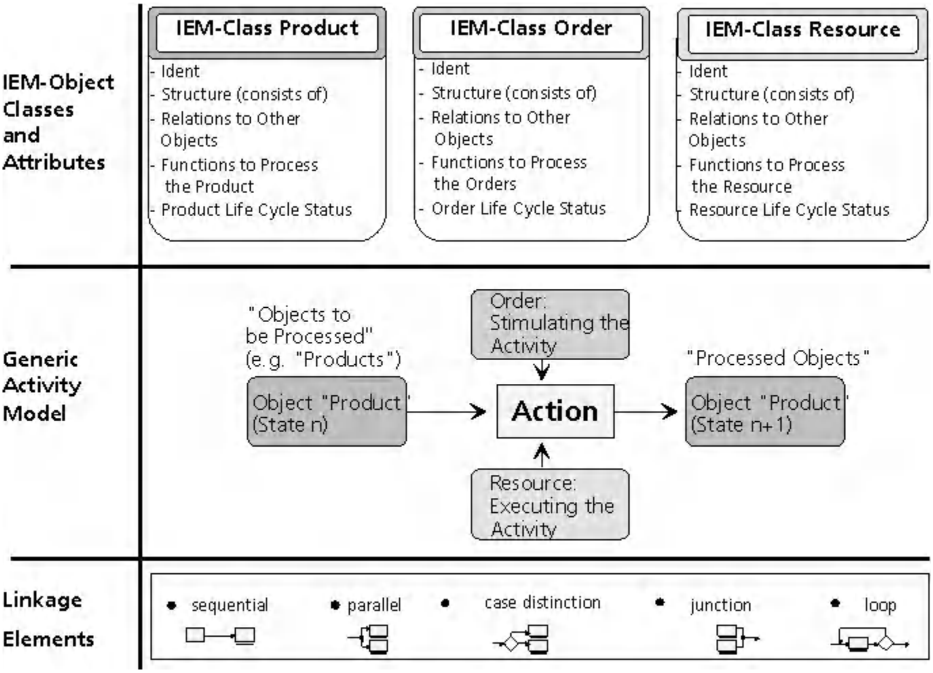

IEM-Object Classes and Attributes
IEM-Class Product
- Ident
- Structure (consists of)
- Relations to Other Objects
- Functions to Process the Product
- Product Life Cycle Status
IEM-Class Order
- Ident
- Structure (consists of)
- Relations to Other Objects
- Functions to Process the Orders
- Order Life Cycle Status
IEM-Class Resource
- Ident
- Structure (consists of)
- Relations to Other Objects
- Functions to Process the Resource
- Resource Life Cycle Status
Generic Activity Model
"Objects to be Processed" (e.g. "Products")
Object "Product" (State n)
Order: Stimulating the Activity
Action
Resource: Executing the Activity
"Processed Objects"
Object "Product" (State n+1)
Linkage Elements
sequential
parallel
case distinction
junction
loop

Figure 31 The Object Classes, Generic Activity Model, and Linking Elements of IEM.

tasks of knowledge management. The scope is then extended towards the analysis of the relations between the knowledge-processing tasks within the business process. This step contains the evaluation of the degree of connectivity inherent in the core process activities of knowledge management within the selected business process. The result shows whether the business processes supporting knowledge management are connected in a coherent manner. The optimization and new design of business processes aim at closing the identified gaps within the underlying core processes and sequencing the core tasks of knowledge management. One design principle is to use available procedures, methods, tools, and results from the process to design the solution. In the last step of the analysis, the focus shifts from the actions towards the resources used and the results produced within the process. The results of the analysis not only demonstrate which kind of knowledge is applied, generated, stored, and distributed but also the other resources, such as employees, databases, and documents. Due to the project aim of the improvement, the user will be able to evaluate whether the required knowledge is explicitly available or available only via the internal expert using the expert's implicit or tacit knowledge. The identified weaknesses and shortcomings in the business process will be addressed by knowledge-management building blocks consisting of process structures. The improvement measures have to integrate not only actions directed to a better handling of explicit knowledge but elements to improve the exchange of implicit knowledge.

5.6. Knowledge-Management Tools

Information technology has been identified as one important enabler of knowledge management. Nevertheless, transfer of information and knowledge occurs primarily through verbal communication. Empirical results show that between 50% and 95% of information and knowledge exchange is verbal (Bair 1998). Computer-based tools for knowledge management improve only a part of the exchange of knowledge in a company. The richness and effectiveness of face-to-face communication should not be underestimated. Computer tools promote knowledge management. The access to knowledge they enable is not subject to time or place. A report can be read in another office a second or a week later.

Therefore, a broad definition of knowledge-management tools would include paper, pencils, and techniques such as brainstorming. According to Ruggles (1997, p. 3), "knowledge management tools are technologies, which automate, enhance and enable knowledge generation, codification and transfer. We do not look at the question if tools are augmenting or automating the knowledge work."

E-mail and computer videoconference systems can also be understood as tools for knowledge management. However, we consider this kind of software to be the basic technology, that is, the building blocks for a knowledge-management system. Initially, groupware and intranets are only systems for the management of information. They become knowledge-management tools when a structure, defined processes, and technical additions are included, such as a means of evaluation by users.

This is not the place for a discussion about whether software can generate, codify, and transfer knowledge alone or can only aid humans in these activities. For the success of knowledge management, the social aspects of its practical use are very important. For example, a sophisticated search engine alone does not guarantee success as long as the user is not able to search effectively. It is not important for this study whether employees are supported in their knowledge management or whether the tool generates knowledge automatically on its own. This is an important point in the artificial intelligence discussion, but we do not need to go into detail here.

Syed (1998) adopts a classification from Hoffmann and Patton (1996) that classifies knowledge techniques, tools, and technologies along the axes complexity–sophistication and intensity along the human–machine continuum, indicating whether certain tools can handle the complexity of the knowledge in question and what kind of workload this means for the user (Figure 32).

5.6.1. Technologies for Knowledge Management

The following is an overview of the basic technologies used in every knowledge-management solution. The following explanation of the basic technologies helps to examine and classify tools more precisely. These are the technologies that we find today in knowledge management (Bottomley 1998). Different knowledge management tasks can be processed using these basic technologies.

Intranet technology: Intranets and extranets are technologies that can be used to build a knowledge-management system. The unified surface and access to various sources of information make this technology perfect for the distribution of knowledge throughout a company.

Groupware: Groupware is a further substantial technology that is used for knowledge-management systems (Tiwana 2000). Groupware offers a platform for communication within a firm and cooperation between employees.

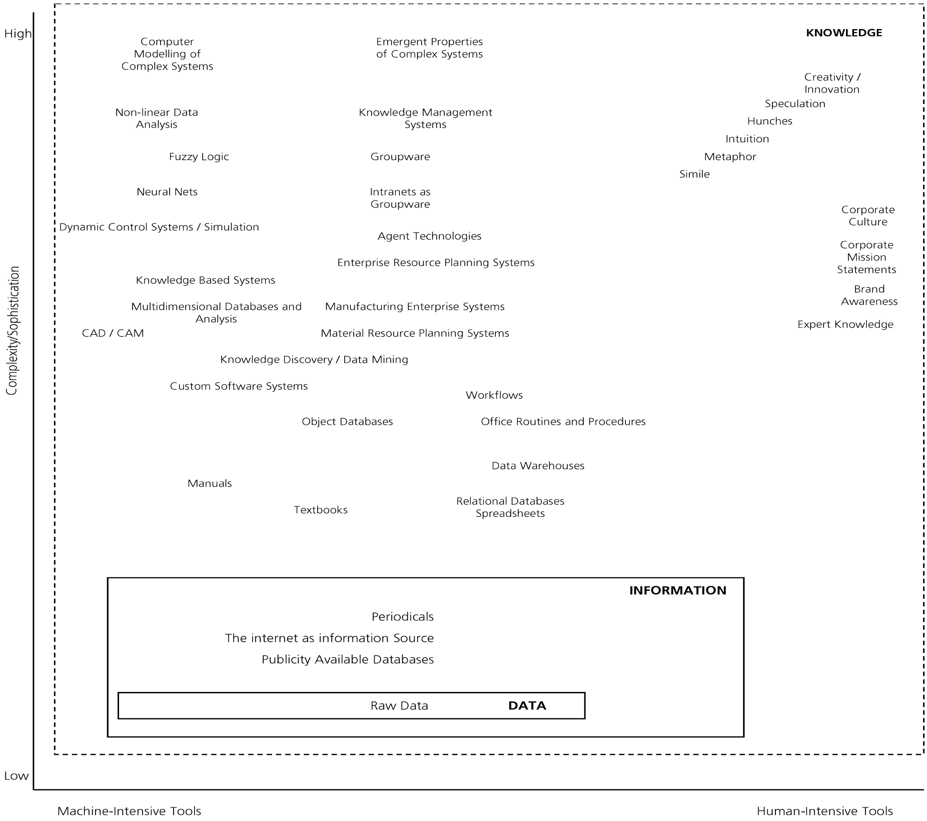

High
Low
Complexity/Sophistication
KNOWLEDGE
Computer Modelling of Complex Systems
Emergent Properties of Complex Systems
Creativity / Innovation
Speculation
Hunches
Intuition
Metaphor
Simile
Non-linear Data Analysis
Knowledge Management Systems
Fuzzy Logic
Groupware
Neural Nets
Intranets as Groupware
Corporate Culture
Dynamic Control Systems / Simulation
Agent Technologies
Corporate Mission Statements
Enterprise Resource Planning Systems
Knowledge Based Systems
Brand Awareness
Multidimensional Databases and Analysis
Manufacturing Enterprise Systems
Expert Knowledge
CAD / CAM
Material Resource Planning Systems
Knowledge Discovery / Data Mining
Custom Software Systems
Workflows
Object Databases
Office Routines and Procedures
Data Warehouses
Manuals
Relational Databases Spreadsheets
Textbooks
INFORMATION
Periodicals
The internet as information Source
Publicity Available Databases
Raw Data
DATA
Machine-Intensive Tools
Human-Intensive Tools

Figure 32 Knowledge Techniques, Tools, and Technologies. (From Syed 1998, p. 65, adapted from Hoffmann and Patton 1996 © 1996 by SRI Consulting, Business Intelligence Program. Reprinted by permission of SRI Consulting Business Intelligence.)

Electronic document management: Documents are a central means of storing and spreading knowledge. Procedures for using and maintaining such documents, such as a check whether an update is overdue, can be easily implemented for knowledge management systems.

Information-retrieval tools: Information retrieval offers a solution to tasks from text searches to the automatic categorization and summation of documents. Advanced search algorithms use thesauri and text mining to discover contexts that could not be found with simple queries. Semantic text analyses can also be implemented.

Workflow-management system: The business processes of a company contains a large part of knowledge. In addition, the integration of knowledge management into business processes is an important factor for success.

Data analysis: Pattern recognition and classification and forecasting are the techniques used for data analysis. Data analysis is a possible method for generating new knowledge.

Data warehousing: A modern database is where data and information are stored. Connections that are not readily apparent can be uncovered with the use of data mining and OLAP. These techniques are part of data analysis.

Agent technology: Software agents based on the essentials of artificial intelligence enable the user to search independently for information according to a personal profile and use various sources and other agents.

Help desks: Help desks are an important application area for case-based reasoning technology based on individual cases. Case knowledge can be quickly put into use in this way.

Machine learning: This technology from the field of artificial intelligence allows new knowledge to be generated automatically. In addition, processes can be automatically optimized with time with little necessity for human intervention.

Computer-based training: This technology is used to pass on knowledge to colleagues. The spread of implicit knowledge is possible with multimedia applications.

5.6.2. Architecture for Knowledge Management

Historical classification explains the special use of a certain product or how the manufacturer understands its use. The following historical roots are relevant (Bair and O'Connor 1998):

- Tools that are further developments of classical information archives or the retrieval of information.
- Solutions from the field of communication and reactivated concepts from the field of artificial intelligence come into play in the analysis of documents and in automated searches.
- Approaches to saving and visualizing knowledge also come from classical information archives.
- Tools for modeling business processes.
- Software that attempts to combine several techniques and support different tasks in knowledge management equally.

The Ovum (Woods and Sheina 1998) approach is an example of a well-structured architectural model. The initial level of the model consists of information and knowledge sources (Figure 33). These are delivered to the upper levels through the infrastructure. Next comes the administration level for the knowledge base where the access control is handled, for example. The corporate taxonomy defines important knowledge categories within the company. The next layer makes services available for the application of knowledge, such as through visualizing tools, and for collaboration, such as through collaborative filtering. The user interface is described as a portal through which the user can access the knowledge to use it in an application.

A further possibility is categorization according to the basic technology from which knowledge-management systems are constructed.

Most knowledge-management tools use existing technologies to provide a collaborative framework for knowledge sharing and dissemination. They are implemented using e-mail and groupware, intranets, and information-retrieval and document-management systems. Applications from data warehousing to help desks can be used to improve the quality of knowledge management (Woods and Sheina 1998).

5.7. Future Trends

Knowledge management is currently a buzzword on the agenda of top management and marketing and sales departments of software providers and consulting companies. Nevertheless, decision maker's awareness is increasing. Knowledge is regarded as one or even the main factor for private and public organizations to gain competitive advantage. First experience from knowledge-management projects show that a win-win situation for companies and employees is possible. By reducing double work, the company saves costs. Moreover, employees increase their experience through continuous learning and their satisfaction through solving new problems and not reinventing the wheel.

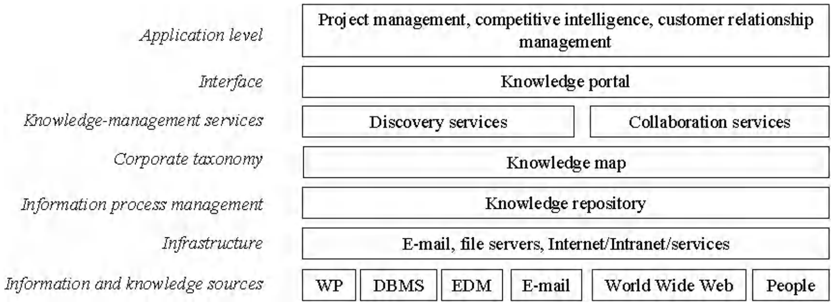

Application level
Project management, competitive intelligence, customer relationship management
Interface
Knowledge portal
Knowledge-management services
Discovery services
Collaboration services
Corporate taxonomy
Knowledge map
Information process management
Knowledge repository
Infrastructure
E-mail, file servers, Internet/Intranet/services
Information and knowledge sources
WP
DBMS
EDM
E-mail
World Wide Web
People

Figure 33 Ovum Knowledge-Management Tools Architectural Model. (From Woods and Sheina 1998, p. 8)

Even if knowledge management goes out of fashion as a buzzword, the essence of the idea of knowledge management—the systematic management of knowledge and experiences of the employees, suppliers, and clients—will definitely never be superfluous. Even in the dynamic new economy, experience and knowhow are still extremely important, as the example of retired experts who are very happy to pass their knowledge as "senior experts" on to young start-up companies shows.

With business process engineering, companies have focused their attention on eliminating non-value-adding process steps in order to improve their competitiveness by means of value-adding process steps. In the future, companies will regard their business processes as knowledge processing processes and enhance their ability to improve their use of their one and only competitive advantage—the knowhow of the people.

In the future, the basic technologies will be standard and increasingly integrated. This will require the integration of knowledge management in everyday business as well as a sense of responsibility from each individual. This will result in knowledge management becoming less discussed and more and more taken for granted.

For the exchange of knowledge to continue to improve, meta-knowledge will become increasingly important. Meta-knowledge helps to describe knowledge and specify its meaning. This form of description will have to be standardized due to the growing need for a global exchange of knowledge. This is apparent in the increased importance of Internet markets, which require a globally accepted description of products and thus knowledge.

In the IT industry, the dominant trend is toward takeovers of small, innovative companies. The market leader can then integrate new technologies into its standard product and advance that product.

Another future trend is toward knowledge management being practiced not only within companies, but between them. The Internet will reinforce the trend toward small companies being able to benefit more from the exchange of knowledge with other companies. However, some companies are becoming disillusioned regarding the use of new technologies. For example, intranet technology is taken for granted as a medium nowadays, although there is still uncertainty about what kinds of information it should be used to transfer to yield the maximum benefits. Despite continuing technological improvements in the future, people will still remain the definitive force.

REFERENCES

Allweyer, T. (1998), "Modellbasiertes Wissensmanagement" in *Information Management,* Vol. 1, pp. 37–45.

Anderl, R. (1989), "Normung von CAD-Schnittstellen," *CIM Management,* Vol. 5, No. 1, pp. 4–8.

Anderl, R. (1993), *CAD-Schnittstellen, Methoden und Werkzeuge zur CA-Integration,* Carl Hanser, Munich.

Bach, V., Vogler, P., and Österle, H., Eds. (1999), "Praxiserfahrungen mit Intranet-basierten Lösungen," Springer, Berlin.

Bair, J. (1998), "Developing a Scalable Knowledge Management Architecture and IT Strategy," *Proceedings of Building the Knowledge Management Framework: The New IT Imperative* (London, July 1–2), Business Intelligence.

Bair, J. H., and O'Connor, E. (1998), "The State of the Product in Knowledge Management," *Journal of Knowledge Management,* Vol. 2, No. 2, pp. 20–27.

Boender, E. (1992), "Finite Element Mesh Generation from CSG Models," Dissertation, Delft University of Technology, Faculty of Technical Mathematics and Informatics.

Bottomley, A. (1998), *Enterprise Knowledge Management Technologies: An Investment Perspective 1998,* Durlacher Research, London.

BRITE-EURAM (1997), Deliverable D2, *Functional Specification for Prototype Stream: Digital Mock-up Modeling Methodologies* (*DMU-MM*) *for Product Conception and Downstream Processes*. Project BPR-CT95-0005.

Bromme, R. (1992), *Der Lehrer als Experte: Zur Psychologie des professionellen Wissens,* Verlag Hans Huber, Bern.

Bullinger, H.-J., and Warschat, J. (1996), *Concurrent Simultaneous Engineering Systems,* Springer, Berlin.

Burghardt, M. (1988), *Projektmanagement,* Siemens AG.

CAD-FEM GmbH (1995), *Infoplaner August 95–Juli 96,* CAD-FEM, Grafing.

Carbon, M., and Heisig, P. (1993), "Verbesserung der Prozeβtransparenz durch konstruktive Veränderungen, Flexibilität durch Erfahrung: Computergestützte erfahrungsgeleitete Arbeit in der Produktion, A. Bolte and H. Martin, Eds., Verlag Institut für Arbeitswissenschaft, Kassel, pp. 71–77.

Davenport, T. H., and Prusak, L. (1998), *Working Knowledge: How Organizations Manage What They Know,* Harvard Business School Press, Boston.

Davenport, T. H., Jarvenpaa, S. L., and Beers, M. C. (1996), "Improving Knowledge Work Processes," *Sloan Management Review,* Summer, pp. 53–65.

De Hoog, R. (1997), "CommonKADS: Knowledge Acquisition and Design Support Methodology for Structuring the KBS Integration Process," in *Knowledge Management and Its Integrative Elements,* J. Leibowitz and L. C. Wilcox, Eds., CRC Press, Boca Raton, FL, 129–141.

Düring, H., and Dupont, P. (1993), "ProSTEP: Baukasten für kompatible und flexible Softwarelösungen," *CAD-CAM Report,* Vol. 12, No. 7, pp. 58–61.

Eck, C. D. (1997), Wissen: ein neues Paradigma des Managements. *Die Unternehmung,* Vol. 3, pp. 155–179.

Eversheim, W., Schuh, G., and Caesar, C. (1989), "Beherrschung der Variantenvielfalt, Methoden und Hilfsmittel," *VDI-Z,* Vol. 131, No. 1, pp. 42–46.

Eversheim, W., Bochtler, W., and Laufenberg, L. (1995), *Erfahrungen aus der Industrie, für die Industrie,* Springer, Berlin.

Fleig, J., and Schneider, R. (1995), *Erfahrung und Technik in der Produktion,* Springer, Berlin.

Forkel, M. (1994), *Kognitive Werkzeuge: ein Ansatz zur Unterstützung des Problemlösens,* Carl Hanser, Munich.

Frei, B. (1993), "Dynamisches Modellieren erleichtert Entwurfsprozeß," *CAD-CAM Report,* Vol. 12, No. 5.

Frepoli, C., and Botta, E. (1996), "Numerical Simulation of the Flow Field in a Turbo-Valve," *Simulation in Industry,* Vol. 1, pp. 444–446.

Gebhardt, A., and Pflug, T. K. (1995), "Innovationsschub mittels Rapid-Prototyping, Teil 2," *CAD-CAM Report,* Vol. 14, No. 9, pp. 78–85.

Göbler, T. (1992), *Modellbasierte Wissensakquisition zur rechnerunterstützten Wissensbereitstellung für den Anwendungsbereich Entwicklung und Konstruktion,* Carl Hanser, Munich.

Gödicke, P. (1992), "Wissensmanagement: aktuelle Aufgaben und Probleme," *Management Zeitschrift,* Vol. 61, No. 4, pp. 67–70.

Golm, F. (1996), "Gestaltung von Entscheidungsstrukturen zur Optimierung von Produktenwicklungsprozessen," Dissertation, Technical University of Berlin, Reihe Berichte aus dem Produktionstechnischen Zentrum Berlin, UNZE, Potsdam.

Grabowski, H., and Anderl, R. (1990), *Produktdatenaustausch und CAD-Normteile,* Expert, Ehingen.

Grabowski, H., and Glatz, R. (1986), "Schnittstellen zum Austausch produktdefinierender Daten," *VDI-Z,* Vol. 128, No. 10, pp. 333–343.

Grätz, J.-F. (1989), *Handbuch der 3D-CAD-Technik: Modellierung mit 3D-Volumensystemen,* Siemens AG, Berlin.

Hammer, M., and Champy, J. (1993), *Reengineering the Corporation,* HarperBusiness, New York.

Hartung, J., and Elpet, B. (1986), *Multivariate Statistik: Lehr-und Handbuch der angewandten Statistik,* Oldenbourg, Munich.

Heisig, P. (2000), "Benchmarking Knowledge Management und wissensorientierte Gestaltung von Geschäftsprozessen. *Organisation: Schlank—Schnell—Flexibel,* R. Bühner, Ed., Verlag Moderne Industrie, Landsberg/Lech, pp. 1–38.

Heisig, P., and Vorbeck, J. (2001), "Benchmarking Survey Results," in *Knowledge Management: Best Practices in Europe,* K. Mertins, P. Heisig, and J. Vorbeck, Eds., Springer, Berlin, pp. 97–123.

Hess, T., and Brecht, L. (1995), *State of the Art des Business Process Redesign: Darstellung und Vergleich bestehender Methoden,* Gabler, Wiesbaden.

Holland, M., and Machner, B. (1995), "Product Data Management on the Basis of ISO 10303 (Produktdatenmanagement auf der Basis von ISO 10303-STEP)," *CIM Management,* Vol. 11, No. 4, pp. 32–40.

Holsapple, C. W., and Whinston, A. B. (1992), "Decision Support Systems," in *Handbook of Industrial Engineering,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 109–141.

ISO 10303-1 (1994), Industrial Automation Systems and Integration—Product Data Representation and Exchange—Part 1: Overview and Fundamental Principles.

Jablonski, S. (1995), *Workflow Management Systems,* International Thomson Publishing, Bonn.

Kiesewetter, T. (1997), "Integrativer Produktentwicklungsarbeitsplatz mit Multimedia- und Breitbandkommunikationstechnik," Dissertation, Technical University of Berlin, Berlin, Berichte aus dem Produktionstechnischen Zentrum Berlin, FhG/IPK, Berlin.

Knothe, K., and Wessels, H. (1992), Finite Elemente, Springer, Berlin.

Kohlhoff, S., Bläßer, S., and Maurer, D. (1994), *Ein rechnerischer Ansatz zur Untersuchung von Fahrzeug-Fahrzeug-Frontalkollisionen zur Entwicklung von Barrierentests,* VDI-Berichte No. 1153, VDI, Düsseldorf.

König, W., Celi, I., Celiker, T., Herfurth, H.-J., and Song, Y. (1994), "Rapid Metal Prototyping" *VDI-Z,* Vol. 136, No. 7/8, pp. 57–60.

Krause, F.-L., Jansen, H., Bienert, M., and Major, F. (1990), "System Architectures for Flexible Integration of Product Development," in *Proceedings of the IFIP WG 5.2/GI International Symposium* (Berlin, November 1989), Elsevier Science Publishers, Amsterdam, pp. 421–440.

Krause, F.-L. (1992), "Leistungssteigerung der Produktionsvorbereitung," in *Proceedings of the Produktionstechnischen Kolloquiums* (*PTK*) *Berlin Markt, Arbeit und Fabrik* (Berlin), pp. 166–184.

Krause, F.-L. (1996), "Produktgestaltung," in *Betriebshütte "Produktion und Management,"* 7th Ed., W. Eversheim and G. Schuh, Eds., Springer, Berlin, pp. 7-34–7-73.

Krause, F.-L., and Pätzold, B. (1992), *Automotive CAD/CAM-Anwendungen im internationalen Automobilbau, Teil 1: Japan,* Gemeinschaftsstudie von IPK-Berlin und Daimler Benz F&E.

Krause, F.-L., Bock, Y., and Rothenburg, U. (1999), "Application of Finite Element Methods for Digital Mock-up Tasks," in *Proceedings of the 32nd CIRP International Seminar on Manufacturing Systems: New Supporting Tools for Designing Products and Production Systems* (Leuven, May 24–26), H. Van Brussel, J.-P. Ruth, and B. Lauwers, Eds., Katholieke Universtiteit Leuven, Hevelee, pp. 315–321.

Krause, F.-L., Ciesla, M., and Stephan, M. (1994), "Produktionsorientierte Normung von STEP," *ZwF,* Vol. 89, No. 11, pp. 536–539.

Krause, F.-L., Jansen, H., and Vollbach, A. (1996), "Modularität von EDM-Systemen," *ZwF,* Vol. 91, No. 3, pp. 109–111.

Krause, F.-L., Ulbrich, A., and Mattes, W. (1993), "Steps towards Concurrent Engineering," in *Proceedings & Exhibition Catalog, ILCE'93* (Montpellier, March 22–26), pp. 37–48.

Kriwet, A. (1995), "Bewertungsmethodik für die recyclinggerechte Produktgestaltung," Dissertation, Technical University of Berlin, Reihe Produktionstechnik—Berlin, Vol. 163, Carl Hanser, Munich.

Kruth, J. P. (1991), "Material Incress Manufacturing by Rapid Prototyping Techniques," *Annals of the CIRP,* Vol. 40, No. 2, pp. 603–614.

Luhn, G. (1999), *Implizites Wissen und technisches Handeln am Beispiel der Elektronikproduktion,* Meisenbach, Bamberg.

Martin, H. (1995), *CeA: Computergestützte erfahrungsgeleitete Arbeit,* Springer, Berlin.

McKay, A., Bloor, S., and Owen, J. (1994), "Application Protocols: A Position Paper," in Proceedings of European Product Data Technology Days, *International Journal of CAD/CAM and Computer Graphics,* Vol. 3, No. 9, pp. 377–338.

Mertins, K., Heisig, P., and Vorbeck, J. (2001), Knowledge Management. Best Practice in Europe. Springer, Berlin.

Mertins, K., and Jochem, R. (1999), *Quality-Oriented Design of Business Processes,* Kluwer Academic Publishers, Boston.

Mertins, K., Schallock, B., Carbon, M., and Heisig, P. (1993), "Erfahrungswissen bei der kurzfristigen Auftragssteuerung," *Zeitschrift für wirtschaftliche Fertigung,* Vol. 88. No. 2, pp. 78–80.

Nonaka, I., and Takeuchi, H. (1995), *The Knowledge-Creating Company,* Oxford University Press, Oxford.

Nowacki, H. (1987), "Schnittstellennormung für gegenstandsdefinierenden Datenaustausch," *DIN-Mitteilungen,* Vol. 66, No. 4, pp. 182–186.

Pahl, G., and Beitz, W. (1993), *Konstruktionslehre: Methoden und Anwendungen,* 3rd Ed., Springer, Berlin.

Pawellek, G., and Schulte, H., "Logistikgerechte Konstruktion: Auswirkungen der Produktgestaltung auf die Produktionslogistik," *Zeitschrift für Logistik,* Vol. 8, No. 9.

Ploenzke A. G. (1994), *Engineering Data Management Systems, 3. Edition,* Technology Report, Ploenzke Informatik—Competence Center Industrie, Kiedrich/Rheingau.

Ploenzke A. G. (1997), "Mit der Schlüsseltechnologie EDM zum Life Cycle Management," in *Proceedings of the CSC PLOENZKE Congress* (Mainz).

Polanyi, M. (1966), *The Tacit Dimension,* Routledge and Kegan Paul, London.

Probst, G., Raub, S., and Romhardt, K. (1998), *Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen,* 2nd Ed., Frankfurt Allgemeine Zeitung GmbH, Thomas Gabler, Frankfurt am Main.

Rainer, G. (1992), "IGES: Einsatz bei firmenübergreifenden Entwicklungskonzepten," *CAD-CAM Report,* Vol. 11, No. 5, pp. 132–140.

Rathnow, P. J. (1993), *Integriertes Variantenmanagement,* Vandenhoeck & Ruprecht, Göttingen.

Reinwald, B. (1995), "Workflow-Management in verteilten Systemen," Dissertation, University of Erlangen, 1993, Teubner-Texte zur Informatik, Vol. 7, Teubner, Stuttgart.

Ruggles, R. L. (1997), *Knowledge Management Tools,* Butterworth-Heinemann, Boston.

Scheder, H. (1991), "CAD-Schnittstellen: ein Überblick," *CAD-CAM Report,* Vol. 10, No. 10, pp. 156–159.

Schönbach, T. (1996), *Einsatz der Tiefziehsimulation in der Prozeßkette Karosserie,* VDI-Berichte No. 1264, VDI, Düsseldorf, pp. 405–421.

Scholz-Reiter, B. (1991), *CIM-Schnittstellen: Konzepte, Standards und Probleme der Verknüpfung von Systemkomponenten in der rechnerintegrierten Produktion,* 2nd Ed., Oldenbourg, Munich.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van de Welde, W., and Wielinga, B. (2000), *Knowledge Engineering and Management: The CommonKADS Methodology,* MIT Press, Cambridge.

Skyrme, D. J., and Amidon, D. M. (1997), *Creating the Knowledge-Based Business,* Business Intelligence, London.

Spur, G., and Krause, F.-L. (1984), *CAD-Technik,* Carl Hanser, Munich.

Spur, G., Mertins, K., and Jochem, R. (1993), *Integrierte Unternehmensmodellierung,* Beuth, Berlin.

Stanek, J. (1989), "Graphische Schnittstellen der offenen Systeme," *Technische Rundschau,* Vol. 81, No. 11, pp. 38–43.

Stephan, M. (1997), "Failure-Sensitive Product Development," in *Proceedings of the 1st IDMME Conference: Integrated Design and Manufactureng in Mechanical Engineering* (Nantes, April 15–17, 1996), P. Chedmail, J.-C. Bocquet, and D. Dornfeld, Eds., Kluwer Academic Publishers, Dordrecht, pp. 13–22.

Syed, J. R. (1998), "An Adaptive Framework for Knowledge Work," *Journal of Knowledge Management,* Vol. 2, No. 2, pp. 59–69.

Szabo, B. A. (1994), "Geometric Idealizations in Finite Element Computations," *Communications in Applied Numerical Methods,* Vol. 4, No. 3, pp. 393–400.

Tiwana, A. (2000), *The Knowledge Management Toolkit,* Prentice Hall, Upper Saddle River, NJ.

VDI Gesellschaft Entwicklung Konstruktion Vertrieb (BDI-EKV) (1992), "Wissensbasierte Systeme für Konstruktion und Arbeitsplanung," Gesellschaft für Informatik (GI), VDI, Düsseldorf.

VDI-Gesellschaft Entwicklung Konstruktion (1993), *VDI-Handbuch Konstruktion, Methodik zum Entwickeln und Konstruieren technischer Systeme und Produkte,* VDI, Düsseldorf.

Victor, B., and Boynton, A. C. (1998), *Invented Here: Maximizing Your Organization's Internal Growth and Profitability,* Harvard Business School Press, Boston.

von Krogh, G., and Venzin, M. (1995), "Anhaltende Wettbewerbsvorteile durch Wissensmanagement." *Die Unternehmung,* Vol. 1, pp. 417–436.

Wagner, M., and. Bahe, F. (1994), "Austausch von Produktmodelldaten: zukunftsorientierte Prozessoren auf Basis des neuen Schnittstellenstandards STEP," in *CAD '94: Produktdatenmodellierung und Prozeßmodellierung als Grundlage neuer CAD-Systeme, Tagungsband der Fachtagung der Gesellschaft für Informatik* (Paderborn, March 17–18), J. Gausemeier, Ed., Carl Hanser, Munich, pp. 567–582.

Warnecke, G., Gissler, A., and Stammwitz, G. (1998), "Wissensmanagement: Ein Ansatz zur modellbasierten Gestaltung wissensorientierter Prozesse," *Information Management,* Vol. 1, pp. 24–29.

Weck, M., and Heckmann, A. (1993), "Finite-Elemente-Vernetzung auf der Basis von CAD-Modellen," *Konstruktion,* Vol. 45, No. 1, pp. 34–40.

Weggeman, M. (1998), *Kenntnismanagement: Inrichtig en besturing van kennisintensieve organisaties,* Scriptum, Schiedam.

Wiig, K. M. (1995), *Knowledge Management Methods: Practical Approaches to Managing Knowledge,* Vol. 3, Schema Press, Arlington, TX.

Wiig, K. M. (1997), "Knowledge Management: Where Did It Come from and Where Will It Go?," *Expert Systems with Applications,* Vol. 13, No. 1, pp. 1–14.

Willke, H. (1998), *Systemisches Wissensmanagement,* Lucius & Lucius, Stuttgart.

Woods, M. S., and Sheina M. (1998), *Knowledge Management: Applications, Markets and Technologies,* Ovum, London.

CHAPTER 7

Computer Networking

LAJOS BÁLINT
HUNGARNET-NIIFI

TAMÁS MÁRAY
Technical University of Budapest

1. INTRODUCTION 228
2. THE ROLE OF COMPUTER NETWORKING 229
  2.1. Information Access 230
  2.2. Information Provision 232
  2.3. Electronic Communication 232
  2.4. Networked Collaboration 234
  2.5. Virtual Environments 234
3. A SHORT HISTORY OF COMPUTER NETWORKING 235
4. THE NETWORKING INFRASTRUCTURE 236
5. INTERNET, INTRANETS, EXTRANETS 237
6. TECHNICAL BACKGROUND 238
  6.1. Architecture of the Internet 238
  6.2. Packet Switching 239
  6.3. Most Important Protocols 239
  6.4. Client–Server Mechanism 240
  6.5. Addressing and Naming 241
7. OVERVIEW OF NETWORKING SERVICES 243
8. OVERVIEW OF NETWORK-BASED APPLICATIONS 243
9. THE WORLD WIDE WEB 244
  9.1. History 244
  9.2. Main Features and Architecture 245
  9.3. HTTP and HTML 245
  9.4. Multimedia Elements 246
10. THE ROLE OF THE WORLD WIDE WEB IN COMMUNICATION, INTEGRATION, AND COLLABORATION 246
  10.1. The World Wide Web as a Means for Universal Information Access 246
  10.2. The World Wide Web as a Tool for Communicating and Exchanging Information 246
  10.3. Collaboration and Integration Supported by the World Wide Web 246
11. CONTENT GENERATION AND CONTENT PROVISION 246
  11.1. Electronic I/O, Processing, Storage, and Retrieval of Multimedia Information 247
  11.2. Classification of Electronically Accessible Information Content 247
  11.3. Rating and Filtering in Content Generation, Provision, and Access 248
12. TRENDS AND PERSPECTIVES 249
  12.1. Network Infrastructure 249
  12.2. Services 250
  12.3. Applications 250
  12.4. Information Content 251
13. NETWORKING IN PRACTICE 252
  13.1. Methodology 252
  13.2. Classification 252
  13.3. Implementation Issues 253

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

14. NETWORKING IN THE PRODUCTION AND SERVICE INDUSTRIES 253

15. SOME PRACTICAL ASPECTS OF INTRODUCING AND USING COMPUTER NETWORKS IN INDUSTRIAL ENGINEERING 254

15.1. Internet Connectivity 254

15.2. LANs, Intranets, WANs, and Extranets 255

15.3. World Wide Web Communication, Integration, and Collaboration 256

16. SUMMARY 257

REFERENCES 257

1. INTRODUCTION

Computer networks serve today as crucial components of the background infrastructure for virtually all kinds of human activities. This is the case with Industrial Engineering and any related activities.

The importance of these computer networks stems from their role in communication and information exchange between humans (as actors in any kind of activities) and/or organizations of these human actors.

In contrast to the situation several decades ago, a number of key attributes of the working environment today serve as basic new supportive elements of working conditions. Some of the most important such elements are:

- Computer technology has moved to the desktop, making possible direct access to advanced techniques of high-speed information processing and high-volume information storage from individual workstations.
- Interconnection of computers has been developing at an exponentially increasing rate, and today high-speed global connectivity is a reality, providing fast and reliable accessibility of high-volume distant information at any connected site worldwide.
- Global addressing schemes, structured information content, and multimedia information handling have become an efficiently functioning framework for computer-to-computer (and thus human-to-human) communication and information exchange, as well as for storing and delivering large amounts of highly sophisticated information at millions of sites, accessible by any other authorized sites anywhere within the global network of networks, the Internet.

Computer networks and services themselves, as well as related applications, are briefly investigated in this chapter. The topic is very broad and there are thousands of sources in books and other literature providing detailed information about all the aspects of computer networking introduced here. This chapter therefore provides a condensed overview of selected key subtopics. However, emphasis is given to every important element of what the network infrastructure looks like, what the main characteristics of the services and applications are, and how the information content is being built and exploited. Some basic principles and methods of computer networking are repeatedly explained, using different approaches, in order to make these important issues clear from different aspects.

This chapter starts with an introductory description of the role of computer networking in information transfer. Then, after a short historical overview, the networking infrastructure is investigated in some detail, including the basic technical elements of how networking operates in practice. Later sections deal with services and applications based on networked computer systems. Separate subsections are devoted to the World Wide Web, which is probably the most important tool at the start of the third millennium for communication, information access, and collaboration. Some questions of content generation and provision are then investigated, after which aspects of future development in computer networking are briefly dealt with. Other subsections deal with a few practical issues, including some related to industrial engineering.

The following major issues are dealt with in the following references at the end of the chapter:

- *Networking in general:* Derfler and Freed (1998); Hallberg (1999); Keshav (1997); Kosiur (1998); Lynch and Rose (1993); Marcus (1999); Taylor (1999)
- *Internet in general:* Minoli and Schmidt (1999); Quercia (1997); Smythe (1995); Young (1999)
- *History of networking:* Salus (1995)
- *Intranets:* Ambegaonkar (1999); Bernard (1997); Hills (1996); Minoli (1996)
- *Extranets:* Baker (1997); Bort and Felix (1997)
- *World Wide Web:* Stout (1999)

- *Multimedia:* Agnew and Kellerman (1996); Buford (1994); Wesel (1998)
- *Virtual reality:* Singhal and Zyda (1999)
- *Quality of service:* Croll and Packman (1999)
- *Security:* Cheswick and Bellovin (1994); Goncalves (1999); Smith (1997)
- *Applications:* Angell and Heslop (1995); Hannam (1997); Kalakota and Whinston (1997); McMahon and Browne (1998); Treese and Stewart (1998)
- *Practical issues:* Derfler (1998); Dowd (1996); Guengerich et al. (1996); Murhammer et al. (1999); Ptak et al. (1998); Schulman and Smith (1997); Ward (1999)
- *Future trends:* Conner-Sax and Krol (1999); Foster and Kesselman (1998); Huitema (1998); Mambretti and Schmidt (1999)

2. THE ROLE OF COMPUTER NETWORKING

Computer networks (see Figure 1) are frequently considered as analogous to the nervous system in the human body. Indeed, like the nerves, which connect the more or less autonomous components in the body, the links of a computer network provide connections between the otherwise stand-alone units in a computer system. And just like the nerves, the links of the network primarily serve as communication channels between the computers. This is also the basis of the global network of networks, the Internet.

The main task in both cases is information transfer. Connected and thus communicating elements of the system can solve their tasks more efficiently. Therefore, system performance is not just the sum of the performance of the components but much more, thanks to the effect of the information exchange between the components.

Computer networks consist of computers (the basic components of the system under consideration), connecting links between the computers, and such additional things as devices making the information transfer as fast, intelligent, reliable, and cheap as possible.

- Speed depends on the capacity of the transmission lines and the processing speed of the additional devices, such as modems, multiplexing tools, switches, and routers. (A short description of these devices is given in Section 4, together with more details about transmission line capacity.)
- Intelligence of the network depends on processing capabilities as well as stored knowledge of these active devices of the network. (It is worth mentioning that network intelligence is different from the intelligence of the interconnected computers.)
- Reliability stems from, first of all, the decreased risk of losing, misinterpreting, or omitting necessary information, as well as from the well preprocessed character of the information available within the system.

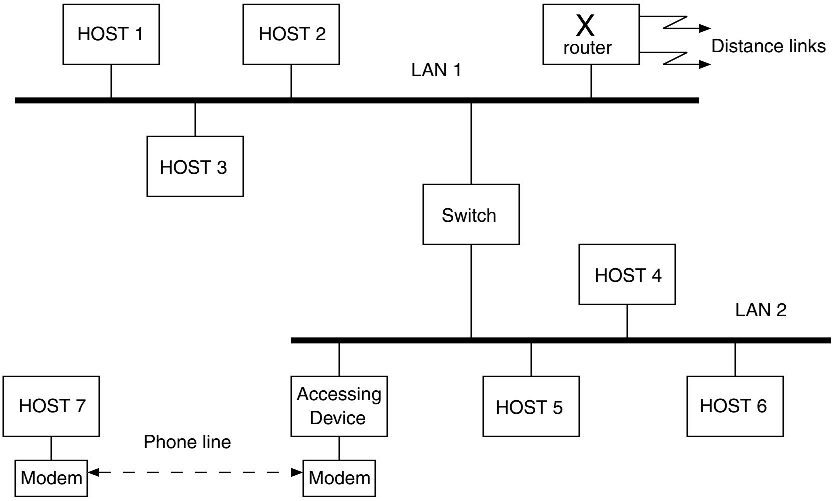


Figure 1 Networks and Network Components.

- Cost is associated with a number of things. The most important direct costs are purchase costs, installation costs, and operational costs. Indirect costs cover a wide range, from those stemming from faulty system operation to those associated with loss of business opportunity. In the planning of the system, the cost/performance ratio must always be taken into consideration.

It is important not to forget that adequate system operation requires adequate operation of all components.

Obviously, adequate operation of each component in a system requires intensive interaction among them. However, direct information transfer between these components makes the interaction much faster and more reliable than traditional means of interaction. This means highly elevated adequacy in the behavior of each component and thus of the system itself.

Another important capability of a system consisting of connected and communicating components is that the intelligence of the individual components is also connected and thus each component can rely not only on its own intelligence but also on the knowledge and problem solving capability of the other components in the system. This means that the connections between the computers in the network result in a qualitatively new level of complexity in the system and its components. The network connections allow each system component to:

- Provide information to its partners
- Access the information available from the other system components
- Communicate while operating (thus, to continuously test, correct, and adapt its own operation)
- Utilize highly sophisticated distant collaboration among the system components
- Exploit the availability of handling virtual environments*

These features are described briefly below. As will be shown below, these basic features pave the way for:

- The systematic construction of network infrastructures
- The continuous evolution of network services
- The rapid proliferation of network-based applications
- The exponentially increasing amount of information content accumulated in the network of interconnected computers

2.1. Information Access

Information access is one of the key elements of what networking provides for the users of a computer connected to other computers. The adequate operation of a computer in a system of interconnected computers naturally depends on the availability of the information determining what, when, and how to perform with it. Traditionally, the necessary information was collected through ‘‘manual’’ methods: through conversation and through traditional information carriers: first paper, and then, in the computer era, magnetic storage devices.

These traditional methods were very inefficient, slow, and unreliable. The picture has now totally changed through computer networking. By the use of the data communication links connecting computers and the additional devices making possible the transfer of information, all the information preprocessed at the partner sites can be easily, efficiently, immediately, and reliably downloaded from the partner computers to the local machine.

The model above works well until individual links are supposed to serve only the neighbor computers in accessing information at each other’s sites. But this was the case only at the very beginning of the evolution of networking. Nowadays, millions of computers can potentially access information stored in all of the other ones. Establishing individual links between particular pairs of these computers would be technically and economically impossible.

The solution to the problem is a hierarchical model (see Figure 2). Very high-speed backbones (core network segments) interconnect major nodes located at different geographical regions, and somewhat lower-speed access lines transfer the information from/to these nodes to/from those servers within the region, which then take care of forwarding/collecting the information traffic in their local

---

* Handling virtual environments means that the individual components don’t communicate continuously with the outside world but systematically use a mirrored synthetic picture about their embedding. This is done by utilizing the information accumulated by their neighbors and thus constructing and using a virtual reality model about the outside world. Of course, direct access to this outside world may provide additional information by which such a virtual reality model can be completed so that the result is a so-called augmented reality model.

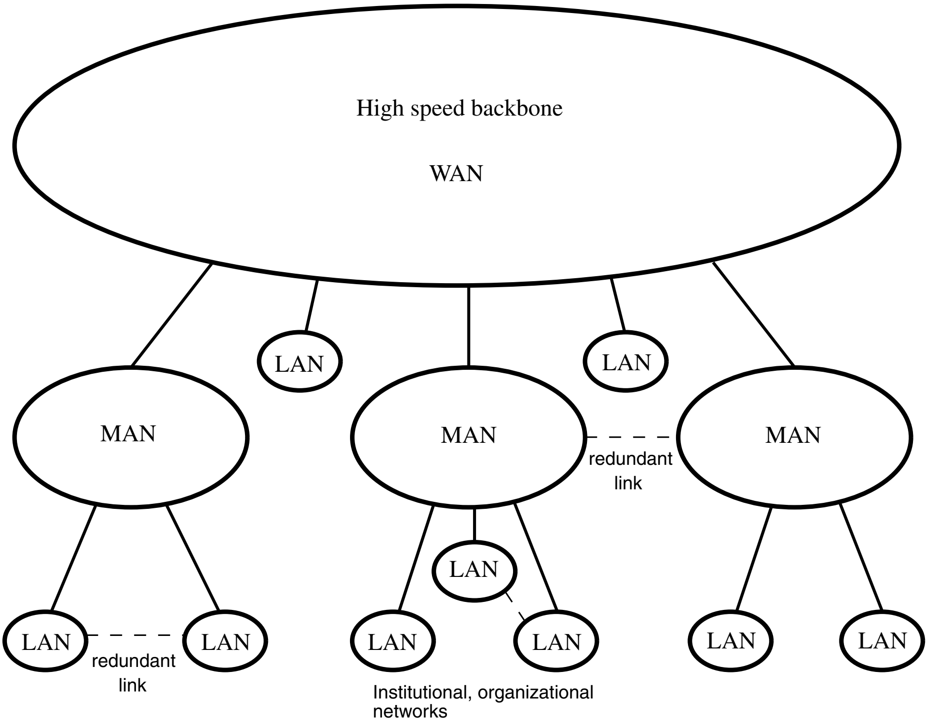


Figure 2 Hierarchical Network Model.

area. Local area networks (LANs) arise from this way of solving the information traffic problem. The hierarchy can be further fragmented, and thus metropolitan area networks (MANs) and wide area networks (WANs) enter the picture. Although this model basically supposes tree-like network configurations (topologies), for sake of reliability and routing efficiency, cross-connections (redundant routes) are also established within this infrastructure.

If a network user wishes to access information at a networked computer somewhere else, his or her request is sent through the chain of special networking devices (routers and switches; see Section 4) and links. Normally this message will be sliced into packets and the individually transferred packets will join possibly thousands of packets from other messages during their way until the relevant packets, well separated, arrive at their destination, recombine, and initiate downloading of the requested information from that distant site to the local machine. The requested information will then arrive to the initiator by a similar process. If there is no extraordinary delay anywhere in the routes, the full process may take just a few seconds or less. This is the most important strength of accessing information through the network.

Note that the described method of packet-based transmission is a special application of time division multiplexing (TDM). The TDM technique utilizes a single physical transmission medium (cable, radiated wave, etc.) for transmitting multiple data streams coming from different sources. In this application of TDM, all the data streams are sliced into packets of specified length of time, and these packets, together with additional information about their destination, are inserted into time frames of the assembled new, higher-bit rate data stream. This new data stream is transmitted through a well-defined segment of the network. At the borders of such network segments, the packets may be extracted from the assembled data stream and a new combination of packets can be assembled by a similar way. By this technique, packets from a source node may even arrive to their destination node through different chains of network segments. More details about the technique will be given in Sections 6.2 and 6.3.

The next subsection describes the basic prerequisites at the site providing the information. First a basic property of the data communication links will be introduced here.

The capability of transmitting/processing high-volume traffic by the connecting lines in the network depends on their speed/bandwidth/capacity. Note that these three characteristic properties are equivalent in the sense that high bandwidth means high speed and high speed means high capacity. Although all three measures might be characterized by well-defined units, the only practically applied unit for them is the one belonging to the speed of the information transfer, bits per second (bps). However, because of the practical speed values, multiples of that elementary unit are used, namely Kbps, Mbps, Gbps, and more recently, Tbps (kilobits, megabits, gigabits, and terabits per second, respectively, which means thousands, millions, billions and thousand billions of bits per second).

2.2. Information Provision

Access to information is not possible if the information is not available. To make information accessible to other network users, information must be provided. Provision of information by a user (individual or corporate or public) is one of the key contributors to the exponential development of networking and worldwide network usage. How the information is provided, whether by passive methods (by accessible databases or, the most common way today, the World Wide Web on the Internet) or by active distribution (by direct e-mail distribution or by using newsgroups for regular or irregular delivery), is a secondary question, at least if ease of accessing is not the crucial aspect.

If the information provided is freely accessible by any user on the network, the information is public. If accessibility is restricted, the information is private. Private information is accessible either to a closed group of users or by virtually any user if that user fulfils the requirements posed by the owner of the information. Most frequently, these requirements are only payment for the access. An obvious exception is government security information.

All information provided for access has a certain value, which is, in principle, determined by the users' interest in accessing it. The higher the value, the more important is protection of the related information. Providers of valuable information should take care of protecting that information, partly to save ownership but also in order to keep the information undistorted. Faulty, outdated, misleading, irrelevant information not only lacks value but may also cause problems to those accessing it in the belief that they are getting access to reliable content.

Thus, information provision is important. If the provider makes mistakes, whether consciously or unconsciously, trust will be lost. Providing valuable (relevant, reliable, and tested) information and taking care of continuous updating of the provided information is critical if the provider wants to maintain the trustworthiness of his information.

The global network (the millions of sites connected by the network) offers an enormous amount of information. The problem nowadays is not simply to find information about any topic but rather to find the best (most relevant and most reliable) sources of the required information. Users interested in accessing proper information should either turn to reliable and trusted sources (providers) or take the information they collect through the Internet and test it themselves.

That is why today the provision of valuable information is separated from ownership. Information brokerage plays an increasingly important role in where to access adequate information. Users looking for information about any topic should access either those sources they know and trust or the sites of brokers who take over the task of testing the relevance and reliability of the related kinds of information. Information brokerage is therefore an important new kind of network-based service.

However, thoroughly testing the validity, correctness, and relevance of the information is difficult or even impossible, especially because the amount of the stored and accessible information increases extremely fast. The final answer to the question of how to control the content (i.e., the information available worldwide) is still not known. One of the possibilities (classifying, rating, and filtering) is dealt with in more detail in a later section.

Besides the difficulty of searching for adequate information, there is a further important aspect associated with the increasing interest worldwide in accessing information. It relates to the volume of information traffic generated by the enormous number of attempts to download large amounts of information.

Traffic volume can be decreased drastically by using the caching technique (see Figure 3). In the case of frequently accessed information sources, an enormous amount of traffic can be eliminated by storing the frequently requested information in dedicated servers that are much closer to the user than the original information provider. The user asking for these kinds of information doesn't even know if the requested information arrives at the requesting site from such a cache server or from the primary source of the information. Hierarchical cache server systems help to avoid traffic congestion on network links carrying intensive traffic.

2.3. Electronic Communication

Providing and accessing information, although extremely important, is just one element of information traffic through the network. Electronic communication in general, as far as the basic function of the network operation is concerned, involves a lot more.

The basic function of computer networks is to make it possible to overcome distance, timing, and capability problems.

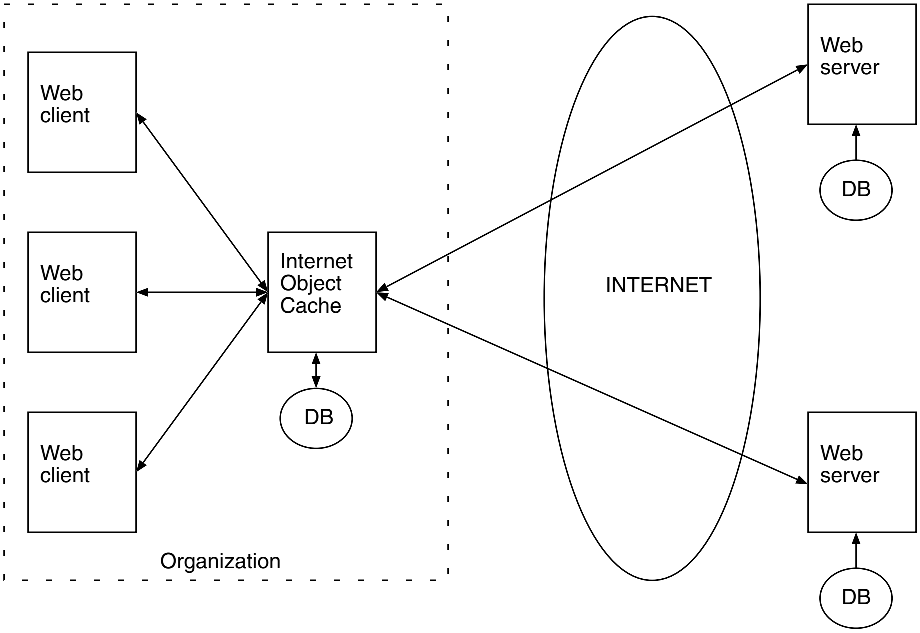


Figure 3 Caching.

- Distance is handled by the extremely fast delivery of any messages or other kinds of transmitted information at any connected site all over the world. This means that users of the network (again individual, corporate, or public) can do their jobs or perform their activities so that distances virtually disappear. Users or even groups of users can talk to each other just as if they were in direct, traditional contact. This makes, for example, remote collaboration a reality by allowing the formation of virtual teams.
- Timing relates to the proper usage of the information transmitted. In the most common case, information arrives at its destination in accordance with the best efforts method of transmission provided by basic Internet operation. This means that information is carried to the target site as soon as possible but without any guarantee of immediate delivery. In most cases this is not a problem because delays are usually only on the order of seconds or minutes. Longer delays usually occur for reasons other than simple transmission characteristics. On the other hand, e-mail messages are most often read by the addressees from their mail servers much later than their arrival. Although obviously this should not be considered the basis of network design, the fact that the recipient doesn't need to be at his or her machine when the message arrives is important in overriding timing problems. However, some information can be urgently needed, or real-time delivery may even be necessary, as when there are concurrent and interacting processes at distant sites. In these cases elevated-quality services should be provided. Highest quality of service (QoS) here means a guarantee of undisturbed real-time communication.
- Capability requirements arise from limited processing speed, memory, or software availability at sites having to solve complex problems. Networked communication helps to overcome these kinds of problems by allowing distributed processing as well as quasisimultaneous access to distributed databases or knowledge bases. Distributed processing is a typical application where intensive (high-volume) real-time (high-speed) traffic is assumed, requiring high-speed connection through all the routes involved in the communication.

The latter case already leads us to one of the more recent and most promising applications of the network. Although in some special areas, including highly intensive scientific computations, distributed processing has been used for some time, more revolutionary applications are taking place in networked remote collaboration. This new role of computer networking is briefly reviewed in the following section.

2.4. Networked Collaboration

Computer networks open new ways of increasing efficiency, elevating convenience, and improving cost and performance of collaboration. The basis of this overall improvement is provided by the properties of computer networks mentioned above, their ability to help overcome distance, timing, and capability problems. The workplace can thus be dislocated or even mobile, the number of collaborators can be theoretically infinite, the cooperating individuals or groups can be involved in joint activities and in shifted time periods (thus even geographical time zone problems can be virtually eliminated), and the collaborating partners can input their information without having to move high-volume materials (books and other documents) to a common workplace. These are all important issues in utilizing the opportunities that networks provide for collaborating when communities are geographically dispersed.

However, the key attribute of networked collaboration is not simply the possibility of avoiding having to collect all the collaborating partners at a well-defined venue and time for a fixed time period. The fact that the network interconnects computers, rather than human users, is much more useful. The reason is that networked computers are not just nodes in a network but also intelligent devices. They not only help exchange information between the collaborating individuals or groups, but also provide some important additional features.

- The information processing capability of the computers at the network nodes of the collaborating parties may also solve complex problems.
- The intelligence of each computer connected and participating in the joint activities is shareable among the collaborating partners.
- The intelligence in these distant computers can be joined to solve otherwise unsolvable problems by utilizing an enormously elevated computing power.

However, appropriate conditions for efficient and successful collaboration assume transmission of complex information. This information involves not only the data belonging to the joint task but also the multimedia information stemming from the networked character of the collaboration. This may be a simple exchange of messages, such as e-mail conversation among the participants, but may also be a true videoconference connection between the cooperating sites. Thus, in many cases networked collaboration requires extremely high bandwidths. Real-time multimedia transmission, as the basis of these applications, supposes multi-Mbps connectivity throughout. Depending on the quality demands, the acceptable transmission speed varies between one or two Mbps and hundreds of megabits per second. The latter speeds have become a reality with the advent of gigabit networking.

If the above conditions are all met, networked collaboration provides multiple benefits:

- Traditional integration of human contributions
- Improvement in work conditions
- Integration of the processing power of a great many machines
- High-quality collaboration attributes through high-speed multimedia transmission

The highest level of networked collaboration is attained by integrated environments allowing truly integrated activities. This concept will be dealt with later in the chapter. The role of the World Wide Web in this integration process, the way of building and using intranets for, among others, networked collaboration, and the role of the different security aspects will also be outlined later.

2.5. Virtual Environments

Network-based collaboration, by using highly complex tools (computers and peripherals) and high-speed networking, allows the participating members of the collaborating team to feel and behave just as if they were physically together, even though the collaboration is taking place at several distant sites. The more the attributes of the environment at one site of the collaboration are communicated to the other sites, the more this feeling and behavior can be supported. The environment itself cannot be transported, but its attributes help to reconstruct a virtual copy of the very environment at a different site. This reconstruction is called a virtual environment.

This process is extremely demanding as far as information processing at the related sites and information transfer between distant sites are concerned. Although the available processing power and transmission speed increase exponentially with time, the environment around us is so complex that full copying remains impossible. Thus, virtual environments are only approximated in practice.

Approximation is also dictated by some related practical reasons:

- Practical feasibility (even if a certain level of copying the target environment is theoretically possible, it would require the implementation of prohibitively complex practical solutions)
- Cost effectiveness (although the practical solution is feasible, the cost is so high that the application doesn't allow the very level of emulation)
- Intended approximation (although a certain level of true copying would be possible and cost effective, the application requires a model of the virtual copy rather than the true original because of experimental applications, trials regarding modifying the environment, etc.)
- Inaccessible environments (the environment to be copied is inaccessible either because of being outside the connected sites or because of transient lack of some prerequisites)

In such cases, virtual environments are substituted for by virtual reality models. These models provide virtual reality by mimicking the supposed environment but not copying it. However, in some applications a mix of virtual copies and virtual models is needed. In *augmented reality,* transmitted information about the attributes of the considered real environment is used together with model parameters about a model environment to build up a virtual construct for the supposed environment.

Virtual constructs are essential tools in a number of applications utilizing the concept of telepresence and/or teleimmersion by users of networked computers. Teleinstruction, telemedicine, and different types of teleworking are examples of how the development of networked communication and collaboration leads to an increasing number of applications using virtual and augmented reality.

3. A SHORT HISTORY OF COMPUTER NETWORKING

Computer networking started in the United States in the late 1960s. Through the rapid development of computing and telecommunications, in 1969 the ARPANet program was launched.

The first voice transmission over cables (1876) and by radio waves (1896), the first radio transfer of video images (1936), the first electronic computer (1944), and the first transistor (1948) preceded this milestone of technical history, but the first microprocessor (1971) and the first IBM personal computer (1981), as well as the first digital radio transmission of voice (1988) and image (1991), came later.

The ARPANet program started with the aim of connecting just a few sites, but the first networking experiments immediately proved that ''Internetting'' was a feasible idea. The program led to the development of the TCP/IP protocol suite (see Section 6), having served as the basis of Internet technology for more than 30 years. These developments resulted in a proliferation in activity, and the number of Internet hosts (networked computers serving several users by Internet services) rapidly increased. As a result of the annual doubling of host numbers, as of the end of the 1990s, dozens of millions of hosts served a worldwide community of an estimated more than 200 million users.

The Internet was established early in the 1980s as a solid infrastructure of networking, mainly in the United States but also with connected subnetworks in Europe and the Asia Pacific region. The organizational, coordinating, and mainly informal controlling frameworks of the Internet were also established by that time. A conscious standardization process also started. These activities resulted in a solid background for an efficiently operating global system of interconnected computers.

Perhaps the most important property of the TCP/IP protocol suite is that it makes possible the interconnection and interworking of subnetworks. The introduction and wide application of the concept of such subnetworks played a fundamental role in the rapid and successful proliferation of the Internet technology. The value of integrating diverse subnetworks into one cohesive Internet is what makes the Internet such a great network tool.

After more than 10 years of exclusively special (defense, and later scientific) applications of the emerging network, commercial use began to evolve. Parallel activities related to leading-edge academic and research applications and everyday commodity usage appeared. This parallelism was more or less maintained during more recent periods, as well, serving as a basis of a fast and continuous development of the technology. A breakthrough occurred when, in the mid-1980s, the commercial availability of system components and services was established. Since then the supply of devices and services has evolved continuously, enabling any potential individual, corporate, or public user simply to buy or hire what it needed for starting network-based activities.

While the leading edge in the early 1980s was still just 56 Kbps (leased line), later in the decade the 1.5 Mbps (T1) transmission speed was achieved in connecting the emerging subnetworks.

Late in the 1980s the basics of electronic mail service were also in use. This was probably the most important step in the application of networking to drive innovation prior to the introduction of the World Wide Web technology.

Early in the 1990s the backbone speed reached the 45 Mbps level, while the development and spread of a wide range of commercial information services determined the directions of further evolution in computer networking.

The 1990s brought at least three new developments.

1. The development of a new-generation Internet was initiated in order to overcome the weaknesses of the traditional Internet technology in handling the exponential growth of the user community, the fast change in quality, reliability, and security requirements, and the increased complexity of the transmitted information. (Strictly speaking, security is outside of the Internet itself. The need for security became a hot topic with the advent of readily accessible information via the Internet. The present TCP/IP protocol suite does not provide security itself.)
2. The speed of transmission (or the speed of the network) and the coverage of the global infrastructure (including worldwide proliferation as well as dense regional network population) reached new records. This development is still far from saturation. As of the turn of the century, the dozens of millions of Internet hosts had, not homogeneously but steadily, spread out through all continents of the world. At the same time, the high-level backbones operated, at least in the more developed areas, at Gbps speed, and the continents were connected by thousands of optical fibers capable of keeping an extremely high volume of traffic.
3. World Wide Web technology has become a standard all over the world, and by the end of the 1990s an enormous amount of information provision and information access had taken place through the Web.

Through these revolutionary advancements, networking has begun to play an important role worldwide in reforming a great many human activities. Parallel to the developments in speed, quality, reliability, manageability, and cost/performance, many international organizations, national governments, and civil initiatives have launched joint projects with the aim of establishing a new and unavoidable worldwide "Information Society." New forms of cooperation among governments, industry, telecom, and network service providers, as well as civil organizations, have emerged, and further development is inevitable.

4. THE NETWORKING INFRASTRUCTURE

As mentioned in Section 2.1, the background of the information transfer is provided by the networking infrastructure. As mentioned, the solution of the problem of how to handle millions of transfers of data simultaneously over the global network in very short time frames is based on a hierarchical model. The infrastructure consists of links and nodes whose role depends on their status in the hierarchy. At the highest level are the largest information exchange centers, which are connected with very high-speed lines. The very high-speed networks of these large exchange nodes and the high-capacity links connecting them are called backbones. The main nodes of the backbones communicate among each other and also collect and distribute the relevant traffic from and to the regional nodes around them. The lines connecting the large exchange nodes to the regional nodes also work at very high speeds and are therefore able to transmit very high-volume traffic.

Lower in the hierarchy, around the regional nodes, somewhat lower-speed access lines transfer the information from and to the area servers (connected to routers and/or switches, i.e., special machines taking care of the direction of the transmitted information packets in the network) within the region. These active devices serve either LANs, MANs, or WANs, depending on the topology and the further hierarchical structure of the network around these area servers. Their task is to collect/distribute the information traffic in their area.

The hierarchical structure might mean a simple (but of course highly segmented and in this respect very complex) tree-like topology. However, as already mentioned, for the purpose of reliability and routing efficiency, cross-connections (redundant routes) are also established within the infrastructure. Obviously, the difficulty of handling the traffic is in relation with the complexity of the network structure, but this is the price for satisfactory global network operation.

The key elements, besides the communication links themselves, in this technique of information transfer through the network are the routers and switches, which take care of directing the information packets through the nodes in this hierarchy so that they finally arrive at their destination.

The speed of the information transfer depends on the bandwidth of the links and the processing speed of the routers and switches in the infrastructure. Taking into account the requirements stemming from the characteristics of the transmitted information and the traffic volume, that is, from the estimated average and peak number of requested transmission transactions, allows the speed of the different sections of the infrastructure to be determined. Some basic figures are as follows:

- Simple alphanumeric messages require a speed that is not too high at the last mile sections, that is, at the lines closest to the transmitters/receivers of the information. Several Kbps is considered satisfactory in such cases of message delivery without special requirements regarding the delivery time.

- Real-time transmission of similarly simple alphanumeric information requires either similar speed, as above, or, in case of higher volume of these kinds of data, a certain multiple of that speed.
- Real-time transmission of sampled, digitized, coded, and possibly compressed information, stemming from high-fidelity voice, high-resolution still video, or high-quality real video signals do require much higher speed. The last-mile speed in these cases ranges from dozens of Kbps to several and even hundreds of Mbps (HDTV). This information is quite often cached by the receiver to overcome short interruptions in the network's delivery of the information at high speed, providing the receiver a smooth uninterrupted flow of information.

This means that although until recently the last-mile line speeds were in the range of several dozens of Kbps, the area access lines were in the range of several Mbps, and the top-level backbones were approaching the Gbps speed, lately, during the 1990s, these figures have increased to Mbps, several hundreds of Mbps, and several Gbps level, respectively. The near future will be characterized by multi-Gbps speed at the backbone and area access level, and the goal is to reach several hundreds of Mbps even at the last-mile sections. Of course, the development of the infrastructure means coexistence of these leading-edge figures with the more established lower-speed solutions. And obviously, in the longer term, the figures may increase further.

As far as the organization of the operation of the infrastructure, the physical and the network infrastructure should be distinguished. Providing physical infrastructure means little more than making available the copper cables or fibers and the basic active devices of the transmission network, while operating the network infrastructure also means managing of the data traffic. Both organizational levels of the infrastructure are equally important, and the complexity of the related tasks has, in an increasing number of cases, resulted in the jobs being shared between large companies specializing in providing either the physical or the network infrastructure.

The first step from physical connectivity to network operation is made with the introduction of network protocols. These protocols are special complex software systems establishing and controlling appropriate network operation. The most important and best-known such protocol is the Internet Protocol.

The elementary services are built up on top of the infrastructure outlined above. Thus, the next group of companies working for the benefit of the end users is the Internet service providers (ISPs), which take care of providing network access points.

Although network services are discussed separately in Section 7, it should be mentioned here that one of the basic tasks of these services is to take care of the global addressing system. Unique addressing is perhaps the most important element in the operation of the global network. Network addresses are associated with all nodes in the global network so that the destination of the transmitted information can always be specified. Without such a unique, explicit, and unambiguous addressing system, it would not be possible even to reach a target site through the network. That is why addressing (as well as naming, i.e., associating unique symbolic alphanumeric names with the numeric addresses) is a crucial component of network operation. The task is solved again by a hierarchy of services performed by the domain name service (DNS) providers. This issue is dealt with in more detail later. Note that it is possible, as a common practice, to proxy the end-node computer through an access point to share the use of a unique address on the Internet. The internal subnetwork address may in fact be used by some other similarly isolated node in a distinctly separate subnetwork (separate intranet).

Everything briefly described above relates to the global public network. However, with the more sophisticated, serious, and sometimes extremely sensitive usage of the network, the need to establish closed subsets of the networked computers has emerged. Although virtual private networks (subnetworks that are based on public services but that keep traffic separated by the use of special hardware and software solutions) solve the task by simply separating the general traffic and the traffic within a closed community, some requirements, especially those related to security, can only be met by more strictly separating the related traffic. The need for this kind of separation has led to the establishment of intranets (special network segments devoted to a dedicated user community, most often a company) and extranets (bunches of geographically distant but organizationally and/or cooperatively connected intranets, using the public network to connect the intranets but exploiting special techniques for keeping the required security guarantees). Although building exclusively private networks, even wide area ones, is possible, these are gradually disappearing, and relying on public services is becoming common practice even for large organizations.

5. INTERNET, INTRANETS, EXTRANETS

The Internet is the world's largest computer network. It is made up of thousands of independently operated (not necessarily local) networks collaborating with each other. This is why it is sometimes called the network of networks. Today it extends to most of the countries in the world and connects dozens of millions of computers, allowing their users to exchange e-mails and data, use online services, communicate, listen to or watch broadcast programs, and so on, in a very fast and cost-effective way. Millions of people are using the Internet today in their daily work and life. It has become part of the basic infrastructure of modern human life.

As was already mentioned above, the predecessor of the Internet was ARPANet, the first wide area packet switched data network. The ARPANET was created within a scientific research project initiated by the U.S. Department of Defense in the late 1960s. Several universities and research institutes participated in the project, including the University of Utah, the University of California at Los Angeles, the University of California at Santa Barbara, and the Stanford Research Institute. The goal of the project was to develop a new kind of network technology that would make it possible to build reliable, effective LANs and WANs by using different kinds of communication channels and connecting different kinds of computers. By 1974, the basics of the new technology had been developed and the research groups led by Vint Cerf and Bob Kahn had published a description of the first version of the TCP/IP (Transmission Control Protocol and Internet Protocol) suite.

The experimental TCP/IP-based ARPANet, which in its early years carried both research and military traffic, was later split into the Internet, for academic purposes, and the MILNet, for military purposes. The Internet grew continuously and exponentially, extending step by step all over the world. Early in the 1990s, the Internet Society (see www.isoc.org) was established, and it has served since then as the driving force in the evolution of the Internet, especially in the technological development and standardization processes.

The invention of the World Wide Web in 1990 has given new impetus to the process of evolution and prompted the commercialization of the Internet. Today the Internet is a general-purpose public network open to anyone who wishes to be connected.

Intranets are usually TCP/IP-based private networks. They may in fact gateway through a TCP/IP node, but this is not common. They operate separately from the worldwide Internet, providing only restricted and controlled accessibility. Although an intranet uses the same technology as the Internet, with the same kinds of services and applications, it principally serves only those users belonging to the organization that owns and operates it. An intranet and its internal services are closed to the rest of the world. Often this separation is accomplished by using network addresses reserved for such purposes, the so-called ten net addresses. These addresses are not routed in the Internet, and a gateway must proxy them with a normal address to the Internet.

Connecting intranets at different geographical locations via the public Internet, results in extranets. If an organization is operating at different locations and wants to interconnect its TCP/IP-based LANs (intranets), it can use the inexpensive public Internet to establish secure channels between these intranets rather than build very expensive, large-scale, wide area private networks. This way, corporate-wide extranets can be formed, allowing internal users to access any part of this closed network as if it were a local area network.

6. TECHNICAL BACKGROUND

To understand the possibilities of the TCP/IP-based networks fully, it is important to know how they work and what kinds of technical solutions lie behind their services.

6.1. Architecture of the Internet

As mentioned in the previous section, the Internet can be described as a network of local area networks (see Figure 4). Therefore, perhaps the most important component of the Internet is LANs. LANs connect computers or, in other words, hosts, and are connected to other LANs by gateways and communication lines. Thus, the four basic logical components of the Internet (or of intranets) are:

1. Hosts
2. Local area networks
3. Gateways
4. Communications lines

The main role of gateways is to provide connections between the communication lines and the local area networks and to route the data packets toward their destinations. This is why they are often called routers. Gateways also play an important role in security by protecting the local area network against external attacks and other illegal or unauthorized access. Quite often, a security zone, referred to as the DMZ, is set up where network traffic is passed through a gateway into a firewall (see Section 15.2). The firewall then provides application-level security to network traffic before sending the information through an internal router to pass to end systems in the intranet.

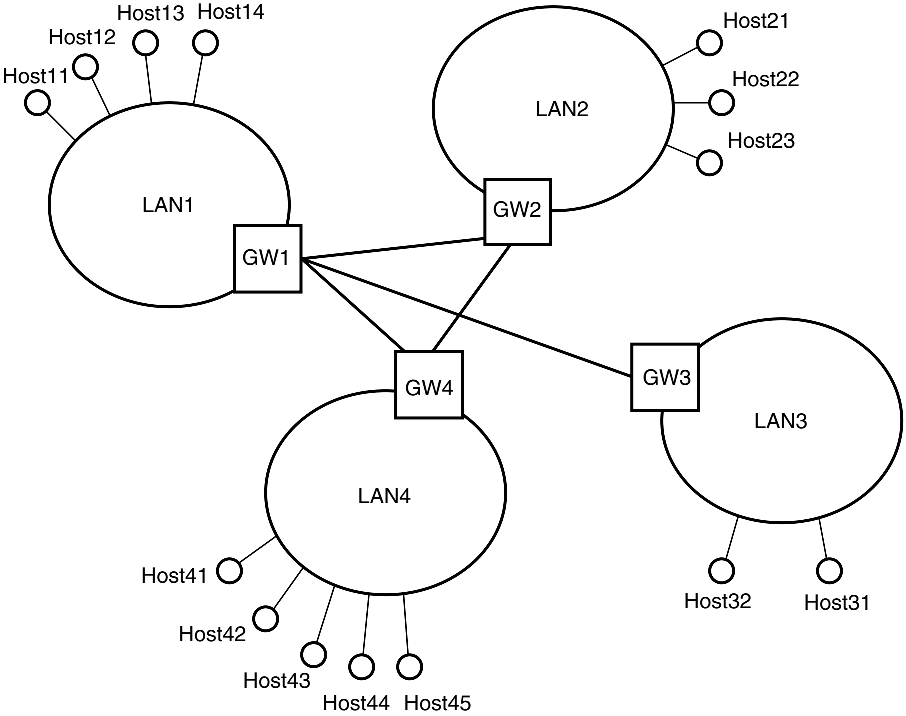


Figure 4 Internet Architecture.

In most cases, communication lines establish steady, 24-hour connection between their endpoints. In order to increase reliability, and sometimes also for traffic load balancing reasons, alternative routes can be built between the local area networks. If any of the lines is broken, the gateways can automatically adapt to the new, changing topology. Thus, the packets are continuously delivered by the alternative routes. Adaptive routing is possible at gateways, particularly for supporting 24 × 7 operation. It commonly occurs in the backbone of the Internet.

6.2. Packet Switching

Packet switching is one of the key attributes of the TCP/IP technology. The data stream belonging to a specific communication session is split into small data pieces, called packets. The packets are delivered independently at the target host. The separated packets of the same communication session may follow different routes to their destination. In contrast to line-switching communication technologies, in packet switched networks there is no need to set up connections between the communicating units before the start of the requested data transmission. Each packet contains all of the necessary information to route it to its destination. This means that packets are complete from a network perspective.

A good example of line-switched technology is the traditional public phone service. In contrast to packet switching, line switching assumes a preliminary connection setup procedure being performed before a conversation starts. After the connection is set up, an individual communication channel (called a circuit) is provided for the data transmission of that communication session. When the data transmission is over, the connection should be closed.

6.3. Most Important Protocols

A network protocol is a set of rules that determines the way of communication on the network. All the attached hosts must obey these rules. Otherwise they won't be able to communicate with the other hosts or might even disturb the communication of the others. For proper high-level communication, many protocols are needed.

The system of protocols has a layered structure. High-level protocols are placed in the upper layers and low-level protocols in the lower layers. Each protocol layer has a well-defined standard interface by which it can communicate with the other layers, up and down. TCP/IP itself is a protocol group consisting of several protocols on four different layers (see Figure 5). TCP and IP are the two most important protocols of this protocol group.

IP is placed in the internetworking layer. It controls the host-to-host communication on the network. The main attributes of IP are that it is connectionless, unreliable, robust, and fast. The most astounding of these is the second attribute, unreliability. This means that packets may be lost, damaged, and/or multiplied and may also arrive in mixed order. IP doesn't guarantee anything about the safe transmission of the packets, but it is robust and fast. Because of its unreliability, IP cannot satisfy the needs of those applications requiring high reliability and guaranteed QoS.

Built upon the services of the unreliable IP, the transport layer protocol, TCP, provides reliable communication for the applications. Reliability is guaranteed by positive acknowledgement and automatic retransmission. TCP performs process-to-process communication. It also checks the integrity of the content. TCP is connection oriented, although the TCP connections are virtual, which means that there is no real circuit setup, only a virtual one. A TCP connection is a reliable, stream-oriented, full-duplex, bidirectional communication channel with built-in flow control and synchronization mechanisms.

There are also routing and discovery protocols that play an important role in affecting network reliability, availability, accessibility, and cost of service.

6.4. Client–Server Mechanism

Any communication on a TCP/IP network is performed under the client–server scheme (see Figure 6). When two hosts are communicating with each other, one of them is the client and the other is the server. The same host can be both client and server, even at the same time, for different communication sessions, depending on the role it plays in a particular communication session. The client sends requests to the server and the server replies to these requests. Such servers include file servers, WWW servers, DNS servers, telnet servers, and database servers. Clients include ftp (file transfer) programs, navigator programs (web browsers), and telnet programs (for remote access). Note that the term *server* does not imply special hardware requirements (e.g. high speed, or large capacity, or continuous operation).

The typical mode of operation is as follows: The server is up and waits for requests. The client sends a message to the server. The requests arrive at particular ports, depending on the type of the expected service. Ports simply address information that is acted upon by the serving computer. Clients provide port information so that server responses return to the correct application. Well-known ports are assigned to well-known services. Communication is always initiated by the client with its first message.

Because multiple requests may arrive at the server at the same time from different clients, the server should be prepared to serve multiple communication sessions. Typically, one communication

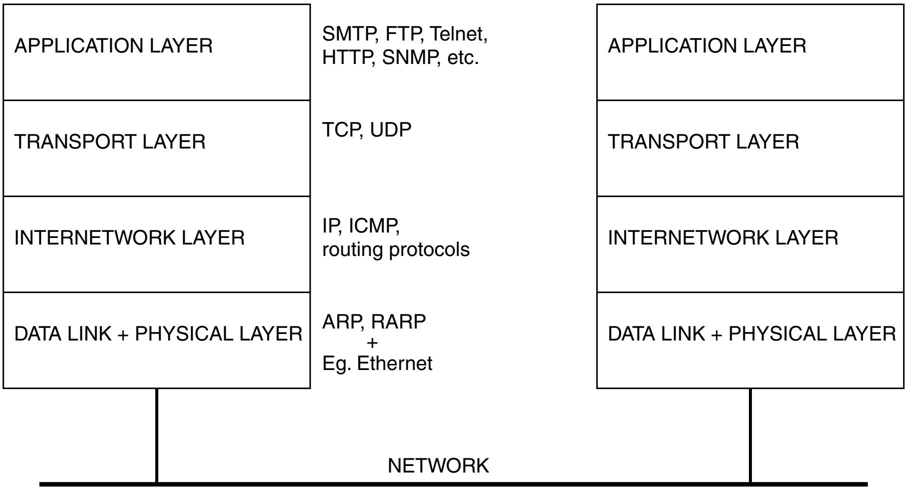


Figure 5 The Layered TCP/IP Protocol Stack.

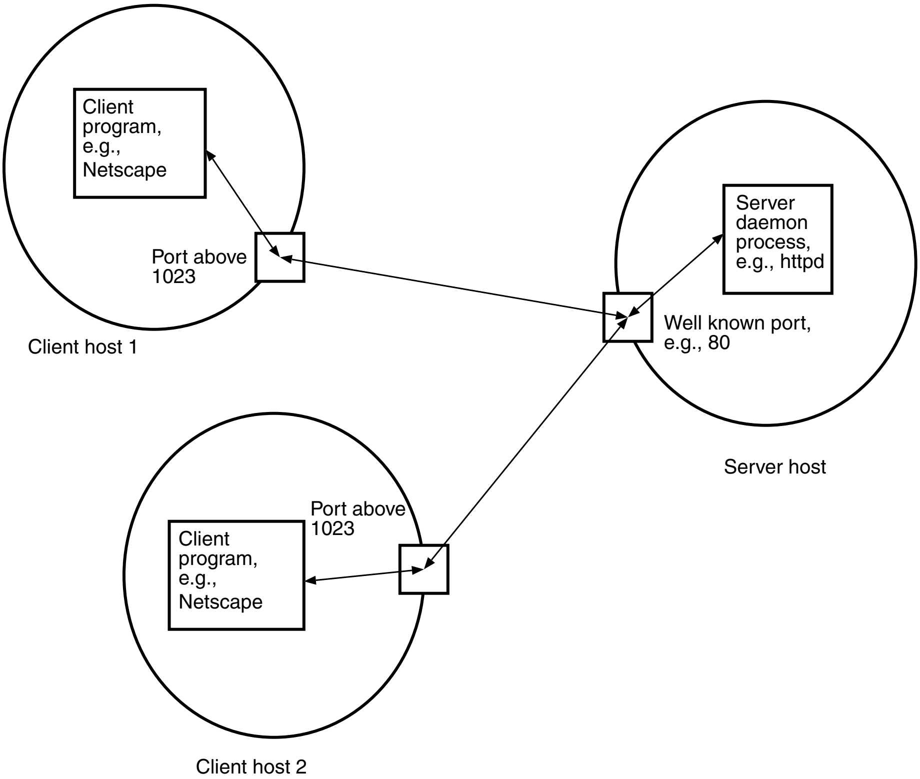

Client program, e.g., Netscape
Port above 1023
Client host 1
Server daemon process, e.g., httpd
Well known port, e.g., 80
Server host
Port above 1023
Client program, e.g., Netscape
Client host 2

Figure 6 Client/Server Scheme.

session is served by one process or task. Therefore, servers are most often implemented on multitasking operating systems. There is no such requirement at the client side, where only one process at a time may be acting as a client program.

One of the main features in the new concept of computing grids is the substitution of the widely used client–server mechanism by realizations of a more general distributed metacomputing principle in future computer networking applications (see Section 12.3).

6.5. Addressing and Naming

In order to be unambiguously identifiable, every host in a TCP/IP network must have at least one unique address, its IP address. IP addresses play an essential role in routing traffic in the network. The IP address contains information about the location of the host in the network: the associated number (address) of the particular LAN and the associated number (address) of the host in the LAN.

Currently, addresses in the version 4 Internet protocol are 32 bits long (IPv4 addresses) and are classified into five groups: A, B, C, D, and E classes. Class A has been created for very large networks. Class A networks are rare. They may contain up to 2^{24} hosts. Class B numbers (addresses) are given to medium-sized networks (up to 65,534 hosts), while class C network numbers are assigned to small (up to 254 hosts) LANs.

The 32-bit-long IP addressing allows about 4 billion different combinations. This means that in principle, about 4 billion hosts can be attached to the worldwide Internet, by using IPv4 addressing. However, intranets, because they are not connected to the Internet or connected through firewalls, may use IP addresses being used by other networks too.

Because the number of Internet hosts at the start of the third millennium is still less than 100 million, it seems that the available IPv4 address range is wide enough to satisfy all the needs. However, due to the present address classification scheme and other reasons, there are very serious limitations in some address classes (especially in class B). The low efficiency of the applied address distribution scheme has led to difficult problems. Because of the high number of medium-sized networks, there are more claims for class B network numbers than the available free numbers in this class.

Although many different suggestions have been made to solve this situation, the final solution will be brought simply by implementing the new version of the Internet protocol, IPv6, intended to be introduced early in the 2000s. It will use 128-bit-long IP addresses. This space will be large enough to satisfy any future address claims even if the growth rate of the Internet remains exponential.

There are three different addressing modes in the current IPv4 protocol:

- Unicast (one to one)
- Multicast (one to many)
- Broadcast (one to all)

The most important is unicast. Here, each host (gateway, etc.) must have at least one class A, class B, or class C address, depending on the network it is connected to. These classes provide unicast addresses.

Class D is used for multicasting. Multicast applications, such as radio broadcasting or video conferencing, assign additional D class addresses to the participating hosts.

Class E addresses have been reserved for future use.

Class A, B, and C addresses consist of three pieces of information: the class prefix, the network number, and the host number. The network number, embedded into the IP address, is the basic information for routing decisions. If the host number part of the address contains only nonzero bits (1s), the address is a broadcast address to that specific network. If all the bits in the host number part are zeroes (0s), the address refers to the network itself.

In order to make them easier to manage, IP addresses are usually described by the four bytes they contain, all specified in decimal notation, and separated by single dots. Examples of IP addresses are:

- Class A: 16.1.0.250
- Class B: 154.66.240.5
- Class C: 192.84.225.2

The human interface with the network would be very inconvenient and unfriendly if users had to use IP addresses when referring to computers they would like access. IP addresses are long numbers, making them inconvenient and difficult to remember and describe, and there is always a considerable risk of misspelling them. They don't even express the type, name, location, and so on of the related computer or the organization operating that computer. It is much more convenient and reliable to associate descriptive names with the computers, the organizations operating the computers, and/or the related LANs. The naming system of the TCP/IP networks is called the domain name system (DNS). In this system there are host names and domains.

Domains are multilevel and hierarchical (see Figure 7). They mirror the organizational/administrative hierarchy of the global network.

At present, top-level domains (TLDs) fall into two categories:

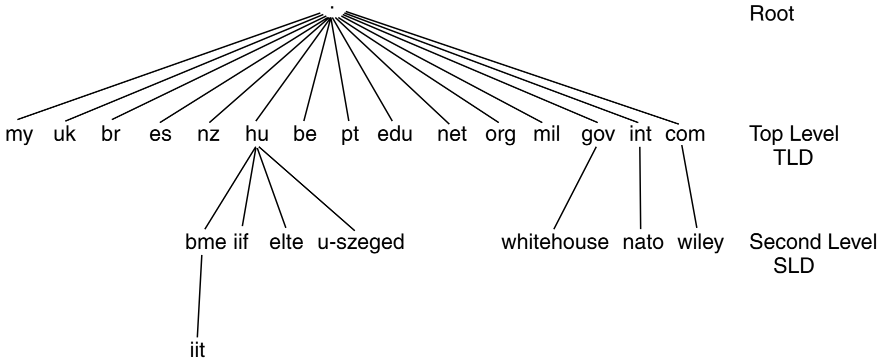


Figure 7 DNS Hierarchy.

1. Geographical (two-letter country codes by ISO)
2. Organizational (three-letter abbreviations reflecting the types of the organizations: edu, gov, mil, net, int, org, com)

Second-level domains (SLDs) usually express the name or the category of the organization, in a particular country. Lower-level subdomains reflect the hierarchy within the organization itself.

The tags of the domain names are separated by dots. The first tag in a domain name is the hostname. The tags are ordered from most specific to least specific addressing.

The domain name service is provided by a network of connected, independently operated domain name servers. A domain name server is a database containing information (IP address, name, etc.) about hosts, domains, and networks under the authority of the server. This is called the domain zone.

Domain name servers can communicate with each other by passing information retrieved from their databases. If a host wishes to resolve a name, it sends a query to the nearest domain name server, which will provide the answer by using its local database, called a cache. However, if necessary, it forwards the query to other servers. The Internet domain name service system is the largest distributed database in the world.

7. OVERVIEW OF NETWORKING SERVICES

Thirty years ago, at the beginning of the evolution of computer networking, the principal goal was to create a communication system that would provide some or all of the following functional services:

- Exchanging (electronic) messages between computers (or their users)—today the popular name of this service is e-mail
- Transferring files between computers
- Providing remote access to computers
- Sharing resources between computers

E-mail is perhaps the most popular and most widely used service provided by wide area computer networks. E-mail provides easy-to-use, fast, cheap, and reliable offline communication for the users of the network. It allows not only one-to-one communication but one-to-many. Thus, mailing lists can be created as comfortable and effective communication platforms for a group of users interested in the same topics.

Of course, file transfer is essential. Computer files may contain different kinds of digitized information, not only texts and programs but formatted documents, graphics, images, sounds, movies, and so on. File transfer allows this information to be moved between hosts.

With the remote access service, one can use a distant computer just as if one were sitting next to the machine. Although not all operating systems allow remote access, most of the multiuser operating systems (UNIX, Linux, etc.) do.

Sharing resources makes it possible for other machines to use the resources (such as disk space, file system, peripheral devices, or even the CPU) of a computer. Well-known examples of resource sharing are the file and printer servers in computer networks.

Besides the above-listed basic functions, computer networks today can provide numerous other services. Operating distributed databases (the best example is the World Wide Web), controlling devices, data acquisition, online communication (oral and written), teleconferencing, radio and video broadcasting, online transaction management, and e-commerce are among the various possibilities.

8. OVERVIEW OF NETWORK-BASED APPLICATIONS

As mentioned at the beginning of this chapter, computer networking has gradually entered all possible fields of applications. Providing an exhaustive overview of the many applications would be practically impossible, so only a brief overview will be given here of what is (and what still will be) available. Classification itself is subjective in this case, so the aspects mentioned below provide only an approach to systematic handling of the wide spectrum of network-based applications.

First, applications can be categorized by what kind of community is interested in them. The two main groups are:

- Public applications, available for any individual or group/organization of individuals
- Private applications, available only to a certain subset of individuals or groups/organizations

Although the realizations of the applications within the two distinct groups are not different in principle from each other, there are at least two basic differences in the background attributes. On one hand, the infrastructure behind an application closely reflects the width of its user community (wide area public network, in contrast to intranet/extranet-type solutions using private or virtual private networks). On the other hand, security issues are definitely different: on top of the general security and privacy requirements, special security and authorization aspects require special firewall techniques in the case of private applications.

Second, applications can be distinguished by how they are utilized. The three main classes are:

- Personal applications (individual and family use)
- Administrative applications (use by governments, municipalities, etc., in many cases with overall, or at least partly authorized, citizen access)
- Industrial applications (practically any other use, from agriculture to manufacturing to commercial services)

The major differences among these three classes are the required levels of availability, geographical distribution, and ease of access. As far as networking is concerned, the classes differ in the required quality of service level/grade.

Third, there is a well-structured hierarchy of the networked applications based on their complexity. This hierarchy more or less follows the hierarchy of those single-computer applications that are enhanced by the introduction of network-based (distributed, dislocated) operation:

- Low-level applications in this hierarchy include distributed word processing, database handling, document generation, computer numerical control, etc.
- Medium-complexity applications include network-integrated editing and publishing, networked logistics and inventory control, CAD (computer aided design), CAM (computer aided manufacturing), etc.
- High-complexity applications include integrated operations management systems, such as integrated management and control of a publishing/printing company, integrated engineering/management of a CAD-CAM-CAT system, etc.

Note that high-complexity applications include lower-complexity applications, while low-complexity applications can be integrated to build up higher-level applications in the hierarchy. Computer networking plays an important role not only in turning from single-computer applications to network-based/distributed applications but also in the integration/separation process of moving up and down in the hierarchy.

Finally, network-based applications can be classified by the availability of the related application tools (hardware and software). The possible classes (custom, semicustom, and off-the-shelf solutions) will be considered later in Section 13.

9. THE WORLD WIDE WEB

Next to e-mail, the World Wide Web is perhaps the most popular application of the Internet. It is a worldwide connected system of databases containing structured, hypertext, multimedia information that can be retrieved from any computer connected to the Internet.

9.1. History

The idea and basic concept of the World Wide Web was invented in Switzerland. In the late 1980s, two engineers, Tim Berners-Lee from the United Kingdom and Robert Cailliau from Belgium, working at CERN (the European Laboratory of Particle Physics) in Geneva, decided to develop a new communication system for Internet-based international collaboration for physicists. The idea became a reality in 1990, when the first version of HTTP (the Hypertext Transfer Protocol), HTML (the Hypertext Markup Language), and URL (the Universal Resource Locator) addressing scheme were introduced and implemented.

For the first three years of its life, the growth of the World Wide Web was quite slow. Only a few dozen web servers were set up, and the Web was used only for what it was originally invented for: scientific communication.

The growth rate of the Web began to accelerate significantly in the second half of 1993 after the introduction of the first portable, graphical web browser, X Mosaic. This was the first browser facilitating the new Internet application, the Web, to spread rapidly and penetrate all over the world. The Web has become a general-purpose technology, today serving not only scientific but all kinds of communication and information presentation needs.

The commercialization of the Internet also started with the birth of the World Wide Web. Companies not previously interested in using the Internet suddenly discovered the new possibilities and became the most active driving force in the development of networking. The first version of Netscape, another important web browser, was released in October 1994. The World Wide Web (W3) Consortium, formed in 1994, drives and leads the technological development and standardization processes of Web technology.

9.2. Main Features and Architecture

As mentioned above, the World Wide Web is a hypertext-based database system of multimedia content. It can be used for building either local or global information systems. It is interactive, which means that the information flow can be bidirectional (from the database to the user and from the user to the database). It integrates all the other traditional information system technologies, such as ftp, gopher (a predecessor of the WWW in organizing and displaying files on dedicated servers), and news.

The content of a Web database can be static or dynamic. Static information changes rarely (or perhaps never), while dynamic content can be generated directly preceding (or even parallel to) actual downloading.

The Web content is made available by web servers, and the information is retrieved through the use of web browsers. The more or less structured information is stored in web pages. These pages may refer to multimedia objects and other web pages, stored either locally or on other Web servers. Thus, the network of the Internet web servers forms a coherent, global information system. Nevertheless, because the different segments of this global system have been created and are operated and maintained independently, there is no unified structure and layout. This fact may result in considerable difficulties in navigating the Web.

The Universal Resource Locator is a general addressing scheme that allows unambiguous reference to any public object (or service), of any type, available on the Internet. It was originally invented for the Word Wide Web, but today URL addressing is used in many other Internet applications. Examples of different URLs include:

- http://www.bme.hu/index.html
- ftp://ftp.funet.fi/
- mailto:maray@fsz.bme.hu

9.3. HTTP and HTML

Hypertext Transfer Protocol (HTTP) is the application-level network protocol for transferring Web content between servers and clients. (Hypertext is a special database system of multimedia objects linked to each other.) HTTP uses the services of the TCP transport protocol. It transfers independently the objects (text, graphics, images, etc.), that build up a page. Together with each object, a special header is also passed. The header contains information about the type of the object, its size, its last modification time, and so on. HTTP also supports the use of proxy and cache servers. HTTP proxies are implemented in firewalls to pass HTTP traffic between the protected network and the Internet. Object caches are used to save bandwidth on overloaded lines and to speed up the retrieval process.

Hypertext Markup Language (HTML) is a document description language that was designed to create formatted, structured documents using hypertext technique (where links are associated to content pieces of the document so that these links can be easily applied as references to other related documents).

As originally conceived, HTML focused on the structure of the document. Later it was improved to provide more formatting features. In spite of this evolution, HTML is still quite simple. An HTML text contains tags that serve as instructions for structuring, formatting, linking, and so on. Input forms and questionnaires can also be coded in HTML. A simple example of an HTML text is as follows:

```
<HTML>
<HEAD>
<TITLE>Example</TITLE>
</HEAD>
<BODY>
<CENTER>
<H1>Example</H1>
</CENTER>
This is just a short <B>HTML</B> example, containing almost nothing.
<P>
<HR>
</BODY>
</HTML>
```

Details about the syntax can be found in the titles related to multimedia and the World Wide Web in the References. The above example will be displayed by the browser like this:

Example

This is just a short **HTML** example, containing almost nothing.

9.4. Multimedia Elements

The World Wide Web is sometimes called a multimedia information system because Web databases may contain also multimedia objects. It is obvious that in a well-designed and formatted document page graphic information and images can be used in addition to text. Web technology allows the inclusion of sound and moving pictures as well. Web pages may also contain embedded programs (e.g., Javascript, a language for designing interactive WWW sites) and thus can have some built-in intelligence.

10. THE ROLE OF THE WORLD WIDE WEB IN COMMUNICATION, INTEGRATION, AND COLLABORATION

The following subsections provide an insight into the role of the World Wide Web in information access, communication, exchanging information, collaboration, and integration.

10.1. The World Wide Web as a Means for Universal Information Access

In the past few years the World Wide Web has become an essential tool of common application in many areas of human life. Today it is a universal, global, and widely used information system that makes it possible for every user of the Internet to access a vast amount of information from every segment of human activity, including science, business, education, and entertainment.

The World Wide Web is universal because it can easily be adapted to almost any kind of information publishing needs. It can substitute for traditional paper-based publications like books, newspapers, newsletters, magazines, catalogs, and leaflets. But it is much more. It can be dynamic, multimedia based, and interactive.

The World Wide Web is global because using the transport provided by the Internet, it can be used worldwide. It allows any kind of information to flow freely to any destination, regardless of physical distance.

The available human interfaces (editors, browsers, etc.) of the Web are friendly and easy to use, making it open for everyone. The Web technology supports different national languages and character sets and thus helps to eliminate language barriers.

10.2. The World Wide Web as a Tool for Communicating and Exchanging Information

The basic role of the Web is to allow users to share (exchange) information. Accessing a particular set of web pages can be free to anybody or restricted only to a group of users.

However, the Web is also a technology for communication. There are Web-based applications for online and offline communication between users, for teleconferencing, for telephoning (Voice over IP [VoIP]), for using collective whiteboards, and so on.

Because the Web integrates almost all other Internet applications (including e-mail), it is an excellent tool for personal and group communication.

10.3. Collaboration and Integration Supported by the World Wide Web

An increasing amount of Web-based applications are available in the market for supporting collaborative and collective work (see the titles related to the World Wide Web and applications in the References). This evolution is based on the fact that the Web technology (including formats and protocols) is standardized, widespread, inexpensive, and platform independent.

Any network user can have access to the Web, regardless of the type of machine he or she works with or the way that machine is connected to the network. Thus, distributing information among collaborating partners is made easier by using the Web.

The Web can also be used for cooperative preparation of documents and/or software by collaborating network users, even if they are far away from each other. That is why Web technology is becoming increasingly popular in intranets and extranets as well. Companies and organizations can thus increase the efficiency of collective work within their staff by using Web-based tools.

Web technology can also be integrated in many different applications or products where there is a need for or a function of sharing or distributing information and/or communicating between different parties and users.

11. CONTENT GENERATION AND CONTENT PROVISION

Generation, provision, distant accessibility, and remote processing of information content can be considered the ultimate goal in introducing computer networks into an organization.

As many systems of information content can be postulated as there are organizations and applications within these organizations. Successful integration of the applications within an organization requires the integration of these separate sets of information content as well. Moreover, the organizations' outside relations can be made most effective if the information content within the organization is correlated to the relevant information content accessible in communicating with partner organizations.

As far as the structure of the correlated individual systems of information content is concerned, widely accepted distributed database techniques, and especially the worldwide-disseminated World Wide Web technology, provide a solid background. However, the structure and the above-mentioned techniques and technologies provide only the common framework for communicating and cooperating with regard to the related systems of information content. Determining what information and knowledge should be involved and how to build up the entire content is the real art in utilizing computer networks. There are no general recipes for this part of the task of introducing and integrating computer and network technologies into the activities of an organization.

The generation and accessibility (provision) of information content are briefly investigated in the following subsections. Also provided is a short overview of the classification of the information content, along with a subsection on some aspects of content rating and filtering.

11.1. Electronic I/O, Processing, Storage, and Retrieval of Multimedia Information

Generation and provision of information content involves a set of different steps that generally take place one after the other, but sometimes with some overlapping and possibly also some iterative refinement. In principle, these steps are independent of the type of the information. However, in practice, multimedia information requires the most demanding methods of handling because of its complexity. (''Multimedia'' means that the information is a mix of data, text, graphics, audio, still video, and full video components.)

Obviously, before starting with the steps, some elementary questions should answered: what information is be to put into the stored content, how, and when is this information to be involved? These questions can be answered more or less independently of the computer network.

If the answers to these questions above are known, the next phase is to start with that part of the task related to the network itself. The basic chain of steps consists of the following elements:

1. Inputting the information electronically into the content-providing server
2. Processing the input in order to build up the content by using well-prepared data
3. Storing the appropriate information in the appropriate form
4. Providing ''anytime accessibility'' of the information

The first step is more or less straightforward using the available input devices (keyboard, scanner, microphone, camera, etc.)

The next steps (processing and storing) are a bit more complicated. However, this kind of processing should be built into the software belonging to the application system. This kind of software is independent of the information itself and normally is related exclusively to the computer network system being under consideration.

More important is how to find the necessary information when it is needed, whether the information is internal or external to the organization.

The task is relatively simple with internal information: the users within an organization know the system or at least have direct access to those who can supply the key to solving any problems in accessing it.

However, the task of looking for external information will not be possible without the most recent general tools. These tools include different kinds of distributed database handling techniques. More important, they include a number of efficient and convenient World Wide Web browsers, search engines, directories, general sites, portals, topical news servers, and custom information services. With these new ways of accessing the required information and a good infrastructure as a basis, there is virtually no search task that cannot be solved quickly and efficiently, provided that the requested information is recognizable.

11.2. Classification of Electronically Accessible Information Content

The more application areas there are (see Section 8 above), the more content classes can be identified. Thus, we can distinguish between:

- Public and private contents (depending on who can access them)
- Personal, administrative, and industrial contents (depending on where they were generated and who accesses them)
- Contents of varying complexity (depending on volume of information, structure of storage and accessibility, kind of information, distribution of storage, etc.)
- Contents of varying commercial availability (off-the-shelf, semi-custom, or custom generated).

These classifications determine the requirements for the contents that are under consideration. The main aspects of these requirements are:

- Availability (where and how the content can be reached)
- Accessibility (what authorization is needed for accessing the content)
- Reliability (what levels of reliability, validity, and completeness are required)
- Adequacy (what level of matching the content to the required information is expected)
- Updating (what the importance is of providing the most recent related information)
- Ease of use (what level of difficulty in accessing and using the information is allowed)
- Rating (what user groups are advised to, discouraged from, prevented from accessing the content)

Some of these aspects are closely related to the networking background (availability, accessibility), while others are more connected to the applications themselves. However, because accessing is always going on when the network is used, users may associate their possibly disappointing experiences with the network itself. Thus, reliability, adequacy, updating, ease of use, and rating (content qualification) are considered special aspects in generating and providing network accessible contents, especially because the vast majority of these contents are available nowadays by using the same techniques, from the same source, the World Wide Web. Therefore, special tools are available (and more are under development) for supporting the care taken of these special aspects.

Networking techniques can't help too much with content reliability. Here the source (provider or broker) is the most important factor in how the user may rely on the accessed content. The case is similar with adequacy and updating, but here specific issues are brought in by the browsers, search techniques/engines, general sites, portal servers (those responsible for supporting the search for adequate information), and cache servers (mirroring Web content by taking care of, for example, correct updating).

Some specific questions related to rating and filtering are dealt with in the next subsection.

11.3. Rating and Filtering in Content Generation, Provision, and Access

Although the amount of worldwide-accessible information content is increasing exponentially, the sophistication, coverage, and efficiency of the tools and services mentioned in the previous two subsections are increasing as well. However, one problem remains: the user can never exactly know the relevance, reliability, or validity of the accessed information. This problem is well known, and many efforts are being made to find ways to select, filter, and, if possible, rate the different available sources and sites (information providers) as well as the information itself. An additional issue should be mentioned: some information content may be hurtful or damaging (to minorities, for example) or corrupting, especially to children not well prepared to interpret and deal with such content when they accidentally (or consciously) access it.

There is no good, easily applicable method available today or anticipated in the near future for solving the problem. The only option at the moment is to use flags as elementary information for establishing a certain level of rating and filtering. This way, special labels can be associated with content segments so that testing the associated labels before accessing the target content can provide information about the content, itself. These flags can provide important facts about the topics, the value, the depth, the age, and so on, of the related content and about the target visitors of the site or target audience of the content.

The flags can be associated with the content by the content provider, the content broker, the service provider, or even the visitors to the related sites. However, because appropriate flags can be determined only if the content is known, the most feasible way is to rely on the content provider. The problem is that if the provider consciously and intentionally wishes to hide the truth about the content or even to mislead the potential visitors/readers, there is practically no way of preventing such behavior. This is a real problem in the case of the above-mentioned hurtful, damaging, or corrupting contents, but also with any other contents as well, as far as their characteristics (validity, value, adequacy, age, etc.) are concerned.

If flags are present, filtering is not a difficult task. It can be performed either ‘‘manually’’ (by a human decision whether to access or not access) or even by an automated process, inhibiting the access if necessary or advisable.

However, the above procedure assumes standardized techniques and standard labeling principles/rules, as well as fair usage. There is still a lot of work to do before these conditions will generally be met (including codification).

More has to be done with respect to future intelligent machine techniques that could automatically solve the tasks of labeling and filtering. Solving this problem is extremely complex but also extremely important and urgent because of the exponentially increasing amount of content accessible through the Internet.

12. TRENDS AND PERSPECTIVES

Computer networking is a continuously developing technology. With the rapid evolution of the theory and techniques behind them, the worldwide infrastructure, the globally available services, the widening spectrum of applications, the exponentially growing amount of accessible content, and the ability and readiness of hundreds of millions of users are all going through an intensive developmental process. With the progression of the networks (sometimes simply called the Internet), together with the similarly rapid evolution of the related computing, communications, control, and media technologies and their applications, we have reached the first phase of a new historical period, the Information Society. The evolving technologies and applications penetrate every segment of our life, resulting in a major change in how people and organizations, from small communities to global corporations, live their lives and manage their activities and operations.

Behind the outlined process is the development of computer networking in the last 30 years. Global connectivity of cooperating partners, unbounded access to tremendous amounts of information, as well as virtually infinite possibilities of integrating worldwide distributed tools, knowledge, resources, and even human capabilities and talents are becoming a reality as we go rapidly down a road that we have only just started to pave. Although the future of networking is impossible to see today, some important facts can already be recognized.

The most important attributes of the evolution in computer networking are openness, flexibility, and interoperability:

- Openness, in allowing any new technologies, new solutions, new elements in the global, regional, and local networks to be integrated into the existing and continuously developing system of infrastructure, services, and applications
- Flexibility, in being ready to accept any new idea or invention, even if unforeseen, that takes global advancement a step ahead
- Interoperability, in allowing the involvement of any appropriate tool or device so that it can work together with the existing system previously in operation

These attributes stem from:

- The more conscious care taken of the hierarchical and/or distributed architecture in networking technology
- The cautious standardization processes going on worldwide
- The carefulness in taking into consideration the requirements for interfacing between evolving new tools and devices

As a result, the progression in networking continues unbroken and the proliferation of network usage unsaturated, even at the start of the third millennium. Some trends in this development process are briefly outlined in the following subsections.

12.1. Network Infrastructure

Major elements of the trends in the development of the network infrastructure can be summarized, without aiming at an exhaustive listing, as follows:

- Transmission speeds are increasing rapidly. The trend is characterized by the introduction of more fiberoptic cables; the application of wavelength division multiplexing and multiple wavelengths in the transmission technique (utilizing the fact that a single fiber can guide several different frequency waves in parallel and sharing these waves among the information coming from different sources and being transmitted to different destinations); the integration of both guided and radiated waves in data communication by combining terrestrial and satellite technologies in radio transmission; the introduction of new techniques for mobile communication; the application of a more structured hierarchy in the infrastructure; and more care being taken of ‘‘last-mile’’ connectivity (to the access speeds at the final sections of the connections, close to the individual PCs or workstations), so that multiple-Mbps cable modems and digital subscriber lines, or cable TV lines, direct satellite connections, or wireless local loops, are applied close to the user terminals.
- Active devices of the infrastructure, together with the end user hardware and software tools and devices, are improving continually, allowing more intelligence in the network.

- The number, capability, and knowhow of the public network and telecom operators is increasing, so market-oriented competition is continuously evolving. The result is lower prices, which allows positive feedback on the wider dissemination of worldwide network usage.
- Governments and international organizations are supporting intensive development programs to help and even motivate fast development. Tools applied include financial support of leading-edge research and technological development, introduction of special rules and regulations about duties and taxes related to information technologies, and, last but not least, accelerating codification on Internet-related issues.

As a result of these elements, the network infrastructure is developing rapidly, allowing similarly fast development in the service sector.

12.2. Services

The trends in the evolution of networking services, available mostly through Internet service providers (ISPs) and covering domain registration and e-mail and Web access provision, as well as further service types related to providing connectivity and accessibility, can be characterized by continuing differentiation.

Although the kinds of services don't widen considerably, the number of users needing the basic services is growing extremely fast, and the needs themselves are becoming more fragmented. This is because the use of the network itself is becoming differentiated, covering more different types of applications (see Section 12.3) characterized by different levels of bandwidth and quality demand.

Quality of service (QoS) is the crucial issue in many cases. From special (e.g., transmission-intensive critical scientific) applications to real-time high-quality multimedia videoconferencing to bandwidth-demanding entertainment applications and to less demanding e-mail traffic, the different quality-level requirements of the service result in multiple grades of QoS, from best-efforts IP to guaranteed transmission capacity services.

Different technologies, from asynchronous transfer mode (ATM) managed bandwidth services (MBS) to new developments in Internet Protocol (IP) level quality management, support the different needs with respect to QoS grades. Lower quality demand may result in extremely low prices of service.

An important trend is that access to dark fiber (optical cables as physical infrastructure elements rather than leased bandwidths or managed connectivity services) and/or just separate wavelengths (specific segments of the physical fiber capacity) is becoming available and increasingly popular among well-prepared users, besides the traditional, mainly leased-line, telecom services. While this new element in the service market assumes higher levels of technological knowhow from the user, the cost aspects may make this possibility much more attractive in some cases than buying the more widespread traditional services. This also means an important new kind of fragmentation in the market, based on buying fiber or wavelength access and selling managed transmission capacity.

Another trend is also evolving with respect to services, namely in the field of providing and accessing content on the Web. It is foreseeable that content provision services such as hierarchical intelligent caching/mirroring, as well as content-accessing services such as operating general and topical portals or general and topical sites, will proliferate rapidly in the future, together with the above-mentioned trend toward separation of content provision and content brokerage.

These trends in networking services will result in a more attractive background for the increasing amount of new applications.

12.3. Applications

An overview of the development trends in network infrastructure and networking services is not difficult to provide. In contrast, summarizing similar development trends in the far less homogeneous applications is almost impossible because of their very wide spectrum. However, some general trends can be recognized here.

First, the areas of applications do widen. Just to mention some key potential trends:

- The most intensive development is foreseeable in the commercial and business applications (e-commerce, e-business).
- Industrial applications (including the production and service industries) are also emerging rapidly; one of the main issues is teleworking, with its economic as well as social effects (*teleworking* means that an increasing number of companies are allowing or requesting part of their staff to work for them either at home or through teamwork, with team members communicating through the network and thus performing their joint activities at diverse, distant sites).

- Applications in the entertainment sector (including home entertainment) are also becoming more common, but here the trends in the prices of the service industry have a strong influence.
- Applications in governmental administration and the field of wide access to publicly available administrative information seem to be progressing more slowly at the start of the new millennium. The reasons are probably partly financial, partly political (lack of market drive and competition results in lack of elevated intensity in the development of the networking applications too).
- Applications related to science and art will also evolve rapidly, but the speed will be less impressive than in the commercial area, although internal interest and motivation complement specific market drive elements here.
- Some special factors are playing an important role in the fields of telemedicine and teleteaching (teleinstruction). Here, the healthy mix of market-driven competitive environments, governmental commitment, and wide public interest will probably result in an extremely fast evolution.

Second, some general features of the applications are showing general trends as well, independently of the application fields themselves. Two of these trends are:

- Multimedia transmission is gradually entering practically all applications, together with exploitation of the above-mentioned techniques of telepresence and teleimmersion (these techniques are characterized by combining distant access with virtual reality and augmented reality concepts; see Section 2).
- The concept of grids will probably move into a wide range of application areas. In contrast to the well-known client–server scheme, this concept exploits the potential of integrating distributed intelligence (processing capability), distributed knowledge (stored information), and distributed resources (computing power) so that a special set of tools (called middleware) supports the negotiation about and utilization of the intelligence, knowledge, and resource elements by goal-oriented integration of them within a grid. (A grid is a specific infrastructure consisting of a mutually accessible and cooperative set of the joined hosting sites, together with the distributed intelligence, knowledge, and resources and the middleware tools.)

The development trends in the field of the applications obviously deeply influence the trends in the area of content generation and content provision.

12.4. Information Content

Forecasting in the area of content generation and content provision is even more difficult than in the field of network applications. Only some basic issues can be mentioned here.

The first important trend is the emergence of a new branch of industry, the content industry. No longer are content generation and content provision secondary to building the infrastructure, providing services, and/or introducing or keeping applications. The related complex tasks (at least in the case of professional activities of a high standard) require large investments and call for adequate return on these investments. Well-organized joint efforts of highly educated and talented staff are required in such demanding activities.

The second trend is directly linked to the amount of work and money to be invested in content generation. Contents will gain more and more value in accordance with such investments. Moreover, while accessing the infrastructure and hiring services will probably become less expensive in the future, the price for accessing valuable information will probably grow considerably. The price for accessing any information (content) will soon be cost based, depending on the size of the investment and the potential number of paying users.

The third trend is also due to the increasing complexity of the tasks related to content generation and content provision. Task complexities have initiated the separation of the activities related to generating contents from those related to organizing services for content provision to the public. This means that special expertise can be achieved more easily in both of these types of demanding activities.

The fourth trend is related to storing the information. Future contents will be distributed: because transmission will get cheaper and cheaper, merging distant slices of information content will be made much cheaper by maintaining the current distance and accessing the required remote sites as needed instead of transporting these slices into a central site and integrating them on-site. However, in order to exploit this possibility, a certain level of distributed intelligence is also required, as well as inexpensive access to the connectivity services. This trend is closely related to those regarding portals and grids discussed in the previous subsection.

XML is an emerging protocol of note. Applied for exchange of structured information, it is increasingly used in e-commerce applications.

13. NETWORKING IN PRACTICE

The following subsections go into some detail about the practical issues of networking. A brief overview is provided of the suggested way of starting activities devoted to advance analysis of the environment, planning the applications and their prerequisites, designing the network itself, and implementing the appropriate hardware and software. Classification of networking solutions and implementation of network technology are investigated separately. As in the other parts of this chapter, the reader should look for a more detailed discussion in the specialized literature, some of which is listed in the References.

13.1. Methodology

Designing a good network is a complex task. It is essential that the planning and implementation phase be preceded by a thorough survey in which the following questions, at least, have to be answered:

- What is the principal goal of building the network?
- What is the appropriate type and category of the network?
- Is there a need for integration with existing networks?
- What kinds of computers and other hardware equipment are to be connected?
- What kinds of applications are to be implemented over the network?
- What kind of protocols should be supported?
- How much traffic should be carried by the network? What is the required bandwidth and throughput of the network? What is the highest expected peak load of the network?
- What is the maximum acceptable delay (latency) on the network?
- What level of reliability should be guaranteed?
- What level of security has to be reached?
- What kinds of enhancements and improvements are expected in the future? What level of scalability is desirable?
- What kinds of physical constraints (distance, trace, etc.) should be taken into consideration?

The actual process of implementation depends on the answers to these questions. The result (the adequacy and quality of the network) will greatly depend on how consistently the rules have been followed about building networks.

Designing the most appropriate topology is one of the key issues. Many other things depend on the actual topology. If the topology is not designed carefully, the network may be far from having optimum adequacy, quality, and cost/performance.

Choosing the right physical solutions and low-level protocols is also very important. Low-level protocols (TCP/IP, etc.) have a strong impact on the functionality of the network. Building a multiprotocol network is advisable only if there is an explicit need for such a selection; this may increase the costs and make the management and maintenance tasks difficult and complicated.

13.2. Classification

Classification of networks by type may help significantly in selecting the solution for a specific application. Obviously, different types of networks are designed to serve different types of application needs.

Networks can be classified by a number of attributes. Important classification factors include:

- Purpose (general, industrial, process control, etc.)
- Size (local area network, metropolitan area network, wide area network)
- Technology (ethernet, fast ethernet, token ring, fiber digital data interface [FDDI], asynchronous transfer mode [ATM], etc.)
- Applied low-level protocols (TCP/IP, IPX/SPX, Decnet, AppleTalk, etc.)
- Speed (low, medium, high, ultrahigh)
- Mode of operation (open, private, intranet, extranet, etc.)

Often more than one kind of technology and/or protocol has to be used in the same network. Especially if the size of the network is large, different segments can be built up by using different technologies and/or protocols. Fortunately, modern network equipment allows the use of different technologies and protocols at the same time within the same network. This equipment takes care of the necessary translations when routing the information traffic.

13.3. Implementation Issues

Like the design process, implementation is a complex task that can involve difficult problems. The design team should understand the available technological solutions well in order to select the most appropriate ones for each part of the system.

As an example, ethernet technology is a widespread, efficiently functioning solution in most cases, but because it does not guarantee appropriate response times, it cannot be used in certain real-time environments. Ethernet is a reliable, effective, inexpensive, and very good technology, but the applied CSMA/CD algorithm is not real time. In practice, the process to be controlled by the network may be much slower than the ethernet technology. Therefore, especially if oversized network capacity is implemented, there will be no problem in such cases. But theoretically it is not the best choice. Of course, the choice and the means of implementation are also a financial question, and all the important factors should be taken into consideration in looking for a good compromise.

Decisions about what and how to implement have a serious impact on the reliability and security of the network, too. For building a high-reliability, high-availability network, redundant components must be used. Communication lines and passive and active network components all have to be multiplied in a well-designed way to get a partly or fully fault-tolerant system. Of course, this may result in significant extra costs. Similarly, if extra-high levels of security must be guaranteed, a number of additional hardware and software components may be needed. Again, a good compromise can be found by taking into account all the reliability and security aspects and cost considerations.

14. NETWORKING IN THE PRODUCTION AND SERVICE INDUSTRIES

The application of computer networks in the production and service industries should be fitted to the required complexity and coverage.

As far as complexity of network usage, two main aspects should be taken into consideration:

1. What kind of infrastructure is built within the organization (including internal structure and technology as well as connectivity towards the outside world)? The spectrum goes from connecting a few workplaces (PCs) to each other, and possibly to the Internet, by the simplest LAN techniques and telephone modems, respectively, to using the most complex high-speed intranet and/or extranet applications with high-end workstations, or even large computer centers, together with broadband connections to the global Internet.
2. What kinds of services are applied? In this respect, the simplest solutions make possible only the simple exchange of messages between workplaces, while the other end is characterized by the exchange of the most complex multimedia information by using World Wide Web techniques, accessing large databases, applying distributed information processing, utilizing virtual environments, exploiting network-based long-distance collaboration, and so on.

Coverage of network usage also involves two important aspects:

1. What kinds of applications within the organization are introduced? The low end is simple word processing with possible electronic document handling. The other extreme is characterized by a complex system of network-based planning, computer-based distributed decision making, network-based integrated management, computer aided design, manufacturing, and testing, distributed financing by the use of computer networking, computerized human resource management, networked advertising, promotion, marketing, retailing, online sales transactions, public relations, and so on.
2. What amount and complexity of information content are generated, processed, and stored by the networked system of the organization? Here the levels are very much company specific, but it can be said that companies in both the production and the service industries can start with applications using elementary sets of information and may get ahead until virtually the entire information and knowledge base about the full spectrum of activities within the company (and also about the related outside world) are appropriately stored and processed by a well-established system of networked services and applications.

The basic principle here is that introducing computers and computer networks amplifies the strengths as well as the weaknesses of the organization. Well-organized companies can gain a lot from computerized and networked applications. However, those with considerable organizational problems mustn't look for a miracle: they will rapidly realize that these tools increase rather than decrease the problems of operating the related production or service activities.

The full process (from the first elementary steps until completing the implementation of the network-based system) is a combination of two basic components:

1. Introducing and integrating network technology (infrastructure and services) inside the company
2. Introducing and integrating networked applications into the activities within the company (together with continuously building up the information content).

If careful introduction and integration of the new technologies and methods is performed in a systematic manner and by appropriately restructuring and re-forming all the related activities within the company, the results will be higher efficiency, better performance, and lower costs, provided that good management exploits the opportunities that are made available.

Moreover, properly introduced and appropriately applied computer techniques and network technologies will not only help in getting ahead with enhancing efficiency and performance as well as cost cutting, but also result in elevated competitiveness. This way, survival against the increasing worldwide competition is supported, too.

15. SOME PRACTICAL ASPECTS OF INTRODUCING AND USING COMPUTER NETWORKS IN INDUSTRIAL ENGINEERING

As mentioned above, applying up-to-date computer networks in the production and service industries requires two basic steps:

1. Integrating network technology (Internet connectivity and accessibility) in the related organization or company
2. Integrating networking services and network-based applications (Internet technology) in the operations and activities of the organization or company.

The most important practical issues with regard to these steps are briefly investigated in the following subsections but some preliminary comments should be made.

First, there are, in principle, two distinct possibilities, based on the top-down and the bottom-up approach, respectively:

- In the case of fast top-down introduction of network technology, an advance analysis (feasibility study) should be performed, in order to avoid the potential risks associated with lack of careful preparation for taking a giant step ahead with networking in the company.
- In the case of applying the bottom-up approach, the speed of ‘‘getting networked’’ company-wide is much lower, but getting ahead step-by-step makes it possible to correct any mistakes and adapt to all the recognized internal circumstances so that the risks are minimized.

In any case, the top-down approach usually requires outside guidance from a professional company that specializes in introducing and integrating network technology in the organization and activities of industrial companies. This is normally not required with the bottom-up approach, although preliminary analysis and outside advice may help a lot there, too.

Another issue is coverage and depth in introducing networked information technology. The basic questions are where to stop with the bottom-up approach and what goals to define with the top-down approach. Coverage here means organizational (what departments to involve) and geographic (which sites to connect to the corporate network) coverage, while depth relates to the level at which network technology is introduced (which tasks to involve and what level of completion to achieve by network integration).

A further issue is to determine whether buying an off-the-shelf solution, buying a semicustom solution and adapting it to the local circumstances, or starting in-house development of a full custom solution is preferable. For small enterprises, the first alternative will be optimum; for high-end corporations, in many cases, the third. Although today the second alternative is the most common way of integrating network technology in a company, each organization should be considered separately, requiring its own specific handling.

15.1. Internet Connectivity

The introduction of Internet technology in industrial engineering starts with a preliminary definition of the company's business requirements and objectives. From these, the potential network applications can be derived. The planning and implementation phases can be started.

The first steps in the process of establishing Internet connectivity at an industrial company are to define the specifications and launch the necessary procedures:

- Consulting with experts about possibilities and goals
- Identifying what Internet applications will be introduced
- Defining human resource needs and deciding on staff issues related to networking
- Contracting an advisor/consultant to support activities related to company networking
- Estimating what bandwidth the selected applications will require
- Determining what equipment (hardware and software devices and tools) will be necessary
- Deciding about security issues and the extra equipment required
- Selecting the most appropriate Internet service provider (which may be the company itself)
- Deciding about the domain name(s) to be applied by the company
- Negotiating with the ISP about services to be bought (provided that the company is not its own ISP)
- Starting purchasing the necessary equipment

Naturally, the above steps do not follow each other in linear order. An iterative–interactive process is to be assumed, in which earlier decisions may sometimes be changed because of later recognition of specific problems, barriers, or difficulties.

The next steps should be devoted to preparations inside the company (or company sites). This means not just planning and building the in-house infrastructure (cabling, server room, etc.), but also deciding about the internal server structure (domain name server, mail server, web server, etc.). Another issue is starting the purchasing of the components (HW and SW) of the internal infrastructure and services that will be needed for covering the target applications. A further task is to prepare the tools and devices for the applications themselves.

Ideally, the full process may only take several weeks, but it may last several months (e.g., in the case of a complex infrastructure covering several sites, each bringing its own complex networking tasks into the picture).

Behind some of the above decisions and steps are several opportunities and choices (bandwidth of connectivity, level of security, server structure, equipment brands, etc.). Some also involve selecting among different technology variants. These alternatives require careful analysis before decisions are made.

It is not surprising that the Internet connection is to be upgraded later. This is normally not a problem at all and can be performed relatively simply, even if the upgrade also means a technological change.

15.2. LANs, Intranets, WANs, and Extranets

Establishing Internet connectivity is just the first step in the process. It makes communication with the outside world possible, but it doesn't yet solve the task of establishing internal connections within the enterprise.

In the case of a small enterprise and a single company site, the only additional task in establishing connectivity is to implement interconnections between the PCs and workstations used by the company staff. This is done by building the LAN (local area network) of the organization so that the resources (servers as well as PCs and workstations) are connected to the in-house infrastructure, in most cases an ethernet network, of the company. The required speed of the internal network depends on the traffic estimates. New ethernet solutions allow even gigabit per second transmission.

Once the internal LAN is established, the next question is how the users within the LAN will communicate with their partners outside the corporate LAN. Here is where the company intranet enters the picture.

The intranet is a network utilizing the TCP/IP protocols that are the basis of the Internet but belonging exclusively to the company and accessible only to the company staff. This means that outside partners can access the machines within the intranet only if they have appropriate authorization to do so. The intranet is in most cases connected to the global Internet through a firewall, so that although the websites within the intranet look and behave just like other websites outside the intranet, the firewall in front of the Intranet prevents unauthorized access.

In companies with separate, distant sites, all these sites may have their own LANs, firewalls, and intranets. In most cases these are connected to each other through the public MAN (high-speed metropolitan area network of a town or city), or sometimes through the regional WAN (extra-high-speed wide area network of a large region), although private network connections can also be applied in case of specific security or traffic requirements.

However, intranets connected through public MANs or WANs may also be interconnected so that, in spite of the geographically scattered locations of the related company sites, they behave as a single intranet. A common solution is to establish extranets so that the separate LANs are connected by a virtual private network over the public MAN or WAN and authorized access is controlled by a set of firewalls taking care of intranet-like traffic between the authorized users (company staff and authorized outside partners). Thus, extranets may also be considered as distributed Intranets accessible not only to company employees, but partially accessible to authorized users outside the organization or company.

LANs, intranets, and extranets provide a means for organizations or companies to utilize best the possibilities stemming from integrating the Internet in their operations and activities without losing security.

Although LANs, intranets and extranets are quite scalable, it is good to think ahead when planning the infrastructure and services of the company so that extension and upgrading do not occur more frequently than really necessary. The network address schema is impacted by the infrastructure, so changing network numbering can be costly.

15.3. World Wide Web Communication, Integration, and Collaboration

Once the LANs are in place, with their servers, PCs, and workstations, and the intranet/extranet infrastructure is working well, by taking care of secure interconnectivity towards the outside world, and by utilizing appropriate routing and firewall equipment, the details of the applications come into the picture. Here the best solutions can be built up by exploiting the possibilities that modern World Wide Web systems provide.

The Web is the global system of specialized Internet servers (computers delivering web pages to machines asking for them). The Web is also a tool integrating:

- Worldwide addressing (and thus accessibility)
- Hypertext technique (allowing navigation through a multitude of otherwise separate websites by following the links associated to web page content pieces)
- Multimedia capabilities (covering a wide spectrum of formats used by the Web contents, from structured text to graphics and from audio to still and real video)

Interacting with the Web content by initiating content-oriented actions, including also customizing them, is thus supported, too.

The Web is an ideal tool not only for worldwide information access, but also for communication, integration, and collaboration inside an enterprise and among enterprises (eXtensible Markup Language [XML]). The key is the combination of Web technology with the intranet/extranet environment. Practically, this means operating several web servers within the enterprise so that some of them serve only internal purposes and thus do not allow outside access, while others support communication outside the enterprise boundaries. While the latter normally differ from any other web servers only in their special property of being able to take over public content from the internally accessible web servers, the structure, linking, and content of the former are constructed so that they directly serve internal communication, integration, and collaboration.

This type of communication differs from general e-mail usage only in providing a common interface and a well-organized archive of what has been communicated through the system. This makes it possible to build the archive of the company continuously by automatically storing all relevant and recorded documents for easy, fast, and secure retrieval. The management of the organization or company can thus be considerably improved.

More important is that such a (Web-based) store of all relevant documents also take care of integrating all the activities resulting in or stemming from those documents. If a document starts its life in the system, any access to it, modification of it, attachment to it—in general, any event related to it—is uniquely recorded by the system. Thus, a reliable and efficient way of integrating the activities within the company, as well as between the company and its partners, can be achieved. Typical examples are e-commerce and e-business.

Communication and integration within an enterprise by the use of computer networks also mean high-quality support for overall collaboration, independent of where, when, and how the collaborating staff members take part in the joint activities. The only prerequisite is that they all take into consideration the rules of how to access the company websites. (Nevertheless, efficiently functioning systems are aware of these rules, too.) Different departments, from design to manufacturing, from sales to marketing, and from service provision to overall logistics, can collaborate in this way so that the company can maintain reliable and efficient operation.

However, as mentioned above, disorder in the operations of a company is amplified when World Wide Web communication, integration, and collaboration are introduced. Thus, in order to achieve a really reliable and efficient system of operations, careful introduction and well-regulated usage of the network-based tools are a must.

All this would be impossible without computer networking. And this is how industrial engineering can benefit fully from the computer network infrastructure and services within an industrial organization or company.

16. SUMMARY

Computer networking in industrial engineering is an invaluable tool for establishing and maintaining reliable and efficient operation and thus competitiveness. By utilizing the possibilities provided by the Internet and the World Wide Web, industrial enterprises can achieve results that earlier were impossible even to imagine.

However, the present level of networking technology is just the starting phase of an intensive global evolution. The development will not be stopping, or even slowing down, in the foreseeable future. New methods and tools will continue to penetrate into all kinds of human activities, from private entertainment to business life, from government to administrations to industrial engineering.

Getting involved in the mainstream of exploiting the opportunities, by getting connected to the network and introducing the new services and applications based on the Net, means being prepared for the coming developments in computer networking. However, missing these chances means not only losing present potential benefits, but also lacking the ability to join the future network-based innovation processes.

In 1999, the worldwide Internet Society announced its slogan: "The Internet is for everyone." Why not for those of us in the industrial engineering community?

REFERENCES

Ambegaonkar, P., Ed. (1997), *Intranet Resource Kit,* Osborne/McGraw-Hill, New York.

Agnew, P. W., and Kellerman, S. A. (1996), *Distributed Multimedia: Technologies, Applications, and Opportunities in the Digital Information Industry: A Guide for Users and Practitioners,* Addison-Wesley, Reading, MA.

Angell, D., and Heslop, B. (1995), *The Internet Business Companion: Growing Your Business in the Electronic Age,* Addison-Wesley, Reading, MA.

Baker, R. H. (1997), *Extranets: The Complete Sourcebook,* McGraw-Hill, New York.

Bernard, R. (1997), *The Corporate Intranet,* John Wiley & Sons, New York.

Bort, J., and Felix, B. (1997), *Building an Extranet: Connect Your Intranet with Vendors and Customers,* John Wiley & Sons, New York.

Buford, J. F. K. (1994), *Multimedia Systems,* Addison-Wesley, Reading, MA.

Cheswick, W., and Bellovin, S. (1994), *Firewalls and Internet Security,* Addison-Wesley, Reading, MA.

Conner-Sax, K., and Krol, E. (1999), *The Whole Internet: The Next Generation,* O'Reilly & Associates, Sebastopol, CA.

Croll, A. A., and Packman, E. (1999), *Managing Bandwidth: Deploying QoS in Enterprise Networks,* Prentice Hall, Upper Saddle River, NJ.

Derfler, F. (1998), *Using Networks,* Macmillan, New York.

Derfler, F., and Freed, L. (1998), *How Networks Work,* Macmillan, New York.

Dowd, K. (1996), *Getting Connected,* O'Reilly & Associates, Sebastopol, CA.

Foster, I., and Kesselman, K., Eds. (1998), *The Grid: Blueprint for a New Computing Infrastructure,* Morgan Kaufmann, San Francisco.

Goncalves, M., Ed. (1999), *Firewalls: A Complete Guide,* Computing McGraw-Hill, New York.

Guengerich, S., Graham, D., Miller, M., and McDonald, S. (1996), *Building the Corporate Intranet,* John Wiley & Sons, New York.

Hallberg, B. (1999), *Networking: A Beginner's Guide,* McGraw-Hill, New York.

Hannam, R. (1997), *Computer Integrated Manufacturing: From Concepts to Realization,* Addison-Wesley, Reading, MA.

Hills, M. (1996), *Intranet as Groupware,* John Wiley & Sons, New York.

Huitema, C. (1998), *Ipv6, the New Internet Protocol,* Prentice Hall, Upper Saddle River, NJ.

Kalakota, R., and Whinston, A. B. (1997), *Electronic Commerce: A Manager's Guide,* Addison-Wesley, Reading, MA.

Keshav, S. (1997), *An Engineering Approach to Computer Networking,* Addison-Wesley, Reading, MA.

Kosiur, D. (1998), *Virtual Private Networks,* John Wiley & Sons, New York.

Lynch, D. C., and Rose, M. T. (1993), *Internet System Handbook,* Addison-Wesley, Reading, MA.

Mambretti, J., and Schmidt, A. (1999), *Next Generation Internet,* John Wiley & Sons, New York.

Marcus, J. S. (1999), *Designing Wide Area Networks and Internetworks: A Practical Guide,* Addison-Wesley, Reading, MA.

McMahon, C., and Browne, J. (1998), *CAD/CAM: Principles, Practice and Manufacturing Management,* Addison-Wesley, Reading, MA.

Minoli, D. (1996), *Internet and Intranet Engineering,* McGraw-Hill, New York.

Minoli, D., and Schmidt, A. (1999), *Internet Architectures,* John Wiley & Sons, New York.

Murhammer, M. W., Bourne, T. A., Gaidosch, T., Kunzinger, C., Rademacher, L., and Weinfurter, A. (1999), *Guide to Virtual Private Networks,* Prentice Hall, Upper Saddle River, NJ.

Ptak, R. L., Morgenthal, J. P., and Forge, S. (1998), *Manager's Guide to Distributed Environments,* John Wiley & Sons, New York.

Quercia, V. (1997), *Internet in a Nutshell,* O'Reilly & Associates, Sebastopol, CA.

Salus, P. H. (1995), *Casting the Net: From ARPANET to INTERNET and Beyond . . . ,* Addison-Wesley, Reading, MA.

Schulman, M. A., and Smith, R. (1997), *The Internet Strategic Plan: A Step-by-Step Guide to Connecting Your Company,* John Wiley & Sons, New York.

Singhal, S., and Zyda, M. (1999), *Networked Virtual Environments: Design and Implementation,* Addison-Wesley, Reading, MA.

Smith, R. E. (1997), *Internet Cryptography,* Addison-Wesley, Reading, MA.

Smythe, C. (1995), *Internetworking: Designing the Right Architectures,* Addison-Wesley, Reading, MA.

Stout, R. (1999), *The World Wide Web Complete Reference,* Osborne/McGraw-Hill, New York.

Taylor, E. (1999), *Networking Handbook,* McGraw-Hill, New York.

Treese, G. W., and Stewart, L. C. (1998), *Designing Systems for Internet Commerce,* Addison-Wesley, Reading, MA.

Ward, A. F. (1999), *Connecting to the Internet: A Practical Guide about LAN-Internet Connectivity, Computer Networks and Open Systems,* Addison-Wesley, Reading, MA.

Wesel, E. K. (1998), *Wireless Multimedia Communications: Networking Video, Voice and Data,* Addison-Wesley, Reading, MA.

Young, M. L. (1999), *Internet Complete Reference,* Osborne/McGraw-Hill, New York.

CHAPTER 8
Electronic Commerce

Soon-Yong Choi
University of Texas

Andrew B. Whinston
University of Texas

1. **INTRODUCTION** **259**

2. **ELECTRONIC COMMERCE FRAMEWORKS** **260**
   - 2.1. Economics of the Digital Economy 261
   - 2.2. Product and Service Customization 261
   - 2.3. Flexible and Responsive Organization 262

3. **ENTERPRISE AND B2B ELECTRONIC COMMERCE** **262**
   - 3.1. Web-Based Procurement 262
   - 3.2. Contract Manufacturing 263
   - 3.3. Logistics Applications 264

4. **ELECTRONIC COMMERCE AND RETAILING** **265**
   - 4.1. Web Storefronts 265
   - 4.2. E-Retailing of Physical Products 266
   - 4.3. E-Retailing of Digital Products 266
   - 4.4. E-Retailing of Services 266

5. **PRICING IN THE INTERNET ECONOMY** **267**
   - 5.1. Security and Privacy in Transaction 267
     - 5.1.1. Types of Anonymity 268
     - 5.1.2. Tools for Privacy 268
   - 5.2. Real-Time Pricing 269
   - 5.3. Digital Product Pricing 270

6. **INTERMEDIARIES AND MARKETS** **271**
   - 6.1. Types of Intermediation 271
   - 6.2. Managing Distributed Commerce 271
   - 6.3. Association and Alliance 272

7. **ONLINE TRADING MARKETS AND AUCTIONS** **273**
   - 7.1. Types of Auctions 273
   - 7.2. B2B Trading Markets 275
   - 7.3. Auctions in Consumer Markets 275
   - 7.4. Reverse Auctions 275
   - 7.5. Emerging Market Mechanisms 276
     - 7.5.1. Double Auction 277
     - 7.5.2. Bundle Trading 277

8. **OUTLOOK AND CHALLENGES** **277**

REFERENCES **278**

1. INTRODUCTION

Unlike applications based on electronic data interchange (EDI) and other previous uses of computer networks, the Internet has brought integration and versatility to existing computer and network technologies to the extent that firms and consumers are said to be in a virtual economic arena. A variety of commercial and economic activities fall within the realm of electronic commerce as long as they are carried out in the electronic marketplace. In this chapter, we present an overview of business and industrial applications of the Internet technologies.

Specific advantages of the Internet over previous closed, proprietary networks are numerous. First, the investment cost necessary to establish an Internet presence is relatively small compared to earlier

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

private value-added networks, which limited EDI applications to large corporations. Lower costs in turn allow small firms and individuals to be connected to the global network. Open TCP/IP protocols of the Internet also ensure that communicating parties can exchange messages and products across different computing platforms and geographic regional boundaries.

In physical markets, geographical distance and political boundaries hinder the free movement of goods and people. Similarly, closed proprietary networks separate virtual markets artificially by establishing barriers to interoperability. This is equivalent to having a railway system with different track widths so that several sets of identical rail cars must be maintained and passengers must be transferred at all exchange points.

Neither computing nor networking is new to businesses and engineers. Large-scale private networks have been an essential ingredient in electronic data interchange, online banking, and automatic teller machines. Business investments in information technology over the past decades have enabled firms to reengineer manufacturing, inventorying, and accounting processes. Nevertheless, the strength of the Internet lies in its nature as an open network. Economically speaking, the open Internet allows easy entry and exit into a market because of lower costs and greater market reach. A store located on a small tropical island can effectively reach global partners and consumers, collaborating and competing with multinational corporations without investing in international branches and sales presence.

While computers and networking technologies have advanced steadily over the past decades, they have lacked the characteristics of a true infrastructure. An infrastructure needs to be open and interoperable so as to allow various private enterprises with differing products and goals to collaborate and transact business in a seamless environment. As an infrastructure, the Internet provides open connectivity and uniformity as the first technological medium of its kind that supports a persistent development of universal applications and practices.

2. ELECTRONIC COMMERCE FRAMEWORKS

An apparent rationale for implementing electronic commerce is to reduce transaction costs related to manufacturing, distribution, retailing, and customer service. Many such uses involve automating existing processes through the use of computers and networks. But more importantly, new technologies now enable economic agents to move from simple automation to process innovation and reengineering. The complex web of suppliers, distributors, and customers doing business on the World Wide Web is allowing businesses to transform traditional markets and hierarchies into a new form called a network organization. Unlike hierarchies and centralized markets common in the physical economy, this structure based on networks allows a high degree of flexibility and responsiveness, which have become two pillars of the digital economy (see Section 2.3).

The Internet-based economy is multilayered. It can be divided into several layers that help us in grasping the nature of the new economy. Barua et al. (1999) have identified four layers of the Internet economy in their measurement of the Internet economy indicators. The first two, Internet infrastructure and Internet applications layers, together represent the IP or Internet communications network infrastructure. These layers provide the basic technological foundation for Internet, intranet, and extranet applications. The intermediary/market maker layer facilitates the meeting and interaction of buyers and sellers over the Internet. Through this layer, investments in the infrastructure and applications layers are transformed into business transactions. The Internet commerce layer involves the sales of products and services to consumers or businesses. According to their measurements, the Internet economy generated an estimated $301 billion in U.S. revenues and created 1.2 million jobs in 1998. Estimates of revenues and jobs contributions by each layer are presented in Table 1.

TABLE 1 Internet Revenues and Jobs in 1998, U.S.

	Estimated Internet Revenues (millions of dollars)	Attributed Internet Jobs
Internet infrastructure layer	114,982.8	372,462
Applications layer	56,277.6	230,629
Intermediary/market maker layer	58,240.0	252,473
Internet commerce layer	101,893.2	481,990
Total	301,393.6	1,203,799

Source: Barua et al. 1999. Reprinted with permission.

2.1. Economics of the Digital Economy

The digital revolution is often viewed as the second industrial revolution. But why does the Internet have such a great effect on business activities and the economy? How is the Internet-driven economy different from the previous industrial economy? Despite its obvious usefulness, a comparison to the industrial revolution is misleading—the digital revolution operates on quite different premises. In many respects, the digital revolution is undoing what we achieved in the previous age of industrial production. For example, the primary commodity of the digital age—information and other knowledge-based goods—behaves quite differently than industrial goods.

Industrial goods and production technologies that can churn out millions of an item with the least unit cost have been the hallmark of the modern economy. From ordinary household goods such as silverware and dishes to mass-produced industrial goods like automobiles and consumer appliances, increasing availability and decreasing price of these goods have brought an unimaginable level of mass consumption to the general public. Nevertheless, mass-produced industrial goods, typified by millions of identical Ford Model Ts, are standardized in an effort to minimize costs and as a result are unwieldy in fitting individual needs.

The characteristics of the industrial economy are summarized in Table 2. Business processes of an industrial firm are optimized for a supply-driven commerce, while the digital economy is geared toward customer demand. The economics of industrial goods has promoted least-cost solutions and a pervasive focus on costs that has become the limiting factor in both product choices offered to customers and manufacturing options open to producers. Values are created not from maximizing user satisfaction but from minimizing costs, not from flexibility in production but from production efficiency, which often disregards what the customers want and need. Value creation in the Industrial Age flows in a linear, rigid, inflexible, and predetermined stage of preproduction research, manufacturing, marketing, and sales. The need to minimize costs is so overwhelming that firms apply the same cost economics to nonmanufacturing stages of their business, such as distribution, inventory management, and retailing.

Partly because of the economic efficiency achieved during the industrial revolution, manufacturing now rarely accounts for more than half of a firm's total operating costs. Product research, marketing and advertising, sales, customer support, and other nonproduction activities have become major aspects of a business organization. This trend towards a nonmanufacturing profile of a firm is reflected in today's focus on business strategies revolving around quality management, information technology, customer focus, brand loyalty, and customization.

The Internet economy departs from the cost-minimization economics of the industrial age, but this transformation is not automatic simply because one is dealing with digital goods. For example, typical information goods such as news and databases are subject to the same economics as industrial goods as long as they are traded as manufactured goods. Cost minimization is still a necessary concern in the newspaper business. Limitations of the industrial age will translate into the Internet economy even when newspapers and magazines are put on the Web if these online products are nothing more than digitized versions of their physical counterparts. Many content producers and knowledge vendors may be selling digital goods but be far from participating in the digital economy if their products still conform to the cost-minimization economics of the industrial age.

2.2. Product and Service Customization

Knowledge is a critical part of economic activities in both industrial and digital economies, but they differ significantly in the way knowledge is utilized. While the main focus in generating and applying knowledge during the industrial age was on maximizing efficient production through lower costs, the use of knowledge in the digital economy focuses on providing customers with more choices. Instead of standardizing products, the digital revolution drives firms to focus on maximizing customer satisfaction by customizing products and meeting consumption needs.

To offer more choices and satisfaction to customers, business processes must be flexible and responsive. Web-based supply chain management, trading through online auctions, targeted marketing

TABLE 2 Industrial Economy vs. Digital Economy

Industrial Economy	Digital Economy
Supply-driven	Demand-driven
Cost minimization	Value maximization
Standardization	Customization
Linear value chain	Nonlinear value Web
Price competition	Service competition

and sales, and interactive customer service create values not simply by reducing costs but by allowing firms to be responsive to customers' needs.

2.3. Flexible and Responsive Organization

Just as new products have been born from the technologies of the Internet, so has a new organizational form. The nonlinear technology of the Web makes it possible to have an organization that represents the highest level of flexibility. In this new form, defining or classifying virtual firms and markets based on traditional organizational structures such as a hierarchy or an M-form can be very difficult. Indeed, a very flexible organization may exist only as a network organization that defies any structural formula.

In physical markets, a firm is organized into a functional hierarchy from the top-level executive to divisions and managers on down the hierarchy. It may structure its divisions following various product groups that rarely intersect in the market. The markets are organized under the natural order of products and producers from materials, intermediate goods, consumption goods, distributors, and retailers. Firms operating in traditional physical markets organize their business activities in a linear fashion. After the planning and product selection stage, the materials and labor are collected and coordinated for manufacturing. The manufactured products are then handled by the distribution and marketing divisions, followed by sales and customer service activities. These functions flow in a chain of inputs and outputs, with relatively minor feedback between stages. The value embedded in the materials is increased in each step of the manufacturing and marketing processes by the added labor, materials, and other inputs. In a linear market process such as this, the concept of the value chain can be used to highlight the links between business processes.

On the other hand, a networked economy is a mixture of firms that is not restricted by internal hierarchies and markets and does not favor controlled coordination like an assembly line. Businesses operating in this virtual marketplace lack incentives to maintain long-term relationships—based on corporate ownership or contracts—with a few suppliers or partners. Increasingly, internal functions are outsourced to any number of firms and individuals in a globally dispersed market.

Rather than adhering to the traditional linear flow, the new digital economy will reward those that are flexible enough to use inputs from their partners regardless of where they are in the linear process of manufacturing. In fact, the linear value chain has become a "value web" where each and every economic entity is connected to everyone else and where they may often function in a parallel or overlapping fashion. Electronic commerce then becomes an essential business tool to survive and compete in the new economy.

3. ENTERPRISE AND B2B ELECTRONIC COMMERCE

Sharing information within business and between businesses is nothing new. Electronic data interchange (EDI) has allowed firms to send and receive purchase orders, invoices, and order confirmations through private value-added networks. Today's EDI now allows distributors to respond to orders on the same day they are received. Still, only large retailers and manufacturers are equipped to handle EDI-enabled processes. It is also common for consumers to wait four to six weeks before a mail order item arrives at their door. Special order items—items not in stock—at Barnes & Noble bookstores, for example, require three to four weeks of delay. In contrast, an order placed on a website at Land's End (a clothing retailer) or an online computer store arrives within a day or two.

The business use of the Internet and electronic commerce enables online firms to reap the benefits of EDI at lower cost. The ultimate fast-response distribution system is instantaneous online delivery, a goal that a few e-businesses in select industries have already achieved. By their very nature, on-demand Internet audio and video services have no delay in reaching customers. In these examples, the efficiency stems from highly automated and integrated distribution mechanisms rather than from the elimination of distribution channels as in more traditional industries.

3.1. Web-Based Procurement

A traditional business's first encounter with e-commerce may well be as a supplier to one of the increasingly common Internet Web stores. Supply chain management is in fact a key, if not a critical, factor in the success of an Internet retailer. The number of products offered in a Web store depends not on available shelf space but on the retailer's ability to manage a complex sets of procurement, inventory, and sales functions. Amazon.com and eToys (http://www.etoys.com), for example, offer 10 times as many products as a typical neighborhood bookstore or toy shop would stock. The key application that enables these EC enterprises is an integrated supply chain.

Supply chain management refers to the business process that encompasses interfirm coordination for order generation, order taking and fulfillment, and distribution of products, services, and information. Suppliers, distributors, manufacturers, and retailers are closely linked in a supply chain as independent but integrated entities to fulfill transactional needs.

In physical markets, concerns about existing investments in warehouses and distribution systems often outweigh the desire and cost to implement fast response delivery. Some retailers still rely 100% on warehouses and distribution centers to replenish its inventory. Other retailers such as Tower Records have moved away from warehousing solutions to drop shipping, which entails shipping directly from manufacturers to retail stores. Drop shipping, just-in-time delivery, and other quick-response replenishment management systems address buyers' concerns about delays in receiving the right inventory at the right time. Many quick-response systems take this a step further and empower suppliers to make shipments on their own initiative by sharing information about sales and market demand.

Applications for a supply chain did not appear with the Internet or intranets. The supply chain concept evolved during previous decades to address manufacturers' needs to improve production and distribution where managing parts and inventories is essential in optimizing costs and minimizing production cycles. EDI on the Internet and business intranets, used by large corporations such as General Electric Information Services and 3M, are aimed at achieving efficient supply management at the lower costs afforded by the Internet. Intermediaries such as FastParts (http://www.fastparts.com) further facilitate corporate purchasing on the Internet (see Section 7.2 below for a detailed discussion of B2B auctions). Traditional retailers such as Dillard's and Wal-Mart have implemented centralized databases and EDI-based supply and distribution systems. The Internet and EC now allow small to medium-sized retailers to implement technologies in their procurement system through a low-cost, open, and responsive network.

An efficient procurement system and software, once installed, can enable firms to cross market boundaries. In a highly scalable networked environment, the size of a retailer depends only on the number of customers it attracts, rather than on capital or the number of outlets it acquires. The infrastructure set up for one product can also be expanded for other products. Amazon.com, for example, uses its integrated back office system to handle not only books but CDs, videos, and gifts. Store designs, product database and search algorithms, recommendation systems, software for shopping baskets and payments systems, security, and server implementation for book retailing have simply been reused for other products. An efficient online retailer and its IT infrastructure can be scaled for any number of products with minimum constraints added.

3.2. Contract Manufacturing

In the digital economy, the trend toward outsourcing various business functions is growing rapidly because it offers a less costly alternative to in-house manufacturing, marketing, customer service, delivery, inventorying, warehousing, and other business processes. This would be consistent with the observation that firms in the physical economy also delegate production activities to external organizations if they find it less costly than to internalize them. As long as the cost of internal production (or service provision) is higher than the cost of contracting and monitoring, firms will prefer to outsource.

Regardless of the logic of this, it appears that the type of cost savings plays a critical role in the outsourcing decision. A study by Lewis and Sappington (1991) argues that when a firm's decision to buy vs. internally produce inputs involves improvements in production technology, more in-house production and less outsourcing is preferred. Their result does not depend on whether the subcontractor's production technology was idiosyncratic (only useful to produce the buyer's inputs) or transferable (the supplier could use its production technology and facility to service other potential buyers). In the case of transferable technology, the supplier would be expected to invest more in production technology, and thus offer lower costs, which may favor more outsourcing. Nevertheless, the buyer still preferred to implement more efficient technology himself internally.

In cases where buyer's production technology is substantially inferior and monitoring costs are significantly lower, we would expect contract manufacturing to be favored. Whether this is the case or not is mostly an empirical question. However, when determining whether to outsource, one must consider not only production cost savings but also savings and other benefits in product development, marketing, and distribution. By delegating manufacturing, a firm may better utilize its resources in nonproduction functions. Because production logistics are taken care of, it may be able to consider more diverse product specifications. In fact, many Internet-based firms are focusing on customer assets and marketing value of their reputation among consumers while delegating manufacturing and distribution to third parties such as Solectron (http://www.solectron.com), who offers global manufacturing networks and integrated supply chains (see Figure 1). The prevalence of outsourcing and subcontracting goes hand in hand with the use of information technology that facilitates horizontal coordination and relationships with suppliers (Aoki 1986).

Manufacturers with a well-recognized brand name and their own manufacturing operations have used manufacturer-pushed logistics management, where manufacturers dictate terms of distribution. On the other hand, manufacturers who are concerned with product development, market competition, and other strategic issues rely on contract manufacturers for efficient bulk manufacturing and dis-

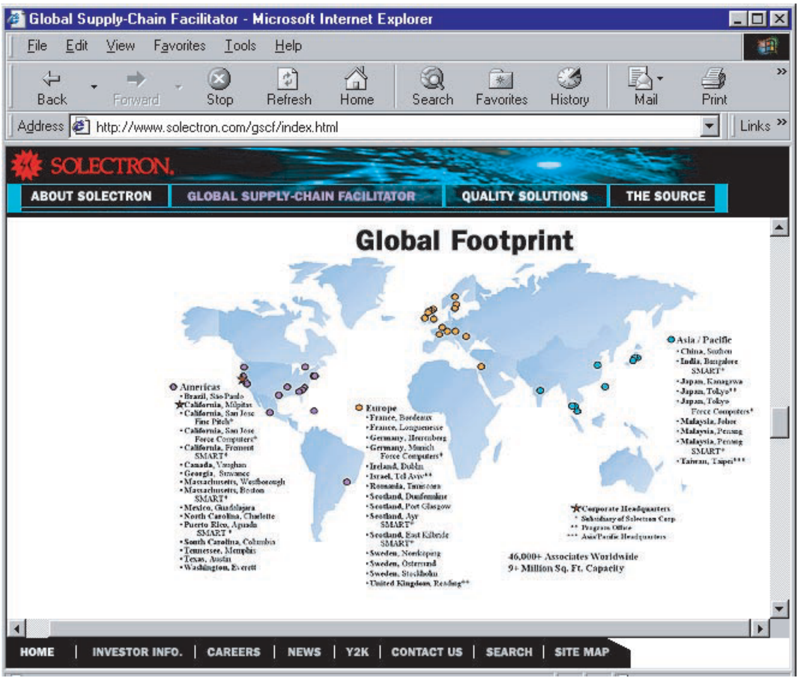

Global Supply-Chain Facilitator - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Address
http://www.solectron.com/gscf/index.html
Links
SOLECTRON.
ABOUT SOLECTRON
GLOBAL SUPPLY-CHAIN FACILITATOR
QUALITY SOLUTIONS
THE SOURCE
Global Footprint
Americas
• Brazil, São Paulo
★California, Milpitas
• California, San Jose Fine Pitch*
• California, San Jose Force Computers*
• California, Fremont SMART*
• Canada, Vaughan
• Georgia, Suwanee
• Massachusetts, Westborough
• Massachusetts, Boston SMART*
• Mexico, Guadalajara
• North Carolina, Charlotte
• Puerto Rico, Aguada SMART*
• South Carolina, Columbia
• Tennessee, Memphis
• Texas, Austin
• Washington, Everett
Europe
• France, Bordeaux
• France, Longuenesse
• Germany, Herrenberg
• Germany, Munich Force Computers*
• Ireland, Dublin
• Israel, Tel Aviv**
• Romania, Timisoara
• Scotland, Dunfermline
• Scotland, Port Glasgow
• Scotland, Ayr SMART*
• Scotland, East Kilbride SMART*
• Sweden, Norrkoping
• Sweden, Ostersund
• Sweden, Stockholm
• United Kingdom, Reading**
Asia / Pacific
• China, Suzhou
• India, Bangalore SMART*
• Japan, Kanagawa
• Japan, Tokyo**
• Japan, Tokyo Force Computers*
• Malaysia, Johor
• Malaysia, Penang
• Malaysia, Penang SMART*
• Taiwan, Taipei***
★Corporate Headquarters
* Subsidiary of Solectron Corp
** Program Office
*** Asia/Pacific Headquarters
46,000+ Associates Worldwide
9+ Million Sq. Ft. Capacity
HOME
INVESTOR INFO.
CAREERS
NEWS
Y2K
CONTACT US
SEARCH
SITE MAP

Figure 1 Global Manufacturing Networks of Contract Manufacturer Solectron. Reproduced with permission.

tributors for the ancillary tasks of moving their products to retailers. They often interact only with a few large distributors, who are expected to push assigned products down to the retail channel. Their partners are more flexible manufacturers and distributors, who have developed closer ties to end customers. For example, channel marketers in the computer industry consist of original equipment manufacturers (OEMs) who provide basic components to distributors, who build computers after orders are received. New players are now taking advantage of information technology to cut costs while delegating most functions to third parties.

Dell Computers and Gateway began a new distribution model based on direct marketing. They now rely on contract manufacturers to distribute their products. Orders received by Dell are forwarded to a third party who assembles and ships the final products directly to consumers. This built-to-order manufacturing model attains distribution objectives by outsourcing manufacturing as well as distribution functions to those who can optimize their specialized functions. Distributors, in addition to delivering products, face the task of integrating purchasing, manufacturing, and supply chain management.

For traditional in-house manufacturers and retailers, integrated distributors offer technological solutions to delegate order fulfillment and shipping tasks to outside contractors. For example, Federal Express (http://www.fedex.com) offers Internet-based logistics solutions to online retailers. FedEx introduced FedEx Ship, which evolved from its EDI-based FedEx PowerShip, in 1995. FedEx Ship is free PC software that customers can download and use to generate shipping orders. In 1996, FedEx launched its Internet version on its website, where customers can fill out pickup and shipping requests, print labels and track delivery status. As an integrated logistics operator, FedEx also offers a turnkey solution to online retailers by hosting retailer's websites or linking their servers to retailer's sites in order to manage warehouses and invoices coming from retailer's websites, then pack and ship products directly to consumers. Retailers are increasingly delegating all warehousing and shipping functions to a third party such as FedEx, while distributors expand their services into all aspects of order fulfillment.

3.3. Logistics Applications

Being ready for the digital economy means more than simply allowing customers to order via the Internet. To process orders from online customers requires seamless, integrated operation from manufacturing to delivery, readiness to handle continuous feedback from and interaction with customers, and the capability of meeting the demand and offer choices by modifying product offerings and services. Clearly, being integrated goes beyond being on the Internet and offering an online shopping basket. Manufacturing, supply chain management, corporate finance and personnel management, customer service, and customer asset management processes will all be significantly different in networked than in nonnetworked firms.

Logistics management, or distribution management, aims at optimizing the movement of goods from the sources of supply to final retail locations. In the traditional distribution process, this often involves a network of warehouses to store and distribute inventoried products at many levels of the selling chain. Manufacturers maintain an in-house inventory, which is shipped out to a distributor who stores its inventory in warehouses until new outbound orders are fulfilled. At each stage the inventory is logged on separate database systems and reprocessed by new orders, adding chances for error and delayed actions. Compaq, for example, which uses an older distribution model, has to allow almost three months for its products to reach retailers, while Dell, a direct-marketing firm, fulfill the order in two to three weeks.

Wholesalers and retailers often suffer from inefficient logistics management. Distributors may have as much as 70% of their assets in inventory that is not moving fast, while retailers receive replenishments and new products long after sales opportunities have disappeared. Optimizing distribution cycles and lowering incurred costs are a common concern for manufacturers, distributors, and retailers.

A conventional logistics management model built around warehouses and distribution centers is an efficient solution when products have similar demand structure and must be moved in the same manner. In this case, distribution centers minimize overall transportation costs by consolidating freight and taking advantage of the scale economy. This practice closely mirrors the hub-and-spoke model of airline transportation. By consolidating passenger traffic around a few regional hubs, airlines can employ larger airplanes on major legs and save on the number of flights and associated costs. Nevertheless, passengers often find that they must endure extra flying time and distance because flights between small but adjacent cities have been eliminated. While the hub-and-spoke system provides many advantages, it is too inflexible to respond to individual flying patterns and preferences.

Similarly, warehousing and distribution centers fail to function when products must move speedily through the pipeline. When market prices change as rapidly as computer components, Compaq's computers, which sit in distribution centers for several months, lose their value by the time they reach consumers. Dell's more responsive pricing is made possible by its fast-moving distribution channel.

More responsive distribution management cannot be achieved by simply removing several layers of warehouses and distribution centers. Rather, distributors in the United States are being integrated into the whole value chain of order taking, supply chain, and retailing. In the traditional logistics management model, distributors remained a disjointed intermediary between manufacturers and retailers. In an integrated logistics model that depends heavily on information technology and Web-based information sharing, distributors are fully integrated, having access to manufacturing and sales data and their partners' decision making process.

4. ELECTRONIC COMMERCE AND RETAILING

The most prominent electronic commerce application is the use of the Internet for consumer retailing that will change a firm's relationship with its customers. With the growth of online retail activities, the Internet challenges not only the survival of established retail outlets but also the very mode of transactions, which in physical markets occur through face-to-face seller–buyer interactions. Online retailers such as Amazon.com, Dell, and Garden.com (http://www.garden.com) cater to customer demands and increase revenues rapidly without opening a single retail outlet. Online shops offer consumers convenience and price advantages, critical decision factors in many retail industries.

Other observers remain skeptical about consumers abandoning their daily trips to stores and shopping malls in favor of online shopping. Physical inspections of goods, personal interactions with sales representatives, the sheer act of going to a mall with friends, and other characteristics of shopping in the physical world may limit the growth of electronic retailing. However, regardless of the future extent of electronic retailing, Internet technologies have shown that they can not only make shopping more convenient but also reorganize the retail industry to meet new demands and new desires of customers.

4.1. Web Storefronts

The Internet is the most efficient information exchange medium and interactivity tool ever to impact the retail industry. The distinguishing factors of online retailing reside in offering customers useful product information and responsive customer service in all phases of their shopping experience. Applications addressing these functions tend to be maximized by new entrants in the market. For example, Amazon.com, having no physical retail outlets to worry about, has designed a store improving the range of product and product information to match customer needs and offering fast and efficient shopping service. In contrast, players in existing retail markets are concerned with protecting their existing channels. Most retailers open their Web stores either to keep up with competitors (28%) or to explore a new distribution channel (31%) (*Chain Store Age* 1998). About 40% of retailers use their Web stores in an effort to extend their business into the virtual market.

The exceptional growth of online retailing in the United States can be traced back to several favorable characteristics. First, U.S. consumers have long and favorable previous experience with catalog shopping. Second, credit cards and checks are widely used as a preferred payment method, making it easy to migrate into the online environment. More importantly, commercial infrastructure and environment of the U.S. markets are transparent in terms of taxation, regulation, and consumer protection rules. Consumers also have access to efficient delivery networks of Federal Express and United Parcel Service in order to receive purchased products on time and in good condition. These auxiliary market factors have been essential in the initial acceptance of Web-based commerce.

4.2. E-Retailing of Physical Products

Online retailing often refers to a subcategory of business that sells ''physical'' products such as computers (Dell online store), automobiles (Auto-by-tel), clothing (Lands' End online), sports equipment (Golfweb), and flowers and garden tools (Garden.com). Currently, books and music CDs fall into this category, although retailers are becoming increasingly aware of their digital characteristics.

Electronic commerce in physical goods is an extension of catalog selling where the Internet functions as an alternative marketing channel. In this regard, online shops compete directly with physical retail outlets, leaving manufacturers to juggle between established and newly emerging distribution channels.

4.3. E-Retailing of Digital Products

Retailers of digital products have a distinct advantage over other sectors in that they deliver their goods via the Internet. Sellers of news (e.g., *The New York Times* [http://www.nytimes.com]), magazines (*BusinessWeek* [http://www.businessweek.com]), textbooks (SmartEcon [http://www.smartecon.com]), information, databases, and software can provide products online with no need for distribution and delivery by physical means. This sector also includes such search services as Yahoo, Excite, and other portals, although these firms currently rely on advertising revenues rather than pricing their digital products.

A number of online retailers are selling physical products that are essentially digital products but currently packaged in physical format. For example, Amazon.com sells printed books, which are in essence digital information products. Likewise, audio CDs and videos sold by Amazon.com and CDNow (http://www.cdnow.com) are digital products. Because digital products can be transferred via the network and are highly customizable, online retailing of these products will bring about fundamental changes that cannot be duplicated in physical markets.

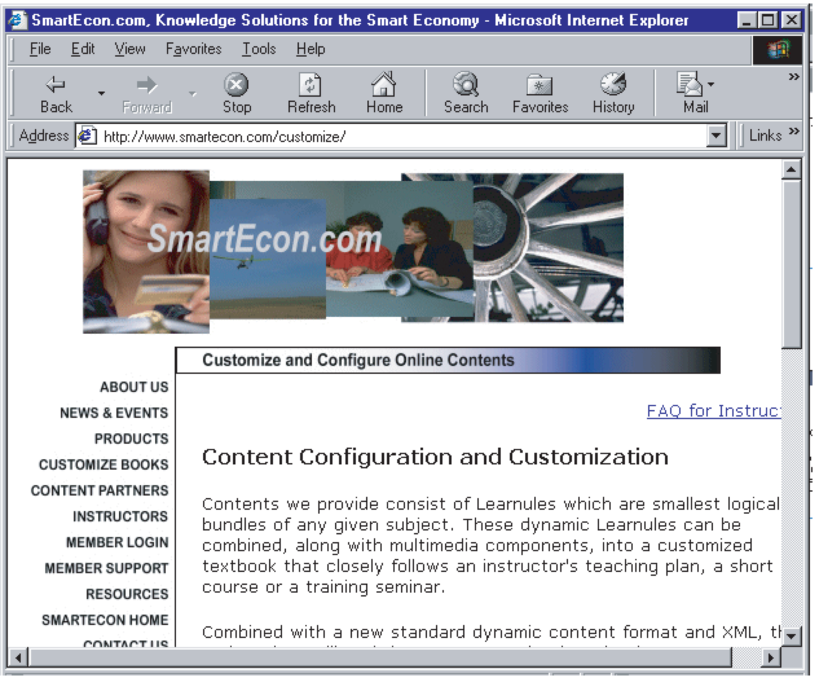
For example, instead of being an alternative distribution channel for books and CDs, online stores can offer sampling through excerpts and RealAudio files. Books and CDs are beginning to be customized and sold to customers in individualized configurations (see Figure 2), downloadable to customer's digital book readers or recordable CDs. Digitized goods can often be delivered in real time on demand. These functions add value by providing consumers more choices and satisfaction. The primary retail function is no longer getting products to customers at the lowest cost but satisfying demand while maximizing revenues by charging what customers are willing to pay.

4.4. E-Retailing of Services

A growing subcategory of electronic retailing deals with intangible services such as online gaming, consulting, remote education, and legal services and such personal services as travel scheduling, investment, tax, and accounting. Online service providers are similar to digital product sellers because their service is delivered via the Internet. Nevertheless, transactions consist of communications and interactions between sellers and buyers, often without an exchange of any final product other than a receipt or confirmation of the transaction.

A much anticipated aspect of online service delivery involves remote services, such as telemedicine, remote education, or teleconsulting. The future of remote services is critically dependent on the maturity of technologies such as 3D and virtual reality software, video conferencing on broader bandwidth, and speech and handwriting recognition. Advances in these technologies are necessary to replicate an environment where physical contact and interaction are essential for providing personal services.

Despite a low representation of remote services on today's Internet, several types of service retailing are currently practiced. For example, 2 of the top 10 online retailing activities—travel and financial services—sell services rather than products. Airline tickets are digital products that simply


SmartEcon.com, Knowledge Solutions for the Smart Economy - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Mail
Address
http://www.smartecon.com/customize/
Links
SmartEcon.com
ABOUT US
NEWS & EVENTS
PRODUCTS
CUSTOMIZE BOOKS
CONTENT PARTNERS
INSTRUCTORS
MEMBER LOGIN
MEMBER SUPPORT
RESOURCES
SMARTECON HOME
CONTACT US

Customize and Configure Online Contents

FAQ for Instruc

Content Configuration and Customization

Contents we provide consist of Learnules which are smallest logical bundles of any given subject. These dynamic Learnules can be combined, along with multimedia components, into a customized textbook that closely follows an instructor's teaching plan, a short course or a training seminar.

Combined with a new standard dynamic content format and XML, t

Figure 2 Online Textbooks Are Customized at SmartEcon.com. Reproduced with permission.

represent future uses of airline services. An airline ticket, therefore, signifies a service to schedule one's trips. Likewise, products in online stock markets and financial services are entitlements to company assets and notational currencies. Online service providers in these markets include those in investment, banking, and payment services.

5. PRICING IN THE INTERNET ECONOMY

In this section, we review some issues in pricing and payment in the Internet economy. In the physical market, prices are fixed for a certain period of time and displayed to potential customers. Such fixed and posted prices are commonly observed for mass-produced industrial goods such as books, music CDs, products sold through catalogs, and other consumer goods—products that are standardized and sold in mass quantity. Posted prices, in turn, allow both sellers and buyers to compare prices and often lead to highly competitive and uniform prices.

In the Internet economy, the menu cost of changing prices is relatively low, allowing sellers to change prices according to fluctuating demand conditions. Per-transaction costs of clearing payments also decline significantly with the use of online payment systems (see Choi et al. [1997, chap. 10] for an overview of online payment systems). In addition, sellers have access to an increasing amount of customer information as Web logs and network service databases are mined and analyzed to yield detailed information about buyers. Such developments raise serious questions about the use of identifiable customer information and how such information is used in setting prices.

5.1. Security and Privacy in Transaction

Complete and absolute anonymity is very rare in commercial transactions because the very nature of trade often requires some form of identification to verify payments or any legally required qualifications. The only transactions that come close to being completely anonymous are cash-based. Even with these, repeat business will convey some information about the buyer to the seller, not to speak of serial numbers of bills, fingerprints, security cameras, and so on. In many cases, anonymity means customer privacy is guaranteed by the seller. Technically, a Web vendor may assign a random number to each customer without linking the information gathered around that number to a physically identifiable consumer. If customer information is fully disclosed for personalization, the vendor may guarantee not to use it for any other purposes or sell it to outsiders. Whether we are concerned with anonymity or privacy, the gist of the matter is the degree to which customer information is conveyed to the seller.

5.1.1. Types of Anonymity

In general, anonymity is an extreme form of privacy in that no information is transferred. But even anonymity comes in a few varieties.

- *Complete anonymity:* With complete anonymity, sellers do not know anything about customers except that a transaction has occurred. Products are not personalized, and payments are made in cash or in an untraceable form of payment. The market operates as if sellers have no means to identify customers or learn about them in any way.
- *Incomplete anonymity:* Some anonymous customers can be traced to an identifiable person. When identities are traceable, customers have incomplete anonymity. For reasons of security and criminal justice, digital files and communications have digital fingerprints that enable authorities to trace them back to their originators, albeit with difficulty. One rationale for such a compromise is that the nature of digital products and their reproducibility threaten the very viability of intellectual property on the Internet unless unauthorized copies can be identified and traced to the culprits. Microsoft has for unknown reasons surreptitiously incorporated such measures in any document created by its Word program. Some computer manufacturers have also proposed implanting serial numbers in microprocessors so as to endow each computer with its own digital identity. This sort of information does not relate directly to any identifiable persons, although they may be traceable through usage and ownership data.
- *Pseudonymity:* A pseudonym is an alias that represents a person or persons without revealing their identity. Pseudonymity operates exactly the same way as anonymity: it can be untraceable and complete or traceable and incomplete. Pseudonymity has particular significance in the electronic marketplace because a pseudonym may have the persistency to become an online persona. In some cases, an online person known only for his or her pseudonym has become a legendary figure with a complete personality profile, knowledge base, and other personal characteristics recognized by everyone within an online community. Persistent pseudonyms are also useful in providing promotional services and discounts such as frequent flyer miles and membership discounts without disclosing identity.

Privacy is concerned with an unauthorized transfer (or collection) of customer information as well as unauthorized uses of that information. Anonymity prevents such abuses by removing identity and is therefore the most effective and extreme example of privacy. Privacy can be maintained while identifiable information is being transferred to the seller, which presents no problem in and of itself. Concerns arise when that information is used without the consent of its owner (the consumer) to affect him or her negatively (e.g., junk e-mails, unwanted advertisements, and marketing messages). What a seller can and cannot do with collected information is often decided by business convention but could ultimately be determined by law.

5.1.2. Tools for Privacy

In a marketplace, whenever a consumer need or concern arises, the market attempts to address it. Here, too, the market has created technologies that address consumers' concerns about lack of privacy. The most effective way to preserve privacy is to remove information that identifies a person by physical or electronic address, telephone number, name, website, or server address. Several models are available on the Internet.

- *Anonymity services:* Anonymity systems have devised a number of ways to strip a user's identity from an online connection. In one system, an anonymous remailer receives an encrypted message from a user. The message shows only the address to which the message should be forwarded. The remailer sends the message without knowing its content or originator. This process may continue a few times until the last remailer in the chain delivers the message to the intended destination—a person, bulletin board, or newsgroup.
- *Proxy server:* A proxy server acts as an intermediary. For example, Anonymizer.com operates a server that receives a user's request for a web page and fetches it using its own site as the originator. Websites providing the content will not be able to identify original users who requested their content.
- *Relying on numbers:* Expanding on this proxy model, a group of consumers may act as a shared computer requesting a web page. Crowds, developed by AT&T Laboratories, relies on the concept that individuals cannot be distinguished in a large crowd. Here, individual requests are randomly forwarded through a shared proxy server, which masks all identifiable information

about the users. It is interesting to note that the system tries to maintain privacy by relying on crowds when the word "privacy" suggests being apart from a crowd.

- *Pseudonyms:* Lucent's Personalized Web Assistant and Zero Knowledge Systems' Freedom rely on pseudonyms. This approach has the advantage of providing persistent online personas or identities. Persistent personas, on the other hand, can be targeted for advertisements based on historical data, just like any real person.

These and other currently available anonymity services are effective in most cases. But whether they are secure for all occasions will depend on the business's technical and operational integrity. For example, the anonymous remailer system requires that at least one of the remailers in the chain discard the sender and destination information. If all remailers cooperate, any message can be traced back to its originator. In addition, software bugs and other unanticipated contingencies may compromise the integrity of any anonymity service. Because technologies create anonymity, they can also be broken by other technologies.

Commercial transactions conducted by anonymous or pseudonymous users still have the advantage of allowing sellers to observe and collect much more refined demand data than in physical markets. For conventional market research, identifiable information is not necessary as long as it results in better demand estimation and forecast. Even without knowing who their customers are or at what address they reside, sellers can obtain enough useful data about purchasing behaviors, product preferences, price sensitivity, and other demand characteristics.

In other cases, identifiable information may be necessary for users to receive the benefits from using online technologies. Customization and online payment and delivery clearly require users' personal information. To some extent, these activities may be handled through online pseudoidentities. However, when a reasonable level of privacy is assured, customers may be willing to reveal their identity just as they do by giving out credit card information over the phone.

Current industry-led measures to protect online privacy are aimed at reducing consumers' unwillingness to come online and mitigate any effort to seek legal solutions. Essentially, these measures require the sellers to disclose clearly to consumers what type of data they collect and what they do with it. Secondly, they also encourage sellers to offer consumers a means to specify what they are willing to agree to. In a sense, sellers and consumers negotiate the terms of information collection and uses. Some websites display logos indicating that they follow these guidelines. P3P (Platform for Privacy Preferences) by the World Wide Web Consortium (W3C; http://www.w3.org) and TRUSTe from CommerceNet (http://www.commercenet.org) are the two main industry-wide efforts toward privacy in this regard.

As the trend toward privacy disclosure indicates, consumers seem to be concerned about selling personal information to third parties. Unsolicited, unrelated junk e-mails from unknown sellers are soundly rejected. They are, however, less upset about receiving advertisements, in the form of product news, from the sellers they visit. This is largely because consumers recognize the need for information in ordinary commerce and the benefit from customization. Privacy is largely an issue of avoiding unwanted marketing messages, as collecting and selling information about consumers has long been standard practice in many industries.

5.2. Real-Time Pricing

Prices change almost instantly in auctions, but posted prices may also change, albeit at a slower pace. Any price that clears the market is determined by demand and supply conditions, which fluctuate. For example, when there is an excess demand (supply), price goes up (down) until buyers and sellers agree to make a transaction. Posted prices may be fixed during a given period, but unless a seller possesses perfect information about consumer valuations, he must rely on his information about costs and the market signal received from previous sales to decide prices. To increase sales, prices must be lowered; when there is a shortage, prices can be raised. Through this trial-and-error process, prices converge on a market clearing level until changes in demand or supply conditions necessitate further price changes.

In this vein, posted price selling is a long-term version of real-time pricing. If we consider sales trends as bids made by market participants, there is little qualitative difference between posted-price selling and auctions in terms of price movements, or how market clearing prices are attained. Nevertheless, fundamental differences exist in terms of the number of market participants, the speed at which prices are determined, or in the transactional aspects of a trade, such as menu costs.

The Web is ideal for personalized product and service delivery that employs flexible, real-time changes in pricing. Vending machines and toll collectors operating on a network may charge different prices based on outdoor temperatures or the availability of parking spaces. A soda may be sold at different prices depending on location or time of day. User-identified smart cards, which also provide relevant consumer characteristics, may be required for payment so that prices can be further differentiated in real time based on their identity.

5.3. Digital Product Pricing

Pricing strategies in the digital economy undergo significant transformation from conventional cost-based pricing. Because pricing strategies are an integral part of the overall production process, it would be folly to assume that existing economic modeling and reasoning can simply be reinterpreted and applied to the digital economy. For instance, digital products are assumed to have zero or nearly zero marginal costs relative to their high initial costs needed to produce the first copy. This implies that competitive (based on the marginal cost) prices will be zero and that all types of knowledge-based products will have to be given away for free. But no firm can survive by giving out its products for free.

For example, a firm invests $1000 to produce a digital picture of the Statue of Liberty. Suppose that it costs additional $1 (marginal cost) to make a copy of the file on a floppy disk. Because of the overwhelming proportion of the fixed cost, the average cost continues to decline as we increase the level of production (see Figure 3). Its average cost will converge toward $1, its marginal cost. If the variable cost is zero, its average cost, along with its market price, will also be zero.

This ever-declining average cost poses serious problems in devising profitable pricing strategies. A leading concern is how to recover fixed cost. For example, suppose that Microsoft spends $10 million dollars for one of its software upgrades. If it finds only 10 takers, Microsoft will lose money unless it sells it at $1 million each. At $100, its break-even sales must be no less than 100,000. Microsoft, with significant market power, will find little difficulty in exceeding this sales figure. On the other hand, many digital product producers face the possibility of not recovering their initial investment in the marketplace.

For digital products or any other products whose cost curves are declining continuously, the average cost or the marginal cost has little meaning in determining market prices. No firm will be able to operate without charging customers for their products and services. Traditional price theory, which relies heavily on finding average or marginal cost, is nearly useless in the digital economy.

As a result, we must consider two possibilities. The first is that the cost structure of a digital product may be as U-shaped as that of any other physical product. In this case, the pricing dilemma is solved. For example, the variable cost of a digital product may also increase as a firm increases its production. At the very least, the marginal cost of a digital product, for example, consists of a per-copy copyright payment. In addition, as Microsoft tries to increase its sales, it may face rapidly increasing costs associated with increased production, marketing and advertising, distribution, and customer service. In this case, traditional price theory may be sufficient to guide e-business firms for pricing.

Secondly, the possibility of extreme product customization implies that each digital product has unique features that have different cost schedules. In this case, the relevant cost schedule cannot be

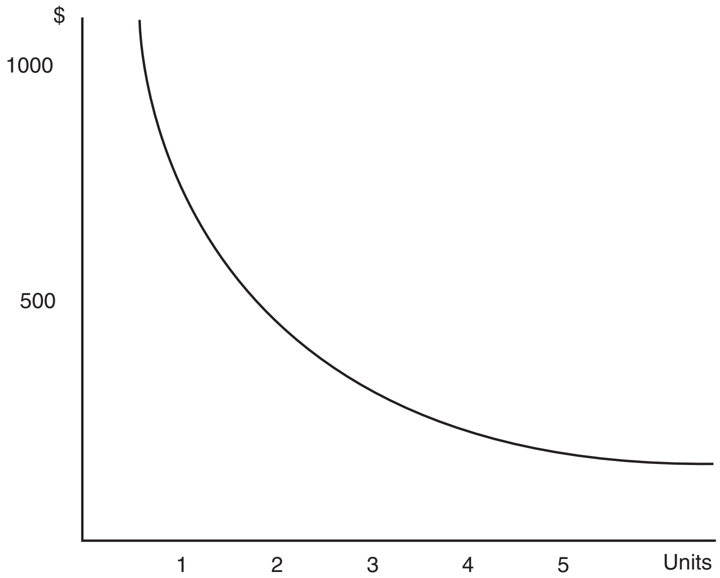


Figure 3 Decreasing Average Cost of a Digital Product Firm.

tabulated on the basis of the number of units reproduced. The marginal cost to consider is not that of reproduction, but that of production quality of a unique, personalized copy. Conventional price theory based on competitive sellers and buyers would be inadequate to deal with customized products, whether digital or physical, because of the heterogeneity of products. In such a market, cost factors have little meaning because the seller has some market power and the buyer has limited capability to substitute vendors. Instead, demand-side factors such as consumers' willingness to pay and relative positions in bargaining and negotiation determine the level of price.

6. INTERMEDIARIES AND MARKETS

The term *digital economy* makes it clear that digital technologies are more than tools that offer an alternative to conventional catalog, TV, and mall shopping activities. Digital technologies provide incentives to reorganize firms, reconfigure products and marketing strategies, and explore novel ways to make transactions. In this section, we review new forms of market mechanisms that are based on networked markets.

6.1. Types of Intermediation

Economic activities revolve around markets where sellers, buyers, and other agents must meet and interact. Therefore, the structure of the electronic marketplace plays an important role in achieving ultimate economic efficiency. How the marketplace is organized is a question that deals with the problem of matching sellers with buyers in the most efficient manner, providing complete and reliable product and vendor information and facilitating transactions at the lowest possible cost.

The three models—portals, cybermediaries, and auction markets—represent different solutions to tackling these problems:

- *Portals:* In portals, the objective is to maximize the number of visitors and to present content in a controlled environment, as in physical malls. Internet portals generate revenues from advertising or from payments by firms whose hypertext links are strategically placed on the portal's web page. Corporate portals, an extended form of intranet, do not produce advertising revenues but offer employees and customers an organized focal point for their business with the firm.
- *Cybermediaries:* Cybermediaries focus on managing traffic and providing accounting and payment services in the increasingly complex Web environment. Their revenues depend on actual sales, although the number of visitors remains an important variable. Many of them provide tertiary functions that sellers and buyers do not always carry out, such as quality guarantees, marketing, recommendation, negotiation, and other services.
- *Electronic auctions:* Electronic auction markets are organized for face-to-face transactions between sellers and buyers. The market maker, although it is a form of intermediary service, plays a limited role in negotiation and transaction between agents. Conceptually, electronic markets are a throwback to medieval markets where most buyers and sellers knew about each other's character and their goods and services. This familiarity is provided by the market maker through information, quality assessment, or guarantee. In this market, negotiations for terms of sales become the most important aspect of trade.

We find that the reason why more intermediaries are needed in electronic commerce is that the complex web of suppliers and customers—and real-time interactions with them—poses a serious challenge to fulfilling transactional requirements. Business relationships in physical markets are often hierarchical, organized along the value chain, and dominated by static arrangements in long-term contracts. Retailers typically rely on distributors and manufacturers, who in turn prefer to maintain steady and reliable relationships with suppliers. In the networked economy, however, the number of potential business partners virtually equals that of all firms and consumers. More importantly, their relationships, with no hierarchical structure, change dynamically. A simple business process such as a payment settlement between a seller and a buyer often involves several agents. For these reasons, matching sellers with buyers requires more than an advertiser-supported portal and a hypertext link. To support interactive purchases on the Internet, the electronic marketplace requires more innovative mechanisms, which we review below.

6.2. Managing Distributed Commerce

The simple act of buying an online news article may involve numerous agents who provide search information, local and out-of-state newspaper publishers who have contractual obligations, reporters and columnists (who actually produce contents), payment service providers, banks, copyright clearing houses and so on. If Yahoo or Microsoft owned most Internet firms (publishers, payment services, search sites, etc.), a customer could possibly buy (or sell) products on the company's website and necessary transactions and payments could be handled by one firm. A retailer in a physical market provides a similar service. In a grocery store, the customer collects all necessary items and makes one payment to the owner.

To enable such a convenient buying experience, a Web store must carry all items that customers need. Otherwise a list of items in a shopping basket may come from a number of different vendors, the customer having collected the items after visiting the vendors' individual websites. Options left to stores and buyers are to:

- Process multiple transactions separately
- Choose one seller, who then arranges payment clearance among the many sellers
- Use an intermediary

In a distributed commerce model, multiple customers deal with multiple sellers. An efficient market will allow a buyer to pay for these products in one lump sum. Such an arrangement is natural if a website happens to carry all those items. Otherwise, this distributed commerce requires an equally flexible and manageable mechanism to provide buyers the convenience of paying once, settling amounts due among various vendors. The intermediary Clickshare relies on a framework for distributed user management.

For example, the Clickshare (http://www.clickshare.com) model (see Figure 4) gives certain control to member sites that operate independently while Clickshare, as the behind-the-scenes agent, takes care of accounting and billing. Going beyond payment settlement, an intermediary or ''cybermediary'' not only provides member firms with accounting and billing functions but also undertakes joint marketing and advertising campaigns and exercises some control over product selection and positioning. In this regard, a cybermediary resembles a retailer or a giant discount store. It differs from an online shopping mall, which offers location and links but little else. Unlike a portal, which is inclined to own component services, this intermediation model allows members to maintain and operate independent businesses—thus it is a distributed commerce—while at the same time trying to solve management issues by utilizing an intermediary.

Amazon.com is a well-known intermediary for publishing firms. As Amazon.com expands its product offerings, it has become an intermediary or a distributor/retailer for other products and services as well. A vertical portal such as Amazon.com or a corporate portal such as Dell is well positioned to expand horizontally and become a cybermediary dealing with their own as well as others' businesses.

It must be pointed out that Yahoo, a portal, may expand in the same way. However, portals focus on either owning a share of other businesses or arranging advertising fees for the referrals. In contrast, a cybermediary's main function is to manage commercial transactions of individuals in the distributed environment in addition to providing users with a convenient shopping location.

6.3. Association and Alliance

Conventional web portals such as Yahoo and Excite try to maximize advertising revenues by increasing the number of visitors to their sites and enticing them to stay longer to view advertisements. But advertisements as a form of marketing pose the age-old question, do the viewers really buy advertised

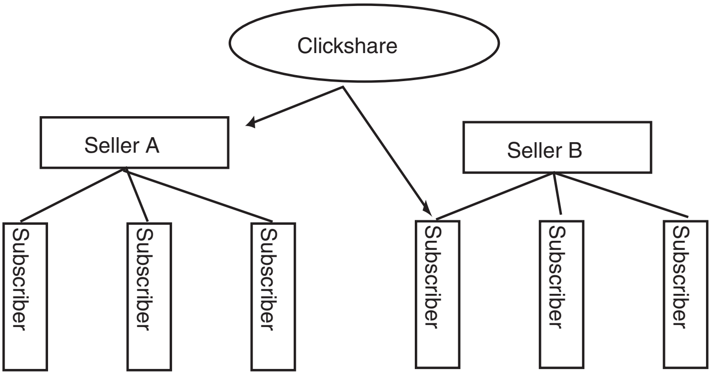


Figure 4 An Intermediary Mediates Cross-market Transactions.

products? To ensure that the message presented on visitors' web pages is producing an effect, some advertisers require click-through responses from viewers. Instead of relying on passive imprints, click-through advertisements are paid only if the viewer clicks on the ad and connects to the advertiser's website. Like hypertext links referring customers to other sites, these click-through ads are aimed at managing visitor traffic. In this way, portals become a customer referral service.

The possibility to refer and manage traffic on the Web is fully realized at Web ventures specifically designed to generate revenues from referrals. For example, Amazon.com has over 100,000 associates on the World Wide Web, who maintain their own websites where visitors may click on hypertext links that directly transport them to the Amazon.com site. Associates—that is, intermediaries—in return receive referral fees based on actual purchase by those referred by their sites. Through the Associates Program, Amazon.com has effectively opened more than 100,000 retail outlets in a couple of years. Referral fees paid to associates have replaced advertising expenses and the costs to open and manage retail stores.

Referral fees can be shared by consumers as well. SmartFrog,* for example, functions as a referrer to various online shops such as Amazon.com and eToys. After becoming a member, a Smart Frog customer visits vendor sites and, when he or she is ready to buy, goes to Smart Frog's website and clicks on the listed vendors, who pay 10% of the purchase price to Smart Frog. Out of this 10%, Smart Frog currently rebates half to its customers and keeps the remaining half.

In this way, Smart Frog, not the vendors, pays for the advertising and marketing efforts needed to generate visitors. In this cybermediated economy, the costs to advertise and attract customers are paid to entrepreneurs and consumers instead of marketing firms and the media who carry advertisements. The intermediary's profit and the benefit for consumers—in the form of lowered prices—both originate from the transaction costs previously marked for advertising and marketing. This new business opportunity and enterprise model is enabled by the distinctive nature of the networked economy.

7. ONLINE TRADING MARKETS AND AUCTIONS

Virtually all types of products are being sold through online auctions such as Onsale.com or eBay (see Figure 5). Firms and governments can implement sophisticated bidding and auction procedures for buying supplies and selling their products. Consumers can search and negotiate for best prices through auctions. By eliminating physical distance and simply increasing the number of products and trading partners available to individual bidders, online auctions offer opportunities unrealized in physical markets. Such a market arrangement, however, may produce price levels—and thus competitiveness and profit levels—that may be fundamentally different from physical markets. In this section, we present an overview of various types of auctions being implemented on the Internet and evaluate their effects on price levels, market competition, and overall economic gains and losses.

7.1. Types of Auctions

Auctions may be for a single object or a package of nonidentical items. Alternatively, auctions may be for multiple units where many units of a homogeneous, standardized good are to be sold, such as gold bullion in the auctions conducted by the International Monetary Fund and the U.S. Treasury in the 1970s and the weekly auctioning of securities by the Treasury.

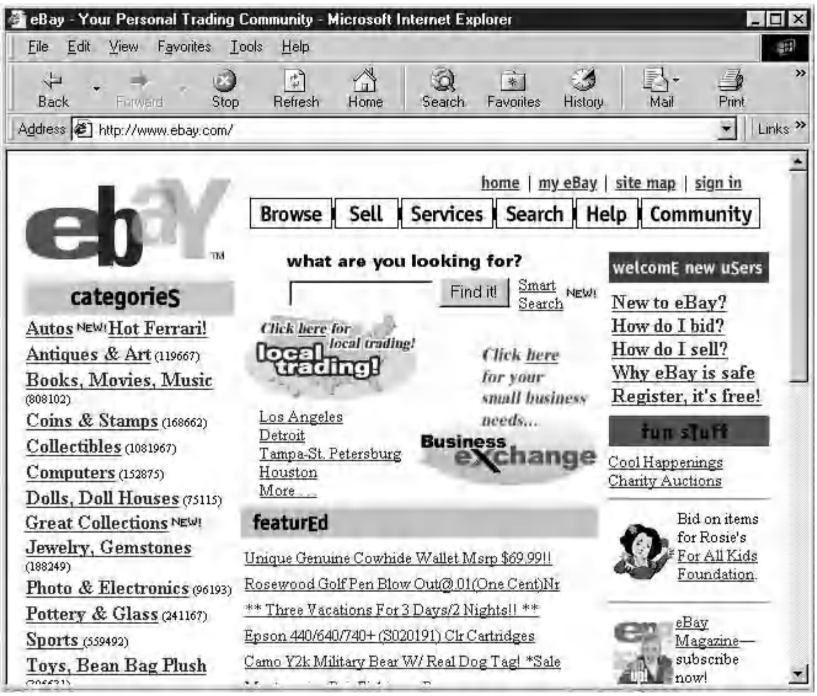
Auctions discussed in this section are single auctions where either an item is offered for sale and the market consists of multiple buyers making bids to buy or an item is wanted and the market consists of multiple sellers making offers to sell. In either case, one side of the market consists of a single buyer or seller. Most online auctions are currently single auctions. On the other hand, multiple buyers and sellers may be making bids and offers simultaneously in a double auction. An example of a double auction is a stock trading pit. Because market clearing level of price may differ substantially between double and single auctions, we discuss double auctions in Section 7.5 below.

Auctions may also be classified according to the different institutional rules governing the exchange. These rules are important because they can affect bidding incentives and thus the type of items offered and the efficiency of an exchange. There are four primary types of auctions:

1. *English auction.* An English auction, also known as an ascending bid auction, customarily begins with the auctioneer soliciting a first bid from the crowd of would-be buyers or announcing the seller's reservation price. Any bid, once recognized by the auctioneer, becomes the standing bid, which cannot be withdrawn. Any new bid is admissible if and only if it is higher than the standing bid. The auction ends when the auctioneer is unable to call forth a

---

* http://www.smartfrog.com—recently purchased by CyberGold (http://www.cybergold.com), who pays customers for viewing advertisements.


eBay - Your Personal Trading Community - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print
Address http://www.ebay.com/
Links »
home | my eBay | site map | sign in
Browse Sell Services Search Help Community
Find it!
Smart Search NEW!

what are you looking for?

categorieS

Autos NEW! Hot Ferrari!
Antiques & Art (119667)
Books, Movies, Music (808102)
Coins & Stamps (168662)
Collectibles (1081967)
Computers (152875)
Dolls, Doll Houses (75115)
Great Collections NEW!
Jewelry, Gemstones (188249)
Photo & Electronics (96193)
Pottery & Glass (241167)
Sports (559492)
Toys, Bean Bag Plush

Click here for local trading!
Los Angeles
Detroit
Tampa-St. Petersburg
Houston
More . . .

Click here for your small business needs...
Business eXchange

featurEd

Unique Genuine Cowhide Wallet Msrp $69.99!!
Rosewood Golf Pen Blow Out@.01(One Cent)Nr
** Three Vacations For 3 Days/2 Nights!! **
Epson 440/640/740+ (S020191) Clr Cartridges
Camo Y2k Military Bear W/ Real Dog Tag! *Sale

welcomE new uSers
New to eBay?
How do I bid?
How do I sell?
Why eBay is safe
Register, it's free!

fun sTuff
Cool Happenings
Charity Auctions

Bid on items for Rosie's For All Kids Foundation.

eBay Magazine—subscribe now!

Figure 5 eBay Online Auction Site. Reproduced with permission of eBay Inc. Copyright © eBay Inc. All rights reserved.

new higher bid, and the item is "knocked down" to the last bidder at a price equal to that amount bid. Examples include livestock auctions in the United States and wool auctions in Australia.

2. *Dutch auction*. Under this procedure, also known as a descending bid auction, the auctioneer begins at some price level thought to be somewhat higher than any buyer is willing to pay, and the price is decreased in decrements until the first buyer accepts by shouting "Mine!" The item is then awarded to that buyer at the price accepted. A Dutch auction is popular for produce and cut flowers in Holland, fish auctions, and tobacco. Price markdowns in department stores or consignment stores also resemble Dutch auctions.
3. *First price auction*. This is the common form of "sealed" or written bid auction, in which the highest bidder is awarded the item at a price equal to the amount bid. This procedure is thus called a sealed-bid first price auction. It is commonly used to sell off multiple units, such as short-term U.S. Treasury securities.
4. *Second price auction*. This is a sealed bid auction in which the highest bidder is awarded the item at a price equal to the bid of the second highest bidder. The procedure is not common, although it is used in stamp auctions. The multiple-unit extension of this second price sealed bid auction is called a competitive or uniform price auction. If five identical items are sold through a competitive price auction, for example, the five highest bidders will each win an item but all will pay the fifth-highest price.

The distinguishing feature of an auction market is the presence of bids and offers and the competitive means by which the final price is reached. A wide variety of online markets will qualify as an auction using this definition. Less than two years after eBay appeared on the Internet, revenues from online auctions have reached billions of dollars in the United States, projected to grow into tens of billions in a few years. Major online auctions are in consumer products such as computers, airline tickets, and collectibles, but a growing segment of the market covers business markets where excess supplies and inventories are being auctioned off.

7.2. B2B Trading Markets

Online auctions for business are an extension of supply chain applications of the Internet, by which firms seek parts and supplies from a large pool of potential business partners. The real-time interaction in these auctions differs significantly from contract-based supply relationships, which may be stable but often inefficient in terms of costs. A more flexible and responsive supply chain relationship is required as the manufacturing process itself becomes more flexible to meet changing demands in real time. Especially for digital products, the business relationship between suppliers and producers may be defined by the requirement for immediate delivery of needed components.

Currently, business-to-business auctions are mainly those that sell excess or surplus inventories. For example, FastParts (http://www.fastparts.com) and FairMarket (http://www.fairmarket.com) offer online auctions for surplus industrial goods, mainly desired by corporate purchasing departments. A significant impediment to widespread B2B auctions is the fact that online auctions allow only a single item to be exchanged at a time. However, corporate purchasing managers typically deal with thousands of items. This will require them to make and monitor bids in thousands of web pages simultaneously, each handling one item. A more efficient auction will allow them to place a bid on a combination of items. Unlike auctions organized around a single item or multiple units of a single item, the new market mechanism poses serious challenge in guaranteeing market clearance because it must be able to match these combinatorial bids and offers, unbundling and rebundling them (see Section 7.5.2).

7.3. Auctions in Consumer Markets

Auctions selling consumer goods are mainly found in used merchandise and collectibles markets. Both in Onsale.com and eBay, all types of consumer goods are offered by individuals for sale. Ordinarily these items would have been sold through classified advertisements in local newspapers or in various "for sale" newsgroups. Online innovations stem from the size of the potential audience and the real-time and interactive bidding process.

While online markets have an inherently larger reach than physical markets in terms of the number of participants the success of eBay and onsale.com is largely due to their allowing potential buyers and sellers an easy way to interact in real time. eBay, for example, is simply an automated classified ad. The list of products offered as well as those who buy and sell at eBay site resembles those found in the classified ad market. Want and for-sale ads are largely controlled by newspapers who generate as much as 30% of their total revenues from individual advertisers. Aware of the increasing threat posed by Internet-based classified ads, several newspapers experimented with online classified advertising based on their print versions. Nevertheless, large-scale initiatives by newspapers have all but failed. Surprisingly, eBay has succeeded in the same product category.

While online classified ads offered by newspapers are nothing more than online versions of their print ads, eBay offers features that consumers find convenient. Classified ads provide buyers only with contact information for purchasing a product. The buyers must make telephone calls and negotiate with the sellers. On eBay's website, all these necessary processes are automated and made convenient. Sometimes the formula of successful online business is simply to maximize the interactivity and real-time capabilities of the online medium itself.

Besides successful online versions of classified ads, another type of online auction deals with perishable goods, which need to be sold quickly. Excess airline tickets and surplus manufacturing products comprise the second type of products sold in online auction markets. For these products, the electronic marketplace offers the necessary global market reach and responsiveness.

Despite the growing interest in online auctions, the majority of consumer goods, except those discussed above, are not suitable for auctions. For these items, conventional selling such as posted price retailing will be more than adequate. Nevertheless, the flexibility offered by online trading may offer innovative market processes. For example, instead of searching for products and vendors by visiting sellers' websites, a buyer may solicit offers from all potential sellers. This is a reverse auction, discussed below. Such a buying mechanism is so innovative that it has the potential to be used for almost all types of consumer goods.

7.4. Reverse Auctions

A reverse auction is where a buyer solicits offers from sellers by specifying terms of trade that include product specification, price, delivery schedule and so on. Once interested sellers are notified and assembled, they may compete by lowering their offers until one is accepted by the buyer. Alternatively, offers may be accepted as sealed bids until one is chosen. In this regard, a reverse auction is more akin to a buyer auction, commonly found in business procurement and government contracting.

But an online implementation of a reverse auction—e.g., Priceline.com—deals with more mundane varieties of consumer goods and services. At Priceline.com website (http://www.priceline.com), consumers can specify the maximum price they are willing to pay for airline tickets, hotel rooms, automobiles and home mortgage (see Figure 6). Then Priceline.com acts as a reverse auction market as it searches and finds a seller who has the good that matches the buyer's requirements and is willing to provide the good or service at the specified terms.

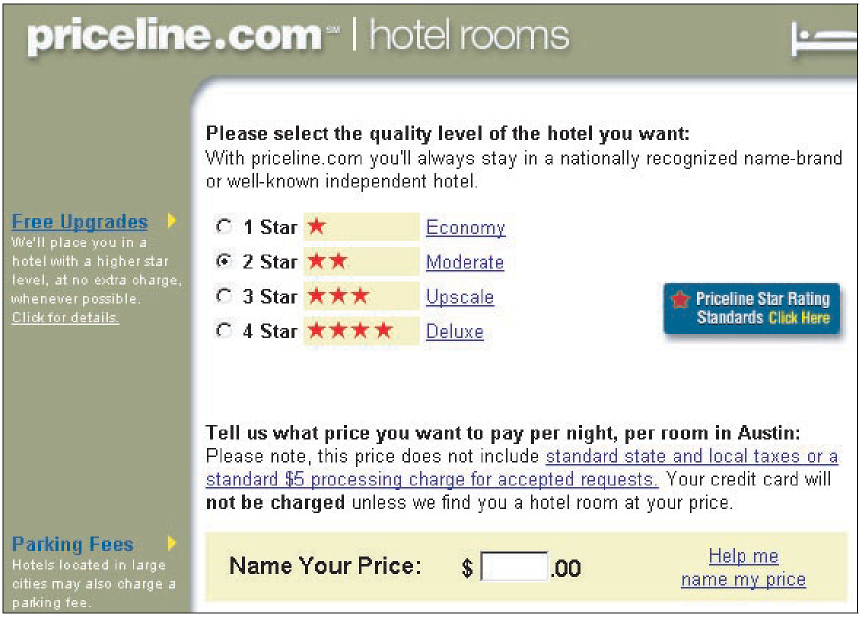
Like classified ads, the reverse auction mechanism is commonly found in physical markets. For example, it is used to determine suppliers and contractors in large-scale projects. In some seller's markets where products are perishable, sellers compete to unload their products before they become spoiled or unserviceable. Not surprisingly, Priceline.com's main source of revenues is the airline industry, where unsold seats are perishable products that cannot be saved and resold at a later date.

As in the case of building contractors and bidders, the advantage of reverse auction hinges on the limited time span of certain products and services and the existence of competition among sellers. However, it is seldom used for manufactured consumption goods. An obvious logistical problem is that there are many more buyers than sellers. This will render reverse auctions almost impossible to handle. On the other hand, online markets may offer an opportunity for reverse auctions to be used more frequently, even for most consumer goods. In a sense, reverse auctions are a form of customer-pulled marketing and an ideal selling mechanism for the digital age.

Consumer-initiated, or pulled, searches may produce some potential vendors who have matching products for sale, but contacts have to be made separately from the search result. Online reverse auctions combine search process with direct contact and negotiation with the sellers. The prospect of using a reverse auction as an online business model depends on its ability to enable consumers to specify various aspects of the good or service they intend to purchase because individual preferences may not be matched by existing products and services. In such cases, interested sellers may engage in ex post manufacturing to satisfy customer requirements. Online searches, reverse auctions, and ex post manufacturing represent the continuing innovation of the digital economy to satisfy personalized needs.

7.5. Emerging Market Mechanisms

Auctions are typically used when sellers are uncertain about market demand but want to maximize selling prices. Conversely, buyers use auctions to obtain lowest prices for contracts and supplies. If this is the case, why would sellers use online auctions when they can obtain higher prices through posed-price selling? Revenues are maximized, because since traditional auctions usually involve one seller with multiple buyers or one buyer with multiple sellers. In each of these cases, the individual who sells an object or awards a contract is always better off as the number of bidders increases


priceline.com | hotel rooms
Free Upgrades
We'll place you in a hotel with a higher star level, at no extra charge, whenever possible.
Click for details.
Parking Fees
Hotels located in large cities may also charge a parking fee.
Priceline Star Rating Standards Click Here
Help me name my price

Please select the quality level of the hotel you want:
With priceline.com you'll always stay in a nationally recognized name-brand or well-known independent hotel.

○ 1 Star ★ Economy
● 2 Star ★★ Moderate
○ 3 Star ★★★ Upscale
○ 4 Star ★★★★ Deluxe

Tell us what price you want to pay per night, per room in Austin:
Please note, this price does not include standard state and local taxes or a standard $5 processing charge for accepted requests. Your credit card will **not be charged** unless we find you a hotel room at your price.

Name Your Price: $ ☐ .00

Figure 6 Customers Name Their Prices at Priceline.com. Reproduced with permission.

(Wang 1993; Bulow and Klemperer 1996). However, if the auction consists of many sellers and buyers simultaneously, the result changes dramatically. Such an auction is called a double auction market.

7.5.1. Double Auction

In a typical auction, a single seller receives bids from multiple buyers or one buyer collects offers from multiple sellers. In a double auction, both multiple sellers and buyers submit bids and offers simultaneously, similar to trading in security markets. Multiple units of different products may be auctioned off at the same time.

A double auction closely resembles supply-and-demand interactions in physical markets. Because of this simple fact, a double auction results in a very different price level from single auctions, described above. In a single auction, the selling price may be far above the competitive level due to competition among the buyers. With many sellers and buyers, however, double auction markets tend to generate competitive outcomes. A double auction is simply an interactive form of market where both buyers and sellers are competitive.

Ideally, any effort to promote competitiveness should include expanding double auctions and similar market mechanisms in the digital economy because they offer an opportunity to raise economic efficiencies unsurpassed by any physical market organizations. For auctioneers, however, single auctions generate substantially more revenues than double auctions. Where will the incentives be for them to participate in such a competitive market? Both sellers and buyers will prefer single auctions. On the other hand, a variant type of double auction may present a unique opportunity in the digital marketplace. For example, when buyers are looking for a bundle of products and services, a double auction with many sellers and buyers is necessary to clear the market.

7.5.2. Bundle Trading

Bundle trading is an essential market mechanism when products are customized and buyer's needs must be satisfied by many sellers. Customized and personalized products often consist of a collection of complementary goods and services. For example, a combination of airline tickets, hotel rooms, a rental car, meals, and amusement park admission tickets can be bundled as a packaged leisure product. Some products that are vertically related, such as a computer operating system and a web browser, may be provided by different vendors, requiring buyers to deal with multiple sellers.

While a purchase that involves multiple sellers may be carried out through a series of transactions or auctions, bundle trading offers a simplified and efficient solution. In addition, products and services are increasingly bundled and integrated rather than being sold as separate units. As a result, a different kind of problem arises, namely how to facilitate markets that allow convenient buying and selling of a wide range of products and services in one basket or transaction.

A technological solution to bundle trading is to provide an auction that allows buyers and sellers to trade any number of goods in any combination. For example, stock trading markets such as NYSE or Nasdaq are double auctions that clear individual assets one by one. On the other hand, investors usually hold their assets in a portfolio that consists of diverse assets, consistent with their investment objectives on the overall returns and values. Nevertheless, physical markets are unable to carry out unbundling and rebundling of assets offered and demanded in the market.

The networked environment on the Internet offers a possibility of allowing such portfolio-based transactions. In essence, the desired auction mechanism must unbundle and rebundle offers and bids presented by both sellers and buyers. A prototype of such a mechanism is a portfolio trading algorithm developed for the investment community (Fan et al. 1999; Srinivasan et al. 1999). Its basic setup, however, can extend into procurement process in manufacturing as well as bundle trading in physical products and personal services.

Alternatively, third-party intermediaries may provide an agent-based solution to trading bundles. Like a travel agent, an intermediary can assemble a package of products and services to match a customer's need. Because of the increasing trend toward integrated products, intermediaries or agent-based service providers will play a greater role in the Internet economy.

8. OUTLOOK AND CHALLENGES

New technologies and applications are continually developed and applied to business processes on the Internet. The basic infrastructure of the digital economy clearly consists of computers and networking technologies. However, underlying component technologies and applications are evolving rapidly, allowing us to make only a haphazard guess as to which will turn out to be most critical or most widely accepted in the marketplace.

But not all technologies are created equal: those that aid commerce in a smart way will become critical in meeting the increasing demand for more flexibility and responsiveness in a networked commercial environment. The drive toward interoperable and distributed computing extends the very advantage of the Internet and World Wide Web technologies. Broadband networking, smart cards, and mobile network applications bring convenience and manageability in the network-centered environment. New technologies and computing models such as XML and multitier distributed computing help firms to implement more efficient management solutions. These technologies, when combined, improve and reinvent existing business processes so as to meet flexibility and responsiveness demanded by customers.

Nevertheless, regulatory as well as technical variables may become an important factor that hinders future growth of the Internet economy. Despite a high level of efficiency enabled by technologies, free markets are sometimes unable to produce efficient results when:

- Lack of information and uncertainty about products and vendors results in market failures.
- Goods and services have characteristics of public goods that private economies do not produce sufficiently.
- An industry is dominated by a monopoly or a few oligopolistic firms where outputs are reduced and prices are higher than competitive markets.
- Market players do not enjoy a transparent and universal commercial environment.

In such cases, a third party such as a government needs to intervene to provide information, promote goods production, and regulate and provide legal and commercial infrastructure. Even in the free-spirited Internet, governments play many essential roles as policy-making bodies. Business and regulatory policies as well as taxation influence firm organization and behaviors, competition, and profit levels, which ultimately determine the level of consumer welfare. (For more in-depth discussion on policies toward electronic commerce, see Choi et al. [1997] and Choi and Whinston [2000]).

Should governments regulate online markets? Various types of inefficient markets can be made efficient through regulation. A primary example is regulation of monopolies, lowering prices and increasing outputs. Health and safety regulations protect consumers. At the same time, however, regulators may be ‘‘captured’’ by those who are being regulated and protect firms’ interests against consumer welfare and market efficiency. Regulation of online commerce may have similarly controversial results. It may protect consumers’ interests and ensure efficiency. On the other hand, it may hinder the growth of online business. Regardless of its final effects, there is a growing list of online business activities that receive increasing attention from government agencies that want to regulate them.

In addition to still-evolving government policies toward e-commerce, technological factors continue to pose challenges to those doing business on the Internet. Recent episodes of distributed denial of service (DDoS) attacks on major e-commerce sites, where web servers are flooded with bogus service requests to the degree that normal customers become unable to connect, show how vulnerable Internet-based commerce can be. Despite heavy interest and large investments in security measures and technologies, DDoS attackers take advantage of the open infrastructure that lies at the bottom of the Internet telecommunications networking and are able to interrupt its normal performance through various types of requests, often using a host of third-party servers who become unwitting partners of the attack.

Technologies alone are unable to provide a sure firewall to prevent such attacks as long as the basic TCP/IP network of the Internet cannot distinguish the type of traffic that runs through or there exist some insecure servers on the network who become free riders on such attacks (Geng and Whinston 2000). Increasingly, economic analysis and tools are becoming an essential ingredient in solving technological problems.

Despite policy and technical barriers, the number of business applications based on the Internet is growing rapidly into teleconferencing, logistics support, online services in banking and medicine, customer service through chat lines, online publishing, distance learning, and broadband/mobile content delivery. In essence, the Internet has become the information infrastructure necessary to carry out various types of business operations, market processes, and transactional requirements in the fully digital economy. Any effort to reengineer business and industrial processes in the 21st century will necessarily involve the use of the Internet and its associated applications. Understanding the economic aspects of the Internet will become the final challenge in promoting and selecting a winning combination of technologies and business models.

REFERENCES

Aoki, M. (1986), ‘‘Horizontal vs. Vertical Information Structure of the Firm,’’ *American Economic Review,* Vol. 76, No. 5, pp. 971–983.

Barua, A., Pinnell, J., Shutter, J., and Whinston, A. B. (1999), ‘‘Measuring the Internet Economy: An Exploratory Study,’’ CREC/Cisco Working Paper, Center for Research in Electronic Commerce, University of Texas at Austin, Austin, TX. Available at http://crec.bus.utexas.edu/works/articles/internet_economy.pdf.

Bulow, J., and Klemperer, P. (1996), "Auctions Versus Negotiations," *American Economic Review,* Vol. 86, No. 1, pp. 180–194.

Chain Store Age (1998), "Sticking to the Web," June 1998, pp. 153–155.

Choi, S.-Y., and Whinston, A. B. (2000), *The Internet Economy: Technology and Practice,* SmartEcon Publishing, Austin, TX.

Choi, S.-Y., Stahl, D. O., and Whinston, A. B. (1997), *The Economics of Electronic Commerce,* Macmillan Technical, Publishing, Indianapolis.

Fan, M., Stallaert, J., and Whinston, A. B. (1999), "A Web-Based Financial Trading System," *IEEE Computer,* April 1999, pp. 64–70.

Geng, X., and Whinston, A. B. (2000), "Dynamic Pricing with Micropayment: Using Economic Incentive to Solve Distributed Denial of Service Attack," Center for Research in Electronic Commerce, University of Texas at Austin, Austin, TX.

Lewis, T. R., and Sappington, D. E. (1991), "Technological Change and the Boundaries of the Firm," *American Economic Review,* Vol. 81, No. 4, pp. 887–900.

Srinivasan, S., Stallaert, J., and Whinston, A. B. (1999), "Portfolio Trading and Electronic Networks," Center for Research in Electronic Commerce, University of Texas at Austin, Austin, TX.

Wang, R. (1993), "Auctions versus Posted-price Selling," *American Economic Review,* Vol. 83, No. 4, pp. 838–851.

CHAPTER 9

Enterprise Modeling

AUGUST-WILHELM SCHEER
FRANK HABERMANN
OLIVER THOMAS
Saarland University

1. MODELING BASICS 280
  1.1. Creating a Model World 280
  1.2. Levels of Abstraction 281
  1.3. Principles of Modeling 283
    1.3.1. Principle of Correctness 284
    1.3.2. Principle of Relevance 284
    1.3.3. Principle of Cost vs. Benefit 284
    1.3.4. Principle of Clarity 284
    1.3.5. Principle of Comparability 284
    1.3.6. Principle of Systematic Structure 284
2. MODELING BENEFITS 284
  2.1. Benefits for Business Administration and Organizational Processes 284
  2.2. Benefits for Developing Information Systems 285
3. MODELING VIEWS AND METHODS 286
  3.1. Process-Oriented Enterprise Modeling 286
    3.1.1. Organization Views 286
    3.1.2. Function Views 287
    3.1.3. Output Views 287
    3.1.4. Data Views 288
    3.1.5. Process View 290
  3.2. Object-Oriented Enterprise Modeling 291
4. MODELING ARCHITECTURES 293
  4.1. Architecture of Integrated Information Systems (ARIS) 293
  4.2. Information Systems Methodology (IFIP-ISM) 300
  4.3. CIM Open System Architecture (CIMOSA) 301
  4.4. Zachman Framework 302
5. MODELING TOOLS 303
  5.1. Benefits of Computerized Enterprise Modeling 303
  5.2. Characterization of Modeling Tools 303
6. MODELING OUTLOOK 306
REFERENCES 307
ADDITIONAL READING 307

1. MODELING BASICS

1.1. Creating a Model World

Enterprises are sociotechnological real-world systems. A system is a composition of elements and their relationships. Examples of elements of the system enterprise are employees, products, activities, machines, computers, and software, which interact in manifold ways. They can be described by their structures (static view) as well as their behavior (dynamic view). Because enterprises are very complex, in many cases analysis cannot be done directly on the real-world-system.

Enterprise modeling aims at reducing the complexity of the system to be studied. Only specific aspects of an enterprise are examined, such as data structures, input–output relationships, and logistic

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

processes. Such subsystems of the real world are called "mini world" or "model domain." Thus, models are descriptions of the most important characteristics of a focused domain. The creative process of building such an abstracted picture of a real-world system is called modeling (see Figure 1).

In addition to the actual purpose of modeling, the modeling methods applied also determine the focus of the model. This is particularly true when a certain type of method has already been defined. Models reproduce excerpts of reality. They are created by abstracting the properties of real objects, whereas their essential structures and behavior remain intact (homomorphy). Not only the content-related purpose of the model but also the permissible illustration methods determine to what extent nonessential characteristics may be abstracted. For example, if an object-oriented modeling approach or a system-theoretic approach is selected, modeling only leads to objects applicable to the syntax or the semantics of these particular methods. In order to avoid lock-in by certain methods, architectures and methodologies are developed independently of any particular method, while supporting a generic business process definition.

The following sections we will discuss modeling methods and architectures in greater detail.

1.2. Levels of Abstraction

Abstraction is the basic concept of modeling. In system theory, we know many perspectives of abstraction. Thus, we can concentrate either on the structure or behavior of enterprises; on certain elements of the enterprise, such as data, employees, software, products; or on a bounded domain, such as sales, accounting, manufacturing. Typically, in enterprise modeling projects, several of these perspectives are combined. For instance, a typical modeling project might be "Describe the sales data structures."

However, the term *level of abstraction* describes the relationships between a model system and the respective real-world system. That is, it describes the steps of abstraction. From low to high abstraction we can distinguish at least three levels in modeling: instances, application classes, and meta classes.

At the instance level, each single element of the real world is represented through one single element in our model. For example, in the case of modeling an individual business process, every element involved in the business process is instantiated by the affixed name, such as customer (no.) 3842, material M 32, or completed order O 4711, etc. (see Figure 2).

Enterprise modeling at the instance level is used for controlling individual business processes. In manufacturing, this is for creating work schedules as the manufacturing process descriptions for individual parts or manufacturing orders. In office management, individual business processes are executed through workflow systems. Therefore, they must have access to information regarding the respective control structure and responsible entities or devices for every single business case.

At the application class level, we abstract from instance properties and define classes of similar entities. For example, all individual customers make up the class "customer," all instances of orders constitute the class "order," and so on (see Figure 2). Every class is characterized by its name and the enumeration of its attributes, by which the instance is described. For example, the class "cus-

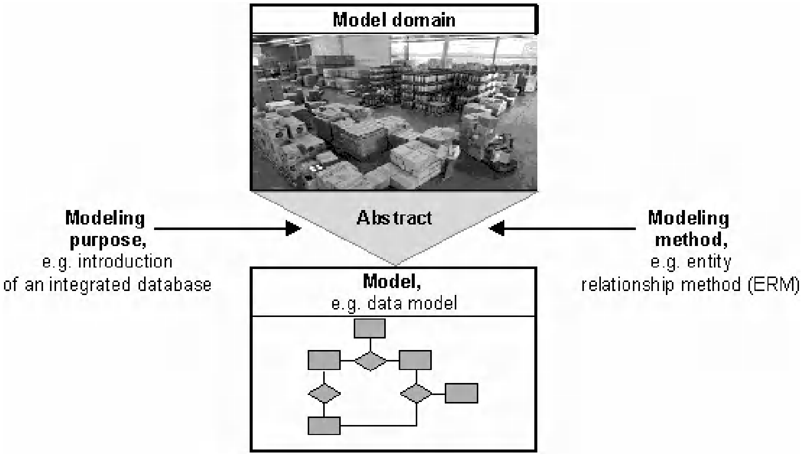

Model domain
Abstract
Modeling purpose,
e.g. introduction of an integrated database
Modeling method,
e.g. entity relationship method (ERM)
Model,
e.g. data model

Figure 1 Modeling.

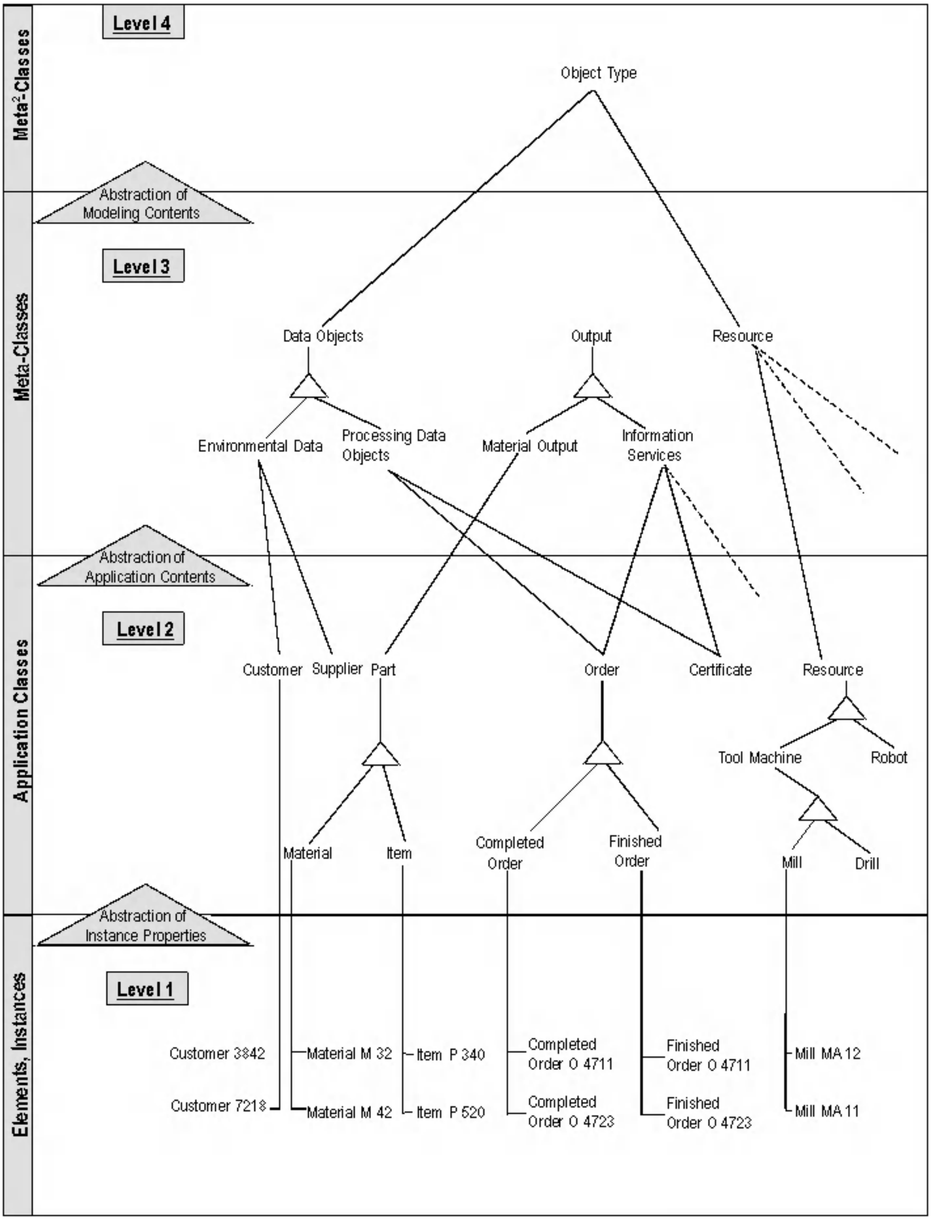

Meta²-Classes
Level 4
Object Type
Abstraction of Modeling Contents
Meta-Classes
Level 3
Data Objects
Output
Resource
Environmental Data
Processing Data Objects
Material Output
Information Services
Abstraction of Application Contents
Application Classes
Level 2
Customer
Supplier
Part
Order
Certificate
Resource
Tool Machine
Robot
Material
Item
Completed Order
Finished Order
Mill
Drill
Abstraction of Instance Properties
Elements, Instances
Level 1
Customer 3842
Customer 7218
Material M 32
Material M 42
Item P 340
Item P 520
Completed Order O 4711
Completed Order O 4723
Finished Order O 4711
Finished Order O 4723
Mill MA 12
Mill MA 11

Figure 2 Levels of Abstraction.

tomer" is characterized by the attributes customer number, customer name, and payment period. The instances of these characteristics are the focus of the description at Level 1.

Finding classes is always a creative task. It thus depends on the subjective perspective of the modeler. Therefore, when defining, for example, order designations, we will only abstract specific properties of cases 4711 or 4723, respectively, leading to the classes "completed order" or "finished order." At Level 2, we will abstract the "completed" and "finished" properties and create the parent class "order" from the subset. This operation is known as generalization and is illustrated by a triangular symbol.

When quantities are generalized, they are grouped to parent quantities. This makes order instances of Level 1 instances of the class "order" as well. The class "order" is designated as the property "order status," making it possible to allocate the process state "completed" or "finished" to every instance. Materials and items are also generalized, making them "parts" and "resources."

Thus, Level 2 contains application-related classes of enterprise descriptions. On the other hand, with new classes created from similar classes of Level 2 by abstracting their application relationships, these are allocated to Level 3, the meta level, as illustrated in Figure 2. Level 2 classes then become instances of these meta classes. For example, the class "material output" contains the instances "material" and "item" as well as the generalized designation "part." The class "information services" contains the class designation "order," along with its two child designations, and the class designation "certificate." The creation of this class is also a function of its purpose. Thus, either the generalized classes of Level 2 or their subclasses can be included as elements of the meta classes.

When classes are created, overlapping does not have to be avoided at all costs. For example, from an output flow point of view, it is possible to create the class "information services" from the classes "order" and "certificate." Conversely, from the data point of view, these are also data objects, making them instances of the class "data objects" as well.

The classes at modeling Level 3 define every object necessary for describing the facts at Level 2. These objects make up the building blocks for describing the applications at Level 2. On the other hand, because the classes at Level 2 comprise the terminology at Level 1, objects at Level 3 are also the framework for describing instances.

This abstraction process can be continued by once again grouping the classes at Level 3 into classes that are then allocated to the meta^2 level. Next, the content-related modeling views are abstracted. In Figure 2, the general class "object type" is created, containing all the meta classes as instances.

1.3. Principles of Modeling

Enterprise modeling is a creative process and can therefore not be completely directed by rules. However, if certain standards are observed, it is indeed possible to classify and understand third-party models. Furthermore, it is also a good idea to establish certain quality standards for enterprise models.

Figure 3 illustrates the relationship between principles and enterprise modeling. A principle is a fundamental rule or strategy that guides the entire process of enterprise modeling. Thus, modeling

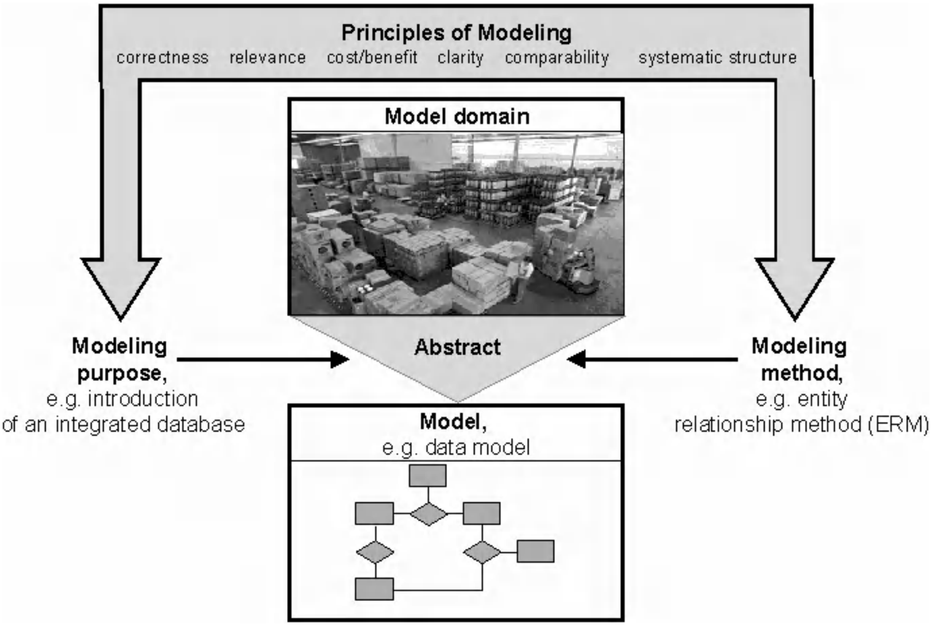


Figure 3 Principles of Modeling.

principles concern the association between the domain and the respective model as well as the selection of appropriate modeling methods and tool support. A modeling method is a set of components and a description of how to use these components in order to build a model. A modeling tool is a piece of software that can be used to support the application of a method. We will discuss modeling methods and tools in more detail in the following chapters.

1.3.1. Principle of Correctness

The correctness of models depends on correct semantics and syntax, that is, whether syntax of the respective metamodel is complete and consistent. Semantic correctness of a model is measured by how closely it complies with the structure and behavior of the respective object system. In real-world applications, compliance with these requirements can be proven only after simulation studies have been carried out or other similar efforts have been made. Some modeling tools provide a simulation function that can be used for this purpose.

1.3.2. Principle of Relevance

Excerpts of the real-world object system should only be modeled provided they correspond with the purpose of the model. Models should not contain more information than necessary, thus keeping the cost vs. benefit ratio down to an acceptable level.

1.3.3. Principle of Cost vs. Benefit

One of the key factors ensuring a good cost vs. benefit ratio is the amount of effort necessary to create the model, the usefulness of modeling the scenario, and how long the model will be used.

1.3.4. Principle of Clarity

"Clarity" ensures that a model is understandable and usable. It also determines how pragmatic the relationship between the model and the user is. Because models contain a large amount of information regarding technical and organizational issues, only specialists are usually able to understand them quickly. Once models are broken down into subviews, individual views are easier to comprehend.

1.3.5. Principle of Comparability

Models created in accordance with a consistent conceptual framework and modeling methods are comparable if the objects have been named in conformity with established conventions and if identical modeling objects as well as equivalent degrees of detailing have been used. In models created with different modeling methods, it is important to make sure that their metamodels can be compared.

1.3.6. Principle of Systematic Structure

This principle stipulates that it should be possible to integrate models developed in various views, such as data models, organizational models, and business process models. This requires an integrated methodology, that is, an architecture providing a holistic metamodel (see Section 4).

2. MODELING BENEFITS

Enterprise models are used as instruments for better understanding business structures and processes. Thus, the general benefit of enterprise modeling is in business administration and organizational processes. Focusing computer support in business, we can use enterprise models additionally to describe the organizational impacts of information technology. Consequently, the second major benefit of enterprise modeling is for developing information systems.

2.1. Benefits for Business Administration and Organizational Processes

Corporate mission statements entail the production and utilization of material output and services by combining production factors. Multiple entities and devices are responsible for fulfilling these tasks. In accordance with corporate objectives, their close collaboration must be ensured. In order for human beings to be able to handle complex social structures such as enterprises, these structures must be broken down into manageable units. The rules required for this process are referred to as "organization."

Structural or hierarchical organizations are characterized by time-independent (static) rules, such as by hierarchies or enterprise topologies. Here, relationships involving management, output, information, or communication technology between departments, just to name a few, are entered. Org charts are some of the simple models used to depict these relationships.

Process organizations, on the other hand, deal with time-dependent and logical (dynamic) behavior of the processes necessary to complete the corporate mission. Hierarchical and process organizations are closely correlated. Hierarchical organizations have been a key topic in business theory for years.

However, due to buzzwords such as business process reengineering (BPR), process organizations have moved into the spotlight in recent years.

Reasons for creating business process models include:

- Optimizing organizational changes, a byproduct of BPR
- Storing corporate knowledge, such as in reference models
- Utilizing process documentation for ISO-9000 and other certifications
- Calculating the cost of business processes
- Leveraging process information to implement and customize standard software solutions or workflow systems (see Section 2.2).

Within these categories, other goals can be established for the modeling methods. In business process improvement, we must therefore identify the components that need to be addressed. Some of the many issues that can be addressed by business process optimization are:

- Changing the process structure by introducing simultaneous tasks, avoiding cycles, and streamlining the structure
- Changing organizational reporting structures and developing employee qualification by improving processing in its entirety
- Reducing the amount of documentation, streamlining and accelerating document and data flow
- Discussing possible outsourcing measures (shifting from internal to external output creation)
- Implementing new production and IT resources to improve processing functions

In these examples, we are referring to numerous modeling aspects, such as process structures, hierarchical organizations, employee qualification, documents (data), and external or internal output as well as production and IT resources. Obviously, an enterprise model, particularly a business process model for the purpose of optimization, must be fairly complex. Moreover, it should address multiple aspects, for which numerous description methods are necessary. These various purposes determine the kind of modeling objects as well as the required granularity.

2.2. Benefits for Developing Information Systems

Information systems can be designed as custom applications or purchased as off-the-shelf standard solutions. After the initial popularity of custom applications, integrated standard solutions are now the norm. With the advent of new types of software, such as componentware (where software components for certain application cases are assembled to form entire applications), a blend between the two approaches has recently been making inroads.

The development of custom applications is generally expensive and is often plagued by uncertainties, such as the duration of the development cycle or the difficulty of assessing costs. Thus, the tendency to shift software development from individual development to an organizational form of industrial manufacturing—in “software factories”—is not surprising.

In this context, multiple methods for supporting the software development process have been developed. They differ according to their focus on the various software development processes and their preferred approach regarding the issue at hand, such as data, event, or function orientation, respectively.

Due to the wide range of methods that differ only slightly from one another, this market is cluttered. In fact, the multitude of products and approaches has actually impeded the development of computer-aided tools based on these methods. We therefore recommend a methodology (study of methods) covering various development methods. The following are typical questions that leverage the framework capabilities of a methodology:

1. Are there really so many totally different ways of designing a computer-aided information system?
2. If not, how similar are these methods? If so, why are there so many different ways?
3. Is there an optimal way of developing an information system?
4. Where does the development process start and where does it end?
5. What does the finished product of the design process look like?
6. How many steps are necessary to obtain a development result?
7. Should only one particular kind of information system be used or are several methods required, each for a different system? According to which criteria should the methods be selected?

The purpose of these questions is to classify and evaluate the various modeling methods. After these issues are addressed, there is, however, a second group of reasons for dealing with information system design methodologies (ISDMs), resulting from the fact that usually several business partners are involved in complex development projects. Sometimes they use different development methods, or the results of their work might overlap. Only an architecture integrating the individual methods, confirming agreement or pointing out any overlap, can lead to mutual understanding.

The alternative to individual software development is to buy a standardized business administration software solution. Such solutions include modules for accounting, purchasing, sales, production planning, and so on. Financial information systems are characterized by a markedly high degree of complexity. Many corporate and external business partners are involved in the implementation of information systems. This becomes apparent in light of seamlessly integrated data processing, where data is shared by multiple applications. Examples include comprehensive IS-oriented concepts implemented in enterprises, CIM in manufacturing companies, IS-supported merchandise management systems for retailers, and electronic banking in financial institutions.

Until the mid-1990s, the ratio between the effort of implementing financial packaged applications in organizations and their purchase price was frequently more than 5:1. This ratio is so high because off-the-shelf systems are more or less easy to install, yet users must also determine which goals (strategies) they wish to reach with the system, how the functionality of the system can achieve this, and how to customize, configure, and technically implement the package.

With hardware and software costs rapidly decreasing, that ratio became even worse. Small and medium-sized enterprises (SMEs) are not able to pay consultants millions of dollars for implementation. Hence, architectures, methods, and tools have become increasingly popular because they can help reduce the cost of software implementation and at the same time increase user acceptance of standard software solutions.

Several modeling approaches are possible:

- Reduce the effort necessary for creating the target concept by leveraging “best-practice case” knowledge available in reference models.
- Create a requirements definition by leveraging modeling techniques to detail the description.
- Document the requirements definition of the standard software by means of semantic modeling methods, making the business logic more understandable.
- Use semantic models to automate reconciliation of the requirements definition of the target concept with the standard software as much as possible, cutting down on the need for specific IS skills.
- Leverage semantic models as a starting point for maximum automation of system and configuration customizing.

3. MODELING VIEWS AND METHODS

3.1. Process-Oriented Enterprise Modeling

Generally speaking, a business process is a continuous series of enterprise activities, undertaken for the purpose of creating output. The starting point and final product of the business process are the output requested and utilized by corporate or external customers. Business processes often enable the value chain of the enterprise as well as focusing on the customer when the output is created.

In the following, we explain the key issues in modeling business processes with a simple example from customer order processing. First, let us outline the scenario:

> A customer wants to order several items that need to be manufactured. Based on customer and item information, the feasibility of manufacturing this item is studied. Once the order has arrived, the necessary materials are obtained from a supplier. After arrival of the material and subsequent order planning, the items are manufactured according to a work schedule and shipped to the customer along with the appropriate documentation.

This scenario will be discussed from various points of view. As we have already seen, in system theory we can distinguish between system structures and system behavior. We will begin by describing the responsible entities and relationships involved in the business process. Then, by means of function flow, we will describe the dynamic behavior. Output flows describe the results of executing the process, and information flows illustrate the interchange of documents involved in the process.

Functions, output producers (organizational units), output, and information objects are illustrated by various symbols. Flows are depicted by arrows.

3.1.1. Organization Views

Figure 4 depicts the responsible entities (organizational units) involved in the business process, along with their output and communication relationships, illustrated as context or interaction diagrams. The

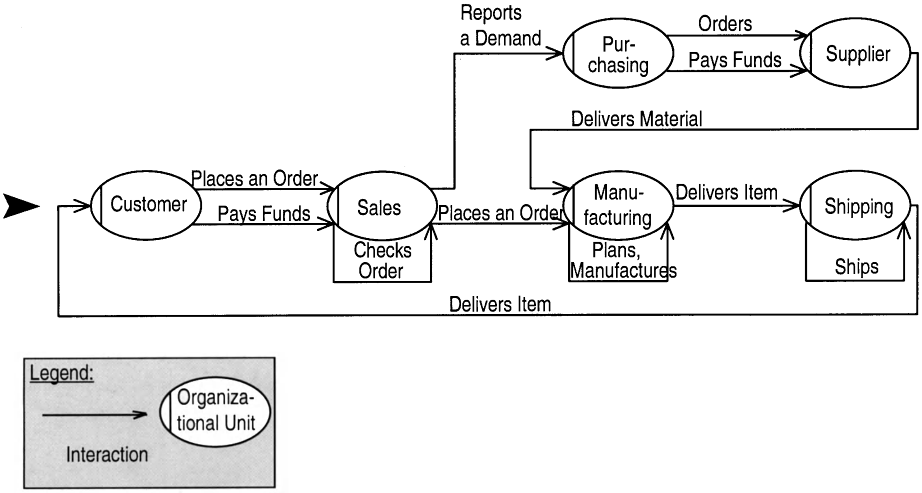


Figure 4 Interaction Diagram of the Business Process "Order Processing."

sequence in which processes are carried out is not apparent. Nevertheless, this provides an initial view of the business process structure. In complex processes, the myriad interchanges among the various business partners can become somewhat confusing. In addition to the various interactions, it is also possible to enter the activities of the responsible entities. This has been done in only a few places.

The class of organization views also includes the hierarchical organization structure (org charts). Org charts are created in order to group responsible entities or devices that are executing the same work object. This is why the responsible entities, responsible devices, financial resources, and computer hardware are all assigned together to the organization views.

3.1.2. Function Views

Figure 5 describes the same business process by depicting the activities (functions) to be executed, as well as their sequence. The main issue is not responsible entities, as with the interaction diagram, but rather the dynamic sequence of activities. For illustration purposes, the organizational units are also depicted in Figure 5. Due to redundancies, their interrelationship with the interaction diagram is not as obvious. As function sequences for creating output, function flows characterize the business process. The output flows themselves will be displayed individually.

The class of function views also includes the hierarchical structure of business activities transforming input into output. According to the level of detail, they are labeled "business processes," "processes," "functions," and "elementary functions."

Because functions support goals, yet are controlled by them as well, goals are also allocated to function views—because of the close linkage. In application software, computer-aided processing rules of a function are defined. Thus, application software is closely aligned with "functions" and is also allocated to function views.

3.1.3. Output Views

The designation "output" is very heterogeneous. Business output is the result of a production process, in the most general sense of the word. Output can be physical (material output) or nonphysical (services). Whereas material output is easily defined, such as by the delivery of material, manufactured parts, or even the finished product, the term *services* is more difficult to define because it comprises heterogeneous services, such as insurance services, financial services, and information brokering services. Figure 6 illustrates this simplified classification of "output"—that is, also "input"—as a hierarchical diagram.

Concerning our business process example, the result of the function "manufacture item" in Figure 7 is the material output, defined by the manufactured item. Likewise, quality checks are carried out and documented during the manufacturing process. All data pertinent to the customer are captured in "order documents," themselves a service by means of the information they provide. After every intercompany function, output describing the deliverable is defined, which in turn is entering the next

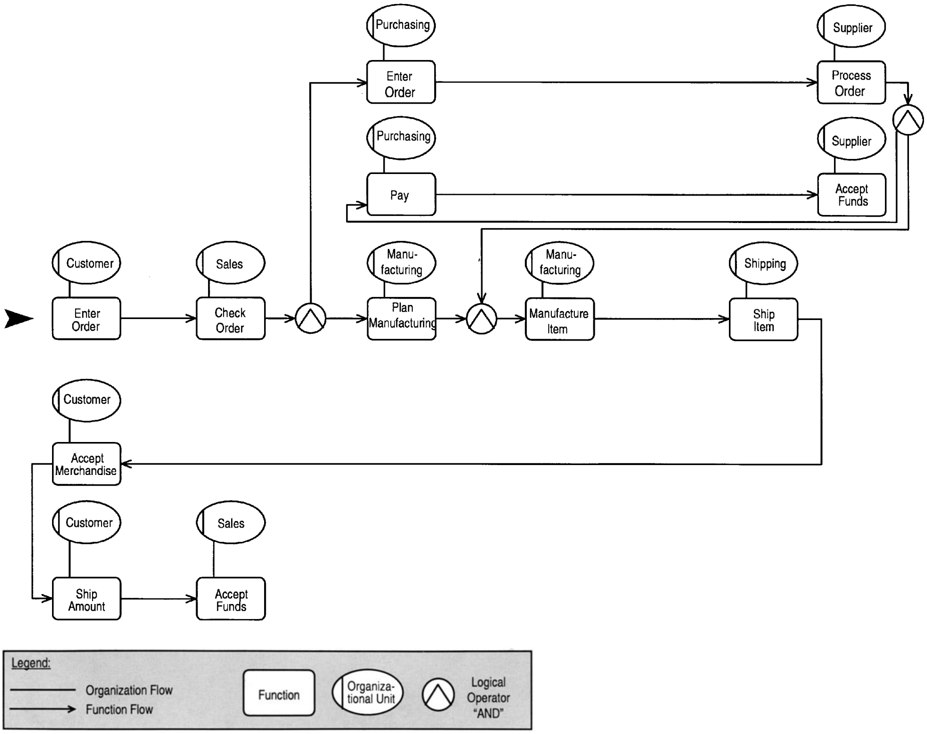

Purchasing
Enter Order
Supplier
Process Order
Purchasing
Pay
Supplier
Accept Funds
Customer
Enter Order
Sales
Check Order
Manu-facturing
Plan Manufacturing
Manu-facturing
Manufacture Item
Shipping
Ship Item
Customer
Accept Merchandise
Customer
Ship Amount
Sales
Accept Funds
Legend:
Organization Flow
Function Flow
Function
Organiza-tional Unit
Logical Operator "AND"

Figure 5 Function Flow of the Business Process "Order Processing."

process as input. To avoid cluttering the diagram, the organizational units involved are not depicted. It is not possible to uniquely derive the function sequence from the illustration of the output flow.

3.1.4. Data Views

The designations "data" and "information" are used synonymously. In addition to information services, other information, used as environment descriptions during the business processes, constitutes

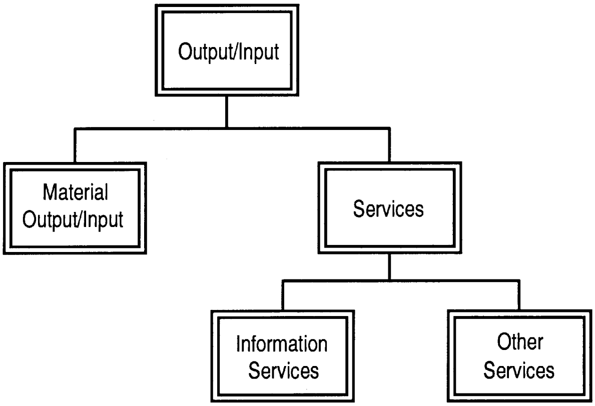

Output/Input
Material Output/Input
Services
Information Services
Other Services

Figure 6 Types of Input–Output.

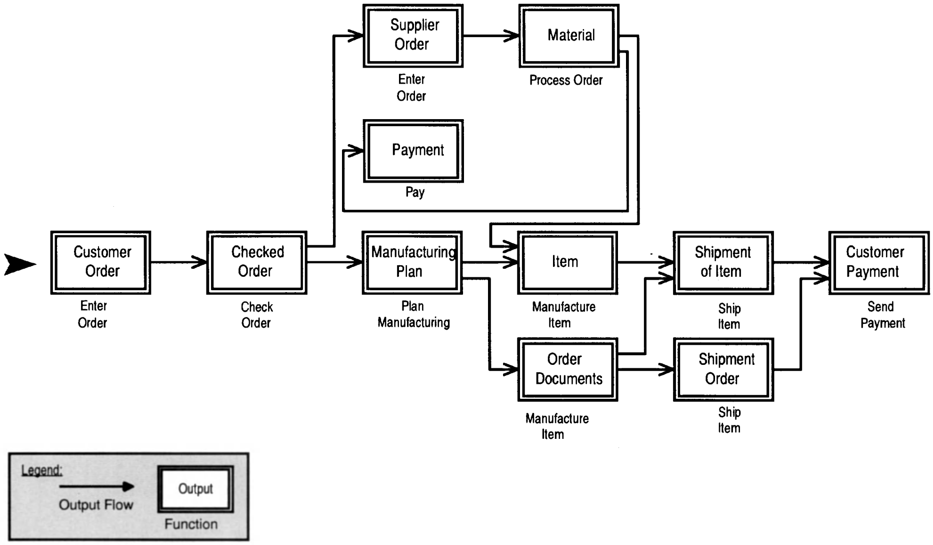

Supplier Order
Enter Order
Material
Process Order
Payment
Pay
Customer Order
Enter Order
Checked Order
Check Order
Manufacturing Plan
Plan Manufacturing
Item
Manufacture Item
Order Documents
Manufacture Item
Shipment of Item
Ship Item
Shipment Order
Ship Item
Customer Payment
Send Payment
Legend:
Output Flow
Output
Function

Figure 7 Output Flow of the Business Process "Order Processing."

process components. Figure 8 illustrates the information objects of our business process example, along with the data interchanged among them. Objects listed as information services have double borders. Information objects describing the environment of the business process are shown as well, such as data regarding suppliers, items, or work schedules. These data are necessary to create infor-

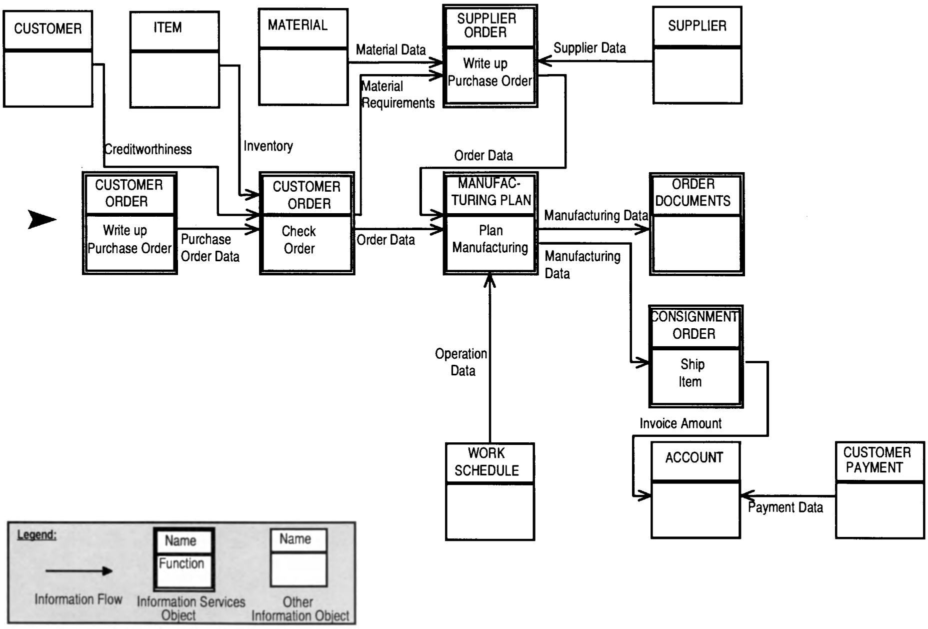

CUSTOMER
ITEM
MATERIAL
SUPPLIER ORDER
Write up Purchase Order
SUPPLIER
Material Data
Supplier Data
Material Requirements
Creditworthiness
Inventory
Order Data
CUSTOMER ORDER
Write up Purchase Order
Purchase Order Data
CUSTOMER ORDER
Check Order
Order Data
MANUFACTURING PLAN
Plan Manufacturing
Manufacturing Data
ORDER DOCUMENTS
Manufacturing Data
CONSIGNMENT ORDER
Ship Item
Operation Data
Invoice Amount
WORK SCHEDULE
ACCOUNT
CUSTOMER PAYMENT
Payment Data
Legend:
Information Flow
Name
Function
Information Services Object
Name
Other Information Object

Figure 8 Information Flow of the Business Process "Order Processing."

mation services. For example, when orders are checked, the customer's credit is checked and inventory is checked for availability.

Because data flow is triggered by the functions that are linked to the information objects, it is more or less possible to read the function flow in Figure 8. However, if multiple functions are applied to an information object or multiple data flows are requested by a function, the function process cannot be uniquely deduced.

Besides information flow modeling, the (static) description of data structures is a very important modeling task. Static enterprise data models are used to develop proper data structures in order to implement a logically integrated database. Chen's entity relationship model (ERM) is the most widespread method for the conceptual modeling of data structures.

3.1.5. Process View

Building various views serves the purpose of structuring and streamlining business process modeling. Splitting up views has the added advantage of avoiding and controlling redundancies that can occur when objects in a process model are used more than once. For example, the same environmental data, events, or organizational units might be applied to several functions. View-specific modeling methods that have proven to be successful can also be used. Particularly in this light, view procedures differ from the more theoretical modeling concepts, where systems are divided into subsystems for the purpose of reducing complexity. In principle, however, every subsystem is depicted in the same way as the original system. This is why it is not possible to use various modeling methods in the same system.

It is important to note that none of the flows (organization, function, output, and information flow, respectively) illustrated above is capable of completely modeling the entire business process. We must therefore combine all these perspectives. To this end, one of the views should be selected as a foundation and then be integrated into the others. The function view is closest to the definition of a business process and is therefore typically used as a starting point. However, in the context of object-oriented enterprise modeling information, flows can serve as a starting point as well.

Figure 9 provides a detailed excerpt of our business process example, focusing the function "manufacture item" with all flows described above.

The method used to describe the process in Figure 9 is called event-driven process chain (EPC). The EPC method was developed at the Institute for Information Systems (IWi) of the Saarland

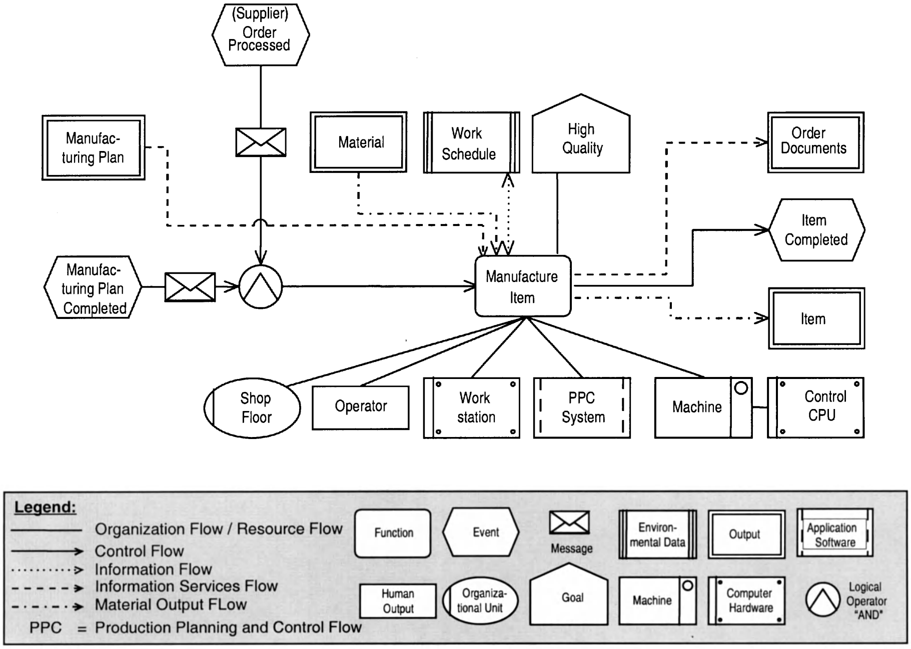

(Supplier) Order Processed
Manufac-turing Plan
Material
Work Schedule
High Quality
Order Documents
Item Completed
Manufac-turing Plan Completed
Manufacture Item
Item
Shop Floor
Operator
Work station
PPC System
Machine
Control CPU
Legend:
Organization Flow / Resource Flow
Control Flow
Information Flow
Information Services Flow
Material Output FLow
PPC = Production Planning and Control Flow
Function
Event
Message
Environ-mental Data
Output
Application Software
Human Output
Organiza-tional Unit
Goal
Machine
Computer Hardware
Logical Operator "AND"

Figure 9 Detailed Excerpt of the Business Process "Order Processing."

University, Germany, in collaboration with SAP AG. It is the key component of SAP R/3's modeling concepts for business engineering and customizing. It is based on the concepts of stochastic networks and Petri nets. Simple versions exclude conditions and messages and include only E(vent)/A(ction) illustrations.

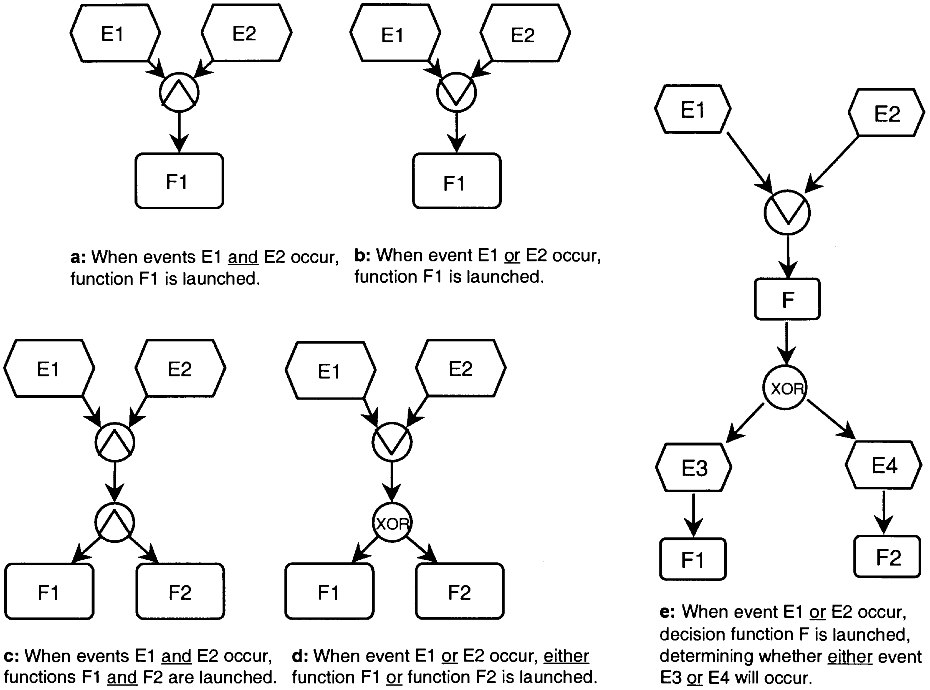
Multiple functions can result from an event. On the other hand, multiple functions sometimes need to be concluded before an event can be triggered. Logical relationships are illustrated by "and" ($\wedge$), "inclusive-or" ($\vee$) and "exclusive-or" (XOR) symbols. Figure 10 gives some typical examples of event relationships. When there are more complex relationships between completed functions and functions that have just been launched (such as different logical relationships between groups of functions), decision tables for incoming and outgoing functions, respectively, can be stored in an event.

3.2. Object-Oriented Enterprise Modeling

Object-oriented enterprise modeling starts with analyzing the entities of the real world. In the object-oriented model, these entities will be described through objects. The object data (attributes) represent characteristics of the real system and are accessible only by means of object methods. By their definition, attributes and methods objects are entirely determined. Objects that share the same characteristics are instances of the respective object class. Classes can be specialized to subclasses as well as generated to superclasses. This is called inheritance—each subclass will inherit the attributes and methods from its superclass.

We can equate the term *method* used in object-oriented analysis with the term *function.* Due to the fact that classes are frequently data classes (such as "customers," "suppliers," "orders," etc.), they represent the link between data and function view. We have already experienced object-oriented class design when we discussed the levels of abstraction and the examples of the modeling views (see Section 1.2 as well as Section 3.1), so we can skim the properties of creating object-oriented classes.

Object-oriented modeling is not based on a standardized method. Rather, a number of authors have developed similar or complementary approaches. The various graphical symbols they use for their approaches make comparisons difficult. The Unified Modeling Language (UML), introduced by



a: When events E1 and E2 occur, function F1 is launched.

b: When event E1 or E2 occur, function F1 is launched.

c: When events E1 and E2 occur, functions F1 and F2 are launched.

d: When event E1 or E2 occur, either function F1 or function F2 is launched.

e: When event E1 or E2 occur, decision function F is launched, determining whether either event E3 or E4 will occur.

Figure 10 Event Relationships in EPC.

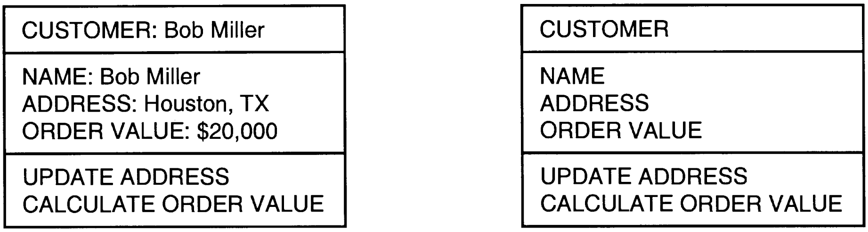

CUSTOMER: Bob Miller
NAME: Bob Miller
ADDRESS: Houston, TX
ORDER VALUE: $20,000
UPDATE ADDRESS
CALCULATE ORDER VALUE
CUSTOMER
NAME
ADDRESS
ORDER VALUE
UPDATE ADDRESS
CALCULATE ORDER VALUE

Figure 11 Object and Object Class.

Rumbaugh, Booch, and Jacobsen aims to streamline and integrate various object-oriented approaches, which is why we will use these symbols.

Objects, each with its own identity and indicated by an ID number, are described by properties (attributes). The functions (methods) that can be applied to the object define their behavior. Objects represent instances and are illustrated by rectangles. Objects with identical attributes, functionality, and semantics are grouped into object classes or regular classes. The quantity of customers thus forms the class "customer" (see Figure 11).

By naming attributes and methods, classes define the properties and the behavior of their instances, that is, objects. Because attributes and methods form a unit, classes realize the principle of encapsulation. In addition to attribute and method definitions for objects, we can also use class attributes and class methods that are valid only for the classes themselves, not for the objects. An example would be "number of customers" and "creating a new customer."

One significant property of the object-oriented approach is inheritance, giving classes access to the properties (attributes) and the behavior (methods) of other classes. Inherited attributes and methods can be overwritten and redefined by the inheriting class. Inheritance takes place within a class hierarchy with two types of classes, namely overriding classes and subordinate classes, as is shown by generalizing and specializing operations in data modeling. A class can also inherit properties from several overriding classes (multiple inheritance), with the resulting class diagram forming a network. Figure 12 gives an example of inheritance among object classes.

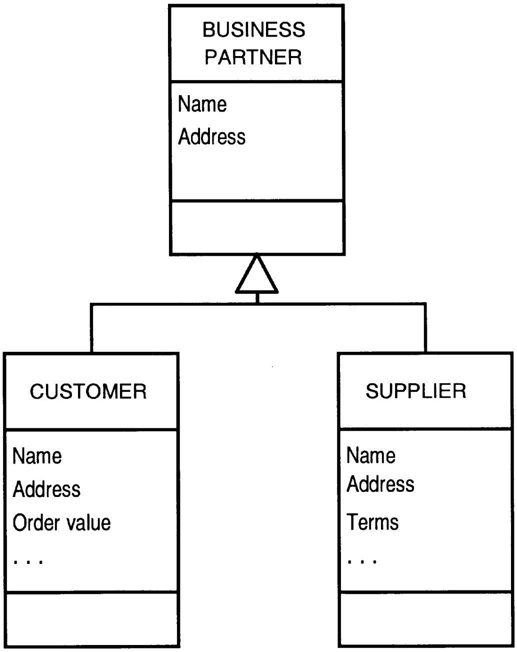

BUSINESS PARTNER
Name
Address
CUSTOMER
Name
Address
Order value
. . .
SUPPLIER
Name
Address
Terms
. . .

Figure 12 Inheritance.

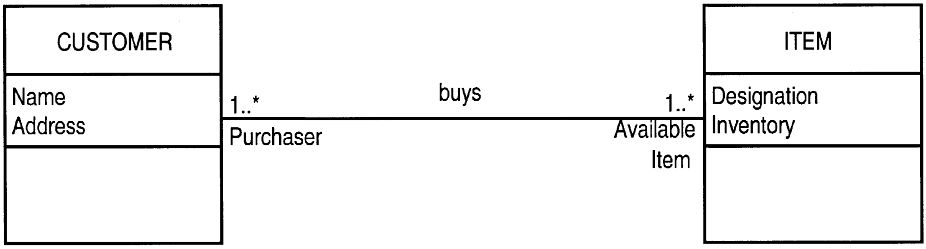


Figure 13 Association.

In addition to the generalizing relationships between classes, there are also relationships (i.e. associations) between objects of equal class ranking or between objects of the same class. These associations equate to relationships in entity relationship models, although here they are illustrated by only one line. These illustrations should be read from left to right. Cardinalities are allocated to an association. At each end of the association, role names can be allocated to the associations. If associations contain attributes, they are depicted as classes (see the class "purchasing process" in Figure 13).

Aggregations are a special kind of association describing the "part of" relationships between objects of two different classes. Role names can be applied to aggregations as well. If attributes are allocated to an aggregation, this leads to a class, as shown by the class "structure," in an aggregation relation for a bill of materials (see Figure 14).

Class definition, inheritance, associations, and aggregations make up the key structural properties of object-oriented modeling. Overall, the UML provides seven methods for describing the various static and dynamic aspects of enterprises. Nevertheless, even with methods such as use case diagrams and interaction diagrams, it is difficult to depict process branching, organizational aspects, and output flows. Therefore, one of the main disadvantages of the object-oriented approach is it does not illustrate business process in a very detailed manner.

4. MODELING ARCHITECTURES

4.1. Architecture of Integrated Information Systems (ARIS)

In the architecture of integrated information systems (ARIS), we can distinguish two different aspects of applications:

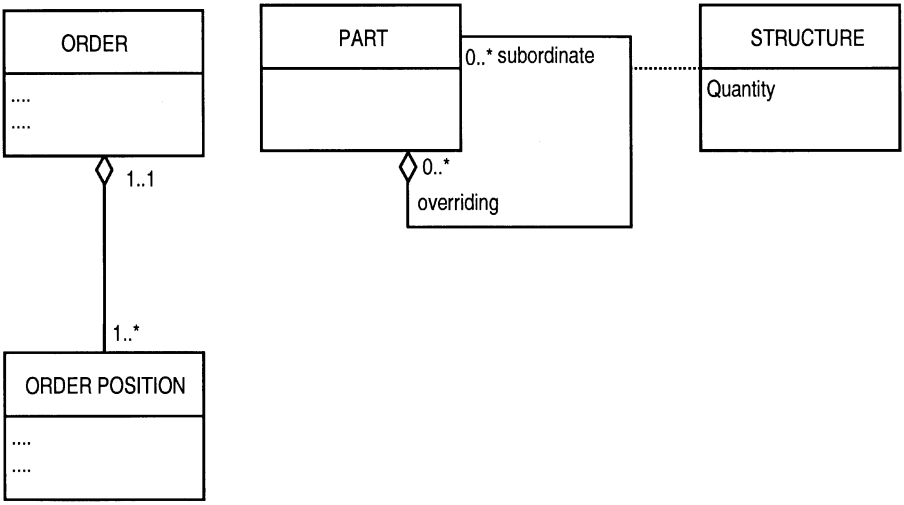


Figure 14 Aggregation.

1. The ARIS concept (ARIS house), an architecture for modeling enterprises, particularly describing business processes
2. The ARIS house of business engineering (HOBE), representing a concept for comprehensive computer-aided business process management.

Both applications are supported by the ARIS Toolset software system, developed by IDS Scheer AG. We will discuss tool support for enterprise modeling in Section 5.

The ARIS concept consists of five modeling views and a three-phase life cycle model. The integration of both leads to 15 building blocks for describing enterprise information systems. For each block the ARIS concept provides modeling methods, meta structures of which are included in the ARIS information model.

As we have already discussed static and dynamic modeling views (see Section 3.1), we can skim over the creation of modeling views. ARIS views are created according to the criterion of semantic correlation similarity, that is, enterprise objects that are semantically connected are treated in the same modeling view. The ARIS modeling views are (see Figure 15):

- *Data view:* This view includes the data processing environment as well as the messages triggering functions or being triggered by functions. Events such as "customer order received,"

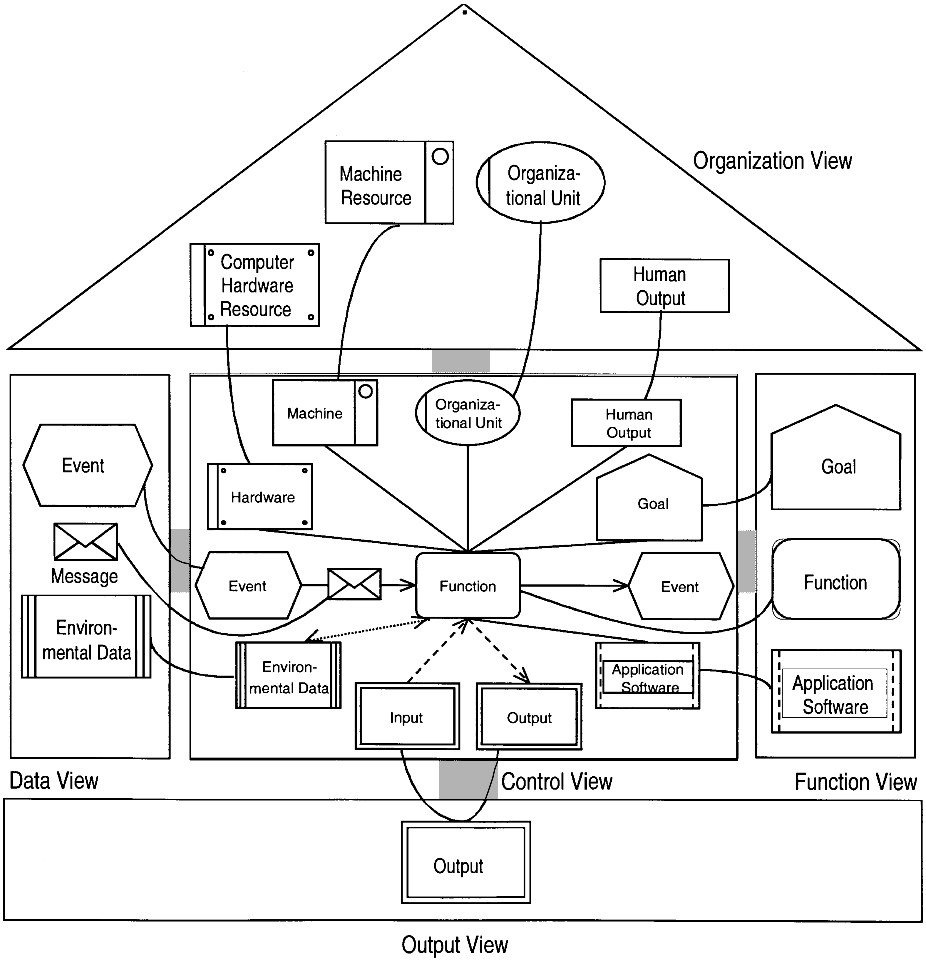

Organization View
Machine Resource
Organizational Unit
Computer Hardware Resource
Human Output
Machine
Organizational Unit
Human Output
Hardware
Goal
Event
Message
Environmental Data
Event
Function
Event
Environmental Data
Input
Output
Application Software
Goal
Function
Application Software
Data View
Control View
Function View
Output
Output View

Figure 15 Views of the ARIS Concept.

"completion notice received," and "invoice written" are also information objects represented by data and therefore modeled in the data view.

- *Function view:* The business activities to be performed and their relationships form the function view. It contains the descriptions of each single activity (function) itself, the associations between super- and subordinate functions, and the logical sequence of functions. Enterprise goals that guide the performance of the functions and the application systems that support their execution are also components of this view.
- *Organization view:* Departments, office buildings, and factories, are examples of organizational units that perform business functions. They are modeled according to criteria such as "same function" and "same work object." Thus, the structure as well as the interaction between organizational units and the assigned resources (human resources, machine resources, computer hardware resource) are part of the organization view.
- *Output view:* This view contains all physical and nonphysical output that is created by business functions, including services, materials, and products as well as funds flows. Because each output is input for another function—even if this function is probably carried out by an external partner—input–output diagrams are also components of this view.
- *Control view/process view:* In the views described above, the classes are modeled with their relationships relative to the views. Relationships among the views as well as the entire business process are modeled in the control or process view. This enables a framework for the systematic inspection of all bilateral relationships of the views and the complete enterprise description using a process-oriented perspective.

The ARIS concept provides a set of methods that can be used to model the enterprise objects and their static and dynamic relationships. For each view at least one well-proven method is provided. Because ARIS is an open architecture, new modeling methods, such as UML methods, can easily be integrated into the meta structure. Figure 16 gives examples of ARIS methods for conceptual modeling.

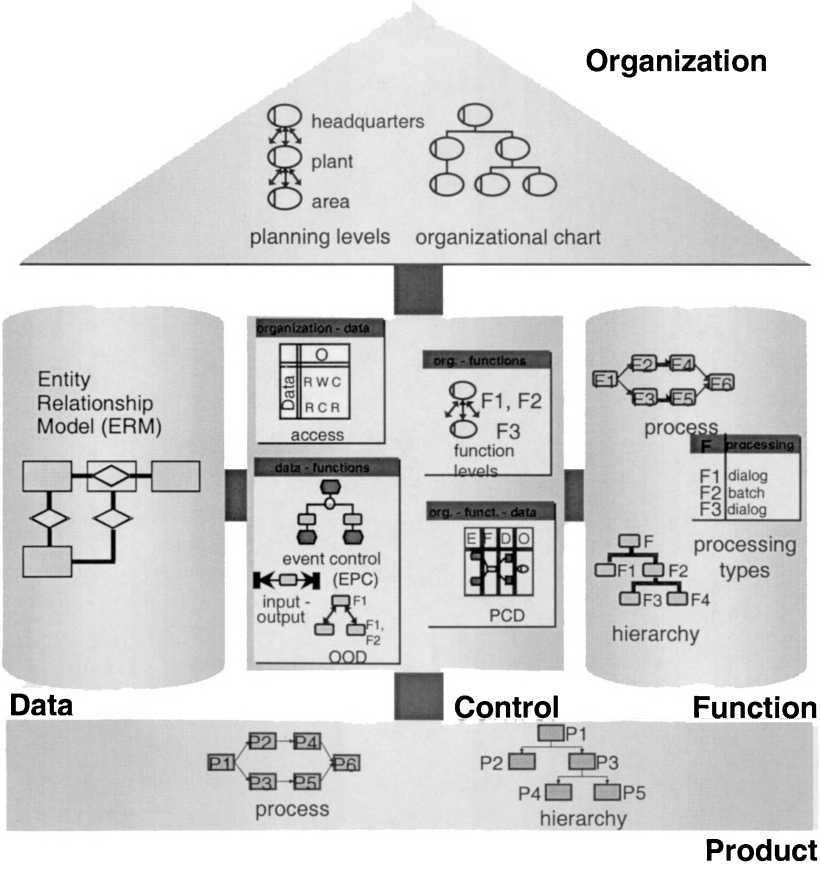
Up to now, we have discussed enterprises, particularly business processes, from a management point of view, that is, without any particular focus on information technology. The aforementioned application programs (components of the function view), computer hardware (a component of the organization view), and data media (components of the data view) contain only system names, not IT descriptions. These are included in the ARIS concept by evaluating the IT support provided by each ARIS view.

- Function views are supported by the application programs, which may be described in more detail by module concepts, transactions, or programming languages.
- Organization views, along with their production resources and the computer resources responsible, may be detailed further by listing network concepts, hardware components, and other technical specifications.
- Data views may be detailed more precisely by listing data models, access paths, and memory usage.
- Output views group the various types of output, such as material output and information services. Here again there is a close alignment with the supporting information technology. In material output (e.g., entertainment technology, automobiles, and machine tools), more and more IT components (e.g., chip technology), along with the necessary hardware, are used. Other service industries, such as airline reservations, are closely linked with IT as well.
- The fact that the respective views can be combined within the control view means that there is a definite link with IT, as demonstrated by the above arguments.

Using a phase model, enterprise models are thus transformed step by step into information and communication technology objects (see Figure 17).

The starting point of systems development is the creation of an IS-oriented initial strategic situation in Phase 1. "IS-oriented" means that basic IT effects on the new enterprise concepts are already taken into account. Some examples of these relationships might be creating virtual companies through communication networks, PC banking, integrated order processing and product development in industry (CIM), or integrated merchandise management systems (MMS) in retail.

Strategic corporate planning determines long-term corporate goals, general corporate activities, and resources. Thus, planning has an effect on the long-term definition of enterprises, influencing corporate goals, critical success factors, and resource allocation. The methods in question are similar to management concepts for strategic corporate planning. Provided actual business processes have already been described, this occurs in a general fashion. At this stage, it is not advisable to split up functions into ARIS views and then describe them in detail.


Organization
headquarters
plant
area
planning levels
organizational chart
Entity Relationship Model (ERM)
organization - data
O
Data
R W C
R C R
access
org. - functions
F1, F2
F3
function levels
data - functions
event control
(EPC)
input - output
F1
F1, F2
OOD
org. - funct. - data
E F D O
PCD
E1
E2
E4
E3
E5
E6
process
F
processing
F1 dialog
F2 batch
F3 dialog
processing types
F
F1
F2
F3
F4
hierarchy
Data
Control
Function
P1
P2
P4
P3
P5
P6
process
P1
P2
P3
P4
P5
hierarchy
Product

Legend:

ERM = Entity Relationship Model
Fn = Function n
E = Event
D = Data
O = Organizational Unit
Pn = Product n
R = Read
W = Write
C = Create
PCD = Process Chain Diagram
EPC = Event-driven Process Chain
OOD = Object Oriented Design

Figure 16 ARIS Methods for Conceptual Modeling.

In Phase 2, the requirements definition, individual views of the application system are modeled in detail. Here as well, business-organizational content is key. Examples of business processes should be included at this level. However, in this phase more conceptual modeling methods should be used than in the strategic approach because the descriptions for the requirements definition are the starting point for IT implementation. Modeling methods that are understandable from a business point of view should be used, yet they should be sufficiently conceptual to be a suitable starting point for a

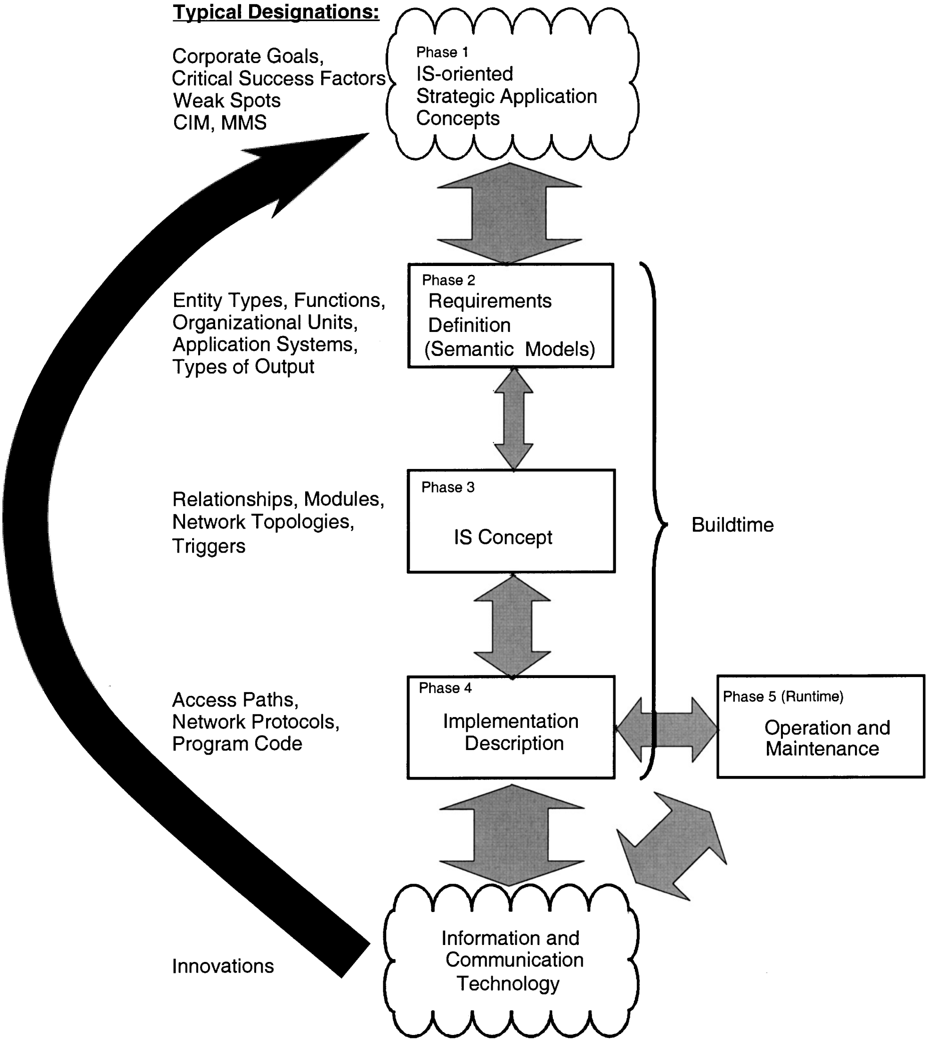

Typical Designations:
Corporate Goals,
Critical Success Factors
Weak Spots
CIM, MMS
Phase 1
IS-oriented
Strategic Application
Concepts
Entity Types, Functions,
Organizational Units,
Application Systems,
Types of Output
Phase 2
Requirements
Definition
(Semantic Models)
Relationships, Modules,
Network Topologies,
Triggers
Phase 3
IS Concept
Buildtime
Access Paths,
Network Protocols,
Program Code
Phase 4
Implementation
Description
Phase 5 (Runtime)
Operation and
Maintenance
Innovations
Information and
Communication
Technology

Figure 17 ARIS Phase Model.

consistent IT implementation. Therefore, it makes sense to include general IT objects, such as databases or programs, at this level.

Phase 3 calls for creating the design specification, where enterprise models are adapted to the requirements of the implementation tool interfaces (databases, network architectures, or programming languages, etc.). At this time, actual IT products are still irrelevant.

Phase 4 involves the implementation description, where the requirements are implemented in physical data structures, hardware components, and real-world IT products.

These phases describe the creation of an information system and are therefore known as buildtime. Subsequently, the completed system becomes operable, meaning it is followed by an operations phase, known as runtime. We will not address the operation of information systems, that is, runtime, in great detail.

The requirements definition is closely aligned with the strategic planning level, illustrated by the width of the arrow in Figure 17. However, it is generally independent of the implementation point of view, as depicted in the narrow arrow pointing to the design specification.

Implementation description and operations, on the other hand, are closely linked with the IT equipment and product level. Changes in the system's IT have an immediate effect on its type of implementation and operation.

The phase concept does not imply that there is a rigid sequence in the development process, as in the waterfall model. Rather, the phase concept also includes an evolutionary prototyping procedure. However, even in evolutionary software development, the following description levels are generally used. Phase models are primarily used because they offer a variety of description objects and methods.

The ARIS concept in Figure 18 is enhanced by the phases of the buildtime ARIS phase model. After a general conceptual design, the business processes are divided into ARIS views and documented and modeled from the requirements definition to the implementation description. These three description levels are created for controlling purposes as well. This makes it possible to create the links to the other components at each of the description levels.

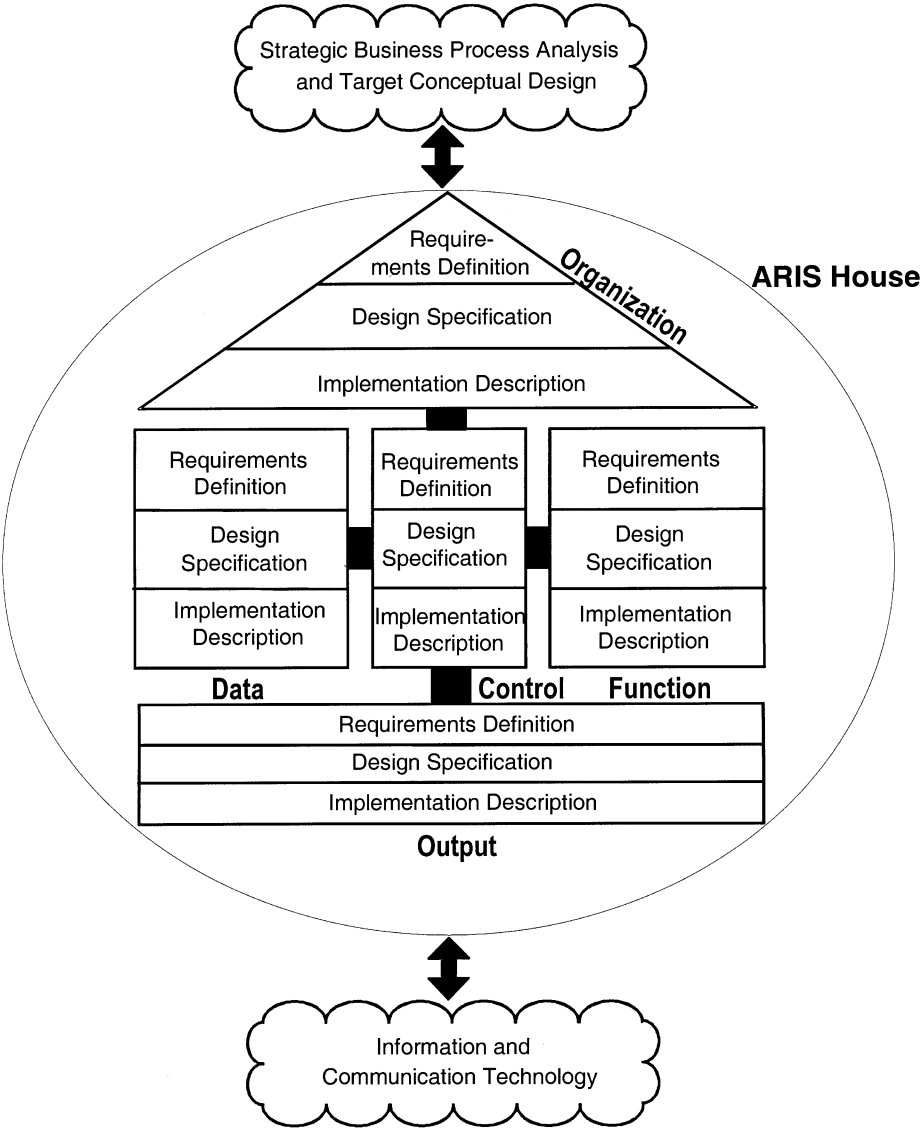

Strategic Business Process Analysis and Target Conceptual Design
Requirements Definition
Design Specification
Implementation Description
Organization
ARIS House
Requirements Definition
Design Specification
Implementation Description
Requirements Definition
Design Specification
Implementation Description
Requirements Definition
Design Specification
Implementation Description
Data
Control
Function
Requirements Definition
Design Specification
Implementation Description
Output
Information and Communication Technology

Figure 18 ARIS Concept with Phase Model.

The ARIS concept paves the way for engineering, planning, and controlling enterprises, particularly business processes. The ARIS house of business engineering (HOBE) enhances the ARIS concept by addressing comprehensive business process management from not only an organizational but an IT perspective. We will outline how ARIS supports business management in the design and development stages, using ARIS-compatible software tools.

Because business process owners need to focus on the ‘‘one-shot’’ engineering and description aspects of their business processes, ARIS HOBE provides a framework for managing business processes, from organizational engineering to real-world IT implementation, including continuous adaptive improvement. HOBE also lets business process owners continuously plan and control current business procedures and devote their attention to continuous process improvement (CPI). Comprehensive industry expertise in planning and controlling manufacturing processes is a fundamental component of HOBE. Objects such as ‘‘work schedule’’ and ‘‘bill of material’’ provide detailed description procedures for manufacturing processes, while production planning and control systems in HOBE deliver solutions for planning and controlling manufacturing processes. Many of these concepts and procedures can be generalized to provide a general process management system.

- At level I (process engineering), shown in Figure 19, business processes are modeled in accordance with a manufacturing work schedule. The ARIS concept provides a framework covering every business process aspect. Various methods for optimizing, evaluating, and ensuring the quality of the processes are also available.
- Level II (process planning and control) is where business process owners’ current business processes are planned and controlled, with methods for scheduling and capacity and (activity-based) cost analysis also available. Process monitoring lets process managers keep an eye on the states of the various processes.
- At Level III (workflow control), objects to be processed, such as customer orders with appropriate documents or insurance claims, are delivered from one workplace to the next. Electronically stored documents are delivered by workflow systems.
- At Level IV (application system), documents delivered to the workplaces are specifically processed, that is, functions of the business process are executed using computer-aided application systems (ranging from simple word processing systems to complex standard software solution modules), business objects, and Java applets.

HOBE’s four levels are linked with one another by feedback loops. Process control delivers information on the efficiency of current processes. This is where continuous adaptation and improve-

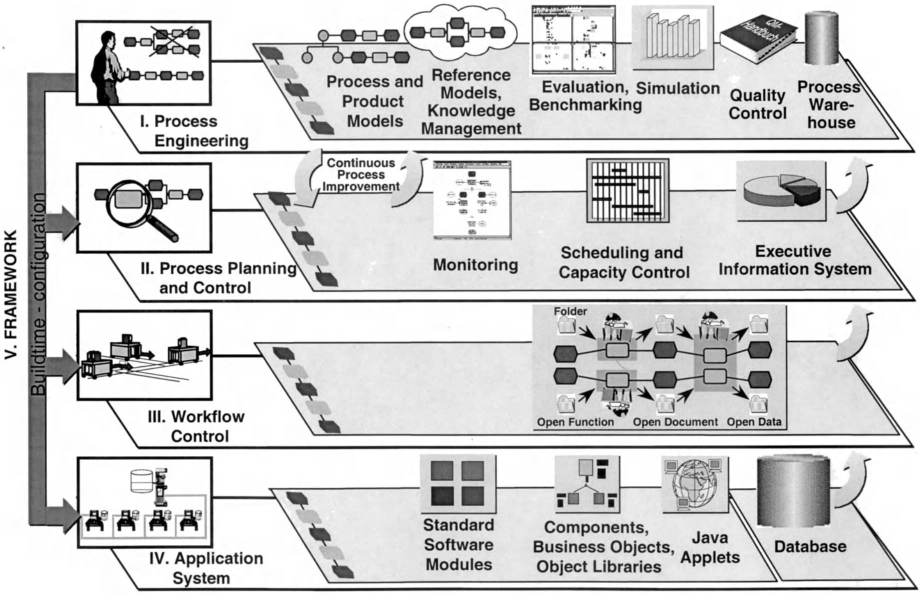

V. FRAMEWORK
Buildtime - configuration
I. Process Engineering
Process and Product Models
Reference Models, Knowledge Management
Evaluation, Benchmarking
Simulation
QM-Handbuch
Quality Control
Process Ware-house
II. Process Planning and Control
Continuous Process Improvement
Monitoring
Scheduling and Capacity Control
Executive Information System
III. Workflow Control
Folder
Open Function
Open Document
Open Data
IV. Application System
Standard Software Modules
Components, Business Objects, Object Libraries
Java Applets
Database

Figure 19 ARIS House of Business Engineering.

ment of business processes in accordance with CPI begins. Workflow control reports actual data on the processes to be executed (amounts, times, organizational allocations) to the process control level. Application supporting modules are then started by the workflow system.

In the fifth component of the HOBE concept, Levels I through IV are consolidated into a framework. Frameworks contain information on the respective architecture and applications, configuring real-world applications from the tools at levels II and III as well as application information from the reference models (level I). Frameworks contain information on the composition of the components and their relationships.

4.2. Information Systems Methodology (IFIP-ISM)

Olle et al. (1991) give a comprehensive methodology for developing more traditional information systems. The designation "methodology" is used at the same level as "architecture." The seven authors of the study are members of the International Federation for Information Processing (IFIP) task group, in particular of the design and evaluation of information systems working group WG 8.1 of information systems technical committee TC 8. The research results of the study are summarized in the guideline "Information Systems Methodology."

The design of the methodology does not focus on any particular IS development methods. Rather, it is based on a wide range of knowledge, including as many concepts as possible: IDA (interactive design approach), IEM (information engineering methodology), IML (inscribed high-level Petri nets), JSD (Jackson system development), NIAM (Nijssen's information analysis method), PSL/PSA (problem statement language/problem statement analyzer), SADT (structured analysis and design technique) as well as Yourdon's approach of object-oriented analysis.

This methodology is described by metamodels of an entity relationship concept. It features the point of view and stages of an information system life cycle, distinguishing data-oriented, process-oriented, and behavior-oriented perspectives (see Figure 20). Creating these perspectives is less a matter of analytical conclusion than simply of reflecting a goal of addressing the key issues typical in traditional IS developing methods.

Entity types and their attributes are reviewed in the data-oriented perspective. The process-oriented perspective describes events (business activities), including their predecessor or successor relationships. Events and their predecessor or successor relationships are analyzed in the behavior-oriented perspective.

From a comprehensive 12-step life-cycle model we will select three steps, information systems planning, business planning, and system design, and then examine the last two in detail in terms of their key role in the methodology.

Information systems planning refers to the strategic planning of an information system. In business analysis, existing information systems of an entire enterprise or of a subarea of the enterprise are

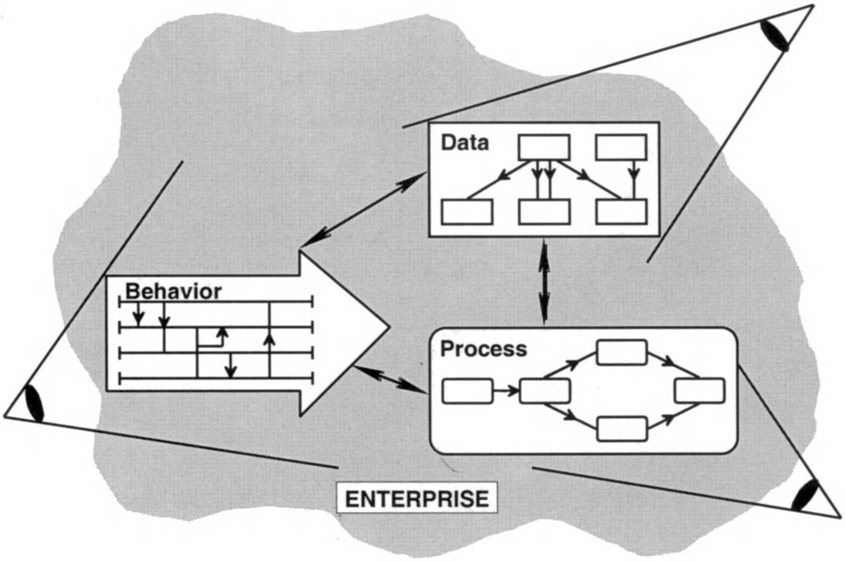


Figure 20 Perspectives of the IFIP Architecture. (From Olle et al. 1991, p. 13)

analyzed. The respective information system is designed in the step system design. This concept also includes a comprehensive procedural model, including a role concept for project organization.

With regard to ARIS, this concept has some overlapping areas. In others there are deviations. What both concepts have in common is their 2D point of view, with perspectives and development steps. There are differences in their instances, however. For example, Olle et al. do not explicitly list the organization view but rather review it along with other activities, albeit rudimentarily. The process definition more or less dovetails with ARIS's function definition. Data and functions or events and functions are also strictly separated from one another. The three perspectives linked together are only slightly comparable to ARIS control view. The step system design blends together the ARIS phases of requirements definition and design specification, with the emphasis on the latter.

4.3. CIM Open System Architecture (CIMOSA)

The ESPRIT program, funded by the European Union (EU), has resulted in a series of research projects for developing an architecture, Computer Integrated Manufacturing Open System Architecture (CIMOSA), for CIM systems. CIMOSA results have been published by several authors, including Vernadat (1996). This project originally involved 30 participating organizations, including manufacturers as the actual users, IT vendors, and research institutes. Although the project focused on CIM as an application goal, its mission was to provide results for general enterprise modeling. One of CIMOSA's goals was also to provide an architecture and a methodology for vendor-independent, standardized CIM modules to be "plugged" together, creating a customer-oriented system ("plug and play").

The CIMOSA modeling framework is based on the CIMOSA cube (see Figure 21).

CIMOSA distinguishes three different dimensions, described by the three axes of the cube. The vertical direction (stepwise derivation) describes the three description levels of the phase concept:

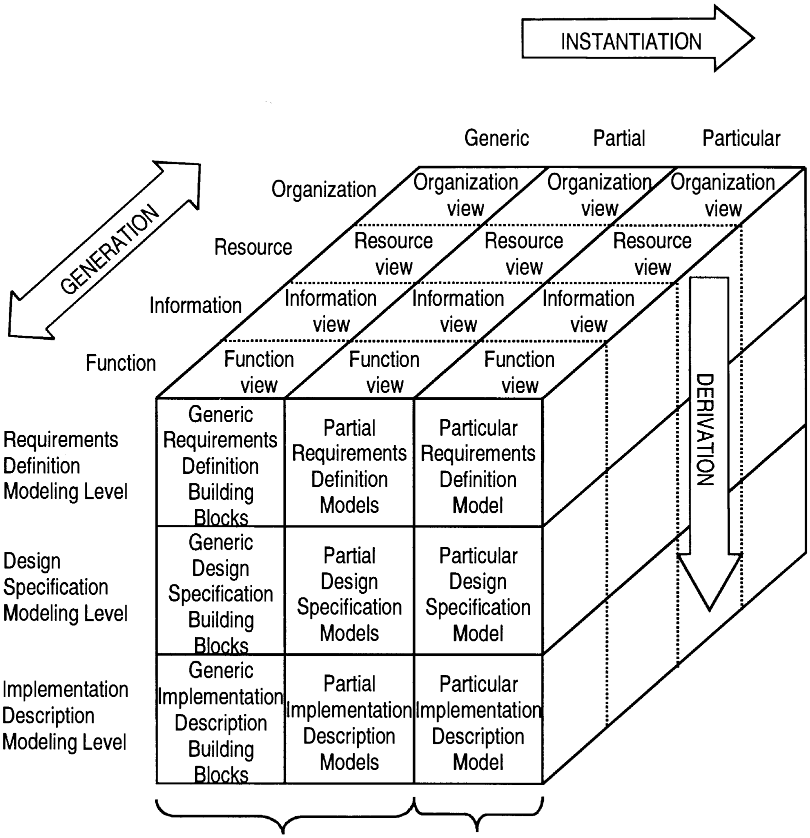

INSTANTIATION
Generic
Partial
Particular
GENERATION
Organization
Resource
Information
Function
Organization view
Resource view
Information view
Function view
DERIVATION
Requirements Definition Modeling Level
Design Specification Modeling Level
Implementation Description Modeling Level
Generic Requirements Definition Building Blocks
Partial Requirements Definition Models
Particular Requirements Definition Model
Generic Design Specification Building Blocks
Partial Design Specification Models
Particular Design Specification Model
Generic Implementation Description Building Blocks
Partial Implementation Description Models
Particular Implementation Description Model

Figure 21 The CIMOSA Modeling Architecture. (CIMOSA cube). (From Vernadat 1996, p. 45)

requirements definition, design specification, and implementation description. These levels are for the most part identical with those of the ARIS life cycle.

In the horizontal dimension (stepwise instantiation), concepts are individualized step by step. First, basic requirements (generic requirements, building blocks) are defined, then particularized in the next step according to industry specific requirements (partial requirements). In Step 3, they are broken up into enterprise-specific requirements (particular requirements).

This point of view makes it clear that initially, according to CIMOSA, general building blocks should be used to define standards, after which the building blocks are grouped into industry specific reference models. In the last step, they are used for developing enterprise-specific solutions. In ARIS, the degree of detailing an information model is defined while the granularity issues are addressed.

By directly entering content-related reference models, the CIMOSA architecture, it becomes clear, combines general methodological issues regarding information systems and application-related CIM domains.

The third dimension, stepwise generation, describes the various views of an information system. This point of view has goals similar to ARIS regarding the creation of views, although not all the results are the same. CIMOSA divides description views into function view, information view, resource view, and organization view. Function view is the description of events, although it also includes a combination of other elements such as events and processes, including performance and exception handling. Information view refers to the data view or object definition. Resource view describes IT and production resources, and organization view implies the hierarchical organization.

CIMOSA also breaks up the entire context into various views, although it lacks a level for reassembling them, as opposed to ARIS with its control and process views. This results in the fact that in CIMOSA, descriptions of the individual views are combined with one another. For example, when resources are being described, they are at the same time also allocated to functions. The CIMOSA modeling concept does not feature an output view.

The CIMOSA concept develops an architecture suitable for describing information systems, into which content in the form of standardized reference models, all the way to actual software generation, can be entered. Despite the above-mentioned drawbacks, it considers important points. Based on this concept, modeling methods are classified in CIMOSA and described by metamodels, all the while adhering to an event-driven, business process-oriented view. Furthermore, enterprises are regarded as a series of multiple agents communicating with one another.

Despite the considerable financial and intellectual efforts spent on CIMOSA, its practical contribution so far has been minimal. Business users involved in the project have so far reported few special applications resulting therefrom, with the exception of the car manufacturer Renault with a repair service application for manufacturing plants and the tooling company Traub AG with an application for optimizing individual development of tools. To date, a CIMOSA-based modeling tool has not been used much in practice.

The main reason for the lack of success in real-world applications is presumably its very theoretical design, which does not incorporate commercially available IT solutions (standard software, for example). Considering the general lack of interest in CIM concepts, the extremely specialized focus of this approach seems to be working to its disadvantage.

4.4. Zachman Framework

A framework for describing enterprises, quite popular in the United States, was developed by J. A. Zachman. This concept is based on IBM's information systems architecture (ISA) but has been enhanced and presented by Zachman in numerous talks and seminars.

This approach (see Figure 22) consists of six perspectives and six description boxes. In the ARIS terminology, Zachman's description boxes would equate to views and perspectives would equate to the levels of the life-cycle model.

Perspectives are listed in brackets along with the respective role designations of the party involved: scope (planner), enterprise model (owner), system model (designer), technology model (builder), components (subcontractor), and functioning system (user).

The areas to be viewed are designated by interrogatives with the respective actions listed in brackets: what (data), how (function), where (network), who (people), when (time), and why (rationale). Perspectives and files to be viewed are at a right angle to one another. Every field in the matrix is described by a method.

Contrary to ARIS, the Zachman framework is not capable of being directly implemented into an information system and the relationships between the description fields are not entered systematically. Furthermore, the relationship of Zachman's framework with the specific creation of output within the business process is not apparent.

Initial approaches for supporting tools are becoming apparent, thanks to cooperation with Framework Software, Inc.

FOCUS

PERSPECTIVE

Generic Framework Element Bond Element	WHAT *(Dato)* Entity Relationship Entity	HOW *(Function)* Process Input-Output Process	WHERE *(Network)* Node Line Node	WHO *(People)* Agent Work Agent	WHEN *(Time)* Event Cycle Event	WHY *(Rationale)* End Means End
SCOPE *(Planner)*	Entity List	Process List	Location List	Organization List	Major Event List	Objective List
ENTERPRISE MODEL *(Owner)*	Enterprise Entity Enterprise Rule Enterprise Entity	Enterprise Process Resource Enterprise Process	Enterprise Location Enterprise Channel Enterprise Location	Organization Work Organization	Enterprise Event Enterprise Cycle Enterprise Event	Objective Strategy Objective
SYSTEM MODEL *(Designer)*	Entity Type Relationship Type Entity Type	System Process User View System Process	Site Link Site	Role Presentation Role	System Event System Cycle System Event	Criterion Choice Criterion
TECHNOLOGY MODEL *(Builder)*	Data Structure Referential Integrity Data Structure	Application Device Format Application	Connection Point Communication Line Connection Point	User Technical Interface User	Technical Event Technical Cycle Technical Event	Condition Action Condition
COMPONENTS *(Sub-contractor)*	Data Container Aquisition Data Container	Module/Object Couple/Message Module/Object	Address Protocol Address	Individual Transaction Individual	Component Event Component Cycle Component Event	Sub-condition Step/Task Sub-condition
FUNCTIONING SYSTEM *(User)*	Information Integrity Information	Procedure Request Procedure	Client/Server Access Client/Server	Worker Work Session Worker	Operating Event Operating Cycle Operating Event	Target Option Target

Figure 22 Zachman Framework. (From Burgess and Hokel 1994, p. 26, © Framework Software, Inc. All rights reserved)

5. MODELING TOOLS

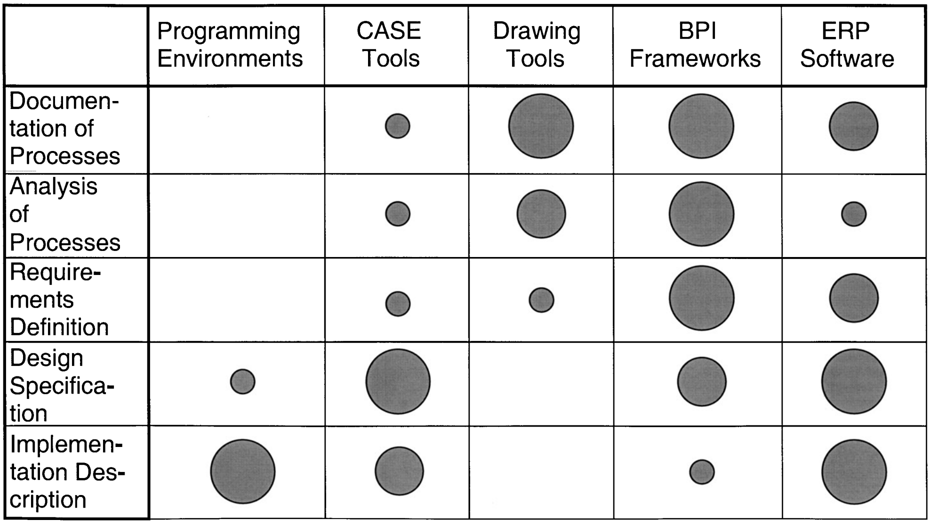
5.1. Benefits of Computerized Enterprise Modeling

Modeling tools are computerized instruments used to support the application of modeling methods. Support in designing enterprise models through the use of computerized tools can play a crucial role in increasing the efficiency of the development process. The general benefits of computer support are:

- All relevant information is entered in a structured, easy-to-analyze form, ensuring uniform, complete documentation. The incorporation of a user-friendly graphical interface offers a comfortable means of creating and modifying the necessary diagrams. Changes in one part of the documentation are automatically updated in other locations and thus need only be performed once.
- Referencing stored design to its creation date enables the different versions to be managed in a comfortable fashion. Versions are overwritten only if the user expressly wishes it; otherwise, each modification is maintained as an independent version.
- The use of a tool facilitates conformance to a fixed procedural model. Consistency checks ensure that modeling activities are not permitted until the preceding level has been completely processed.
- The internal consistency of the draft model can be validated by using rule-based systems. The general system requirements must also be entered within the framework of tool application in order to ensure external consistency. This necessitates the use of a powerful tool that provides structured support for the entire life cycle.
- Generation algorithms enable graphical representations to be derived from existing database structures. Likewise, the graphical tool interface enables the information objects to be categorized in a meaningful, content-related manner. However, since the aim of providing application-oriented data structuring is again based on content-related perspectives, it would be difficult for a tool to provide further support. In particular, reclassifying basic information units such as attributes into information objects that are more meaningful from a subject-oriented perspective is the virtual equivalent of redesigning the model.

5.2. Characterization of Modeling Tools

We can distinguish different groups of tools for enterprise modeling (see Figure 23). In relation to the different tasks of enterprise modeling, each group shows its special support profile.



	Programming Environments	CASE Tools	Drawing Tools	BPI Frameworks	ERP Software
Documentation of Processes		•	⬤	⬤	●
Analysis of Processes		•	●	⬤	•
Requirements Definition		•	•	⬤	●
Design Specification	•	⬤		●	⬤
Implementation Description	⬤	●		•	⬤

Figure 23 Modeling Tools.

The first group that can be seen as modeling tools in a broader sense are programming environments. Programming environments such as Borland JBuilder and Symantec Visual Café for Java clearly emphasize the phase of implementation description. They describe a business process by the language of the software application that is used to support the process performance. In few cases, programming languages provide limited features for design specification.

CASE tools such as ADW, on the other hand, support a specific design methodology by offering data models and function models (e.g., the entity relationship method (ERM) and structured analysis and design technique [SADT]). Their metastructure has to reflect this methodology. Thus, they automatically provide an information model for the repository functionality of this system. Because CASE tools cannot really claim to offer a complete repository, they are also called encyclopedias.

In contrast, drawing tools such as VISIO and ABC FlowCharter support the phases of documentation and, to a certain extent, analysis of business structures and processes. The emphasis is on the graphical representation of enterprises in order to better understand their structures and behavior. In most of the cases, drawing tools do not provide an integrated meta model, that is, repository. Consequently, drawing tools cannot provide database features such as animating and simulating process models, analyzing process times, and calculating process costs.

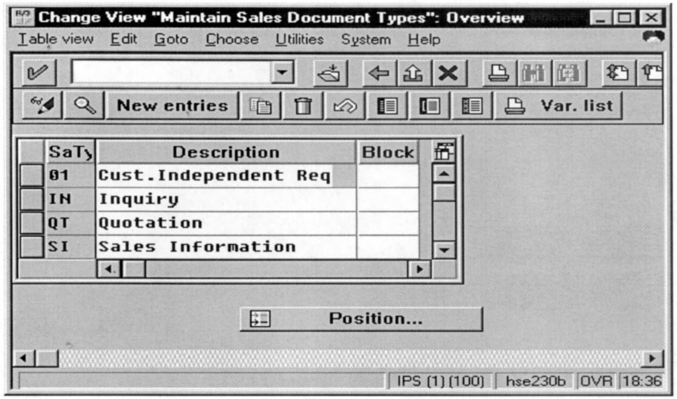
Business process improvement (BPI) frameworks such as ARIS offer a broader set of modeling methods, which are described together in one consistent metamodel. As a result, each model is stored in the framework's repository and thus can be retrieved, analyzed, and manipulated and the model history can be administered (see Section 4.1).

The last group encompasses standard software systems, such as enterprise resource planning (ERP) systems like those from SAP, Baan, or Oracle. ERP systems offer greater flexibility with respect to customization, provide support for data management, and include functionality for creating and managing a (sub)repository. The main weakness is in active analysis of business structures—that is, ERP tools typically do not offer tools for simulation process alternatives.

Because all the modeling tools presented focus on different aspects, none of the systems is suitable for handling the entire systems development process. However, ERP systems are opening up more and more to the modeling and analysis level, endeavoring to support the whole system life cycle.

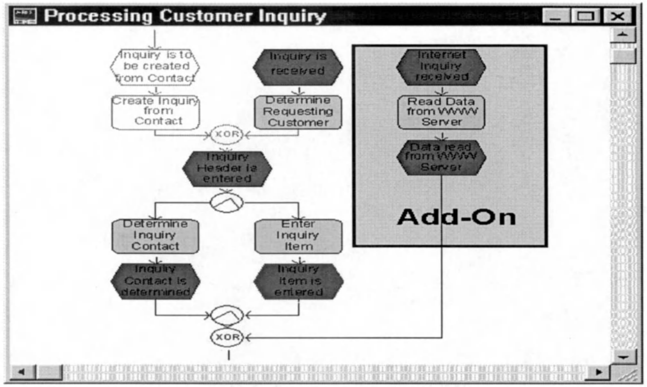
The integration between ARIS and SAP R/3, for example, demonstrates seamless tool support from the documentation to the implementation phase. In an excerpt from the SAP R/3 reference model, Figure 24 shows a customizing example with the ARIS Toolset. For clarification purposes, the four windows are shown separately. In a real-world application, these windows would be displayed on one screen, providing the user with all the information at once.

The upper-right window shows an excerpt from the business process model in the ARIS modeling tool, illustrating the part of the standard software process that can be hidden. An additional process branch, not contained in the standard software, must be added.

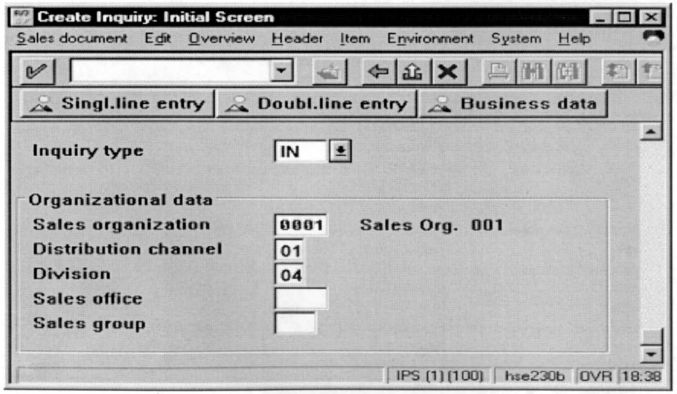

Change View "Maintain Sales Document Types": Overview
Table view Edit Goto Choose Utilities System Help
New entries
Var. list
Position...
IPS (1) (100) hse230b OVR 18:36

SaTy	Description	Block
01	Cust.Independent Req	
IN	Inquiry	
QT	Quotation	
SI	Sales Information	

Customizing: Maintain Sales Document Types


Processing Customer Inquiry
Inquiry is to be created from Contact
Create Inquiry from Contact
Inquiry is received
Determine Requesting Customer
XOR
Inquiry Header is entered
Determine Inquiry Contact
Inquiry Contact is determined
Enter Inquiry Item
Inquiry Item is entered
XOR
Internet Inquiry received
Read Data from WWW Server
Data read from WWW Server
Add-On

Process Model


Create Inquiry: Initial Screen
Sales document Edit Overview Header Item Environment System Help
Singl.line entry
Doubl.line entry
Business data
IPS (1) (100) hse230b OVR 18:38

Inquiry type: IN

Organizational data

Sales organization: 0001 Sales Org. 001

Distribution channel: 01

Division: 04

Sales office:

Sales group:

Function „Create Inquiry“

Process: Processing Customer Inquiry

Function	As-is/Target	Unresolved Issues	Inter-face	In Charge	Date	Effort
1. Determine Ordering Customer	From now on, ordering Customers will be queried in Accordance with ISP Country Codes	CPD Customer necessary	Customer Master Data (internal)	C. Jones	May 29	Standard
2. Determine Inquiry Contact	Define third-party as new Partner Type in customized Version	None	Customer Master Data (internal)	P. Miller	May 29	Standard
3. Enter Inquiry Item	Use AFN Item Type as Standard	None	Customer Master Data (internal)	P. Miller C. Jones	May 30	Standard

Documentation of Results

Figure 24 Customizing SAP R/3 with ARIS Toolset.

The function "create inquiry" asks users which screen is being used in SAP R/3. Using the process model as a modeling tool and by clicking on the function (or starting a command), users can seamlessly invoke SAP R/3. This screen is shown at the bottom left.

The Implementation Management Guide (IMG) customizing tool is activated for customizing the function. In the upper-left hand window, it depicts the function parameters that are available.

With the modeling tool, results of discussions, parameter decisions, unresolved issues, and the like are stored in the function, as depicted in the bottom-right window. This enables detailed documentation of business and IT-specific business process engineering. This documentation can be used at a later point in time for clarifying questions, using this knowhow for follow-up projects, and monitoring the project.

6. MODELING OUTLOOK

In the mid-1990s, the BPR debate drew our attention from isolated business activities to entire value chains. Yet "entire" process management in most of the cases focused on the information flow within departmental, corporate, or national boundaries. Obstacles within those areas appeared hard enough to cope with. Therefore, interorganizational communication and cooperation were seldom seriously put on the improvement agenda. Consequently, enterprise models, particularly business process models were also restricted to interorganizational aspects.

In the meantime, after various BPR and ERP lessons learned, companies seem to be better prepared for business scope redefinition. More and more, they sense the limitations of interorganizational improvement and feel the urge to play an active role in the global e-business community. That means not only creating a company's website but also designing the back-office processes according to the new requirements.

Obviously, this attitude has effects on business application systems, too. While companies are on their way to new business dimensions, implemented ERP systems cannot remain inside organizational boundaries. On the technical side, ERP vendors are, like many other software vendors, forced to move from a traditional client–server to a browser–web server architecture in order to deliver e-business capabilities. Hence, for their first-generation e-business solutions almost all big ERP vendors are using a mixed Java/XML strategy. On the conceptual side, ERP vendors are facing the even bigger challenge of providing instruments for coping with the increasing e-business complexity. Business process models appear to be particularly useful in this context. While e-business process models are fundamentally the same as interorganizational process models, integration and coordination mechanisms become even more important:

- Due to increasing globalization, e-business almost inevitably means international, sometimes even intercultural, business cooperation. While ERP systems were multilingual from the very first, the human understanding of foreign business terms, process-management concepts, legal restrictions, and cultural individualities is much more difficult. Because models consist of graphic symbols that can be used according to formal or semiformal grammars, they represent a medium that offers the means to reduce or even overcome those problems.
- Many improvement plans fail because of insufficient transparent business processes and structures. If people do not realize the reengineering needs and benefits, they will not take part in a BPR project and accept the proposed or made changes. While this is already a serious problem within organizational boundaries, it becomes even worse in the case of interorganizational, that is, geographically distributed, cooperation. In the case of business mergers or virtual organizations, for example, the processes of the partners are seldom well known. Yet, in order to establish a successful partnership, the business processes have to be designed at a very high level of detail. Thus, business process models can help to define the goals, process interfaces, and organizational responsibilities of interorganizational cooperation clearly.
- Up to now, we have mainly discussed strategic benefits of modeling. However, at an operational level of e-business, models are very useful, too. In business-to-business applications such as supply chain management we have to connect the application systems of all business partners. This means, for example, that we have to fight not only with the thousands of parameters of one single ERP system but with twice as many, or even more. Another likely scenario is that the business partners involved in an interorganizational supply chain are using software systems of different vendors. In this case, a business process model can first be used to define the conceptual value chain. Secondly, the conceptual model, which is independent from a certain software, can be connected to the repositories of the different systems in order to adopt an integrated software solution.

These few examples already demonstrate that enterprise models play a major role in the success of e-business. Some software vendors, such as SAP and Oracle, have understood this development and are already implementing their first model-based e-business applications.

REFERENCES

Burgess, B. H., and Hokel, T. A. (1994), *A Brief Introduction to the Zachman Framework,* Framework Software.

Olle, T. W. et al. (1991), *Information System Methodologies: A Framework for Understanding,* 2nd ed., Addison-Wesley, Wokingham.

Vernadat, F. B. (1996), *Enterprise Modeling and Integration: Principles and Applications,* Chapman & Hall, London.

ADDITIONAL READING

Scheer, A.-W., *Business Process Engineering: Reference Models for Industrial Enterprises,* 2nd ed., Springer, Berlin, 1994.

Scheer, A.-W., *ARIS—Business Process Frameworks,* 2nd Ed., Springer, Berlin, 1998.

Scheer, A.-W., *ARIS—Business Process Modeling,* 2nd Ed., Springer, Berlin, 1999.

II.B

Manufacturing and Production Systems

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 10

The Factory of the Future: New Structures and Methods to Enable Transformable Production

HANS-JÜRGEN WARNECKE
Fraunhofer Society

WILFRIED SIHN
Fraunhofer Institute for Manufacturing Engineering and Automation

RALF VON BRIEL
Fraunhofer Institute for Manufacturing Engineering and Automation

1. THE CURRENT MARKET SITUATION: NOTHING NEW 311
2. TRANSFORMABLE STRUCTURES TO MASTER INCREASING TURBULENCE 314
3. CORPORATE NETWORK CAPABILITY 314
  3.1. Internal Self-Organization and Self-Optimization 315
4. NEW METHODS FOR PLANNING AND OPERATING TRANSFORMABLE STRUCTURES 317
  4.1. Process Management through Process Modeling, Evaluation, and Monitoring 317
  4.2. Integrated Simulation Based on Distributed Models and Generic Model Building Blocks 320
  4.3. Participative Planning in the Factory of the Future 320
  4.4. The Integrated Evaluation Tool for Companies 321
5. CONCLUSION 322
REFERENCES 322
ADDITIONAL READINGS 323

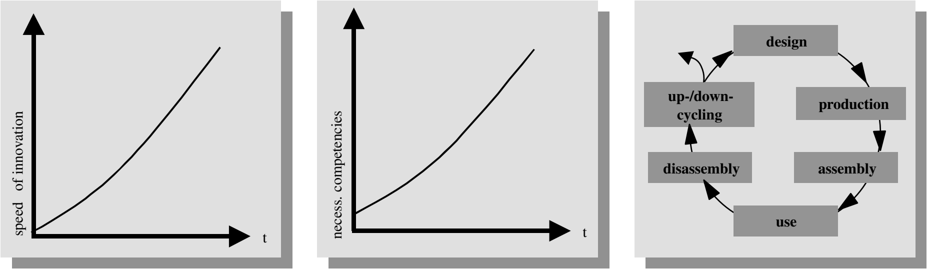
1. THE CURRENT MARKET SITUATION: NOTHING NEW

The renowned American organization scientist Henry Mintzberg has discovered that companies and their managers have been complaining of market turbulence and high cost and competition pressures for more than 30 years (Mintzberg 1998). Market turbulence is thus not a new phenomenon, he concludes, and companies should be able to cope with it.

However, there are only a few practical examples to back Mintzberg's claim. Practical experience and research results differ not because market turbulence forces enterprises to adapt but because the speed of change adds a new dimension to market turbulence. The results of the latest Delphi study show that the speed of change has increased considerably in the last 10 years (Delphi-Studie 1998).

The five charts in Figure 1 show the primary indicators used to measure the growing turbulence in the manufacturing environment. The change in mean product life shows that innovation speed has increased considerably in the last 15 years. This change can be validated with concrete figures from several industries. A recent Siemens survey shows that sales of products older than 10 years have dropped by two thirds over the last 15 years and now amount to only 7% of the company's total turnover. In contrast, the share of products younger than 5 years has increased by more than 50% and accounts for almost 75% of the current Siemens turnover (see Figure 2).

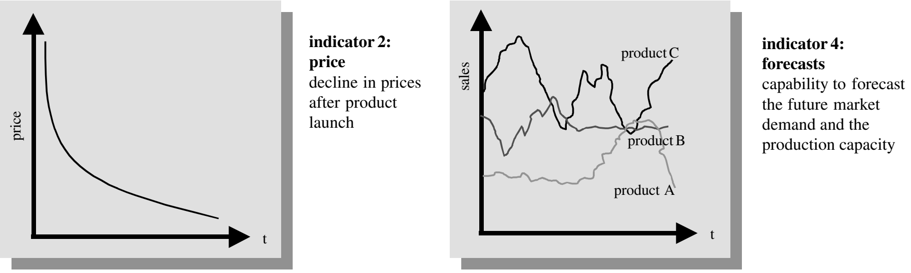
Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.


speed of innovation
t
necess. competencies
t
design
production
assembly
use
disassembly
up-/down-cycling

indicator 1: innovation
growing speed of innovation

indicator 3: complexity
growing product complexity through integration of functions or higher quality standards

indicator 5: importance of life cycle aspects
change in life cycle relevance for product and production


price
t
sales
product C
product B
product A
t

indicator 2: price
decline in prices after product launch

indicator 4: forecasts
capability to forecast the future market demand and the production capacity

Figure 1 Indicators Measuring Market Turbulence.

However, long-range studies are not the only means to identify and verify the changes expressed by the indicators. A short-term analysis, for example, can also serve to prove that change is the main cause of turbulence. The aim of one such analysis is to predict the fluctuations in sales for a medium-sized pump manufacturer. During the first half of the year, the sales trend of some product groups

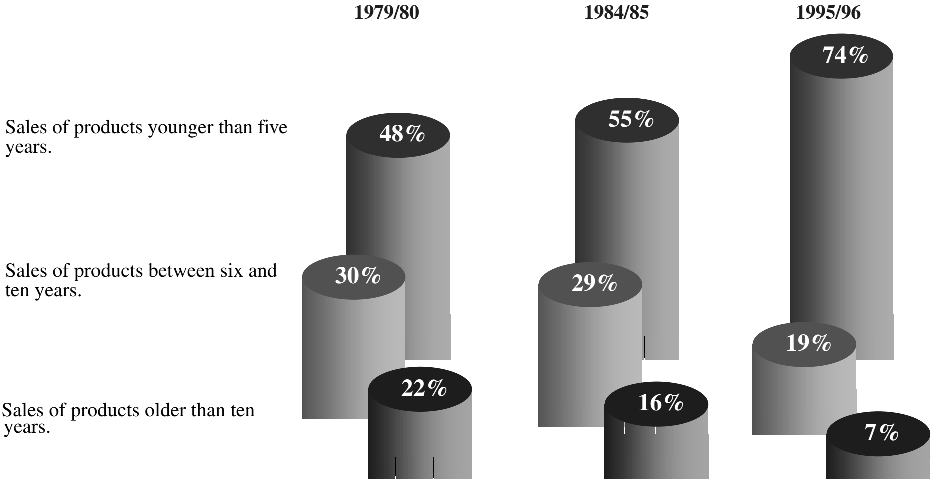


Figure 2 The Changing Life Cycle of Siemens Products. (From Kuhnert 1998)

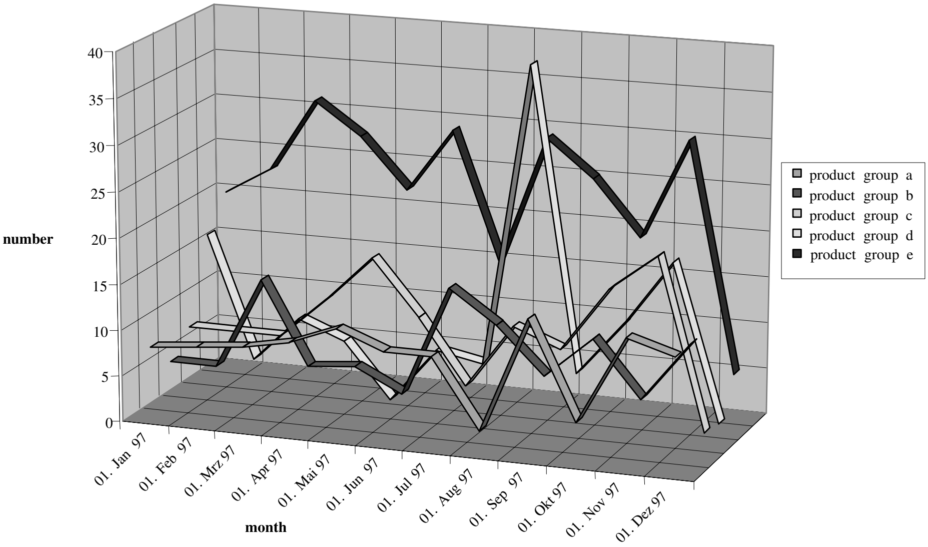


Figure 3 Fluctuations in the Order Entry of a Medium-Sized Pump Manufacturer.

could be predicted based on past results. The forecast for the second half of the year, however, though based on the same procedure, was impossible due to the divergence of monthly sales volume in certain product divisions. Therefore, an enormous adaptation effort on the part of the company was required (see Figure 3).

The overall objective is to raise a company's or factory's *transformability,* the new determinant of corporate market success. Transformability is the ability to adjust rapidly to increased turbulence and recognize changing indicators early enough to initiate proactive adaptation measures.

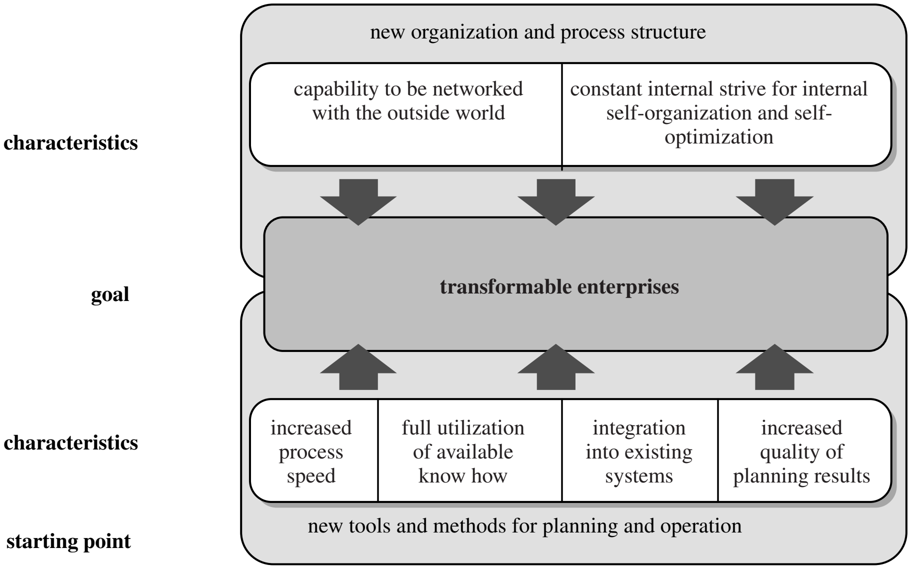

new organization and process structure
characteristics
capability to be networked with the outside world
constant internal strive for internal self-organization and self-optimization
goal
transformable enterprises
characteristics
increased process speed
full utilization of available know how
integration into existing systems
increased quality of planning results
starting point
new tools and methods for planning and operation

Figure 4 Contributing Factors of a Transformable Company.

2. TRANSFORMABLE STRUCTURES TO MASTER INCREASING TURBULENCE

Recent restructuring projects have shown that a company's competitiveness can be enhanced through improved technology and, in particular, through the efficient combination of existing technology with new company structures.

Factories of the future must have transformable structures in order to master the increased turbulence that began in the 1990s. Factories must also possess two important features reflecting two current trends in structural change: they must be able to enable for external structure networking and self-organize and self-optimize structures and processes internally. External networking helps to dissolve company borders that are currently insurmountable and integrate individual enterprises into company networks. Self-organization and self-optimization are intended to enhance the competencies of the value-adding units in accordance with the corporate goal and thus speed up the decision making and change processes of the enterprise.

3. CORPORATE NETWORK CAPABILITY

Corporate network capability describes the capacity of an enterprise to integrate both itself and its core competencies into the overall company network. This cooperation is continually optimized to benefit both the individual company and the network.

Companies with this capacity find that their transformability increases in two respects. First, they can focus on their core competencies and simultaneously profit from the company network's integrated and comprehensive service range. Second, the information flow in the company network is sped up by the continual networking of suppliers and customers. The advantage of the latter is that companies are provided with information concerning market changes and adjustments in consumer and supplier markets in due time. They can thus respond proactively to technological and market changes. These company networks can take four forms, as shown in Figure 5.

Companies are not limited to only one of the four types. Instead, each company can be involved in several company network types on a horizontal as well as a vertical level. In the automobile industry, for example, it is common practice for companies to cooperate in regional networks with companies on the same value-adding level and at the same time be part of the global supply chain of an automobile manufacturer. In other words, companies do not necessarily have to focus on one cooperative arrangement. Different business units might establish different cooperative arrangements.

Behr is an example of an automobile supplier with high network capability. The company's focus is vehicle air conditioning and motor cooling. Its customers include all the large European car manufacturers. In order to offer its customers a comprehensive service range, Behr is involved in nu-

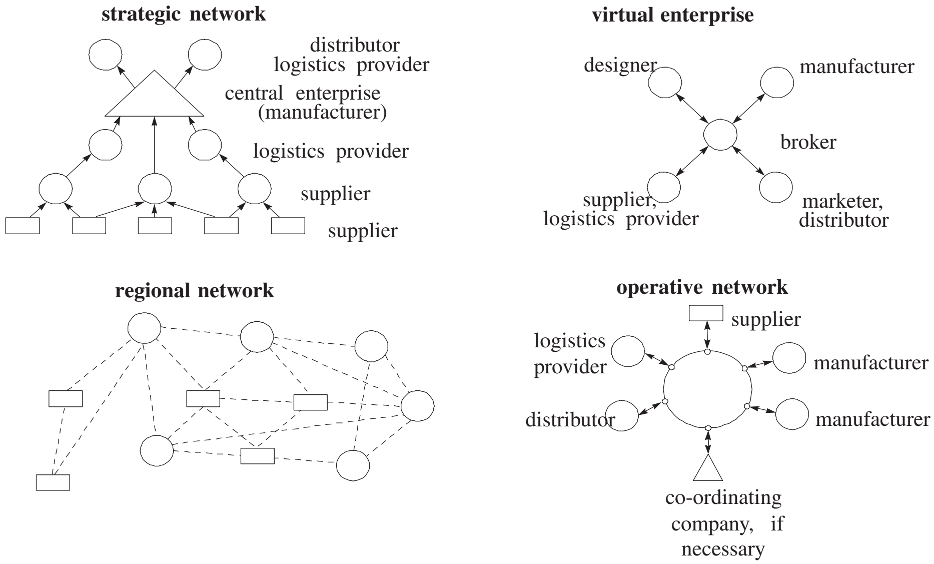

strategic network
distributor
logistics provider
central enterprise
(manufacturer)
logistics provider
supplier
supplier
virtual enterprise
designer
manufacturer
broker
supplier,
logistics provider
marketer,
distributor
regional network
operative network
supplier
logistics
provider
manufacturer
distributor
manufacturer
co-ordinating
company, if
necessary

Figure 5 Basic Types of Company Networks.

merous global and local networks. In 1997 and 1998, Behr joined six cooperation projects with other automotive suppliers in which the cooperation partners were distinguished by their expertise, flat hierarchies, and accompanying short information and communication paths as well as similar corporate culture. The aim of the resulting expert network was to pool the necessary system technology and thus improve the position in the core business as well as to prepare the basis for entering new markets for components.

The project management is assumed by the partner with the greatest market share or the most comprehensive business logistics. It is also important that a continual communication flow exist, accompanied by rapid decision making processes and clearly defined interfaces that are well managed and intelligible for the customer. Accordingly, Behr and one of its cooperation partners have taken the initiative to jointly develop and supply front-end modules. The module is composed of the cooling system's heat transfer, the air-conditioning condenser, cooling fan, air deflector, lighting, bumpers, and radiator grill. Together, the cooperation partners are able to provide 75% of the module components. Through this network, Behr was able to increase its market presence and revenue.

3.1. Internal Self-Organization and Self-Optimization

In corporate self-organization and self-optimization, the authority to make decisions is shifted directly to the value-adding units, creating semiautonomous organizational units. Therefore, apart from the actual output process, semiautonomous units have to integrate management and planning functions as well as coordinate functions to ensure global optima and avoid local optima.

The advantages of semiautonomous structures with regard to improved labor productivity have been confirmed by numerous studies and accepted by most experts (Westkämper et al. 1998). Advantages include increased corporate transformability, which can be achieved through a quicker information flow in semiautonomous units, and the involvement of the process owners in the decision making process from an early stage. The latter also leads to better staff motivation, which positively affects transformability.

An important task for semiautonomous organizational units is to coordinate themselves with the common target system in the course of self-optimization (see Figure 7). Making a profit is the main goal of this target system. However, the system must still be coordinated with the interests of the employees. In addition, the goals must affect wages in order to increase the credibility of the goal system.

The Fraunhofer Institute for Production and Automation (IPA) has carried out more than 100 industrial projects in which it applied the principle of the fractal company to introduce semiautonomous organization structures coordinated by means of a common target system. More than 20 of

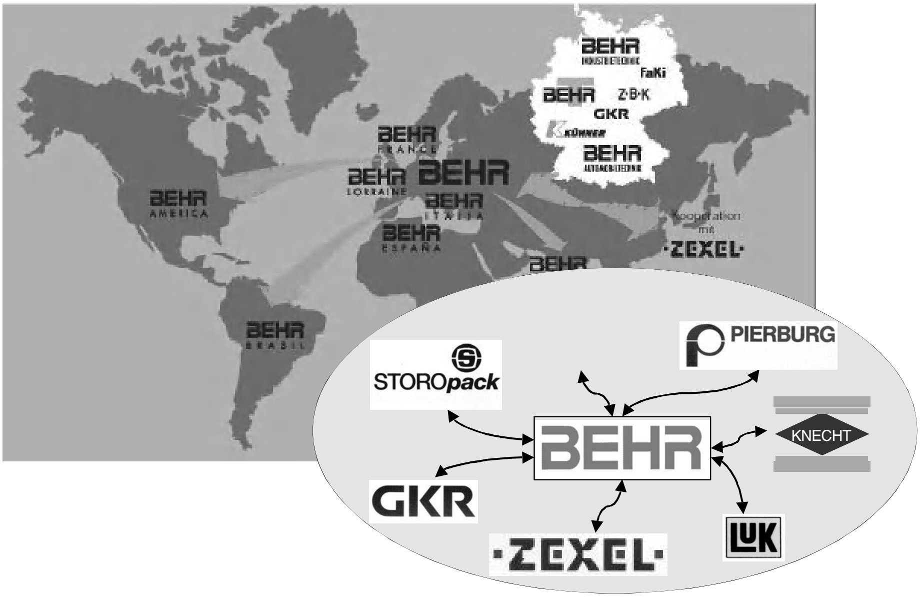


Figure 6 The Behr Company as an Example of a Networked Structure.

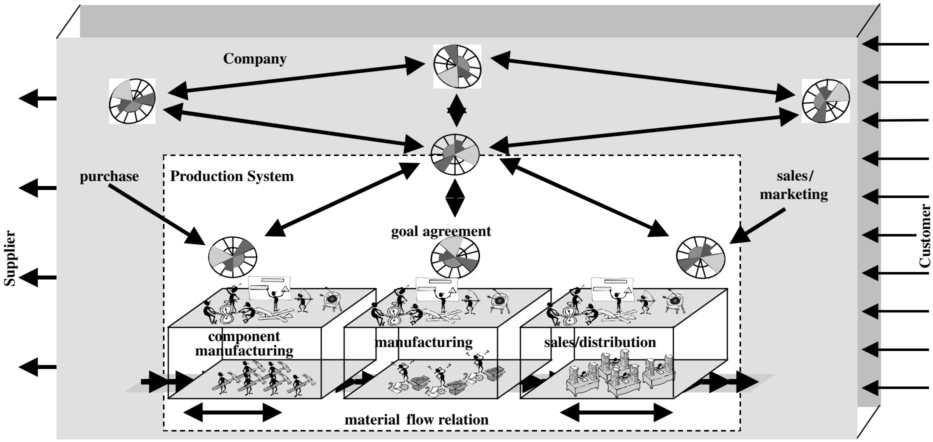

Company
purchase
Production System
sales/
marketing
goal agreement
Supplier
Customer
component
manufacturing
manufacturing
sales/distribution
material flow relation

Figure 7 Common Target System to Coordinate Semiautonomous Organizational Units.

these projects have been further analyzed, showing increased corporate transformability and a significant change in organizational parameters.

A manufacturer of packaging machinery was among the companies analyzed. It exemplifies the changes in the organizational structure and the transition to self-optimizing structures. Before the organizational changes were implemented, the company had a classical functional organization structure (see Figure 9). This structure led to a high degree of staff specialization, which was unsatisfactory in terms of the value-adding processes. For example, six departments and a minimum of 14 foremen and their teams were involved in shipping a single packaging machine.

In the course of the change process, semiautonomous units were created based on the idea of the fractal company. These units performed all functions necessary for processing individual customer orders. Part of the new structure included organizational units that focused on product groups for customized assembly processes. The interfaces of the latter no longer serve to carry out individual

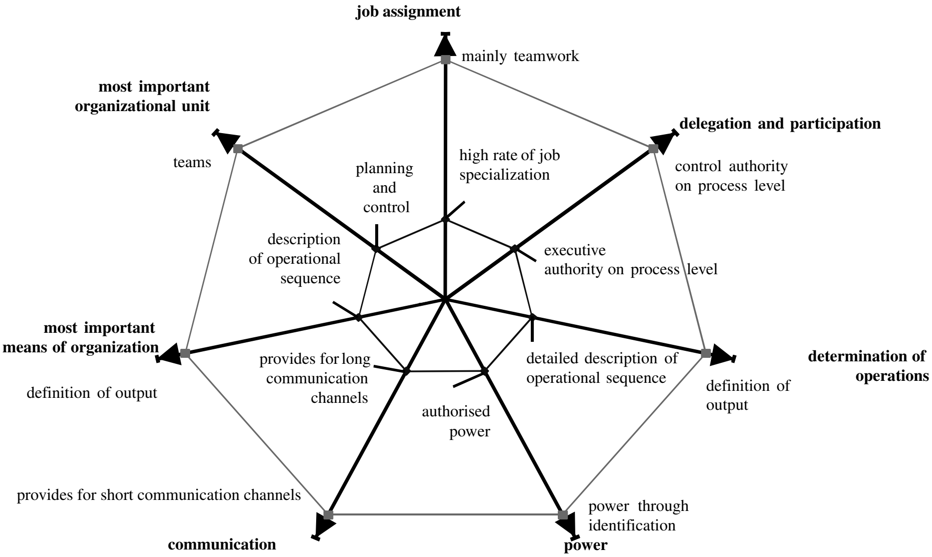

job assignment
mainly teamwork
most important
organizational unit
delegation and participation
teams
planning
and
control
high rate of job
specialization
control authority
on process level
description
of operational
sequence
executive
authority on process level
most important
means of organization
determination of
operations
definition of output
provides for long
communication
channels
detailed description of
operational sequence
definition of
output
authorised
power
provides for short communication channels
power through
identification
communication
power

Figure 8 Basic Change in Self-optimizing Structures. (From Kinkel and Wengel 1998)

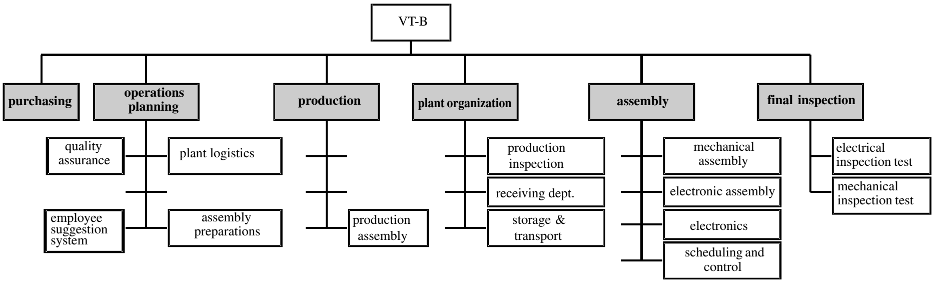

VT-B
purchasing
operations planning
production
plant organization
assembly
final inspection
quality assurance
plant logistics
employee suggestion system
assembly preparations
production assembly
production inspection
receiving dept.
storage & transport
mechanical assembly
electronic assembly
electronics
scheduling and control
electrical inspection test
mechanical inspection test

Figure 9 Traditional Corporate Structure Exemplified by a Manufacturer of Packaging Machinery.

customer orders but are instead needed for long-term planning and maintaining the production process (see Figure 10).

Reducing the interfaces in the planning process and at the same time adjusting the planning and control principles to the new corporate responsibility structure cut the cycle time for producing a machine by 60%.

4. NEW METHODS FOR PLANNING AND OPERATING TRANSFORMABLE STRUCTURES

In the future, companies will need new planning methods to increase their transformability and adaptability to a changing business environment. These methods and tools will speed up processes but also increase performance quality and integrate existing methods. The following four methods allow to the transformability of an enterprise to be increased. They differ in the extent of their market penetration, their market maturity, and their operative range. All methods are based on the new corporate structure and attempt to improve the related processes (see Figure 11).

4.1. Process Management through Process Modeling, Evaluation, and Monitoring

If future enterprises are no longer structured according to Tayloristic functions but focus on added value, then new methods and procedures will be required for mapping and evaluating these processes.

To achieve increased transformability, it is necessary to check the process efficiency continuously with regard to the current speed of change. If processes are to be successfully and comprehensively

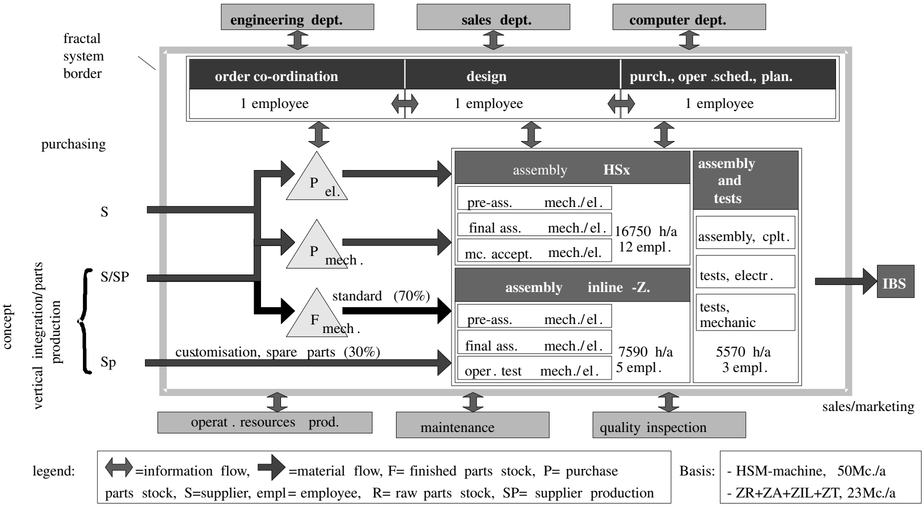

engineering dept.
sales dept.
computer dept.
fractal system border
order co-ordination
1 employee
design
1 employee
purch., oper .sched., plan.
1 employee
purchasing
S
S/SP
Sp
concept
vertical integration/parts production
P el.
P mech .
F mech .
standard (70%)
customisation, spare parts (30%)
assembly HSx
pre-ass. mech./el.
final ass. mech./el.
mc. accept. mech./el.
16750 h/a
12 empl.
assembly inline -Z.
pre-ass. mech./el.
final ass. mech./el.
oper. test mech./el.
7590 h/a
5 empl.
assembly and tests
assembly, cplt.
tests, electr.
tests, mechanic
5570 h/a
3 empl.
IBS
sales/marketing
operat . resources prod.
maintenance
quality inspection
legend:
⇔ =information flow, ➡ =material flow, F= finished parts stock, P= purchase parts stock, S=supplier, empl= employee, R= raw parts stock, SP= supplier production
Basis:
- HSM-machine, 50Mc./a
- ZR+ZA+ZIL+ZT, 23Mc./a

Figure 10 Functional Integration after Transformation.

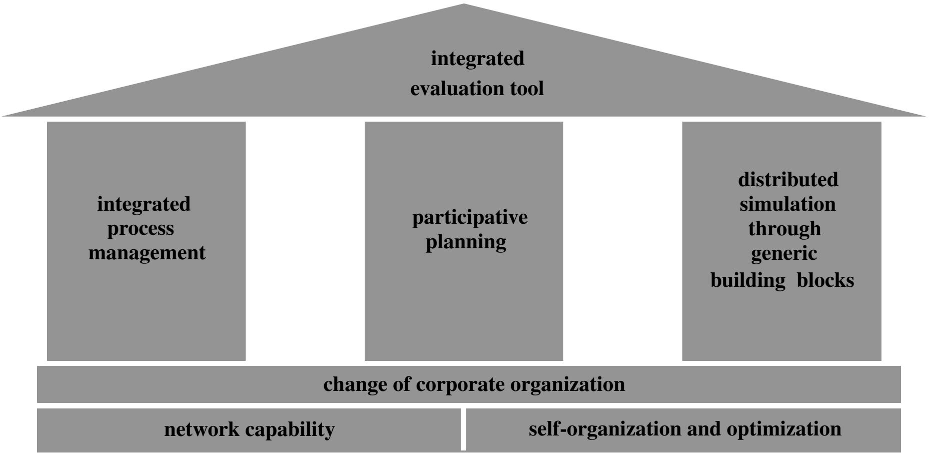

integrated
evaluation tool
integrated process management
participative planning
distributed simulation through generic building blocks
change of corporate organization
network capability
self-organization and optimization

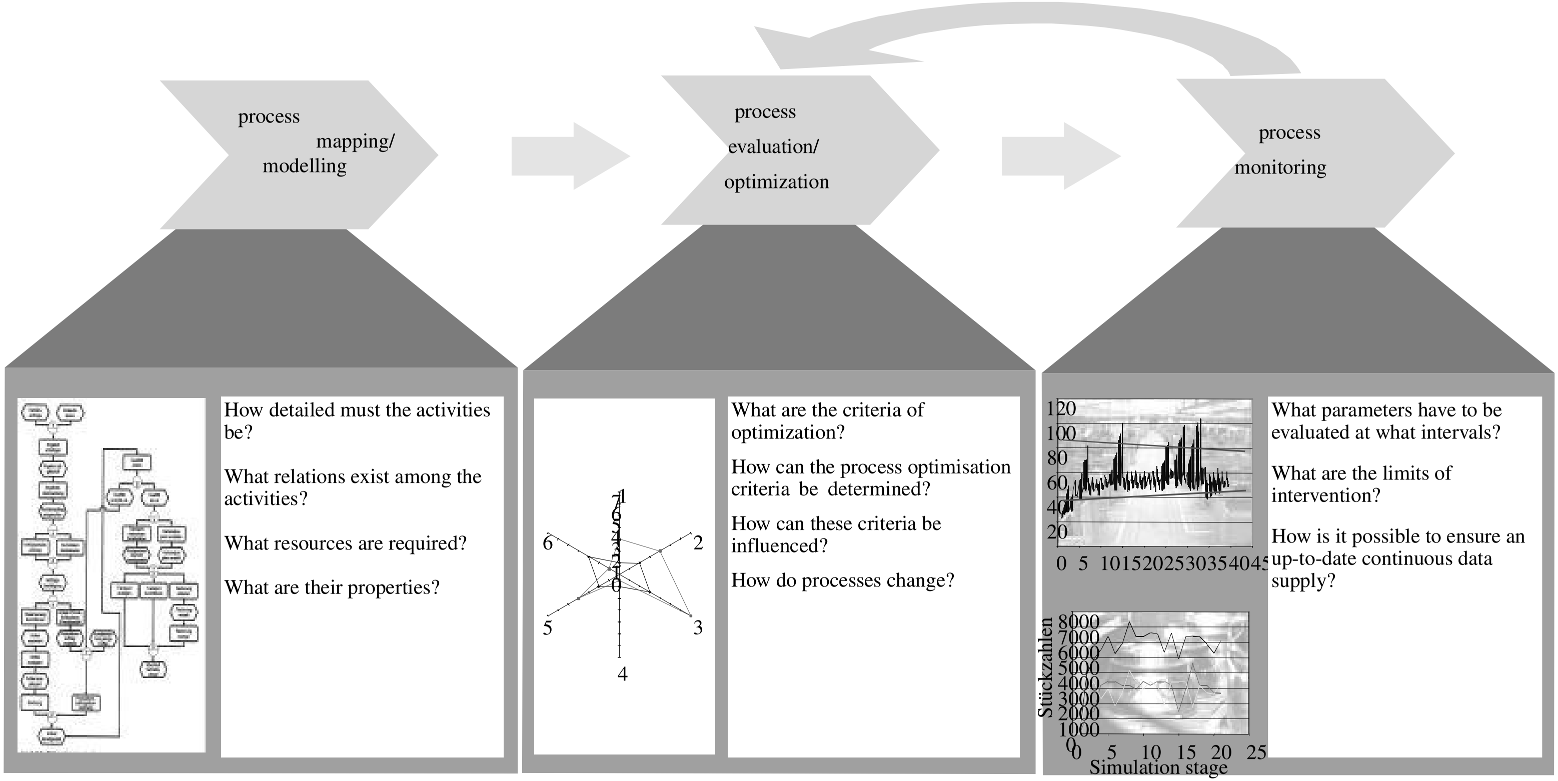
Figure 11 The Structure of the Methods Presented.

managed, it is not sufficient to map the process once but rather use an integrated and holistic approach. It is necessary that process modeling be understood by all those involved in the process. At the process evaluation and optimization stage, the structures of the process model are evaluated and optimized from different points of view before being passed on to the continuous process monitoring stage (see Figure 12).

A great number of modeling tools for process modeling are currently available on the market and applied within companies. These tools all have the capacity to map the corporate process flow using predefined process building blocks. Depending on the necessary degree of specification, the processes can be depicted in greater detail over several levels. Each process module can be equipped with features such as frequency, duration, and specific conditions to enable evaluation of the work flow as a whole. In addition, the process steps can be linked through further modules such as required resources and required and created documents. However, none of the tools can provide specific instructions as to the extent of the models' detail. The specification level depends, for the most part, on the interest of the persons concerned, including both the partners in the corporate network and the semiautonomous organizational units within the individual companies. In any case, the interfaces between the organizational units and the network partners must be adequately specified so that the required results to be delivered at the interfaces are known. However, a detailed description of individual activities in the organizational units is not necessary.

The evaluation and optimization of processes can be based on the performance results provided at the interfaces. Performance results means primarily to the expected quality, agreed deadlines, quantities and prices, and maximal costs. So that the process efficiency and the process changes with regard to the provided results can be evaluated, the immediate process figures have to be aggregated and evaluated in terms of the figures of other systems. This leads to new cost tasks. An important leverage point in this respect is the practical use of process cost calculation and its integration into process management(von Briel and Sihn 1997).

At the monitoring stage, it is necessary to collect up-to-date process figures. Based on these key figures, cost variance analyses can be carried out that allow the relevant discrepancies between desired and actual performance to be determined and illustrated by means of intelligent search and filter functions. The relevant figures are then united and aggregated to create global performance characteristics. Due to the locally distributed data and their heterogeneous origins, monitoring processes that encompass several companies is problematic. To solve this problem, the Fraunhofer Institute developed the supply chain information system (SCIS), which is based on standard software components and uses the Internet to enable the continuous monitoring of inventory, quality, and delivery deadlines in a multienterprise supply chain. The person responsible for the product stores the data of each supplier separately in a database. This product manager is therefore in a position to identify and mark critical parts on the basis of objective criteria (ABC analyses, XYZ analyses) and practical experience. It is then possible to determine dates (e.g., inventory at a certain point in time) for critical parts and carry out time analyses (e.g., deadline analyses for a certain period of time, trend analyses). These analyses are used to identify logistical bottlenecks in the supply chain. Moreover, the analyses can be passed on to the supply chain partners, thus maintaining the information flow within the supply chain.


process mapping/ modelling
process evaluation/ optimization
process monitoring
1
2
3
4
5
6
7
0
120
100
80
60
40
20
0 5 1015202530354045
8000
7000
6000
5000
4000
3000
2000
1000
0
0 5 10 15 20 25
Stückzahlen
Simulation stage

How detailed must the activities be?

What relations exist among the activities?

What resources are required?

What are their properties?

What are the criteria of optimization?

How can the process optimisation criteria be determined?

How can these criteria be influenced?

How do processes change?

What parameters have to be evaluated at what intervals?

What are the limits of intervention?

How is it possible to ensure an up-to-date continuous data supply?

Figure 12 Model of Holistic Process Optimization.

4.2. Integrated Simulation Based on Distributed Models and Generic Model Building Blocks

Simulation is another important method for increasing the transformability of systems. With the help of simulation, production systems can be evaluated not only during standard operation but during start-up and fluctuation periods. Simulation is therefore an ideal planning instrument for a turbulent environment characterized by specific phenomena. However, in practice, the use of simulation is increasing only slowly due to the need for accompanying high-performance computer systems and reservations regarding the creation of simulation models, pooling of knowhow, and maintenance of the models. The main problem remains the nonrecurring creation of simulation modules and the need to create a problem-specific overall model.

Two simulation methods can be used to speed up the planning process. The first method generates company- and industry-specific modules that can be put together to create individual simulation models. This is achieved by simply changing the parameters without having to build a completely new planning model. The modules can be continuously created until the user no longer needs to adapt the individual modules but merely has to modify the parameters of the overall model (see Figure 13).

The model of a production system developed by the Fraunhofer Institute shows that two basic building blocks suffice to create a structured model of 10 different types and 25 pieces of machinery. Within two weeks, 25 variations of the production system could be analyzed in collaboration with the production system staff.

The second method enhances the creation of simulation models by allowing distributed models to be built that interact through a common platform or protocol. Thus, various problem-specific tools can be put into action so that there is no need to apply a single tool for all problems. Moreover, the modeling effort and maintenance costs are shared by all participants. Accordingly, every semiautonomous corporate unit and partner in the corporate network can maintain and adapt its own model. On the other hand, because a common platform ensures that the models are consistent and executable, corporate management doesn't lose control over the overall model. The introduction of HLA communication standards for an increasing number of simulation systems fulfills the principal system requirements for networked simulation.

4.3. Participative Planning in the Factory of the Future

The participative planning method is used to reduce the number of interfaces that arise when a complex planning task is solved—that is, one that involves many planning partners and their individual knowhow. It also improves information flow at the interfaces, allows planning processes to be carried out simultaneously, and prevents double work in the form of repeated data input and data transfer. Participative planning is based on the theory that cooperation considerably increases efficiency in finding solutions for complex problems.

The basic principle of participative planning is not new. It was used in factory and layout planning long before the introduction of computer assistance. Previously, team meetings used paper or metal

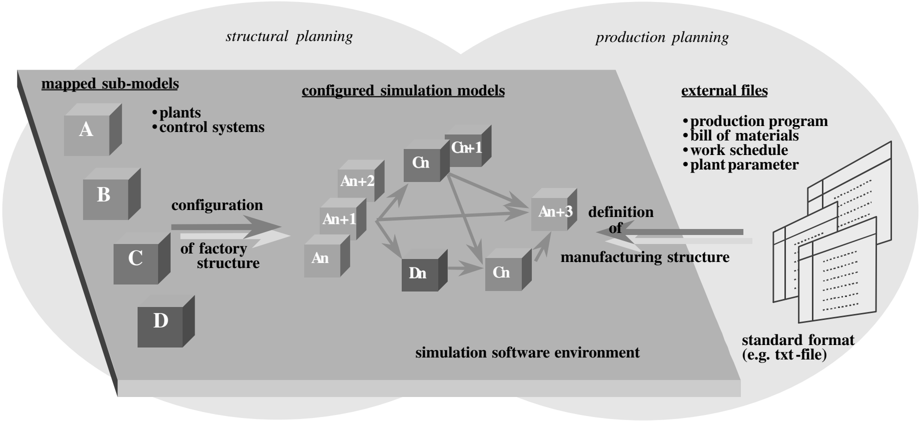

structural planning
production planning
mapped sub-models
• plants
• control systems
A
B
C
D
configuration
of factory structure
configured simulation models
An+2
An+1
An
Cn
Cn+1
Dn
Cn
An+3
definition of manufacturing structure
external files
• production program
• bill of materials
• work schedule
• plant parameter
standard format (e.g. txt-file)
simulation software environment

Figure 13 Modeling with Predefined Building Blocks.

components to represent machinery or production areas. These items could be fastened or glued to the factory's layout plan. This procedure has remained more or less the same in participative planning, the only difference being that interactive factory planning uses virtual elements that exist solely in the computer model and do not need to be glued or pinned. These objects not only possess proper geometrical characteristics but also include additional information concerning object values.

The basic tool of participative planning, the interactive planning table, is based on an idea from the Institute for Hygiene and Applied Physiology and the Institute for Design and Construction at the ETH Zurich. A projector is used to project a 2D image of the planning area via a mirror onto an ordinary table. Simultaneously, a 3D model of the planning area is presented on the wall. The same mirror returns the picture on the table to a camera mounted beside the projector. Thus, feedback on the changes performed on the table is made available. The interactive mechanism of the planning table works by means of metal building bricks with reflective upper surfaces that make it possible for the camera to register the movements of the bricks. Two main forms of interaction exist. One uses one of two brick sizes and the other, the classical way, uses a mouse and keyboard (see Figure 14).

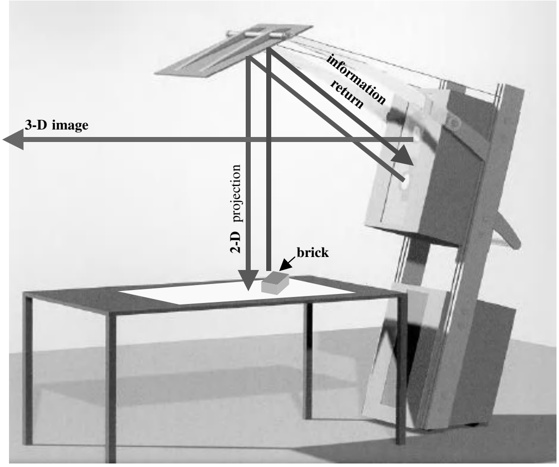
The interactive planning method leads to a significant reduction of planning time while maintaining planning quality. Accordingly, this planning model especially applies to planning cases in which the highest possible planning quality is imperative due to low profit margins and high investment costs, when several partners with different knowhow levels participate in the planning and the planning takes place under high time pressure. Given the increased speed of change, this will apply to an increasing number of planning cases.

4.4. The Integrated Evaluation Tool for Companies

Many enterprises argue that decisions must affect the revenue figures more quickly and in turn influence the corporate profit situation and business value. Hardly any company data are more controversial than the accounting figures. The main criticism of these figures is that for the most part, they are based on past results. Moreover, because the results are aggregated and based on many different accounting parameters, it is not possible to derive clear-cut measures in due time. This means there is too much time between the occurrence and recognition of weak spots in the corporate revenue figures.

However, an integrated evaluation tool has to include accounting figures, which form the basis for corporate decision making and also determine the key data that allow changes to be registered before they take effect in accounting. The balanced scorecard (Kaplan and Norton 1996), a method developed for practical use by Kaplan, provides an easy-to-understand method that enables accounting figures to be combined with other process data (see Figure 15).

The balanced scorecard identifies vital figures for answering the following questions:



2-D projection
Abeamer is used to project an image of the planning area on to an ordinary table.

3-D projection
In addition, a three-dimensional view of the planning area will be projected on the wall.

Image return
The picture reflected on the table is returned via a camera mounted beside the beamer.

Figure 14 The Technology Behind the Factory Planning Table.

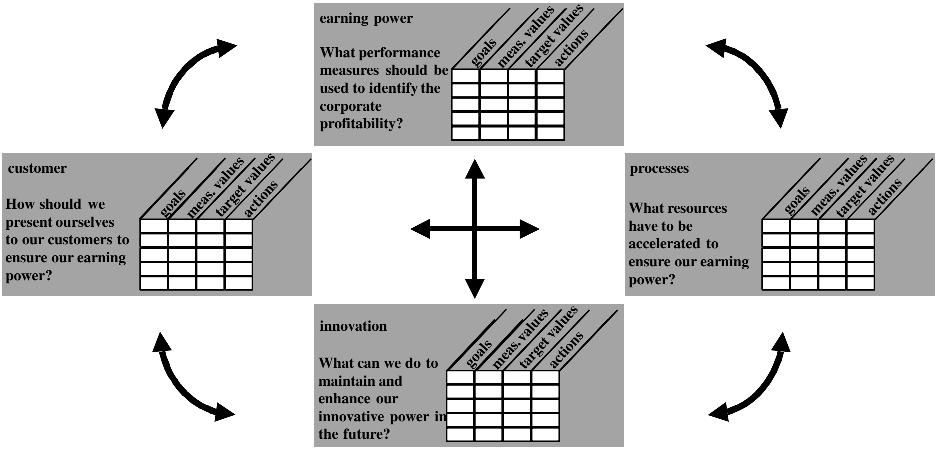

earning power
What performance measures should be used to identify the corporate profitability?
goals
meas. values
target values
actions
customer
How should we present ourselves to our customers to ensure our earning power?
goals
meas. values
target values
actions
processes
What resources have to be accelerated to ensure our earning power?
goals
meas. values
target values
actions
innovation
What can we do to maintain and enhance our innovative power in the future?
goals
meas. values
target values
actions

Figure 15 The Principal Structure of the Balanced Scorecard.

- How does the current profit situation of the company compare to the market situation?
- How is the company's market attractiveness rated by its customers?
- What is the current situation in the company's value-adding processes?
- What improvements can be made in corporate processes and products, and what is the innovative power of the company?

This tool helps a company recognize changes in the corporate situation at an early stage and initiate appropriate actions due to the depth and variety of the questions and their differing time horizon. The balanced scorecard thus serves as an early warning system for a company so that it can actively respond to market changes. The tool also allows the current profit and market situation of the company to be taken into consideration. Thus, the balanced scorecard helps the company avoid the short-range view of the profit situation and expand it to include vital aspects of corporate management.

5. CONCLUSION

Based on Peter Drucker's now generally recognized thesis that only the uncertain is certain in the future markets of manufacturing enterprises (Drucker 1992), it appears safe to forecast that manufacturing structures will change fundamentally. In this context, the transformability of future manufacturing structures will become an important feature, enabling quick and proactive response to changing market demands.

In the scenario of transformable manufacturing structures, the focus will no longer be on the "computer-directed factory without man" of the 1980s but on the accomplishment of factory structures that employ of the latest technology. Human beings, with their unique power for associative and creative thinking, will then take a crucial part in guaranteeing the continuous adaptation of factory structures.

Implementing transformability requires the application of new manufacturing structures in factories and enterprises. These structures are distinguished by their capacity for external networking and increased internal responsibility. Combining these new structures with new design and operation methods will lead to factories exceeding the productivity of current factory structures by quantum leaps. However, this vision will only become reality if both strategies—new structures and methods—are allowed to back up each other, not regarded as isolated elements.

REFERENCES

Delphi-Studie (1998), *Studie zur globalen Entwicklung von Wissenschaft und Technik.*

Drucker, P. (1992), *Managing for the Future,* Dutton, New York.

Kaplan, R. S., and Norton, D. P. (1996), *The Balanced Scorecard: Translating Strategy into Action,* Harvard Business School Press, Boston.

Kinkel, S., and Wengel, J. (1998), ‘‘Produktion zwischen Globalisierung und regionaler Vernetzung,’’ in *Mitteilung aus der Produktionsinnovationserhebung,* Fraunhofer-Institut für Systemtechnik und Innovationsforschung, No. 10.

Kuhnert, W. (1998), ‘‘Instrumente einer Erfolgsorganisation,’’ *Proceedings of 6th Stuttgarter Innovationsforum: Wettbewerbsfaktor Unternehmensorganisation* (October, 14–15, 1998).

Mintzberg, H. (1994), ‘‘That’s not ‘Turbulence,’ Chicken Little, It’s Really Opportunity,’’ *Planning Review,* November/December, pp. 7–9.

von Briel, R., and Sihn, W. (1997), ‘‘Process Cost Calculation in a Fractal Company,’’ *International Journal of Technology Management,* Vol. 13, No. 1, pp. 68–77.

Westkämper, E. (1997), ‘‘Wandlungsfähige Unternehmensstrukturen,’’ *Logistik Spektrum,* Vol. 9 No. 3, pp. 10–11 and No. 4, pp. 7–8.

Westkämper, E., Hüser, M., and von Briel, R. (1997), ‘‘Managing Restructuring Processes,’’ in *Challenges to Civil and Mechanical Engineering in 2000 and Beyond, Vol. 3: Proceedings of the International Conference* (Wroclaw, Poland, June 2–5, 1997), Breslau.

ADDITIONAL READINGS

Preiss, K., Ed., *Agile Manufacturing: A Sampling of Papers Presented at the Sixth National Agility Conference,* San Diego, CA, (March 5–7, 1997).

Warnecke, H. J., *Aufbruch zum fraktalen Unternehmen,* Springer, Berlin, 1995.

CHAPTER 11

Enterprise Resource Planning Systems in Manufacturing*

Mary Elizabeth A. Algeo
Edward J. Barkmeyer
National Institute of Standards and Technology

1. **INTRODUCTION** **325**
  - 1.1. Major Business Functions in Manufacturing Enterprises 326
  - 1.2. Manufacturing Operations Planning 327
  - 1.3. Partitioning the Domain of Manufacturing 329
    - 1.3.1. Nature of Process 329
    - 1.3.2. Nature of the Business in Terms of Customer Orders 330
    - 1.3.3. Combining Nature of Process and Nature of Business in Terms of Customer Orders 331

2. **AN INTERNAL VIEW OF ERP SYSTEMS** **331**
  - 2.1. Scope of ERP Systems in Manufacturing Enterprises 331
  - 2.2. Transaction Management and Basic Decision Support: The Core of ERP 332
    - 2.2.1. Materials Inventory 332
    - 2.2.2. Materials Acquisition 332
    - 2.2.3. Order Entry and Tracking 333
    - 2.2.4. Manufacturing Management 333
    - 2.2.5. Process Specification Management 333
    - 2.2.6. Maintenance Management 334
    - 2.2.7. Warehousing 334
    - 2.2.8. Transportation 335
    - 2.2.9. Human Resource Management 335
    - 2.2.10. Finance Management and Accounting 336
  - 2.3. Interaction Points 336
    - 2.3.1. Contracts Management 336
    - 2.3.2. Supplier Relationship Management 337
    - 2.3.3. Customer Relationship Management 337
    - 2.3.4. Product Configuration Management 338
    - 2.3.5. Product Data Management 338
    - 2.3.6. Supply Chain Execution 338
    - 2.3.7. Supply Chain Planning 338
    - 2.3.8. Manufacturing Execution 338
    - 2.3.9. Human Resource Management 339
    - 2.3.10. Finance 339
  - 2.4. Elements of ERP Implementations 339
    - 2.4.1. Core ERP—Transactions 339
    - 2.4.2. Packaged Decision Support Applications 339
    - 2.4.3. Extended Applications 340
    - 2.4.4. Tools 340
  - 2.5. ERP Architectures 341
  - 2.6. ERP and the Internet 342

---

* Official contribution of the National Institute of Standards and Technologies; not subject to copyright in the United States.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2.6.1. Internal User-to-ERP Interfaces 343
2.6.2. External User-to-ERP Interfaces 343
2.6.3. B2B Supply Chain Operations Interfaces 343
2.6.4. Joint Supply Planning (Advanced Planning and Scheduling) Interfaces 344

3. AN EXTERNAL VIEW OF ERP SYSTEMS 344
3.1. ERP and the Economy 344
3.2. ERP, Supply Chains, and Electronic Commerce 347
3.2.1. Electronic Commerce 347
3.2.2. Supply Chain Management 348

4. ERP CHALLENGES AND OPPORTUNITIES 348
4.1. Research and Technology Development 348
4.1.1. Decision Support Algorithm Development 348
4.1.2. Component Decomposition Analysis 349
4.2. Standards Development 349
4.2.1. ERP–PDM Interfaces 349
4.2.2. ERP–MES Interfaces 349
4.2.3. Supply Chain Operations Interfaces 350
4.3. Establishing Context, Coordination, and Coherence for Achieving Interoperability 350

5. CONCLUSIONS 351

ACKNOWLEDGMENTS 352

DISCLAIMER 352

REFERENCES 352

ADDITIONAL READING 352

1. INTRODUCTION

Enterprise resource planning (ERP) is a class of commercially developed software applications that integrate a vast array of activities and information to support tactical-level operations and operations planning for an industrial enterprise. The term *ERP* refers to the software and not to the related business processes. However, as software, it enables better execution of certain processes. Although often presented as a single package, an ERP system is an envelope around numerous applications and related information. For manufacturers, those applications typically support the operations processes of materials sourcing, manufacturing planning, and product distribution. To its end users, an individual application of an ERP system may appear seamless; however, to those who procure, implement, and/or maintain ERP systems, they are complex software systems that require varying levels of customization and support both centrally and across applications. While ERP systems are commercial applications developed by individual vendors, they can hardly be considered off-the-shelf. They are part of a continuing trend of outsourcing IT solutions in which part of the solution is bought, part is configured, and part is built from scratch. In general, as the scope and complexity of integrated applications have increased from systems supporting a single business unit to systems supporting an entire enterprise and its relationships with business partners, the portions of an IT solution that are bought and configured have increased while the percentage of custom-built software has decreased. Given their broad organizational and functional scope, ERP systems are unlike any other contemporary commercial manufacturing applications. They provide "transaction management," both from the business perspective and from a database perspective. Additionally, they provide a basic level of decision support. Optionally, they enable development of software for higher levels of decision support, which may be offered by ERP vendors or third-party vendors. It is clear that ERP, as a subject, is very complex. Its use marries technology, business practices, and organizational structures. The purpose of this chapter is to present a high-level view of ERP in order to frame a discussion of technological challenges and research opportunities for improving ERP interoperability. Although ERP is relevant to many types of industries (e.g., goods and services) and organizations (e.g., for-profit and not-for-profit), the discussion in this chapter is limited to ERP in manufacturing enterprises. More specifically, the focus of this chapter is ERP that supports the principal operations of a manufacturing enterprise: planning, procuring, making, and delivering products. An ERP system may support other enterprise functions, such as finance management, human resource management, and possibly sales and marketing activities. Detailed analysis of those functions is beyond the scope of this chapter; however, the linkages of those functions with manufacturing-specific functions are not.

This overview looks at ERP by itself and as part of a larger entity (Figure 1). Section 2 discusses ERP internals such as core functions, implementation elements, and technology issues. Additionally,

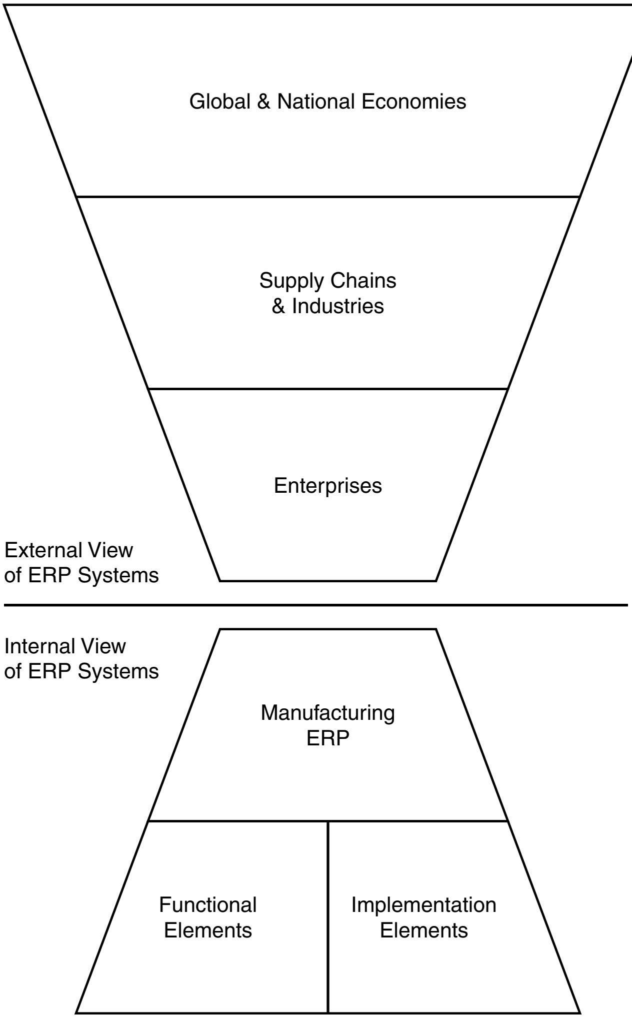

Global & National Economies
Supply Chains & Industries
Enterprises
External View of ERP Systems
Internal View of ERP Systems
Manufacturing ERP
Functional Elements
Implementation Elements

Figure 1 External and Internal Views of ERP.

Section 2 identifies critical integration points for ERP and other applications within manufacturing enterprises. Section 3 discusses ERP and its relationship to three larger entities, namely the U.S. economy, supply chains, and individual manufacturers. Section 4 presents issues and possible resolutions for improving ERP performance and interoperability.

This chapter is the result of a two-year study funded by two programs at the National Institute of Standards and Technology: the Advanced Technology Program's Office of Information Technology and Applications and the Manufacturing Systems Integration Division's Systems for Integrating Manufacturing Applications (SIMA) Program.

The concepts presented in this chapter were gathered from a variety of sources, including literature reviews, manufacturing industry contacts, ERP vendor contacts, consultants specializing in the applications of IT to manufacturing, relevant professional and trade associations, and standards organizations.

1.1. Major Business Functions in Manufacturing Enterprises

Manufacturers typically differentiate themselves from competitors along the three major business functions through which they add value for their customers. Customer relationship management

(CRM), the first dimension of competitive advantage, seeks to add value for customers through those processes that involve direct contact with customers before, during, and after sales. The idea is to understand the prevalent needs and concerns of individual customers and groups of customers. Product development, the second dimension of competitive advantage, focuses on product—what and how to produce an object to satisfy the customer's *want.* Operations, the third dimension of competitive advantage, focuses on satisfying *demand*—how much to make, when to make, and where to make—by producing and delivering products in an effective and efficient manner.

Needs, wants, and *demands* are basic concepts underlying modern, market-based economies (Kotler and Armstrong 1999). Needs beget wants, which beget demand. Needs are states of felt deprivation. They are a basic part of our human condition and are physical, social, and individual in nature. The customer's needs include product capabilities, product servicing, user instruction, and business relationships. Wants are the forms taken by human needs as shaped by culture and individual personality. They are described in terms of objects that will satisfy needs. Demands are human wants that are backed by buying power. Identifying needs and translating them into wants in terms of product and process definitions are the objectives of product development. Satisfying demand, given supply conditions as well as product and process definitions, is the objective of operations. This high-level partitioning of manufacturing business functions into CRM, product development, and operations has growing acceptance in manufacturing and related industries (Hagel and Singer 1999). This acceptance has been fostered by the realization that the underlying activities of these high-level functions are those that add value for the customer.

The complex activities of product development seek to satisfy customer want by translating the abstract to the physical through the product development process. As such, in commercial manufacturing enterprises, product development typically starts with an analysis of market opportunity and strategic fit and, assuming successful reviews through intermediate phases, ends with product release. Among other things, product release serves as a signal to operations to begin production and distribution as the necessary design and manufacturing engineering specifications are ready for execution in a production environment.

Operations, on the other hand, consists of processes for satisfying customer demand by transforming products—in raw, intermediate, or final state—in terms of form, location, and time. To accomplish this objective both effectively and efficiently—and thus meet specific, customer- focused, operational objectives—a manufacturing enterprise must have timely and accurate information about expected and real demand as well as expected and real supply. A manufacturer then considers this information on supply and demand with the current and expected states of its enterprise. It is ERP that allows a manufacturer to monitor the state of its enterprise—particularly the current and near-term expected states. In fact, ERP systems often serve as the cornerstone in the emerging information architectures that support balancing external and internal supply and demand forces. ERP systems play both direct and indirect roles in this trend among manufacturing enterprises towards a synchronized, multilevel, multifacility supply chain planning hierarchy.

1.2. Manufacturing Operations Planning

Figure 2 illustrates the emerging synchronized, multilevel, multifacility supply chain planning hierarchy with ERP as its foundation. The goal of this architecture is to enable more efficient and effective execution across plants, distribution systems, and transportation systems. Independently, these planning activities focus on the strategic, the tactical, and the operational (i.e., execution) levels. Collectively, they support the translation of strategic objectives into actions on the plant floor, in warehouses, and at shipping points throughout the extended enterprise. In addition, they provide top management with up-to-date, synchronized information regarding the state of the entire enterprise.

This synchronization is accomplished by transforming information in a meaningful way from one level within the supply chain planning hierarchy to the next. At the strategic level, top management evaluates numerous factors to determine the design or redesign of the supply chain network as well as time-independent sourcing, production, deployment, and distribution plans. These factors typically include the enterprise's business philosophy as well as company, market, technological, economic, social, and political conditions. Supply chain planning at the strategic level involves "what if" analysis particularly with respect to the first three factors: business philosophy, company conditions, and market conditions. A business philosophy might specify maximizing net revenues or return on assets. Assessment of company conditions considers existing and potential aggregates of fixed (i.e., plant), financial, and human resources. When evaluating market conditions, top management analyzes aggregate product/part demand as well as the anticipated capacity of suppliers and transportation channels—also in aggregate terms. Optimization at this level, which usually employs mathematical programming methods, typically yields the location, size, and number of plants, distribution centers, and suppliers as well as product and supply volumes.

Supply chain operations planning at the tactical level determines the flow of goods over a specific time horizon. Mathematical programming methods yield time-dependent integrated sourcing, pro-

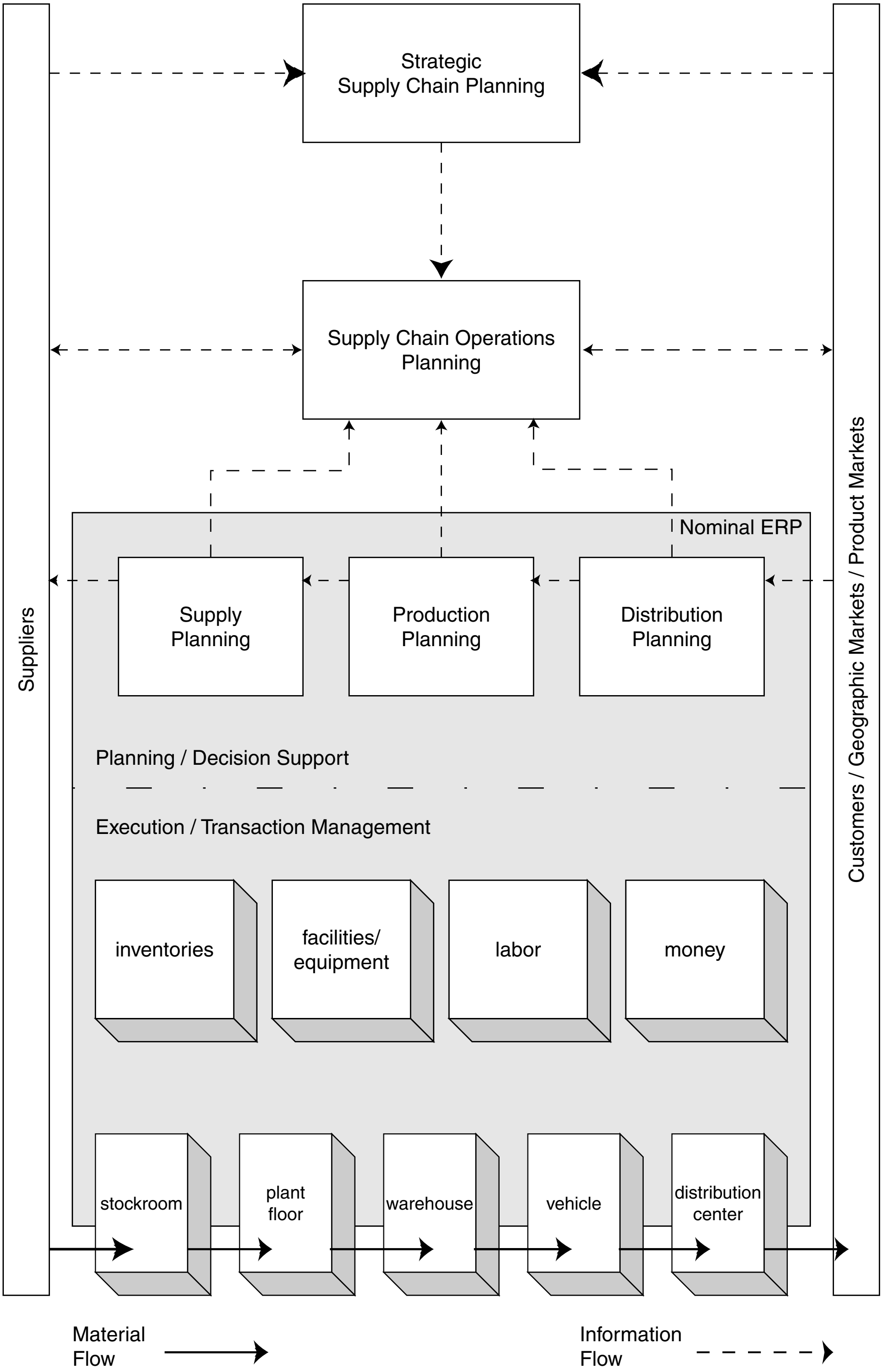

Strategic
Supply Chain Planning
Supply Chain Operations Planning
Nominal ERP
Supply Planning
Production Planning
Distribution Planning
Suppliers
Customers / Geographic Markets / Product Markets
Planning / Decision Support
Execution / Transaction Management
inventories
facilities/ equipment
labor
money
stockroom
plant floor
warehouse
vehicle
distribution center
Material Flow
Information Flow

Figure 2 Intraenterprise View of Supply Chain Planning Hierarchy.

duction, deployment, and distribution plans typically designed to satisfy a financial management objective such as minimizing total supply chain costs or maximizing net revenues by varying product mix. Often, once these plans are established, a more detailed level of tactical planning occurs to optimize supply, production, and distribution independently. Frequently, the objective at this lower level of tactical planning is to minimize costs related to inventories and/or major equipment changeovers.

Supply chain planning at the operational level is, in essence, supply scheduling that occurs on a facility-by-facility basis. As such, separate but coordinated schedules are generated for plants, warehouses, distribution centers, and vehicle systems. Planning at this level differs from tactical and strategic levels in that demand actually exists—that is, orders have been placed. These orders need to be scheduled based on the immediate state of resources (i.e., materials, equipment, and labor). The diverse nature of facilities means that the specifics of optimization vary widely at this level, but the objective typically is to maximize throughput in a given facility.

Individually, these layers serve to separate concerns and enable the definition of tractable planning problems for which mathematical and managerial solutions can be obtained. Collectively, these layers of supply chain planning enable manufacturing enterprises more effectively and efficiently to balance supply, resources, and demand. The upper layers buffer the lower layers from sudden shifts in the market, thus allowing for smoother changes in the enterprise's plants, distribution channels, and transportation systems.

1.3. Partitioning the Domain of Manufacturing

The domain of manufacturing is in fact an aggregate of many subdomains of many types. There is no single correct method for decomposing that complex and dynamic aggregate. The method of decomposition depends on the particular objective at hand. Generally speaking, there are four common approaches to partitioning the manufacturing domain. Each looks at a different aspect of the manufacturing enterprise:

1. *Nature of the product:* This approach categorizes manufacturing industries by the general nature of the product itself—fertilizers, pharmaceuticals, metals, automotive parts, aircraft, etc. This is the approach used by industry classification systems such as the North American Industry Classification System (NAIC) (Office of Management and Budget 1997) and its predecessor, the Standard Industrial Classification (SIC) (Office of Management and Budget 1988). In general, this approach is a good mechanism for characterizing market communities, and thus economic estimators, but it is not a particularly good mechanism for characterizing ERP requirements or planning approaches.
2. *Nature of the customer:* Because most manufacturing output is consumed by other industries, many manufacturers are part of the supply chains ending in original equipment manufacturers (OEMs) in a single major industry: automotive, aerospace, shipbuilding, household appliances, computers, etc. The members of the chain produce different kinds of products, using different processes, with different business behaviors, but the behavior of the supply chain itself often is dominated by the demands of the OEMs.
3. *Nature of the process:* This approach characterizes a domain by the organization of the manufacturing facility and the general nature of the manufacturing processes it employs: continuous process, assembly line, discrete batch, job shop, and construction. There is some correlation between the process type and the product type, in that most manufacturers of a given product type tend to use a particular process organization. In general, process type strongly influences the manufacturing-specific aspects of ERP, including both information capture and planning approaches. On the other hand, large manufacturers often use several different process styles for different products and different components of larger products.
4. *Nature of the business in terms of customer orders:* This characterization includes make-to-stock, make-to-order, assemble-to-order, and engineer-to-order. It has a great deal to do with what the detailed business operations are and how operational and tactical planning is done. Clearly this categorization has a tremendous influence on the ERP requirements and on the behavior of the enterprise in its supply chain. More than any other, this characterization determines the nature of the delivery activities and the dependence on supplier relationships.

These last two categories, as differentiators for ERP, warrant more detailed discussion.

1.3.1. Nature of Process

Continuous process refers to a facility in which products are made by an essentially continuous flow of material through some set of mixing, state transformation, and shaping processes into one or more final products. The final form may itself be intrinsically discrete, or it may be discretized only for packaging and shipment. Examples are wet and dry chemicals, foods, pharmaceuticals, paper, fibers, metals (e.g., plate, bar, tubing, wire, sheet), and pseudocontinuous processes such as weaving, casting, injection molding, screw machines, and high-volume stamping.

Assembly line refers to a facility in which products are made from component parts by a process in which discrete units of product move along an essentially continuous line through a sequence of installation, joining, and finishing processes. Examples are automobiles, industrial equipment, small and large appliances, computers, consumer electronics, toys, and some furniture and clothing.

Discrete batch, also called *intermittent,* refers to a facility in which processes are organized into separate work centers and products are moved in lots through a sequence of work centers in which each work center is set up for a specific set of operations on that product and the setup and sequence is specific to a product family. This describes a facility that can make a large but relatively fixed set of products but only a few types of product at one time, so the same product is made at intervals. This also describes a facility in which the technology is common—the set of processes and the ordering is relatively fixed, but the details of the process in each work center may vary considerably from product to product in the mix. Examples include semiconductors and circuit boards, composite parts, firearms, and machined metal parts made in quantity.

Job shop refers to a facility in which processes are organized into separate work centers and products are moved in order lots through a sequence of work centers in which each work center performs some set of operations. The sequence of work centers and the details of the operations are specific to the product. In general, the work centers have general-purpose setups that can perform some class of operations on a large variety of similar products, and the set of centers used, the sequence, the operations details, and the timing vary considerably over the product mix. Examples include metal shops, wood shops, and other piece-part contract manufacturers supporting the automotive, aircraft, shipbuilding, industrial equipment, and ordnance industries.

Construction refers to a manufacturing facility in which the end product instances rarely move; equipment is moved into the product area and processes and component installations are performed on the product in place. The principal examples are shipbuilding and spacecraft, but aircraft manufacture is a hybrid of construction and assembly line approaches.

1.3.2. Nature of the Business in Terms of Customer Orders

Make-to-stock describes an approach in which production is planned and executed on the basis of expected market rather than specific customer orders. Because there is no explicit customer order at the time of manufacture, this approach is often referred to as a push system. In most cases, product reaches retail outlets or end customers through distribution centers and manufacturing volumes are driven by a strategy for maintaining target stock levels in the distribution centers.

Make-to-order has two interpretations. Technically, anything that is not made-to-stock is made-to-order. In all cases there is an explicit customer order, and thus all make-to-order systems are described as pull systems. However, it is important to distinguish make-to-demand systems, in which products are made in batches, from option-to-order systems, in which order-specific features are installed on a product-by-product basis. The distinction between make-to-demand batch planning and on-the-fly option selection using single setup and prepositioning is very important to the ERP system.

A *make-to-demand* manufacturer makes fixed products with fixed processes but sets up and initiates those processes only when there are sufficient orders (i.e., known demand) in the system. This scenario may occur when there is a large catalog of fixed products with variable demand or when the catalog offers a few products with several options. The distinguishing factor is that orders are batched and the facility is set up for a run of a specific product or option suite. The planning problem for make-to-demand involves complex trade-offs among customer satisfaction, product volumes, materials inventories, and facility setup times.

Option-to-order, also called *assemble-to-order,* describes an approach in which production is planned and executed on the basis of actual (and sometimes expected) customer orders, in which the product has some predefined optional characteristics which the customer selects on the order. The important aspects of this approach are that the process of making the product with options is predefined for all allowed option combinations and that the manufacturing facility is set up so the operator can perform the option installation on a per-product basis during manufacture. This category also applies to a business whose catalog contains a family of fixed products but whose manufacturing facility can make any member of the family as a variant (i.e., option) of a single base product. The option-to-order approach effects the configuration of production lines in very complex ways. The simplest configurations involve prepositioning, in which option combinations occur on the fly. More complex configurations involve combinations of batching and prepositioning.

Engineer-to-order describes an approach in which the details of the manufacturing process for the product, and often the product itself, must be defined specifically for a particular customer order and only after receipt of that order. It is important for other business reasons to distinguish contract engineering, in which the customer defines the requirements but the manufacturer defines both the product and the process, from contract manufacturing, in which the customer defines the product details and the manufacturer defines the process details. But the distinction between contract engineering and contract manufacturing is not particularly important for ERP, as long as it is understood that both are engineer-to-order approaches. In these scenarios, some set of engineering activities must take place after receipt of customer order and before manufacturing execution, and then certain aspects of manufacturing planning can begin.

1.3.3. Combining Nature of Process and Nature of Business in Terms of Customer Orders

Many attempts to characterize manufacturing roll up some combination of the four major categorization approaches (product, customer, process, business) into a single categorization scheme in order to make certain useful generalizations. But because resource planning and ERP systems must deal with important details of the organization's business process, these generalizations do not provide good guidelines for studying the variations in planning and execution information. In particular, no generalization about industries applies to all manufacturing organizations in that industry, and there is no way to roll up the nature of the process with the nature of the business in terms of customer orders. For example, all four business patterns can be observed in the continuous (and pseudocontinuous) processing industries. Even though engineer-to-order (customer-specific recipe) is fairly rare in the chemical and raw metal industries, it is the norm (customer-specific mold) in the injection-molding and casting industries. It does make sense, however, to cross the nature of the process with the nature of the business in terms of customer orders to identify those combined characteristics that have the most significant influence on ERP system requirements. Table 1 identifies 14 distinct categories, out of a possible 20, for which there are known manufacturers in specific industries.

2. AN INTERNAL VIEW OF ERP SYSTEMS

This section describes what ERP systems do and how they do it. Sections 2.1 through 2.3 describe the core functional elements of ERP systems: human resource management, finance management and accounting, contracts management, materials acquisition, materials inventory, maintenance management, order entry and tracking, manufacturing management, process specification management, warehousing, and transportation. Sections 2.4 through 2.6 describe implementation aspects, particularly ERP systems architecture, configuration management tools, and Internet interfaces.

2.1. Scope of ERP Systems in Manufacturing Enterprises

As illustrated in the portion of Figure 2 bounded by the gray box, the functionality of ERP systems encompasses certain interactions among the following elements:

- Four categories of resources (inventories, facilities/equipment, labor, and money)
- Two generic activities (planning/decision support and execution/transaction management)
- Three types of manufacturing operations activities (supply, production, and delivery)
- Five major types of physical facilities (stockroom, plant floor, warehouse, vehicle/depot, and distribution center) through which material flows in manufacturing enterprises

Because there is considerable variation among manufacturers as to which functions in which facilities fall within the scope of an ERP implementation, the gray box does not envelop all physical facilities. Typically, in some way or another, all transactions—all changes of state and many deci-

TABLE 1 Examples per Process and Customer Order Characteristics

	Make-to-Stock	Make-to-Demand	Option-to-Order	Engineer-to-Order
Continuous	refineries	solvents, plastics, alloys	fuels	casting, injection molding
Assembly line	appliances	electric motors, valves	autos, computers	aircraft
Discrete batch	electronic components	windows, auto parts	*no known*	semiconductors, circuit boards
Job shop	*none*	*none*	*none*	metal parts, composites
Construction	*none*	*no known*	aircraft	ships

sions—are captured in an ERP system. Each of these facilities may use additional systems for planning and analysis, execution level scheduling, control, and automated data capture. Some of the planning, analysis, scheduling, and management systems are part of, or tightly connected to, the ERP systems; others are more loosely connected. This variation results from two major factors: systems that are closely coupled to equipment, most of which are highly specialized, and systems that manage information and business processes specific to a particular industry that the ERP vendor may not offer. Because of the historic emphasis on reducing inventory costs, the management of stockroom is, in almost all cases, an intrinsic part of an ERP system. To the contrary, plant floor activity control is almost never a part of ERP. Management of execution activities within warehouses, vehicles/depots, and distribution centers may be handled in a centralized fashion by an ERP system or in a decentralized fashion by an applications specific to those facilities.

2.2. Transaction Management and Basic Decision Support: The Core of ERP

Transactions are records of resource changes that occur within and among enterprises. Through the use of a logically (but not necessarily physically) centralized database, it is the management of these transactions that constitutes the core of an ERP system. More specifically, this transaction database captures all changes of state in the principal resources of the manufacturing enterprise. It also makes elements of the current state available to personnel and software performing and supporting the operations of the enterprise. This scope encompasses all of the numerous resources (i.e., materials inventories, facilities/equipment, labor, and money) and product inventories of all kinds. It also includes the states and results of many business processes, which may not be visible in physical instances (e.g., orders, specifications). The detailed breakdown of this broad scope into common separable components is a very difficult technical task, given the many interrelationships among the objects, significant variations in the business processes, and the technical origins of ERP systems. Nonetheless, it is possible to identify general elements from that complexity and diversity. The functions of a manufacturing enterprise that are supported by transaction management correspond to the major types of resources as follows:

- For inventories, materials inventory and materials acquisition
- For facilities/equipment, manufacturing management, process specification management, maintenance management, warehousing and transportation
- For labor, human resource management
- For money, financial management and accounting
- For product, order entry and tracking

These functions, in whole or in part, make up the core of ERP. The following sections describe each function and its relationship to core ERP. These functions are then discussed in terms of finer-grain execution and planning activities.

2.2.1. Materials Inventory

This function is made up of all information on stores of materials and allocations of materials to manufacturing and engineering activities. That includes information on materials on hand, quantities, locations, lots and ages, materials on order and in transit, with expected delivery dates, materials in inspection and acceptance testing, materials in preparation, and materials allocated to particular manufacturing jobs or product lots (independent of whether they are in stock).

ERP systems routinely capture all of this information and all transactions on it. It is sometimes combined with materials acquisition information in their component architecture.

Execution activities supported by materials inventory include receipt of shipment, inspection and acceptance, automatic (“low-water”) order placement, stocking, stores management, internal relocation, issuance and preparation, and all other transactions on the materials inventory. Except for some stores management functions, all of these are revenue producing.

Planning activities supported by materials inventory include supply planning, manufacturing planning, and manufacturing scheduling.

2.2.2. Materials Acquisition

This function deals primarily with information about suppliers and materials orders. It includes all information about orders for materials, including recurring, pending, outstanding, and recently fulfilled orders, and long-term order history. Orders identify internal source and cost center, supplier, reference contract, material identification, quantity, options and specifications, pricing (fixed or variable), delivery schedule, contacts, change orders, deliveries, acceptances, rejects, delays, and other notifications. It may also include invoices and payments. This also includes special arrangements, such as consignment and shared supply schedules.

ERP systems support part of an enterprise's materials acquisition function by handling all of this information for active suppliers and orders and journalizing all transactions. But they regularly purge closed orders and relationships, using the journal or some other export mechanism to move this information to an archival or data warehouse system. ERP systems generally maintain simple materials specifications directly but typically carry only references to more complex specification documents in some other product data management (PDM) or document management system.

Execution activities supported by materials acquisition include internal requests, placement of external orders and changes, receipt and acceptance of materials, and all other transactions against materials orders. All of these are revenue-producing activities.

Planning activities supported by materials acquisition include supply chain development, supplier identification and qualification, supply planning, manufacturing planning, and cash-flow projections.

2.2.3. Order Entry and Tracking

These functions focus on the customer order through its life cycle. Order entry is the mechanism by which the decision to make product enters the ERP system. It begins with the capture of customer order, including all specifications and options, quantities, packaging requirements, and delivery requirements. It ends with the creation of one or more corresponding manufacturing orders and delivery orders. At that point it becomes order tracking, which follows the customer order through the fulfillment processes (for the surrogate orders) and finally through the payment processes. It is important to note that although manufacturing may be driven directly by customer order, there is a decision point between the entry of the customer order and the release of the associated manufacturing orders, and in most ERP systems they are maintained as separate objects. While this release is often automated, it is a critical business control point and the automation reflects the business rules for the release.

The execution activities supported by order entry and tracking include the revenue-producing activities of customer order capture, production start, and delivery start.

The planning activities supported by order entry and tracking include tactical planning for all of engineering, manufacturing, and delivery, according to the nature of the business as described in Section 1.3.2.

2.2.4. Manufacturing Management

This function deals primarily with the tracking of work through the manufacturing facilities and the management of manufacturing resources that are used in performing that work.

The manufacturing resources include personnel, equipment, and materials. Because each of these is also the subject of another ERP domain (human resources, maintenance, inventory), there is *always* overlap among the concerns. And because many ERP systems developed from manufacturing resource planning (MRP II) systems, which dealt with various aspects of those resources listed above, there is no agreement about where the boundaries are. The one concern that is clearly unique to manufacturing management is the assignment of resources to specific work items. But at some level of planning that depends on resource availability and resource capabilities, which are the boundary areas.

The tracking of work begins with tentative and actual placement of manufacturing orders through the order entry component described above. In general, the manufacturing order information is a part of the manufacturing management component. Planning processes determine which resources (materials, equipment, labor) will be assigned to fulfilling these orders in which time frames and these assignments are captured. Execution processes draw materials (usually tracked as lots) and use equipment and personnel to perform the work. These usages and the flow of work through the facility are captured. Finished goods leave the manufacturing domain for some set of distribution activities, and at this point the completion of the manufacturing orders is tracked.

The execution processes supported by manufacturing management are the revenue-producing processes that convert materials into finished goods, but that support is limited to tracking those processes.

The planning processes supported by manufacturing management are all levels of manufacturing resource planning and scheduling, except for detailed scheduling, as noted above.

2.2.5. Process Specification Management

This function deals with the information associated with the design of the physical manufacturing processes for making specific products. As such, it is an engineering activity and like product engineering, should be almost entirely out of the scope of core ERP systems. But because several information items produced by that engineering activity are vital to resource planning, ERP systems maintain variable amounts of process specification data. In all cases, the materials requirements for a product lot—the ''manufacturing bill of materials''—are captured in the ERP system. And in all cases, detailed product and materials specifications, detailed equipment configurations, detailed operations procedures, handling procedures, and equipment programs are outside the core ERP information bases. These information sets may be maintained by ERP add-ons or third-party systems, but the core contains only identifiers that refer to these objects.

For continuous-process and assembly-line facilities, the major process engineering task is the design of the line, and that is completely out of scope for ERP systems. What ERP systems maintain for a product mix is the line configurations (identifiers) and equipment resources involved, staffing and maintenance requirements, the set of products output, the production rates and yields, and the materials requirements in terms of identification and classification, start-up quantities, prepositioning requirements, and feed rates.

For batch facilities, the major process engineering tasks are the materials selection, the routing (i.e., the sequence of work centers with particular setups), and the detailed specifications for operations within the work centers. The materials requirements, the yields, and the routings for products and product mixes are critical elements of the ERP planning information. The detailed work center operations are unimportant for planning, and all that is captured in the ERP core is the external references to them, the net staffing and time requirements, and the assigned costs of the work center usages.

For job shop facilities, the major process engineering tasks are materials selection, the routing, and the detailed specifications for operations within the work centers. The materials requirements and yields for specific products are critical elements of the ERP planning information. The routing is often captured as a sequence of work center operations—unit processes, each with its own equipment, staffing and time requirements, associated detail specification identifiers, and assigned cost. The detailed unit process specifications—operator instructions, setup instructions, equipment control programs—are kept in external systems.

No execution processes are directly supported by process specification management. All levels of resource planning are directly and indirectly supported by this information.

2.2.6. Maintenance Management

This function includes all information about the operational status and maintenance of equipment, vehicles, and facilities. *Operational status* refers to an availability state (in active service, ready, standby, in/awaiting maintenance, etc.), along with total time in service, time since last regular maintenance, and so on. The ERP system tracks maintenance schedules for the equipment and actual maintenance incidents, both preventive and remedial, and typically an attention list of things that may need inspection and refit. Any maintenance activities include both technical data (nature of fault, repair or change, parts installed, named services performed, etc.) and administrative data (authorization, execution team, date and time, etc.). In addition, this component tracks the schedules, labor, and work assignments of maintenance teams, external maintenance contracts and calls, and actual or assigned costs of maintenance activities.

In those organizations in which machine setup or line setup is performed by a general maintenance engineering group rather than a setup team attached to manufacturing operations directly, it is common to have such setups seen as part of the maintenance component rather than the manufacturing component. Similarly, operational aspects of major upgrades and rebuilds may be supported in the maintenance component of the ERP system. These are areas in which the behavior of ERP systems differs considerably.

This component supports sourcing, manufacturing, and delivery activities indirectly. Maintenance, per se, is purely a support activity.

Planning activities supported by maintenance management include all forms of capacity planning, from manufacturing order release and shipment dispatching (where immediate and expected availability of equipment are important) up to long-term capacity planning (where facility age and statistical availability are important).

2.2.7. Warehousing

This function deals with the information associated with the management of finished goods and spare parts after manufacture and before final delivery to the customer.

For products made-to-stock, this domain includes the management of multiple levels of distribution centers, including manufacturer-owned/leased centers, the manufacturer's share of concerns in customer-owned/leased centers, and contracted distribution services. The primary concerns are the management of space in the distribution centers and the management of the flow of product through the distribution centers. Thus, there are two major elements that are sometimes mixed together: the management of the distribution center resources (warehouse space, personnel, and shipping and receiving facilities) and the management of finished product over many locations, including manufacturing shipping areas, distribution centers per se, and cargo in-transport. Distribution center and product (family) includes tracking actual demand experience, projected demand and safety stocks, units on hand, in-flow and back-ordered, and units in manufacture that are earmarked for particular distribution centers. The primary object in product distribution tracking is the shipment because that is the unit of product management at the factory, in transportation, and through all distribution centers, except possibly the one nearest the final customers. For shipments, what is tracked is the product content, the ultimate recipient, the current location, and the associated delivery/shipping orders.

For certain products made-to-order (both made-on-demand and option-to-order), the distribution center approach is used because it facilitates planning and use of delivery resources and usually because a sizable part of the manufacturer's product line (such as spare parts) is made to stock. In these cases, the information managed is similar to the made-to-stock case, but actual demand and safety stock concerns are replaced by tracking specific customer orders. Customer orders are part of shipments up to the final distribution center, and final delivery to customer is tracked from there.

For most products made-to-order, the warehousing component manages information only for finished goods in factory holding areas awaiting shipment and shipments that are in transportation to the customer. In this case, each shipment is associated with a particular customer at creation, and it may be associated with one or more manufacturing orders even before manufacturing starts. For shipments, what is tracked is the product content, the customer order, the current location, and the associated delivery/shipping orders. In many make-to-order cases, the manufacturing holding areas are managed as manufacturing resources instead of warehousing resources, and the shipments are managed as part of order tracking, thus eliminating the warehousing component.

Execution activities supported by warehousing are the revenue-producing delivery of finished goods via distribution centers and the support activities of distribution center management. The primary planning activities supported by warehousing are distribution planning and distribution requirements planning.

2.2.8. Transportation

This function includes all aspects of movement of parts and finished goods among manufacturing facilities and distribution centers as well as final delivery to customers. It can be subdivided into the management of vehicle fleets, transportation service contracts, and shipping orders.

All manufacturers manage shipping orders—the decision to move shipments of parts and finished goods from one facility to another in the process of fulfilling customer orders. What is captured for the order is the associated shipments, the starting and ending locations, the means of transfer, the nominal pickup and delivery times, and the associated authorizations. The means of transfer can be by owned vehicles or transportation service contracts, or a combination.

For transportation service contracts, what is tracked is the contractual information, the shipping and housing orders placed under those contracts, the states of those orders and corresponding states of the internal shipping orders, and the shipments themselves. In addition, the system tracks other actions under the contract, including payment authorizations, change orders, delays, misdeliveries, damaged and misplaced goods, and shipments not accepted.

The management of vehicle fleets, for enterprises that have their own, entails the capture of maintenance and spare parts information and the capture of vehicle staffing, routes and schedules, and current orders, states, and locations. In general, the activities are the same as those of contract shipping organizations, but the manufacturer's transportation fleet has only one customer and usually a small and mainly predefined set of destinations.

Execution activities supported by transportation include movement of parts between manufacturing centers and movement of spare parts and finished products to customers and distribution centers, all of which are revenue producing. The supporting activities of managing transportation fleets and services are also supported.

Planning activities supported by transportation include delivery planning, transportation resource planning, and transportation route planning.

2.2.9. Human Resource Management

The human resource management (HRM) component includes the management of all information about the personnel of the manufacturing enterprise—current and former employees, retirees and other pensioners, employment candidates, and possibly on-site contractors and customer representatives. For employees, the information may include personal information, employment history, organizational placement and assignments, evaluations, achievements and awards, external representation roles, education, training and skills certification, security classifications and authorizations, wage/salary and compensation packages, pension and stock plan contributions, taxes, payroll deductions, work schedule, time and attendance, leave status and history, company insurance plans, bonding, and often medical and legal data. For contract personnel, some subset of this information is maintained (according to need), along with references to the contract arrangement and actions thereunder.

The HRM system is often also the repository of descriptive information about the organizational structure because it is closely related to employee titles, assignments, and supervisory relationships.

The execution activities supported by the HRM system are entirely support activities. They include the regular capture of leave, time, and attendance information and the regular preparation of data sets for payroll and other compensation actions and for certain government-required reports. They also include many diverse as-needed transactions, such as hiring and separation actions of various kinds, and all changes in any of the above information for individual personnel. But the HRM system supports *no* revenue-producing function directly, and it often plays only a peripheral role in strategic planning.

2.2.10. Finance Management and Accounting

This function includes the management of all information about the monies of the enterprise. The primary accounting elements are grouped under accounts payable (all financial obligations of the organization to its suppliers, contractors, and customers); accounts receivable (all financial obligations of customers, suppliers, and other debtors to this organization); and general ledger (the log of all real and apparent cash flows, including actual receipts and disbursements, internal funds transfers, and accrued changes in value). In reality, each of these is divided into multiple categories and accounts. While smaller organizations often do finance management under the heading *general ledger,* larger ones, and therefore ERP systems, usually separate the finance management concerns from general ledger transactions. They include fixed asset management (acquisition, improvement, amortization, depreciation of plants, facilities, and major equipment); financial asset management (cash accounts, negotiable instruments, interest-bearing instruments, investments, and beneficial interests [e.g., partnerships]); and debt management (capitalization, loans and other financing, and ''assignments of interest'' [e.g., licenses, royalties]).

The major enterprise execution activities supported by the finance management component are contracting, payroll, payment (of contractual obligations), invoicing, and receipt of payment. Payroll is a supporting activity, but payment and receipt are revenue producing.

The primary financial planning activities supported are investment planning, debt planning, and budget and cash flow planning and analysis.

2.3. Interaction Points

It is the intent of many ERP vendors to provide the information systems support for all the business operations of the manufacturing enterprise, and ERP systems have gone a long way in that direction. But there are still several areas in which the manufacturing organization is likely to have specialized software with which the ERP system must interface. The primary mission of the ERP system is to provide direct support to the primary operations activities (materials acquisition, manufacturing, product delivery) and to the planning and management functions for those operations. The software environment of a large manufacturing enterprise includes many other systems that support nonoperations business functions. This includes product planning and design, market planning and customer relations, and supply chain planning and development. Additionally, software that supports the detailed manufacturing processes and the control of equipment is so specialized and therefore so diverse that no ERP provider could possibly address all customer needs in this area.

On the other hand, the wealth of data managed within the ERP core, as well as its logical and physical infrastructure and the demand for those data in many of these related processes, open the door for integrating these other functions with the ERP system. This is the situation that leads to demands for open ERP interfaces.

Moreover, as the ERP market expands to medium-sized enterprises, the cost of the monolithic ERP system has proved too high for that market, leading ERP vendors to adopt a strategy using incremental ERP components for market penetration. This strategy in turn requires that each component system exhibit some pluggable interface by which it can interact with other component systems as they are acquired. While ERP vendors have found it necessary to document and maintain these interfaces, thus rendering them open in a limited sense, none of the vendors currently has an interest in interchangeable components or standard (i.e., public open) interfaces for them. But even among components there are ''forced'' ERP boundaries where the medium-sized enterprise has longer-standing enterprise support software (e.g., in human resources and finance). These also offer opportunities for standardization.

At the current time, the most significant ERP boundaries at which standard interfaces might be developed are depicted in Figure 3.

2.3.1. Contracts Management

Contractual relationships management deals with the information associated with managing the formal and legal relationships with suppliers and customers. It consists of tracking the contract development actions; maintaining points of contact for actions under the agreements; and tracking all formal transactions against the agreements—orders and changes, deliveries and completions, signoffs, invoices and payments, disputes and resolutions, and so on. ERP systems rarely support the document management function, tracking the contract document text through solicitation, offer, counteroffer, negotiation, agreement, amendments, replacement, and termination. They leave that to a legal or

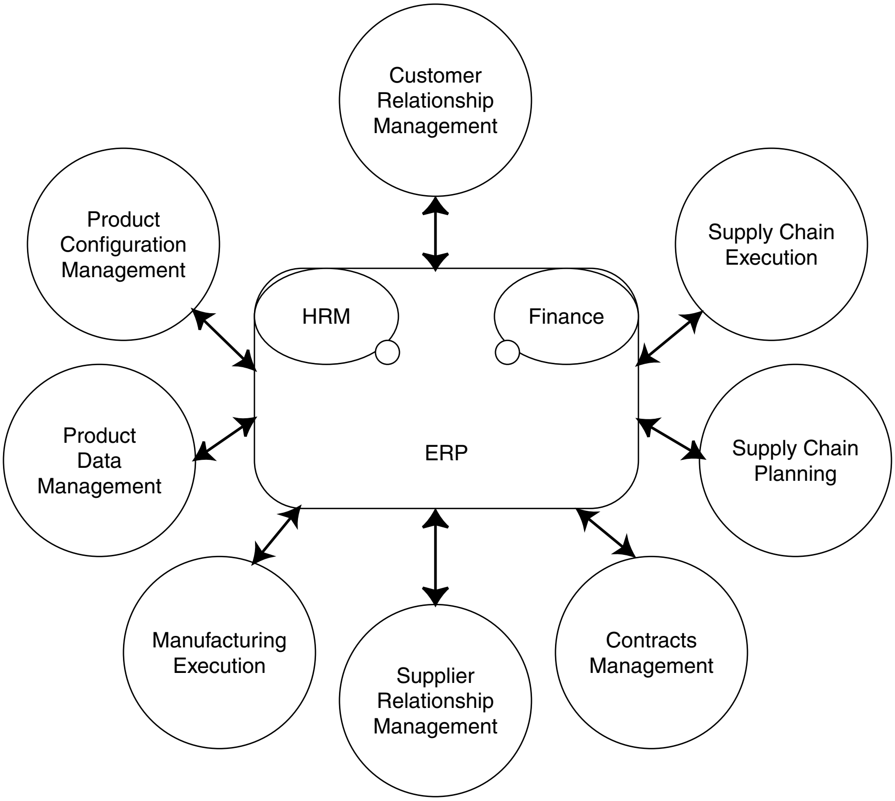

Customer Relationship Management
Product Configuration Management
Supply Chain Execution
HRM
Finance
ERP
Product Data Management
Supply Chain Planning
Manufacturing Execution
Supplier Relationship Management
Contracts Management

Figure 3 Opportunities for Standard ERP Interfaces.

business document management system and carry only references into that system where needed. They do routinely capture all transactions against agreements, but they often capture them in different places. Only a few centralize all these transactions under contracts management.

2.3.2. Supplier Relationship Management

This activity includes information on contractual arrangements with suppliers and points of contact, relationship history (orders, fulfillments, disputes, resolutions), business evaluations, and technical evaluations of products and capabilities, including certifications for specific materials. For specific materials (or product families), there are approved supplier lists that identify suppliers from whom the organization may order that material, often with preference ranking or ranking criteria.

2.3.3. Customer Relationship Management

As mentioned previously, the objective of CRM is to add value for customers through those processes that involve direct contact with customers before, during, and after sales. This function encompasses marketing and sales activities related to the identification and characterization of markets, the characterization of product opportunities within those markets that are consistent with the strategies and expertise of the enterprise, and the development of those markets into a customer base that generates recurring demand for the products of the enterprise. ERP systems may support demand planning activities that make projections for existing products with target volumes and time requirements as well as projections for new products or product modifications. ERP systems may also support customer inquiries as to product lines and company capabilities as well as inquiries and negotiations for alternative supply arrangements. As discussed in Section 2.2.3, ERP systems always support customer orders for existing products, customer order changes and cancellations, and inquiries about customer order status.

2.3.4. Product Configuration Management

A product configurator tracks the design of product options from desired features to manufacturing specifications. It captures product planning, pricing, and engineering decisions about option implementations and interoption relationships. The sales configurator component tells the sales staff what option combinations a customer can order and how to price them. The manufacturing configurator converts the option set on a customer order to a specification for bill of materials, station setup, prepositioning requirements, batching requirements, and process selections.

In "to-order" environments, an ERP system may include a product configuration function. A product configurator captures customer-specified product options in make-to-demand, option-to- order, and engineer-to-order environments. In those environments, product configurators connect the front office with the back office. In make-to-demand and option-to-order environments, product configurators link sales with manufacturing operations. In engineer-to-order environments, product configurators are a conduit between sales and engineering.

2.3.5. Product Data Management

While ERP systems are the principal repository for all operations data, they contain only fragments of product and process engineering data. One of the reasons for this is that the ERP core is short transactions with concise data units, while engineering data management requires support for long transactions with large data files. As ERP systems have grown over the last 10 years, PDM systems have grown rapidly as the product engineering information management system, especially in mechanical and electrical parts/product industries. The rise of collaborative product and process engineering in the automotive and aircraft industries has led to increasing capture of process engineering information in the PDM. Product engineering software tools, particularly CAD systems, are used to design tooling and other process-specific appliances, and these tools often have modules for generating detailed process specifications from the product definitions (e.g., exploded bills of materials, numerical control programs, photomasks). These tools expect to use the PDM as the repository for such data. Moreover, increased use of parts catalogs and contract engineering services has led to incorporation of a significant amount of part sourcing information in the PDM. Many ERP vendors are now entering the PDM product market, and the interface between PDM and ERP systems is becoming critical to major manufacturers and their software providers.

2.3.6. Supply Chain Execution

Although ERP systems may offer a one-system solution to supporting the operations of a given enterprise, one cannot expect that solution to extend beyond the walls. The information transactions supporting the materials flows from suppliers to the manufacturing enterprise and the flows of its products to its customers are becoming increasingly automated. Although basic electronic data interchange (EDI) transaction standards have been in existence for 20 years, they are not up to the task. They were intentionally made very flexible, which means that the basic structure is standard but most of the content requires specific agreements between trading partners. Moreover, they were made to support only open-order procurement and basic ordering agreements, while increased automation has changed much of the behavior of open-order procurement into automated catalogs and automated ordering and made several other supplier–customer operation techniques viable in the last several years. Thus, there is a need for ERP systems to operate, via upgraded e-commerce interfaces, with the ERP systems of the partners in the supply chain.

2.3.7. Supply Chain Planning

Until recently, ERP-supported planning algorithms focused on the internal behavior of the enterprise in managing its production and distribution, treating both customers and suppliers largely as black boxes with documented behaviors. The new concept is resource and market planning that focuses on the participation of the enterprise in various supply chains; thus, it can be effective only if it is part of a joint planning effort of multiple partners in those chains—the enterprise, its peers in the chain, its customers in the chain, and its suppliers. The joint planning activity must be supported by information interchanges between the decision-support software in (or linked to) the separate ERP systems of the partners. Algorithms for performing such joint planning are emerging, and first-generation software to support those algorithms is now available under the title advanced planning and scheduling (APS). Further development of these algorithms and interfaces is a necessary element of the future of ERP systems.

2.3.8. Manufacturing Execution

At some point, the gathering of manufacturing resource status information and work-in-process information becomes specific to the resource and the particular manufacturing task. It requires specialized systems to implement the specialized data-capturing technology and convert those data into resource planning, job planning, and tracking information. Further, particularly in discrete batch and job shop environments, the resource scheduling process itself becomes deeply involved with the details of the manufacturing tasks and the resource setups. Finally, a great deal of information gathered on the manufacturing floor is used to improve the process and product engineering as well as the characterization of machine capabilities, process yields, product quality, and so on. This domain is now loosely called manufacturing execution systems. Such systems deal with the details of data gathering, conversion, and assessment for specific purposes and industries. Future ERP systems must expect to interface with such companion factory management systems in a significant number of customer facilities. The need is to share resource planning information, resource status information, and order/job/lot release and status information.

2.3.9. Human Resource Management

Although it is considered a part of ERP, human resource management (HRM) systems have already penetrated the medium-sized enterprise market in many industries, of which manufacturing is only a subset. As ERP systems grow out of the manufacturing industry to other business areas, the need for interfacing with established HRM systems becomes apparent. And this makes standard interfaces to the HRM component more attractive to ERP and HRM vendors and customers.

2.3.10. Finance

In a similar way, most businesses, large and small, have long since built or acquired financial management software to support their business practices. Moreover, those practices and the related legal requirements vary significantly from country to country and, to a lesser extent, from state to state. For ERP systems, this means no "one size fits all" customers or even all business units of a single customer. Thus, interfaces to specialized and in-place financial software packages will continue to be a requirement.

2.4. Elements of ERP Implementations

The previous section addressed the functional aspects—the "what"—of ERP systems. This section deals with the "how" of ERP, specifically the generic software elements of current commercial ERP systems. It does not address the rationale used by a specific manufacturing enterprise to manage its own ERP selection, deployment, and upkeep. However, it covers briefly some of the tools for managing an ERP system. Additionally, it describes the generic architectures used by ERP vendors.

The basic elements of an ERP implementation include the core transaction system, packaged decision support applications provided by the ERP vendor, in-house or third-party extended applications, and a collection of tools for managing various aspects of the system (Figure 4). Each of these software elements reside in a computing environment that is typically distributed and possibly multiplatform.

2.4.1. Core ERP—Transactions

The ERP core consists of one or more transaction databases as well as transaction services. As described in Section 2.2, these services include capturing, executing, logging, retrieving, and monitoring transactions related to materials inventories, facilities/equipment, labor, and money.

2.4.2. Packaged Decision Support Applications

In addition to transaction management, ERP vendors provide decision support applications that offer varying degrees of function-specific data analysis. The terms *decision support application* and *decision support systems* (DSS) refer to software that performs function-specific data analysis irrespective of enterprise level. That is, decision support includes applications for supply, manufacturing, and distribution planning at the execution, tactical, and strategic levels of an enterprise. There is considerable variability among ERP vendors regarding the types of decision support applications that they include as part of their standard package or as add-ons. At one end of the spectrum, some vendors provide very specific solutions to niche industries based on characteristics of the operations environment (i.e., process and business nature) as well as enterprise size in terms of revenue. For example, an ERP vendor at one end of the spectrum might focus on assembly line/engineer-to-order environment with decision support functionality limited to manufacturing and distribution planning integrated with financials. At the other end of the spectrum, an ERP vendor could offer decision support functionality for supply, manufacturing, and distribution planning at all enterprise levels for a host of operations environments for enterprises with varying revenues. While such a vendor offers an array of decision support tools, one or more tactical level, function-specific applications (i.e., supply, manufacturing, distribution) are typically part of a standard ERP implementation (Figure 2). The others

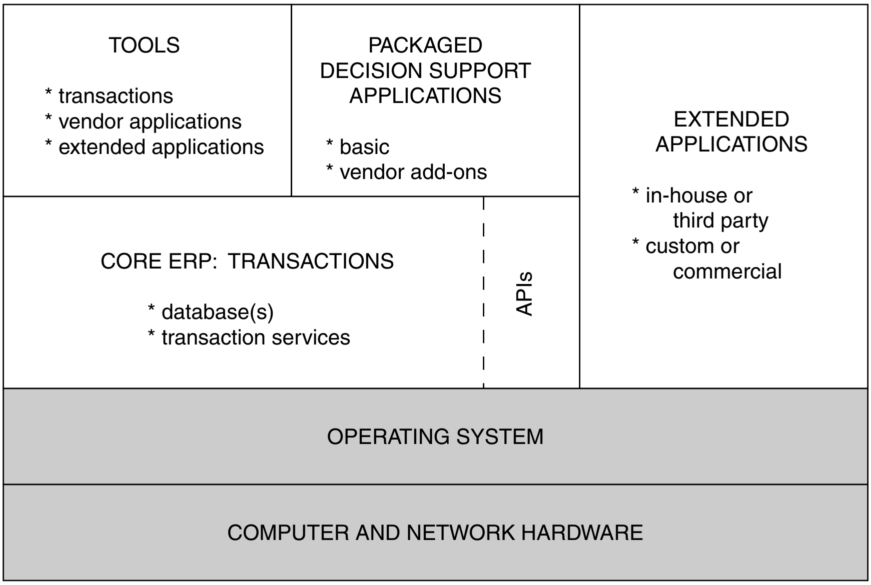

TOOLS
* transactions
* vendor applications
* extended applications
PACKAGED DECISION SUPPORT APPLICATIONS
* basic
* vendor add-ons
EXTENDED APPLICATIONS
* in-house or third party
* custom or commercial
CORE ERP: TRANSACTIONS
* database(s)
* transaction services
APIs
OPERATING SYSTEM
COMPUTER AND NETWORK HARDWARE

Figure 4 Basic Implementation Elements of ERP Systems.

tend to be considered add-ons. While a vendor may offer these add-ons, a manufacturer may opt to forgo the functionality they offer altogether or implement them as extended applications, either developed in-house or procured from third-party software vendors.

2.4.3. Extended Applications

The wealth of data in ERP systems allows many manufacturing enterprises to use ERP as an information backbone and attach extended applications to them. The motivation for such applications is that manufacturers typically see them as necessary to achieve differentiation from competitors. These applications may be developed in-house or by a third party. A third party may be a systems integrator who develops custom software, or it may be a vendor who develops specialized commercial software. As discussed in Section 2.3, these add-ons may provide additional functionality for customer and supplier relationship management, product data management, supply chain planning and execution, and human resource management. Regardless of the source, these applications must integrate with the ERP. Application programmer interfaces (APIs) are the common mechanism for integrating extended applications with the ERP backbone, and specifically the ERP core. Even among their partners and strategic allies (i.e., certain systems integrators and third-party software vendors), ERP vendors discourage the practice of integrating their standard applications with extended applications because of potential problems with upward compatibility. Because APIs to the ERP core have a longer expected lifetime, APIs are presently the most common approach to accessing information in the ERP system.

2.4.4. Tools

Given the enormous scope of the system, ERP can be a challenge to set up and manage. As such, ERP and third-party vendors provide software tools for handling various aspects of these complex systems. These tools generally fall into two categories: application configuration tools, which support the setup and operation of the ERP system itself, and enterprise application integration (EAI) tools, which support integration of the ERP system into the enterprise software environment.

2.4.4.1. ERP Configurators In an effort to decrease the amount of time and effort required to install, operate, and maintain an ERP system, vendors may provide a variety of application configuration tools. These tools vary in complexity and sophistication, but they all perform essentially the following tasks:

- *Define the computing topology:* Query the manufacturing administrator about the enterprise computing environment, the operating platforms, and the number, locations, and kinds of user workstations. Direct the setup program to install the appropriate versions of the ERP core and packaged modules on the server and workstation platforms.
- *Define the information base:* Query the manufacturing administrator about the specifics of their business information environment and assist in developing specialized information models (based on the vendor's generic ERP models) to support those business specifics. Automatically configure the ERP databases and transaction formats accordingly.
- *Define tasks and workflows:* Query the manufacturing administrator about the specifics of their business processes, map individual tasks to users across organizations, applications, and systems, and assist in developing specialized workflow models and decision support models. Then automatically configure the ERP workflow/task models, decision support invocations, and report formats.
- *Define security and control requirements:* Query the manufacturing administrator for user classifications by task responsibilities and authorizations. Define the user privileges, activity logging requirements, and security controls.

2.4.4.2. Enterprise Application Integration Unlike application configuration tools, which are part of an ERP vendor's suite and centered on the ERP system as the integrator of business processes, enterprise application integration (EAI) tools are actually a non-ERP-specific category of software tools whose primary function is to support business processes by linking up distinct software systems in the overall enterprise computing environment. As such, they view ERP as one part of a manufacturer's entire IT solution. But many of these tools come with predefined interfaces to specific ERP systems (some are provided by ERP vendors) and are designed primarily to link other software systems and applications with the ERP system.

This is a relatively new and fragmented class of software, employing a number of techniques of varying sophistication and providing quite variable capabilities. One view of EAI is as ''a selection of technologies to address a number of applications integration problems'' (Gold-Bernstein 1999). In this perspective, EAI technologies include platform integration solutions (messaging, message queueing, publish-and-subscribe, and object request brokers), message brokers (translations and transformation, intelligent routing, and application adapters), some graphical user interface (GUI) tools to define routing and mapping rules, and process automation and workflow.

A simplified view of a typical EAI architecture is provided in Figure 5. The EAI package consists of a number of application-specific adapters, each of which is capable of extracting data from and providing data to a particular application software package in some form convenient to that application. The adapter may also communicate directly with a central EAI engine that moves data sets between the adapters in some message form. In many cases the message is just a file, while in some cases the sole function of the adapter is to command the application to input or output a particular file. In some architectures, the adapter is responsible for converting the application information to or from a common interchange model and format used by the EAI package. In others, each adapter produces information in a convenient form and the engine is responsible for the conversion of the messages or files between the adapters, using a common reference model or an application-to-application-specific translation, or both.

Most EAI packages are designed primarily to solve intraenterprise communication problems. Many provide some mechanisms for Web-based access as a means of interaction with customers and suppliers.

2.5. ERP Architectures

For purposes of illustration, the ERP community refers to tiers when describing the general logical architectures of ERP systems. While the notion of tiers generally implies a hierarchy, such is not the case with tiers in present-day ERP architectures. Advances in distributed computing technologies enable more open communication among components. Instead, these tiers are the basic types of logical elements within an ERP implementation and thus provide a means for describing various ERP execution scenarios.

Figure 6 illustrates the five tiers of ERP architectures: core/internal (often called data), application, user interface, remote application, and remote user interface. A common intraenterprise scenario involves the data, application, and user interface tiers. This scenario does not preclude application-to-application interaction. Similarly, common interenterprise scenarios include an internal user or application requesting information from an external application or an external user or application requesting information from an internal application. The Internet is the conduit through which these internal/external exchanges occur. Additionally, in many ERP systems web browsers have emerged as the platform for both local and remote user interfaces.

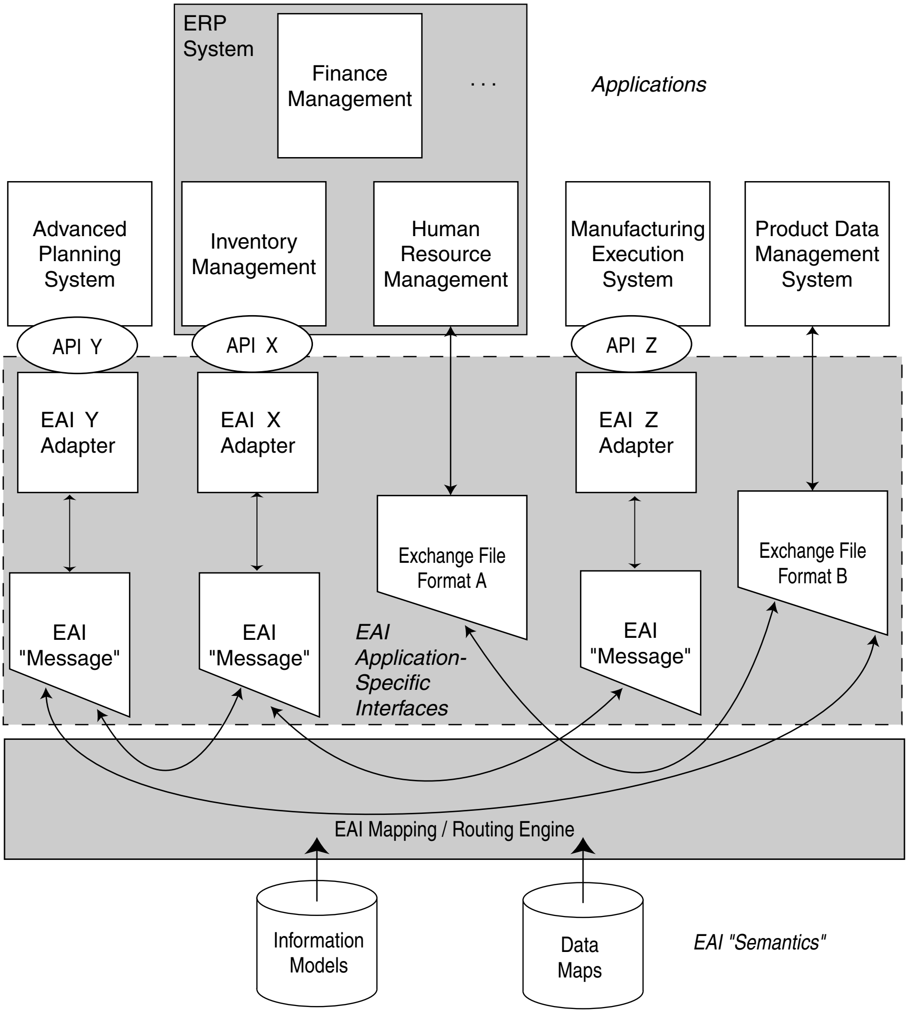

ERP System
Finance Management
. . .
Applications
Advanced Planning System
Inventory Management
Human Resource Management
Manufacturing Execution System
Product Data Management System
API Y
API X
API Z
EAI Y Adapter
EAI X Adapter
EAI Z Adapter
Exchange File Format A
Exchange File Format B
EAI "Message"
EAI "Message"
EAI "Message"
EAI Application-Specific Interfaces
EAI Mapping / Routing Engine
Information Models
Data Maps
EAI "Semantics"

Figure 5 EAI Systems Architectures.

2.6. ERP and the Internet

The emergence of the Internet as the primary conduit for exchange of ERP-managed information among trading partners has spawned the term *Internet-based ERP.* This term does not convey a single concept but may refer to any of the following:

- Internal user-to-ERP-system interfaces based on web browsers
- External user-to-orders interfaces based on web browsers
- Interfaces between decision support software agents of different companies that support supply chain operations
- Interfaces between decision support software agents of different companies that support joint supply chain planning

The increasingly commercial nature of the Internet and the development of communication exchange standards have had significant impact on ERP systems. In describing this impact, the term *tier* is used in the context of ERP architecture (Section 2.5).

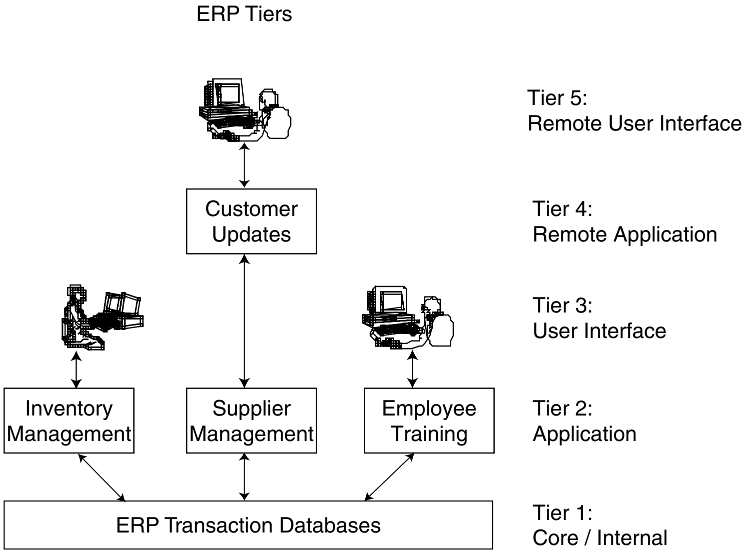


Figure 6 Tiers in ERP Architectures.

2.6.1. *Internal User-to-ERP Interfaces*

Also called application hosting, this approach employs user interfaces (Tier 3) based on Internet/Web technologies, notably Java, the Hypertext Markup Language (HTML), and the Extensible Markup Language (XML). In this scenario, ERP vendors as well as systems integrators take on a new role as application service providers (ASPs). This is an important change in the product architecture of ERP, in that the ERP vendor no longer has to maintain the dedicated workstations or the workstation software for users so connected. It also means that the ERP vendor cannot price that part of its services by station but instead uses transaction volume metrics. This differs from other impacts by being purely intranet-based (i.e., all such connections are from within the enterprise and thus subject to alternative controls and security mechanisms).

2.6.2. *External User-to-ERP Interfaces*

In this scenario, a remote user (Tier 5), via an interface based on Web technologies, accesses application service modules (Tier 2). This is a widely used approach among ERP vendors and CRM vendors. It differs from the first by having many electronic business technical concerns, including access authorization and data filtering, secure sockets, and contract and payment references. The critical question here is whether the functions supported include order entry or just order tracking and how the Web services are related to internal orders management. The CRM system may act as a staging system with no direct connect between the Web interface and the ERP system itself.

2.6.3. *B2B Supply Chain Operations Interfaces*

This scenario involves communication between a remote application and a local application (i.e., Tier 4 and Tier 2). The actual exchange is usually based on EDI or some XML message suites,* using file transfers, electronic mail, or some proprietary messaging technology to convey the messages. This scenario is significantly different from the above in that neither of the communicating agents is a user with a browser. Rather, this is communication between software agents (decision support modules) logging shipment packaging, release, transportation, receipt, inspection, acceptance, and possibly payment on their respective ERP systems (with some separate Tier 3 user oversight at both ends). Special cases of this include vendor-managed inventory and consignment management, which illustrate the use of the Internet in direct support of an operations process.

---

*E.g., RosettaNet (RosettaNet 2000), CommerceNet (CommerceNet 2000), Electronic Business XML (Electronic Business XML 2000), Open Applications Group Interface Specification (OAGIS) (Open Applications Group 2000).

2.6.4. Joint Supply Planning (Advanced Planning and Scheduling) Interfaces

This scenario also involves communication between a remote application and a local application (i.e.,Tier 4 and Tier 2) with the exchange based on a common proprietary product or perhaps an XML message suite. Again the communicating agents are software systems, not users with browsers, but their domain of concern is advanced planning (materials resource planning, distribution resource planning, etc.) and not shipment tracking. There are very few vendors or standards activities in this yet because this represents a major change in business process. This scenario illustrates use of the Internet in direct support of a tactical planning process.

3. AN EXTERNAL VIEW OF ERP SYSTEMS

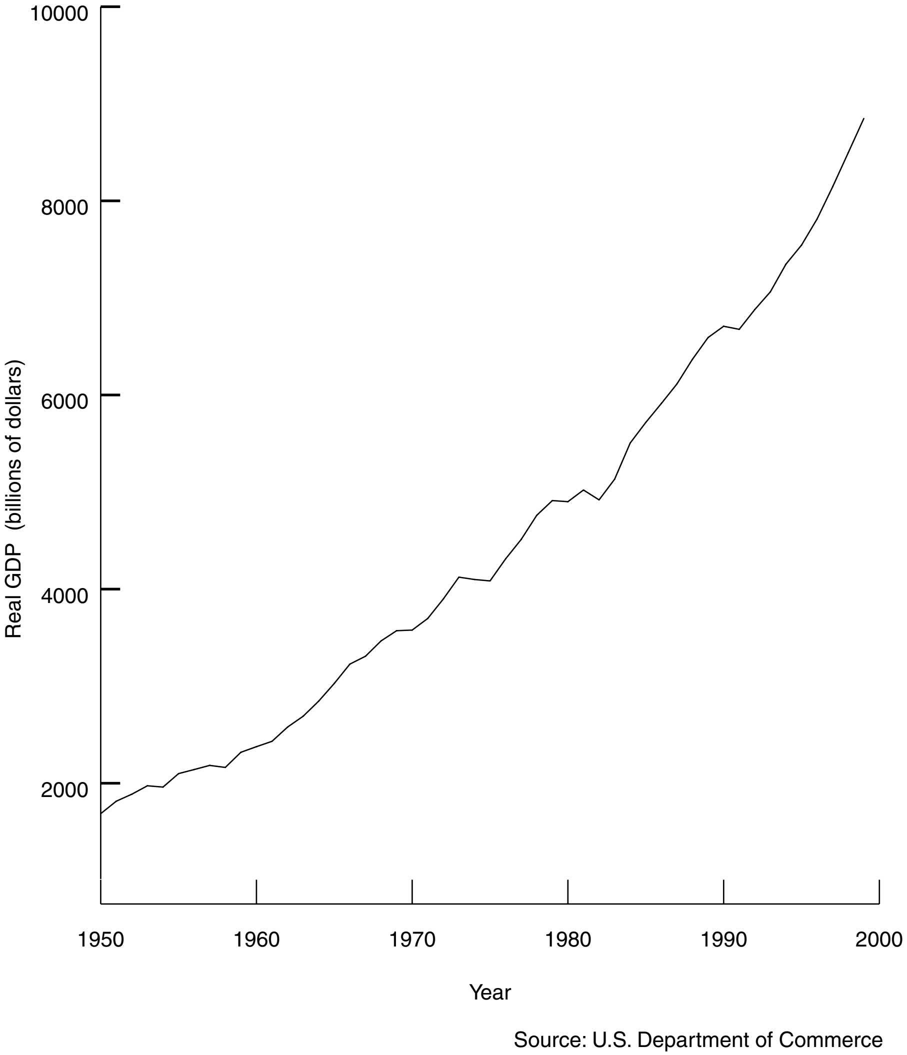
This section illustrates ERP systems in a larger context. Section 3.1 provides some current thinking on the apparent macroeconomic impacts of IT, with a yet-to-be-proven hypothesis specific to ERP systems. Section 3.2 describes the relationship of ERP specifically with respect to electronic commerce, supply chains, and individual manufacturing enterprises.

3.1. ERP and the Economy

Much has been written and said about the emerging digital economy, the information economy, and the New Economy. It is the authors' view that the information economy must support and coexist with the industrial economy because certain fundamental needs of mankind are physical. However, while these economies coexist, it is clear that the sources of wealth generation have changed and will continue to change in fundamental ways. In the last 50 years, the U.S. gross domestic product (GDP), adjusted for inflation, has grown more than 500% (Figure 7). While each major industry has grown considerably during that period, they have not grown identically. Confirming that the economy is a dynamic system, the gross product by industry as a percentage of GDP (GPSHR) saw significant changes in the last half of the 20th century (Figure 8). GPSHR is an indication of an industry's contribution (or its value added) to the nation's overall wealth. While most goods-based industries appear to move towards a kind of economic equilibrium, non-goods industries have seen tremendous growth. The interesting aspect of ERP systems is that they contribute to both goods- and non-goods-based industries in significant ways. In fact, for manufacturers, ERP plays a critical role in extending the existing industrial economy to the emerging information economy. In the information economy, ERP accounts for a significant portion of "business applications" sales, not to mention the wealth generated by third parties for procurement, implementation, integration, and consulting. While these are important, in this chapter we focus on the use of ERP in manufacturing. Therefore, the following sections describe how ERP and related information technologies appear to impact the goods-producing sectors of the current U.S. economy.

Macroeconomics, the study of the overall performance of an economy, is a continually evolving discipline. Still, while economists debate both basic and detailed macroeconomic theory, consensus exists on three major variables: output, employment, and prices (Samuelson and Nordhaus 1998). The primary metric of aggregate output is the gross domestic product (GDP), which is a composite of personal consumption expenditures, gross private domestic investment, net exports of goods and services, and government consumption expenditures and gross investment. The metric for employment is the unemployment rate. The metric for prices is inflation. While these variables are distinct, most macroeconomic theories recognize interactions among them. It appears that the use of information technology may be changing economic theorists' understanding of the interactions among economic variables, particularly for predicting gross domestic product, unemployment, and inflation. It is important to gain a deeper understanding of these changes because their impact would affect government policy decisions, particularly those involving monetary policy and fiscal policy.

In the current record domestic economic expansion, real output continues to increase at a brisk pace, unemployment remains near lows not seen since 1970, and underlying inflation trends are subdued. During this period, inflation has been routinely overpredicted while real output has been underpredicted. Conventional economic theory asserts that as real output increases and unemployment decreases, significant pressures mount and force price increases. Yet, in this economic expansion, inflation remains in check, *apparently* due in part to IT-enabled growth in labor productivity (Greenspan 1999). In the early 1990s, the labor productivity growth rate averaged less than 1% annually. In 1998, that rate had grown to approximately 3%. So what has happened in this decade? In the last 10 years, information technology has enabled companies, most notably manufacturers, to change the way they do business with access to better information (often in real time) and better decision-support technologies. These changes have improved the way manufacturers respond to market wants (i.e., for products) and market demands (wants for products backed by buying power). ERP systems play a significant part in satisfying the latter by enabling better planning and execution of an integrated order fulfillment process. In short, ERP software enables these improvements by providing decision makers in an enterprise with very accurate information about the current state of the enterprise.

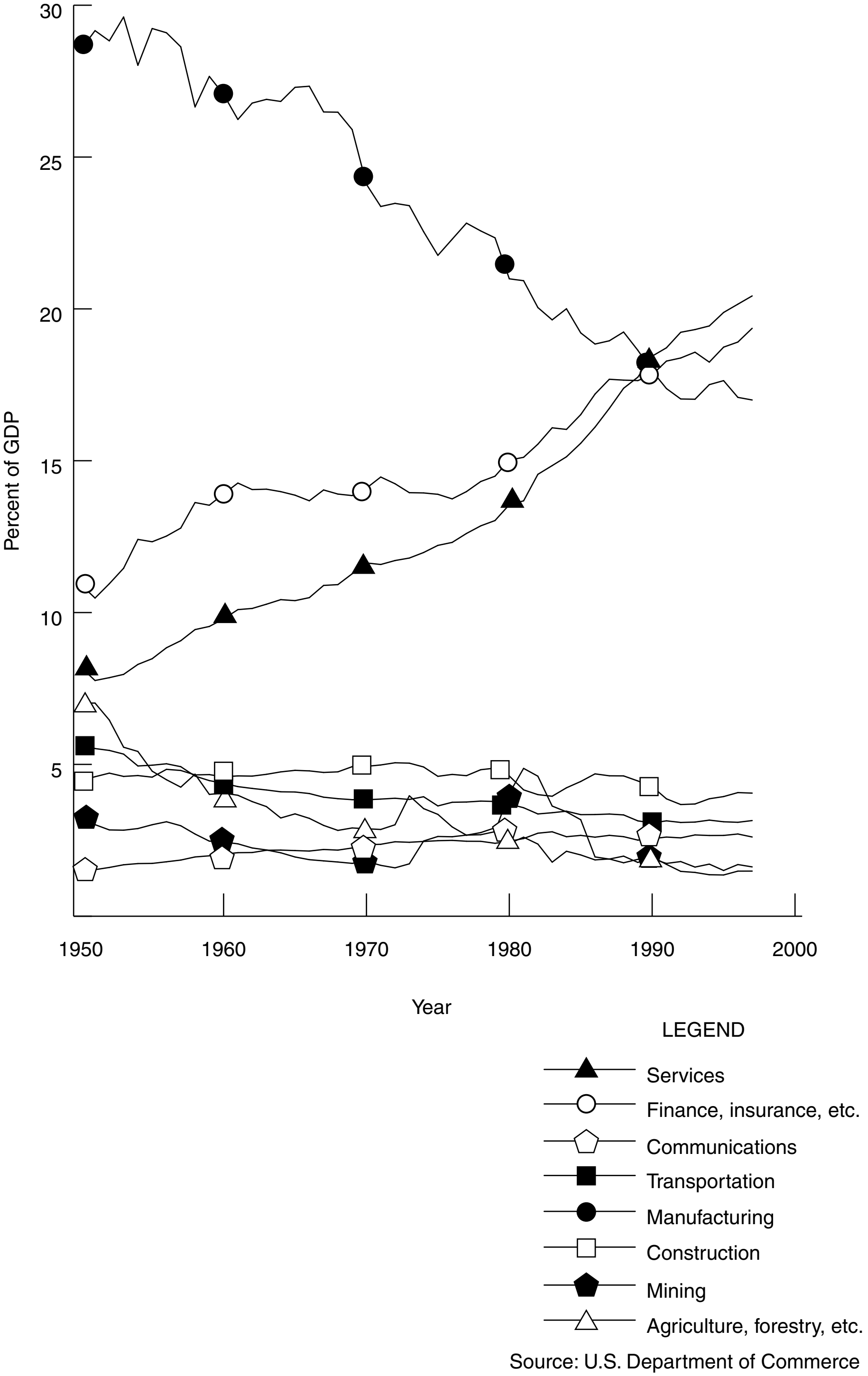


Source: U.S. Department of Commerce

Figure 7 GDP Growth, 1950–1999.

Moreover, an increasing number of manufacturers have direct access to demand information from their customers and resource information from their suppliers. In many cases, the customer's demand information and the supplier's resource information originate in their respective ERP systems. Just as these data are more accurate for the customer/manufacturer/supplier enterprise, so too is the resulting information flowing up and down the supply chain. For the manufacturer between them, this information enables decision makers to base decisions on accurate external information, as well as accurate internal information. The following remarks by Federal Reserve Chairman Alan Greenspan perhaps best capture the essence of this phenomenon (Greenspan 1999):

> As this century comes to an end, the defining characteristic of the current wave of technology is the role of information. Prior to the advent of what has become a veritable avalanche of IT innovations, most of twentieth century business decision-making had been hampered by limited information. Owing to the paucity of timely knowledge of customers' needs and of the location of inventories and materials flows throughout complex production systems, businesses required substantial programmed redundancies to function effectively.



Source: U.S. Department of Commerce

Figure 8 Gross Product Originating by Industry Share of GDP, 1950–1997.

> Doubling up on materials and people was essential as backup to the inevitable misjudgments of the real-time state of play in a company. Judgments were made from information that was hours, days, or even weeks old. Accordingly, production planning required adequate, but costly, inventory safety stocks, and backup teams of people to maintain quality control and for emergency response to the unanticipated and the misjudged.
>
> Large remnants of information void, of course, still persist and forecasts of future events on which all business decisions ultimately depend are still inevitably uncertain. But the recent years' remarkable surge in the availability of real-time information has enabled business management to remove large swaths of inventory safety stocks and work redundancies. . . .
>
> Moreover, information access in real-time resulting from processes such as, for example, checkout counter bar code scanning and satellite location of trucks, fostered mark reductions in delivery lead times on all sorts of goods, from books to capital equipment. This, in turn, has reduced the overall capital structure required to turn out our goods and services, and, as a consequence, has apparently added to growth of multifactor productivity, and thus to labor productivity acceleration.
>
> Intermediate production and distribution processes, so essential when information and quality control were poor, are being bypassed and eventually eliminated.

ERP systems, in part, enable those activities described by Chairman Greenspan by providing two core functions: transaction management and near-term decision support. The objective of transaction management is to track the effect of execution activities on inventories, resources, and orders, while the objective of intermediate-term decision support is to use that and other information to generate accurate plans for sourcing, production, and delivery.

3.2. ERP, Supply Chains, and Electronic Commerce

ERP systems do not provide a complete solution for supply chain management (SCM) or electronic commerce. However, especially for manufacturers, the functionality provided by ERP is a necessary (although by no means sufficient) element of both SCM and, therefore, electronic commerce. This section provides definitions of electronic commerce and SCM and explains the relationships among these concepts and ERP.

3.2.1. Electronic Commerce

Ask 10 people to define electronic commerce and you'll likely get 10 different definitions that reflect the particular biases of those asked. Recognizing the existence of these broad interpretations, this chapter uses an inclusive definition developed by the Gartner Group (Terhune 1999):

> Electronic commerce is a dynamic set of technologies, integrated applications, and multi-enterprise business processes that link enterprises together.

The concept of electronic commerce centers on the use of technology, and those technologies tend to be infrastructural in nature. Some obvious current technological issues include network-related subjects (the Internet, the Web, and extranets), security, interoperability of applications software, and the exchange of application-based information within and across enterprises. These integrated applications, which collectively comprise an enterprise's electronic business or (e-business) environment, include EDI software, supply chain management, ERP, customer relationship management, and enterprise application integration software. Issues in multienterprise business processes revolve around the different interactions that occur across enterprises. Electronic commerce changes, in part or in whole, the mode of these interactions from paper and voice to networked, digital information flows. The nature of these interactions, and the relationships among trading partners in general, range from coercive to collaborative depending on the general structure and practices of a given industry, the goods and/or services produced, and the impact of information technology on the distribution channels through which those goods and services flow. Recognizing and understanding these distinctions are critical for evaluating the existing and potential impact of electronic commerce across industries.

Manufacturing industries face particular challenges in realizing the benefits of electronic commerce because of the coupling of goods and information and the coordination required across those domains. Information-intensive industries (e.g., banking, traveling, advertising, entertainment) experience the effects of electronic commerce before materials-intensive industries such as manufacturing, construction, and agriculture. In information-intensive industries, products and services lend themselves to the technology. In many of these cases, electronic commerce technology simply becomes a new distribution channel for the information product or service. The manufacturing situation is significantly more complex because it requires reconciliation of goods *and* information. The objective is not to develop new distribution channels per se (the information network does not move goods); the objective is to improve the flow of goods by using the information technology to improve the business practices. Optimizing the flow of goods through distribution channels is one particular type of electronic commerce improvement. By so doing, trading partners can root out the inefficiencies within channels and make them more adaptive to changes in the market. It is precisely those inefficiencies and adaptability that are the foci of SCM.

3.2.2. Supply Chain Management

Supply chain management is one of several electronic commerce activities. Like electronic commerce, SCM has acquired buzzword status. Nonetheless, a common understanding of SCM has emerged through the work of industry groups such as the Supply Chain Council (SCC), the Council on Logistics Management (CLM), and the APICS organization, as well as academia.

> SCM is the overall process of managing the flow of goods, services, and information among trading partners with the common goal of satisfying the end customer. Furthermore, it is a set of integrated business processes for planning, organizing, executing, and measuring procurement, production, and delivery activities both independently and collectively among trading partners.

It is important to note a critical distinction here, especially since that distinction is not explicit in the terminology. While *SCM* is often used synonymously with *supply chain integration* (SCI), the two terms have connotations of differing scopes. As stated previously, SCM focuses on planning and executing trading partner interactions of an operations nature—that is, the flow of goods in raw, intermediate, or finished form. SCI is broader; it includes planning and executing interactions of any kind among trading partners, and it refers in particular to the development of cooperating technologies, business processes, and organizational structures.

The operations-specific objectives of SCM can only be achieved with timely and accurate information about expected and real demand as well as expected and real supply. A manufacturer must analyze information on supply and demand along with information about the state of the manufacturing enterprise. With the transaction management and basic decision support capabilities described earlier, ERP provides the manufacturer with the mechanisms to monitor the current and near-term states of its enterprise. As depicted in the synchronized, multilevel, multifacility supply chain planning hierarchy of Figure 2, ERP provides the foundation for supply chain management activities. Section 1 described the various levels of the hierarchy, and Section 2 described the details of transaction management and decision support.

4. ERP CHALLENGES AND OPPORTUNITIES

While evidence suggests that ERP systems have brought about profound positive economic effects by eliminating inefficiencies, there is still room for substantial improvement. The technical challenges and opportunities for ERP systems arise from how well these systems satisfy the changing business requirements of the enterprises that use them. Case studies in the literature highlight the unresolved inflexibility of ERP systems (Davenport 1998; Kumar and Van Hillegersberg 2000). The monolithic nature of these systems often hampers, or even prevents, manufacturers from responding to changes in their markets. Those market changes relate to ERP systems in two important ways: the suitability of decision support applications to an enterprise's business environment and the degree of interoperability among applications both within and among enterprises.

These two issues lead to three specific types of activities to improving ERP interoperability: technology development, standards development, and business/technology coordination.

4.1. Research and Technology Development

4.1.1. Decision Support Algorithm Development

Manufacturing decision support systems (DSSs), especially those that aid in planning and scheduling resources and assets, owe their advancement to progress in a number of information technologies, particularly computational ones. The early versions of these applications—namely materials requirements planning (MRP) and manufacturing resource planning (MRP II)—assumed no limitations on materials, capacity, and the other variables that typically constrain manufacturing operations (e.g., on-time delivery, work-in-process, customer priority). The emergence of constraint-based computational models that represent real-world conditions has enabled manufacturers to better balance supply and demand. In spite of the improvement in DSSs, significant opportunities still exist. Different models apply to different business environments. With the rapid pace of change, the commercial viability of the Internet, and the push to "go global," there are new variables to examine and new models to develop. Instead of convergence to any single optimization function, there will likely be specialization to classes of functions. That specialization has begun as illustrated in Figure 2 and described in Section 3.2 Continued specialization (not simply customization) is likely and necessary. Those classes will account for the variety of ways that manufacturing enterprises choose to differentiate their operations from those of their competitors. While cost reduction has been the focus of the current generation of DSSs, models that also address revenue enhancement and response time are expected to be the focus of future generations.

4.1.2. Component Decomposition Analysis

As decision support technology evolves, so too does ERP system functionality. While advances in software functionality affect an enterprise's continuous IT improvement strategy, they do not comprise all of that strategy. Migration is another critical aspect of continuous IT improvement. That is, replacing and upgrading an existing system—in whole or in part—must be possible for an enterprise to move from one functional state to another. In general, ERP vendors have focused on providing functionality at the expense of replaceability and upgradability (Sprott and Wilke 1998). Consequently, lock-in effects are major concerns for manufacturing and nonmanufacturing enterprises alike.

Possessing the concepts of services and encapsulation, components are touted as the solution to this problem. A component delivers functionality to a user or another software entity through one or more well-defined interfaces. Encapsulation means that a component is a separate entity, thus making it easier with which to manage, upgrade, and communicate. Yet components alone do not guarantee interoperability, especially in complex e-business and electronic commerce environments. For ERP systems to be interoperable, there must be widespread agreement on the services that ERP applications provide and on the boundaries of encapsulation. Numerous approaches to defining services and interfaces exist, including vendor-specific conventions, intraenterprise conventions, industry-specific standards, and technology-specific standards.

These approaches all lack broad perspective and thus do not meet the challenges of enabling ERP interoperability in e-business and electronic commerce environments. To achieve interoperability with the vendor-specific conventions, a single ERP vendor must dominate not just the ERP system market but the markets of all the other systems that (need to) interoperate with ERP. Intraenterprise conventions work up to the bounds of the enterprise, as long as there are no major internal changes (such as mergers or acquisitions). The approach based on industry-specific standards fails to realize that the focal point of operations is business process and that information varies according to the characteristics of those processes. Industries obviously are defined by product, not business process, and commonality of product does not translate into commonality of business process. Operations-wise, what makes sense for one supply chain member (e.g., an original equipment manufacturer [OEM]) does not necessarily make sense for another (e.g., a lower-tier supplier). The fourth approach, technology-specific standards, tends to yield limited solutions because it focuses on syntax and not semantics. Without agreement on the meaning of the information to be exchanged and the function that the information supports, the challenges of reconciliation persist. Many technology-focused efforts fail to perform sufficient domain-specific information requirements analysis.

Because electronic commerce dramatically changes the nature of power and control in supply chains, component decomposition analysis must address the three emerging models of business-to-business (B2B) markets. The first approach is an OEM-controlled model such as those recently announced by U.S. automakers and defense contractors (Stoughton 2000). The second is supplier-controlled model such as those in metals, chemical, and paper industries. The third is an open model that allows control to be shared among all supply chain players. Many think that the open model will prevail in the long run. However, industries will not collectively reach that point at the same rate or at the same time. To realize this model, it is necessary to look beyond industry-specific exchange of operations data. It is necessary to analyze business processes and characterize them at appropriate levels of detail. These characterizations would highlight the different kinds of components and thus the information that needs to be exchanged in different scenarios.

4.2. Standards Development

Standards play an important role in achieving interoperability. With respect to ERP systems, opportunities exist for establishing standard interface specifications with other manufacturing applications.

4.2.1. ERP–PDM Interfaces

As discussed in Section 2.3.5, there is an obvious interaction point between ERP and PDM systems. Thus, there is a need for interfaces between the two systems to share separately captured engineering and sourcing specifications. In the longer run, the goal should be to have PDM systems capture all the product and process engineering specifications and to extract resource requirements information for use in ERP-supported planning activities. Conversely, sourcing information, including contract engineering services, should be captured in the ERP system. To do this, one needs seamless interactions as seen by the engineering and operations users.

4.2.2. ERP–MES Interfaces

As presented in Section 2.3.8, future ERP systems must expect to interface with such companion factory management systems in a significant number of customer facilities. There is the need to share resource planning information, resource status information, order/job/lot release, and status information. However, developing such interfaces is not a straightforward exercise. The separation of responsibilities and the information to be exchanged vary according to many factors both at the enterprise level and the plant level. Prestandardization work is necessary to identify and understand those factors.

4.2.3. Supply Chain Operations Interfaces

Supply chain information flows between the ERP systems of two trading partners have been the subject of standardization activities for 20 years, with a spate of new ones created by Internet commerce opportunities. Most of these changes concentrate on basic ordering agreement and open procurement mechanisms. Requirements analysis of this information is necessary before actual standards activities commence. As the business practices for new trading partner relationships become more stable, standards for interchanges supporting those practices will also be needed. Changes in business operations practices as well as in decision support systems have changed the information that trading partners need to exchange. This includes shared auctions, supply schedule, vendor-managed inventory, and other operational arrangements, but the most significant new area is in joint supply chain planning activities (i.e., advanced planning and scheduling).

4.3. Establishing Context, Coordination, and Coherence for Achieving Interoperability

Several developments in the past decade have combined to extend the locus of information technology from research labs to boardrooms. The commercialization of information technology, the pervasiveness of the Internet, and the relatively low barriers to market entry for new IT companies and their technologies all serve to create an environment of rapid growth and change. The ERP arena, and electronic commerce in general, suffer from a proliferation of noncooperative standards activities, each aimed at creating interoperability among a handful of manufacturers with specific software tools and business practices. There is an imperative to reduce competition among standards efforts and increase cooperation.

Understanding the complex environment that surrounds ERP and other e-business and electronic commerce applications is a critical challenge to achieving interoperability. Topsight is a requirement for meeting this challenge. The objective of topsight is to establish context, coordination, and coherence among those many activities that seek standards-based interoperability among manufacturing applications. While the hurdles that exist in the current environment are considerable, there is significant need—as well as potential benefit—for an industry-led, multidisciplinary, and perhaps government-facilitated effort to provide direction for the development and promulgation of ERP and related standards.

The notion of topsight for improving interoperability among specific applications is not new. The Black Forest Group, a diverse assembly of industry leaders, launched the Workflow Management Coalition (WfMC), which has produced a suite of specifications for improving interoperability among workflow management systems. A similar ERP-focused standards strategy effort would strive to understand better the diversity of operations and operations planning in order to improve interoperability among ERP and related systems.

For a topsight effort to succeed in an arena as broad as ERP, particularly one that is standards-based, there must be a cross-representation of consumers, complementors, incumbents, and innovators (Shapiro and Varian 1999).

As consumers of ERP systems, manufacturers and their trading partners face the risk of being stranded when their systems do not interoperate. The lack of interoperability in manufacturing supply chains can create significant costs (Brunnermeier and Martin 1999), and those costs tend to be hidden. More accurate cost structures must be developed for information goods, particularly for buy-configure-build software applications. Unlike off-the-shelf software applications, ERP systems are more like traditional assets, in the business sense, with capital costs and ongoing operational costs.

Complementors are those who sell products or services that complement ERP systems. Given the role that ERP plays in electronic commerce, this group is very large. It includes both software vendors and systems integrators and management consultants. Some of the software that complements ERP was discussed previously in Section 2.3. Others include additional categories of software necessary for achieving e-business: EDI/e-commerce, business intelligence, knowledge management, and collaboration technologies (Taylor 1999).

Incumbents are the established ERP vendors, and they make up a market that is very dynamic and diverse (Table 2). Achieving consensus of any kind with such a diverse market is a considerable challenge. To achieve ERP interoperability requires, among others, deeper understanding of common elements. That understanding can be acquired by detailed functional and information analysis of the ERP systems.

The notion of innovators in the standards process focuses on those who collectively develop new technology. While many individual technology development activities associated with ERP and electronic commerce might be considered innovative, there have been few explicit collective development

TABLE 2 The Current State of ERP Market

ERP Vendor Categories	Total Annual Revenue	Cross-Functional Scope	Manufacturing Environment Scope	Targeted Industry Scope
Tier 1 ERP Vendors	more than $2 billion	broad, manufacturing, materials, human resource, financial SCM and CRM	broad, continuous, assembly line, discrete batch, job shop, construction	broad, larger and mid-market manufacturers across numerous industries
Tier 2 ERP Vendors	between $250 million and $2 billion	moderate, but expanding beyond one or two functional areas, some adding SCM and/or CRM	variable, some support one or two manufacturing environments while others support more	moderate, larger and mid-market manufacturers across several industries
Tier 3 ERP Vendors	less than $250 million	narrow, fill void of Tier 1 and Tier 2	narrow, typically supports one type of manufacturing environment	narrow, smaller manufacturers in niche industries

efforts. Most ERP vendor partnerships tend to be confined to making existing products work together through application programmer interfaces. They generally do not involve the joint creation of new (and often complementary) technologies. The dynamics that compel banks and credit card companies to pursue smart card development do not exist in the ERP environment. However, there are others who meet this definition of innovator. Academia is one group of innovators whose relationship with ERP vendors tends not to make headlines. Still, many of the technologies in today's ERP systems have academic roots. The perspective of university researchers across numerous disciplines would add significant value to a topsight effort.

5. CONCLUSIONS

ERP is a very complex subject. It marries technology, business practices, and organizational structures. ERP systems are commercially developed software applications that integrate a vast array of activities and information necessary to support business operations and operations planning at the tactical level. ERP is software and not a business process or a set of business processes. However, as software, it enables better execution of certain processes. Although often presented as a single package, an ERP system is an envelope around numerous applications and related information. For manufacturers, those applications typically support the operations processes of materials sourcing, manufacturing planning, and product distribution. To its end users, an individual application of an ERP system may appear seamless; however, to those who procure, implement, and/or maintain ERP systems, they are complex software systems that require varying levels of customization and support both centrally and across each application. While ERP systems are commercial applications developed by individual vendors, they can hardly be considered off-the-shelf. They are part of a continuing trend of outsourcing IT solutions in which part of the solution is bought, part is configured, and part is built from scratch. Given their broad organizational and functional scope, ERP systems are unlike any other contemporary commercial manufacturing applications. They provide transaction management from both the business perspective and a database perspective. Additionally, they provide a basic level of decision support. Optionally, they enable higher levels of decision support that may be offered by ERP vendors or a third-party vendor.

This chapter presented an overview of ERP from the outside and from the inside. The outside view clarified the connection between ERP, electronic commerce, and supply chain management. The inside view describe the functional and implementation elements of ERP systems, particularly in the context of manufacturing enterprises, and identified the points at which ERP interacts with other software applications in manufacturing enterprises. Finally, we looked at open research problems surrounding ERP and identified those that are important to fitting ERP systems into current and future business processes.

Acknowledgments

This work was performed at NIST under the auspices of the Advanced Technology Program's Office of Information Technology and Applications and the Manufacturing Engineering Laboratory's Systems for Integrating Manufacturing Applications (SIMA) Program in the Manufacturing Systems Integration Division. We thank the many colleagues who shared their knowledge, experience and patience with us: James M. Algeo, Sr., Bruce Ambler, Tom Barkmeyer, Jeff Barton, Mary Eileen Besachio, Bruce Bond, Dave Burdick, Neil Christopher, David Connelly, Maggie Davis, Paul Doremus, Chad Eschinger, Jim Fowler, Cita Furlani, Hideyoshi Hasegawa, Peter Herzum, Ric Jackson, Arpan Jani, Al Jones, Amy Knutilla, Voitek Kozaczynski, Mary Mitchell, Steve Ray, Michael Seubert, and Fred Yeadon.

Disclaimer

Commercial equipment and materials are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

Brunnermeier, S., and Martin, S. (1999), *Interoperability Costs Analysis of the U.S. Automotive Supply Chain,* Research Triangle Institute, Research Triangle Park, NC.

CommerceNet (2000), web page, http://www.commerce.net.

Davenport, T. H. (1998), "Putting the Enterprise into the Enterprise System," *Harvard Business Review,* Vol. 76, No. 4, pp. 121–131.

Electronic Business XML (2000), web page, http://www.ebxml.org.

Gold-Bernstein, B. (1999), "From EAI to e-AI," *Applications Development Trends,* Vol. 6, No. 12, pp. 49–52.

Greenspan, A. (1999), Keynote Address, "The American Economy in a World Context," in *Proceedings of the 35th Annual Conference on Bank Structure and Competition* (Chicago, May 6–7, 1999), Federal Reserve Bank of Chicago, Chicago.

Hagel, J., III, and Singer, M. (1999), "Unbundling the Corporation," *Harvard Business Review,* Vol. 77, No. 2, pp. 133–141.

Kotler, P., and Armstrong, G. (1999), *Principles of Marketing,* Prentice Hall, Upper Saddle River, NJ, pp. 1–10.

Kumar, K., and Van Hillegersberg, J. (2000), "ERP Experiences and Evolution," *Communications of the ACM,* Vol. 43, No. 4, pp. 23–26.

Office of Management and Budget (1988), *Standard Industrial Classification Manual,* Statistical Policy Division, U.S. Government Printing Office, Washington, DC.

Office of Management and Budget (1997), *North American Industry Classification System* (*NAICS*)—*United States,* Economic Classification Policy Committee, U.S. Government Printing Office, Washington, DC.

Open Applications Group (2000), web page, http://www.openapplications.org.

RosettaNet (2000), web page, http://www.rosettanet.org.

Samuelson, P. A., and Nordhaus, W. D. (1998), *Macroeconomics,* Irwin/McGraw-Hill, Boston.

Shapiro, C., and Varian, H. R. (1999), *Information Rules,* Harvard Business School Press, Boston.

Sprott, D. (2000), "Componentizing the Enterprise Applications Packages," *Communications of the ACM,* Vol. 43, No. 4, pp. 63–69.

Stoughton, S. (2000), "Business-to-Business Exchanges Are the New Buzz in Internet Commerce," *The Boston Globe,* May 15, 2000, p. C6.

Taylor, D. (1999), *Extending Enterprise Applications,* Gartner Group, Stamford, CT.

Terhune, A. (1999), *Electronic Commerce and Extranet Application Scenario,* Gartner Group, Stamford, CT.

ADDITIONAL READING

Allen, D. S., "Where's the Productivity Growth (from the Information Technology Revolution)?" *Review,* Vol. 79, No. 2, 1997.

Barkmeyer, E. J., and Algeo, M. E. A., *Activity Models for Manufacturing Enterprise Operations,* National Institute of Standards and Technology, Gaithersburg, MD (forthcoming).

Berg, J., "The Long View of U.S. Inventory Performance," *PRTM's Insight,* Vol. 10, No. 2, 1998.

Bond, B., Pond, K., and Berg, T., *ERP Scenario,* Gartner Group, Stamford, CT, 1999.

Bond, B., Dailey A., Jones, C., Pond, K. and Block, J., *ERP Vendor Guide 1997: Overview and Reference,* Gartner Group, Stamford, CT, 1997.

Cox, J. F., III, Blackstone, J. H., and Spencer, M. S., Eds., *APICS Dictionary,* American Production & Inventory Control Society, Falls Church, VA, 1995.

Dilworth, J. B., *Production and Operations Management: Manufacturing and Services,* McGraw-Hill, New York, 1993.

Ginsberg, B. C., ''Material Acquisition Costs Have Declined 63% in the Past Decade,'' *PRTM's Insight,* Vol. 11, No. 1, 1999.

Gormley, J. T., Woodring, S. D., and Lieu, K. C., ''Supply Chain Beyond ERP, Packaged Application Strategies,'' *The Forrester Report,* Vol. 2, No. 2, 1997.

Haeckel, S. H., *Adaptive Enterprise: Creating and Leading Sense-and-Respond Organizations,* Harvard Business School Press, Boston, 1999.

Hammer, M., and Stanton, S., ''How Process Enterprises Really Work,'' *Harvard Business Review,* Vol. 77, No. 4, pp. 108–118, 1999.

Hubbard, T. N., *Why Are Process Monitoring Technologies Valuable? The Use of On-Board Information Technology in the Trucking Industry,* National Bureau of Economic Research, Cambridge, MA, 1998.

Jones, C., and Andren, E., *Manufacturing Applications Strategies Scenario,* Gartner Group, Stamford, CT, 1997.

Lapide, L., '' Supply Chain Planning Optimization: Just the Facts,'' in *The Report on Supply Chain Management,* Advanced Manufacturing Research, Inc., Boston, 1998.

McGrath, M. E., Ed., *Setting the PACE in Product Development,* Butterworth-Heinemann, Boston, 1996.

Mesenbourg, T. L., *Measuring Electronic Business Definitions, Underlying Concepts, and Measurement Plans,* Bureau of the Census, Washington, DC.

Meyer, L. H., ''Q&A on the Economic Outlook and the Challenges Facing Monetary Policy,'' Remarks before the Philadephia Council for Business Economics, Federal Reserve Bank of Philadelphia, Philadelphia, 1999.

Rivlin, A. M., ''On Sustaining U.S. Economic Growth,'' Remarks before the Federal Reserve Board, Minnesota Meeting Group, Minneapolis, 1999.

Scott, B., *Manufacturing Planning Systems,* McGraw-Hill, London, 1994.

Shapiro, J. F., *Bottom-up vs. Top-down Approaches to Supply Chain Management and Modeling,* Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 1998.

Supply Chain Council, *Supply Chain Operations Reference Model,* Supply Chain Council, Pittsburgh, PA, 1998.

CHAPTER 12

Automation and Robotics

Rolf Dieter Schraft
Jens-Günter Neugebauer
Stefan Schmid
Fraunhofer Institute for Manufacturing Engineering and Automation (IPA)

1. INTRODUCTION AND DEFINITION 355

2. CLASSIFICATION OF ASSEMBLY TYPES AND CHOICE OF ASSEMBLY METHODS 356

3. DESIGN OF MANUAL ASSEMBLY SYSTEMS 358

4. DESIGN OF AUTOMATIC ASSEMBLY SYSTEMS 358

5. PERFORMANCE EVALUATION AND ECONOMIC JUSTIFICATION IN SELECTING THE ASSEMBLY SYSTEM 362

6. ASSEMBLY: SCOPE FOR RATIONALIZATION 364

7. ASSEMBLY ENGINEERING: ESSENTIAL TECHNOLOGIES 367

- 7.1. Design for Assembly (DFA) and Assemblability Evaluation 367
  - 7.1.1. Design for Assembly 367
  - 7.1.2. Assemblability Evaluation 368
  - 7.1.3. Boothroyd–Dewhurst DFA Method 369
- 7.2. Simultaneous Engineering 369
- 7.3. Connecting Technologies 371
  - 7.3.1. Screws 371
  - 7.3.2. Rivets 372
  - 7.3.3. Self-Pierce Riveting 372
  - 7.3.4. Press-fitting 372
  - 7.3.5. Clinching 373

8. INDUSTRIAL ROBOTS 373

- 8.1. Definitions 373
- 8.2. Classification and Types of Robots 374
- 8.3. Major Robot Components 374
  - 8.3.1. Power Supply 374
  - 8.3.2. Measuring Equipment 376
  - 8.3.3. Control System 376
  - 8.3.4. Gripper 377
- 8.4. Programming and Robot Simulation 377
  - 8.4.1. Programming 377
  - 8.4.2. Robot Simulation 378
- 8.5. New Applications 379
  - 8.5.1. Courier and Transportation Robots 379
  - 8.5.2. Cleaning Robots 380
  - 8.5.3. Refueling by Robot 381
  - 8.5.4. Medical Robot 381
  - 8.5.5. Assistance and Entertainment 381

9. TECHNOLOGIES FOR ASSEMBLY OPERATIONS 381

- 9.1. Feeding Systems 381
- 9.2. Magazines 383
- 9.3. Fixturing 384
- 9.4. Sensors and Vision Systems 384
  - 9.4.1. Tactile Sensors 385
  - 9.4.2. Force/Torque Sensors 385
  - 9.4.3. Video-optical Sensors 385
  - 9.4.4. Ultrasound Sensors 386

10. COMPUTER-AIDED METHODS FOR ASSEMBLY SYSTEMS 386

- 10.1. Layout Planning and Optimization 386
- 10.2. Simulation of Material Flow 388

11. ASSEMBLY IN INDUSTRY: APPLICATIONS AND CASE STUDIES 388

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

11.1. Automotive Assembly 388
11.2. Assembly of Large Steering Components 389
11.3. Automatic Removal of Gearboxes from Crates 389
11.4. Electronic Assembly 392
  11.4.1. Assembly of an Overload Protector 392
  11.4.2. Assembly of Measuring Instruments 392
  11.4.3. Assembly of Luminaire Wiring 394
  11.4.4. Assembly of Fiberoptic Connectors 395
11.5. Microassembly 395
11.6. Food Industry 396
11.7. Pharmaceutical and Biotechnological Industry 398
REFERENCE **398**
ADDITIONAL READINGS **398**

1. INTRODUCTION AND DEFINITION

Automation has many facets, both in the service industry and in manufacturing. In order to provide an in-depth treatment of the concepts and methods of automation, we have concentrated in this chapter on assembly and robotics. Industrially produced, finished products consist mainly of several individual parts manufactured, for the most part, at different times and in different places. Assembly tasks thus result from the requirement of putting together individual parts, dimensionless substances, and subassemblies into assemblies or final products of higher complexity in a given quantity or within a given unit of time. Assembly thus represents a cross-section of all the problems in the production engineering field, very different activities and assembly processes being performed in the individual branches of industry.

Assembly is "the process by which the various parts and subassemblies are brought together to form a complete assembly or product, using an assembly method either in a batch process or in a continuous process" (Turner 1993). VDI Guideline 2860 defines assembly as "the sum of all processes needed to build together geometrically determined bodies. A dimensionless substance (e.g. slip materials and lubricants, adhesives, etc.) can be applied additionally."

The production system in a company can be divided into different subsystems. The assembly area is linked to the other subsystems by the material and information flow (see Figure 1).

Assembly is concerned with bringing together individual parts, components, or dimensionless substances into complex components or final products. DIN 8593 differentiates the following five main functional sections:

1. Supply
2. Adjustment
3. Inspection
4. Assembly work
5. Help functions

The composition of these activities varies greatly, depending on the industry and product. The structure of assembly systems is based on different organization types, depending on the batch size, the product, and the projected turnover. Assembly is divided into:

- Point assembly (often called site assembly)
- Workshop assembly
- Group assembly
- Line assembly

The evaluated scope for automation is expressed by this list in ascending order.

This chapter provides a general overview of the assembly field. Special emphasis is given to the selection and design of the appropriate assembly approach, assembly design, and assembly techniques.

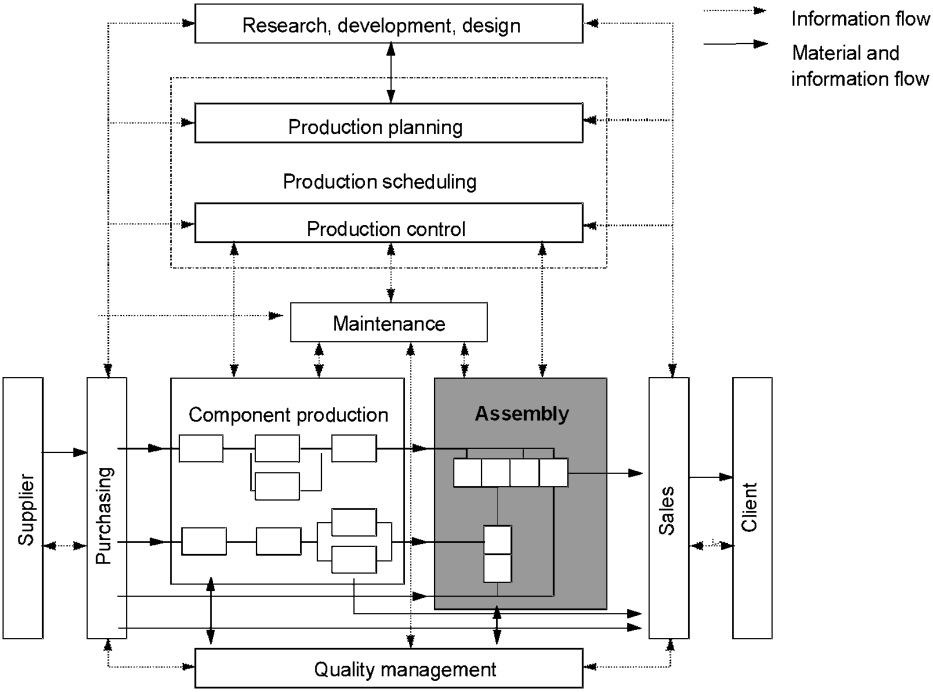

Information flow
Material and information flow
Research, development, design
Production planning
Production scheduling
Production control
Maintenance
Supplier
Purchasing
Component production
Assembly
Sales
Client
Quality management

Figure 1 Assembly as a Subsystem of the Production System.

2. CLASSIFICATION OF ASSEMBLY TYPES AND CHOICE OF ASSEMBLY METHODS

Investigations into automation assembly procedures should first try to ascertain whether and in what scope automation assembly systems can be applied efficiently. This requires a very careful analysis of the assembly tasks as well as a thorough examination of possible alternatives and their profitability.

The total assembly cost of a product is related to both the product design and the assembly system used for its production. Assembly costs can be reduced to a minimum by designing the product so that it can be assembled economically by the most appropriate assembly system. Assembly systems can be classified into three categories:

1. Manual assembly
2. Special-purpose assembly
3. Flex-link (programmable) assembly

In manual assembly, the control of motion and the decision making capability of the assembly operator, assuming that the operator is well trained, are far superior to those of existing machines or artificial intelligence systems. Occasionally, it does make economic sense to provide the assembly operator with mechanical assistance, such as fixtures or a computer display detailing assembly instructions, in order to reduce the assembly time and potential errors.

Special-purpose assembly systems are machines built to assemble a specific product. They consist of a transfer device with single-purpose work heads and feeders at the various workstations. The transfer device can operate on a synchronous indexing principle or on a free-transfer nonsynchronous principle.

Flex-link assembly systems, with either programmable work heads or assembly robots, allow more than one assembly operation to be performed at each workstation and provide considerable flexibility in production volume and greater adaptability in designing changes and different product styles.

Capital-intensive assembly, such as when automatic assembly systems are applied, produces the required low unit costs only when relatively large quantities are produced per time unit and a long amortization time is selected. Due to its flexibility, however, an assembly system able to assemble various products is more economical despite its larger investment volume. A flexible assembly system is also capable of assembling smaller batches cost-effectively, thus meeting the increasing market demands for product variety and shorter product life.

The type of automated assembly system is essentially determined by the workpieces to assemble. The most influential factors are the workpiece's total weight, size, amount, and number of types and variants.

Assembly systems are mainly automated for small-volume products or components with a total weight of a few kilograms. The cycle times for automated assembly systems vary between tenths of seconds and minutes.

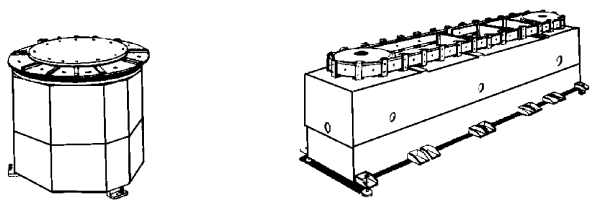
The various types of automated assembly systems arise due to the combination of the different types of assembly stations and work transfer devices, which, in turn, are dependent on the requirements of the workpieces to assemble. Flexibility, that is, being adaptable to different conditions and assembly tasks, should also be a characteristic of assembly systems. An adequate configuration of the assembly station as well as respective work transfer devices will help meet the different flexibility requirements.

Assuming that the product design is compatible with automated assembly, there are several different ways to characterize the operation and configuration of the automated system. One way consists of classifying the system by the type of work transfer used by the system. These types are shown in Figure 2.

In transfer systems linked to rigid work transfer systems, the workpieces and/or partly assembled components run through the assembly process on workpiece carriers in clocked cycles. The assembly stations are placed on a circuit or next to each other parallel to the transfer device. Rigid work transfer systems have the following features:

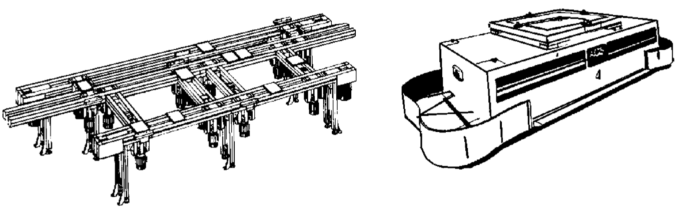
- Joint or jointly controlled handling and/or passing-on devices
- A uniform cycle for all assembly stations set by the pace of the station with the longest cycle time
- Assembly of the manufacturing equipment in specified uniform intervals
- The shutdown of one assembly station results in the stoppage of the whole assembly installation

Work transfer devices for rigid work transfer systems



Cycle circular transfer system Cycle longitudinal transfer system

Work transfer devices for free work transfer systems



Longitudinal transfer system Automated guide vehicle

Figure 2 Alternative Work Transfer Systems for Different Assembly Configurations.

For transfer systems to flex-link systems, flexibility is achieved in the way the individual stations of an assembly system are linked—for example, for a junction in the material flow or the evening out of capacity fluctuations and/or technical malfunctions of the individual assembly stations.

Flex-link systems also allow a different level of automation at each of the assembly stations. Depending on the respective assembly tasks, the automation level can be stipulated according to technical and economic factors. In addition, step-by-step automation of the assembly system is possible because individual manual assembly equipment can still be replaced by automatic assembly station later on.

Some features of flex-link assembly stations are:

- Independent handling and manufacturing equipment that is linked to the interfaces for control purposes
- Single disengagement of the processing and handling cycles that allows different cycle times at each of the assembly stations
- Magazines as malfunction buffers between the assembly stations for bridging short stoppage times
- Freedom of choice in the placement and sequence of the assembly stations.

For longitudinal transfer systems, the transfer movement usually results from the friction between the workpiece carrier and the transfer equipment (belt, conveyor, etc.). The workpiece carriers transport the workpieces through the assembly system and also lift the workpieces in the assembly stations. The workpiece carrier is isolated and stopped in the assembly station in order to carry out the assembly process. The coding of the workpiece is executed by mechanical or electronic write/read units on which up-to-date information about the assembly status or manufacturing conditions can be stored.

Due to flex-linking, the number of linked assembly stations can be very high without impairing the availability of the whole system. Flexible modular assembly systems with well over 80 assembly stations for automated assembly of complex products are therefore not uncommon.

3. DESIGN OF MANUAL ASSEMBLY SYSTEMS

The manual assembly systems are divided into two main categories, manual single-station assemblies and assembly lines.

The manual single-station assembly method consists of a single workplace in which the assembly work is executed on the product or some major subassembly of the product. This method is generally used on a product that is complex or bulky and depends on the size of the product and required production rate. Custom-engineered products such as machine tools, industrial equipment, and prototype models of large, complex consumer products make use of a single manual station to perform the assembly work on the product.

Manual assembly lines consist of multi-workstations in which the assembly work is executed as the product (or subassembly) is passed from station to station along the line (see Figure 3).

At each workstation, one or more human operators perform a portion of the total assembly work on the product by adding one or more components to the existing subassembly. When the product comes off the final station, work has been completed.

Manual assembly lines are used in high-production situations where the sequences to be performed can be divided into small tasks and the tasks assigned to the workstations on the line. One of the key advantages of using assembly lines is the resulting specialization of labor. Because each worker is given a limited set of tasks to perform repeatedly, the worker becomes a specialist in those tasks and is able to perform them more quickly and consistently.

4. DESIGN OF AUTOMATIC ASSEMBLY SYSTEMS

The automation scope of assembly systems includes the following priorities:

- Reduction in the cost of assembly
- Increasing productivity
- Improved product quality

Improved motivation of the staff, improved clearness, shorter processing times, and improved ergonomical workstations as well as better organization are further advantages.

The movement towards automation of manufacturing and assembling sequences takes place progressively. Mechanized facilities can be automated only if specific prerequisites are fulfilled and if employing complex industrial goods seems profitable. Because assembly systems require maximum flexibility, the degree of automation is still relatively low.

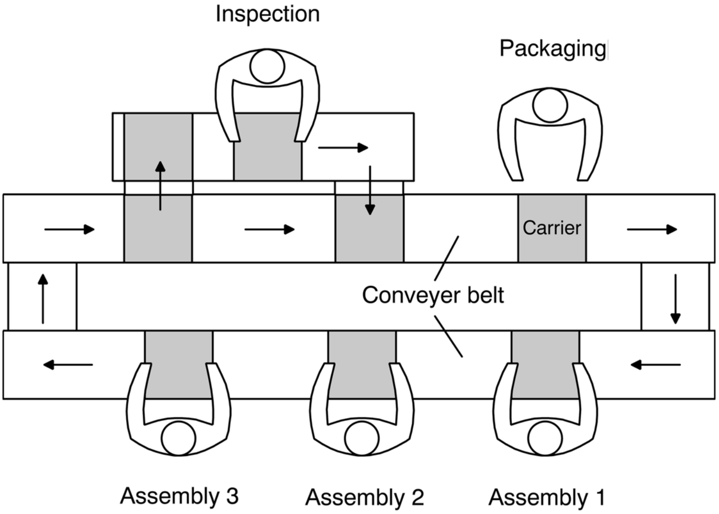


Figure 3 Example of a Manual Assembly Line.

In principle, automatic assembly systems can be further classified into the following subcategories:

- Short-cycle assembly machines
- Flexible, modular assembly systems
- Flexible assembly systems with industrial robots

Short-cycle assembly machines are actually single-purpose machines because each of them is specially designed for one assembly task only. They work in cycles ranging from less than 1 sec up to approximately 5 sec and are also used for mass-production assembly (over 1 million units annually).

Short-cycle assembly machines are often constructed modularly with standardized components for the work transfer devices and the handling and joining equipment. This allows the existing assembly machines to be partially adjusted to the new situation by modifying, converting, or retooling for changes to the assembly product.

Assembly processes are monitored by means of integrated control devices (mechanical keys, inductive feeders, etc.) in the assembly stations or partly in separate stations. This is a very important function from a quality-assurance point of view. For short-cycle assembly machines, either force-moved or pneumatic-driven motion facilities are used in the assembly stations.

The essence of flexible, modular assembly systems is that all subsystems are constructed in a modular way. The modularity pertains not only to the actual assembly stations and the sequencing devices, stands, and so on, but to the work transfer devices, control and regulation. The variety of combinations made possible by the module allows the automation of a wide range of assembly tasks. The modular conception of these systems ensures

- Easy modification and supplementary possibilities
- The capability of changing to adjust to different sets of circumstances
- A high level of reusability
- Expansion of the control technology

The investment risk is thus reduced.

Flexible, modular assembly systems (see Figure 4) usually work in cycles of a few seconds and are often used for the automated assembly of product variants for large to medium numbers of units. The principle of the flex-link assembly stations is applied for flexible, modular assembly systems because this is the only way to realize flexibility in the linking of individual assembly stations. Longitudinal transfer systems are used as work transfer devices. In order to prevent individual cycle times, technical malfunctions, capacity fluctuations, and other factors of certain assembly stations from affecting the next stations, a buffer function in the transfer line between the assembly sta-

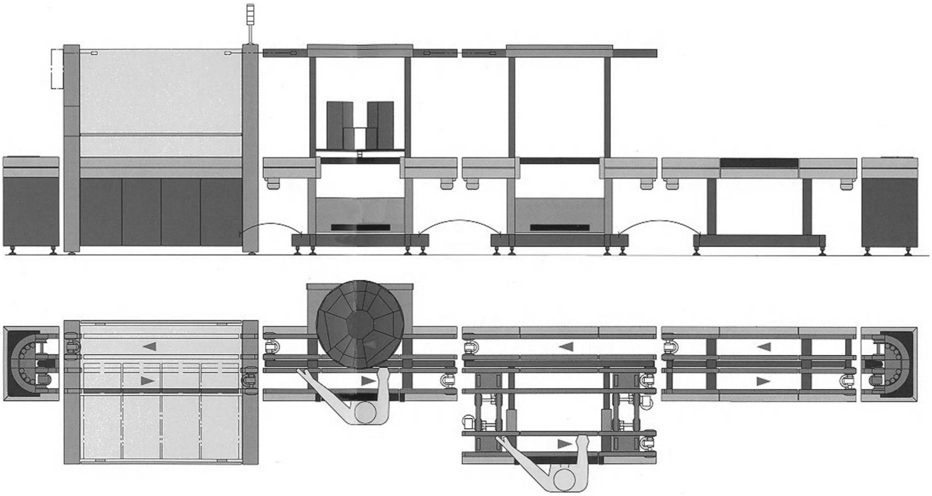


Figure 4 Flexible, Modular Assembly System. (*Source:* teamtechnik)

tions allows several workpiece carriers to stack up. The distance between the individual assembly stations is not determined by the transfer device but by an optimal buffer capacity.

Flexible assembly systems with industrial robots can be divided into three principal basic types with specific system structures:

1. Flexible assembly lines
2. Flexible assembly cells
3. Flex-link assembly systems

The largest number of industrial assembly robots is used in flexible assembly lines (see Figure 5). Flexible assembly lines are more or less comparable to flexible, modular assembly systems with regard to construction, features, and application. The cycle times for flexible assembly lines generally vary between 15 and 30 sec. The main area of application for these systems is the automated assembly of products with annual units of between 300,000 and 1 million.

The application of flexible assembly lines is economically viable especially for assembling products with several variants and/or short product lives because the flexibility of assembly robots can

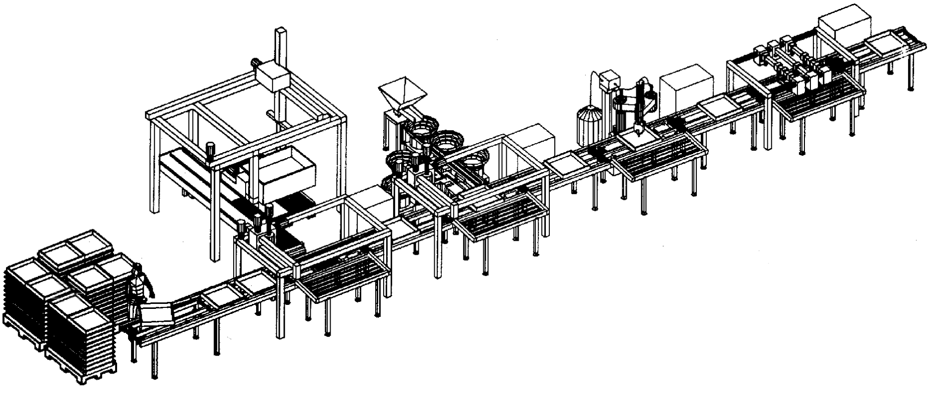


Figure 5 Flexible Assembly Line.

be best utilized for the execution of different assembly procedures for the individual product variations or follow-up products.

Assembly robots can also execute several assembly processes within an assembly cycle. This allows a high level of utilization for the assembly system, in particular for smaller amounts of workpieces, which is important for economic reasons. Depending on the cycle time range, an assembly robot can be allocated with a maximum of five to six assembly procedures in a flexible assembly line.

Flexible assembly cells (see Figure 6) are complex automated assembly stations with one or two assembly robots for large work loads (larger than for assembly stations or flexible assembly lines) where an exact limitation is very difficult and seldom expedient. A certain number of periphery devices are necessary to carry out several assembly processes of complete assembly sequences or large portions of them. Task-specific assembly devices such as presses and welding stations are also integrated if required. These periphery devices significantly limit the possible workload of the flexible assembly cells, to a maximum of 8–10 different workpieces per assembly robot. There are several reasons for this:

- Only a limited number of periphery devices can be placed in the assembly robot's workspace.
- The greater the number of subsystems, the greater the risk of interference to the whole system.
- A large number of periphery devices reduces the accessibility of the system for maintenance, repair, and troubleshooting work.

A large number of the industrial applications of flexible assembly cells are conducted as island solutions, that is, without being linked to other assembly stations. Flexible assembly cells as assembly islands are used for the automated assembly of subassembly components or simple products with usually less than 20 different parts. This system works best technically and economically with cycle times from 25–120 sec. This allows it to be applied in situations requiring an annual number of units to be assembled of between 50,000 and 500,000, depending on the scope of the assembly tasks and the number of shifts with which the system will be run.

The application of flexible assembly cells is not possible when the product or component consists of more than 20 parts. Industrial assembly very often concerns products or components composed

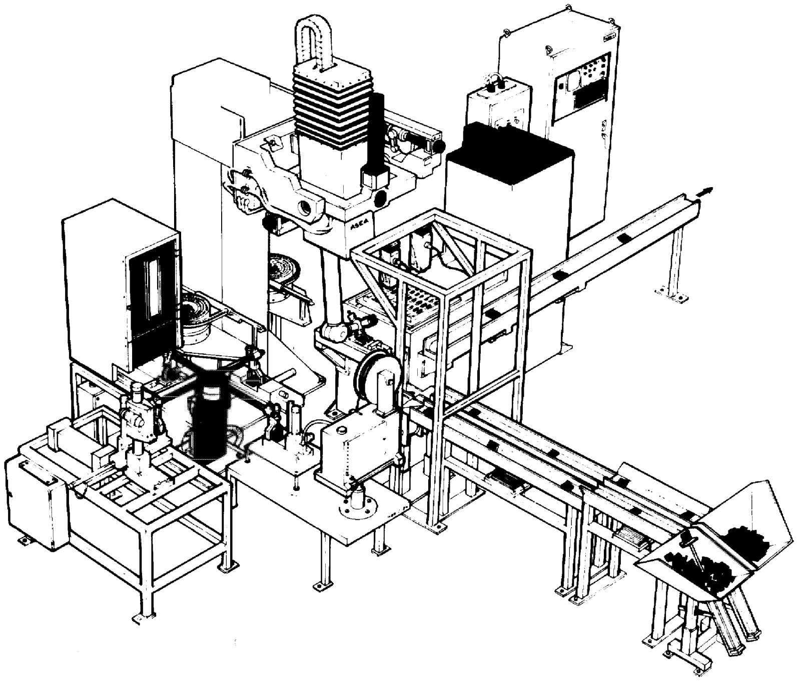


Figure 6 Flexible Assembly Cell.

of many more parts with annual units below 500,000. In order to assemble these products automatically, it is possible to distribute the assembly procedures among several flex-link assembly systems.

There are two typologies of linking flexible assembly systems: permanent and flex-link sequences. Linking with a permanent linking sequence means that each assembly system is linked in a set sequence to each other, such as with longitudinal transfer systems (see Figure 7). The permanent linking sequence allows these assembly systems to execute only a smaller number of assembly tasks with different assembly sequences. They are therefore suitable for the automated assembly of variants and types with comparable assembly sequences and several similar products or components with comparable assembly sequences.

Particularly for small annual workpiece amounts, it is economically necessary to assemble several different products or components in one flexible automated assembly system in order to maintain a high level of system utilization. Because sometimes very different assembly processes must be realized, a linking structure (flex-link sequence), independent from certain assembly processes, is necessary. Assembly systems of this type are able to meet the increasing demands for more flexibility.

In flex-link assembly systems, the workpieces are usually transported by workpiece carriers equipped with workpiece-specific lifting and holding devices. The transfer movement usually results from the friction between the workpiece carrier and the transfer equipment (belt, conveyor, plastic link chain, etc.). The transfer device moves continuously. However, the workpiece carriers are stopped during the execution of the assembly processes. In the assembly station, each of the workpiece carriers is stopped and indexed. If high assembly forces are required, the workpiece carriers must be lifted from the transfer device.

The coding of the workpiece carrier is executed by means of mechanical or electronic write/read units. In this way, it is possible to record up-to-date information about the assembly status or manufacturing procedure on every workpiece carrier and to process this information through the assembly control system. Figure 8 shows a workpiece carrier with a coding device.

5. PERFORMANCE EVALUATION AND ECONOMIC JUSTIFICATION IN SELECTING THE ASSEMBLY SYSTEM

Various methods have been developed and proposed for performance evaluation and system selection. Simplified mathematical models have been developed to describe economic performance, and the effects of several important variables have been analyzed:

1. Parts quality (PQ), represented by this factor, is the average ratio of defective to acceptable parts.
2. Number of parts in each assembly (NA).
3. Annual production volume per shift (VS).

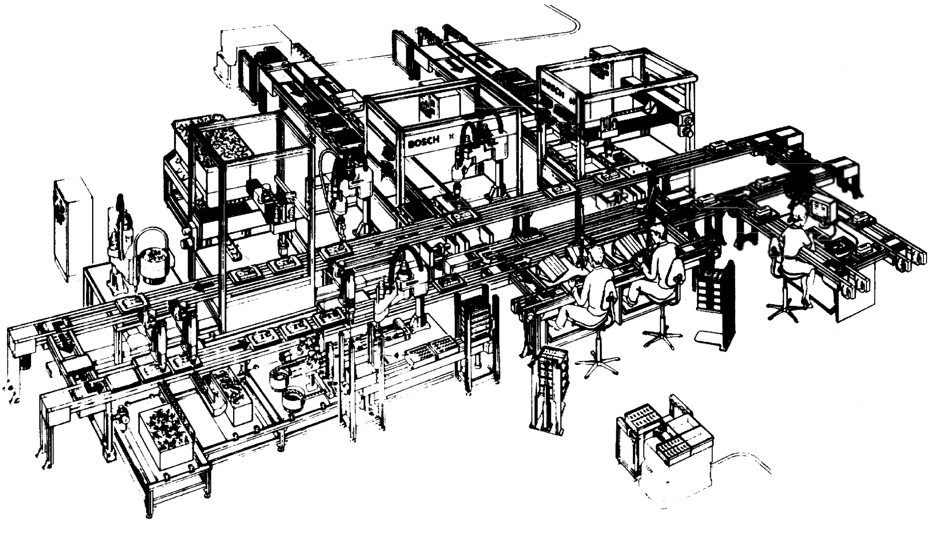


Figure 7 Flex-link Assembly System.

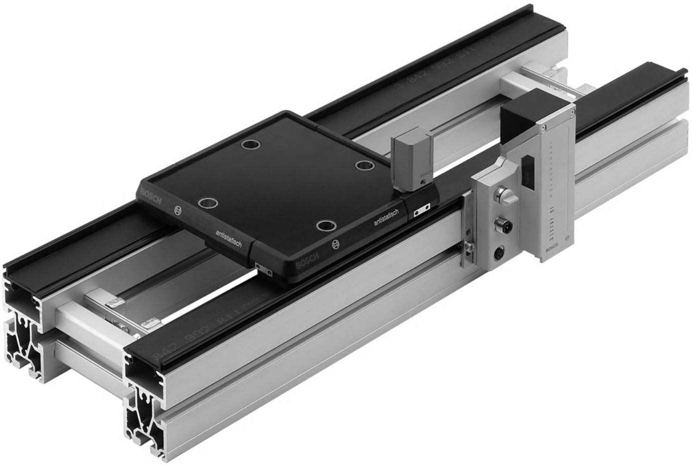


Figure 8 Example of a Workpiece Carrier with a Coding Device. (*Source:* Robert Bosch GmbH)

4. Product style variation (SV) is defined as the ratio of the total number of parts available (NT) to the number NA actually used in each product.
5. Design changes (ND) is the number of parts that are changed (necessitating new feeders and workheads, etc.) during the life of the assembly system.
6. Number of products to be assembled (NP).
7. Economic climate (RI). Economic market conditions are expressed by the capital cost of the equipment relative to an equivalent operator rate.
8. QE is defined as the cost of the capital equipment that can economically be used to do the work of one operator on one shift per year.

The economic climate for investment in automation can be expressed by the ratio RI:

$$\mathrm{RI} = \mathrm{SH}\,\frac{\mathrm{QE}}{\mathrm{WA}} \tag{1}$$

where SH = number of shifts
WA = annual cost of one assembly operator

When this ratio is low, investment in automation is less likely to be profitable.

In addition to the eight variable factors above, it should be remembered that manual assembly systems will be necessary if special fitting or adaptation of parts is required or if there are wide fluctuations in demand for the product, such that its automated system will not be able to respond in time to the new requirements. In addition, intangible benefits of automation, which are difficult to quantify (e.g., shortage of qualified workers) or high consistency requirements must be taken into account.

The cost of assembly (CA) for a complete assembly is given by

$$\mathrm{CA} = \mathrm{TP}\left(\mathrm{WT} + \frac{\mathrm{CE} \times \mathrm{WA}}{\mathrm{SH} \times \mathrm{QE}}\right) \tag{2}$$

where TP = average time between delivery of complete assemblies for a fully utilized system
WT = total rate for the machine operators
CE = total capital cost for all equipment including engineering setup and debugging cost

For the purpose of comparing the economics of assembly systems, the cost of assembly per part will be used and will be nondimensionalized by dividing this cost per part by the rate for one assembly time per part, TA. Thus, the dimensionless assembly cost per part (CD) is given by:

$$CD = \frac{CA}{NA \times WA \times TA} \tag{3}$$

Substituting Eq. (2) into (3) gives

$$CD = \frac{TP}{TA}\left(WR + \frac{CE/NA}{SH \times QE}\right) \tag{4}$$

where

$$WR = \frac{WT}{WA \times NA} \tag{5}$$

and is the ratio of the cost of all operators compared with the cost of one manual assembly operator and expressed per part in the assembly. The dimensionless assembly cost per part for an assembly operator working without any equipment will be one unit, which forms a useful basis for comparison purposes. For a particular assembly system Eq. (4) holds true only if the required average production time (TQ) for one assembly is greater than or equal to the minimum production time (TP) obtainable for the system. This means that if TP $\leq$ TQ (because the system is not fully utilized), then TQ must be substituted for TP.

6. ASSEMBLY: SCOPE FOR RATIONALIZATION

The early history of assembly process development is closely related to the history of the development of mass production methods. Thus, the pioneers of mass production are also considered the pioneers of modern assembly. Their ideas and concepts significantly improved the manual and automated assembly methods employed in large-volume production. In the past decade, efforts have been directed at reducing assembly costs by the application of flexible automation and modern techniques.

In the course of the development of production engineering, mechanization and automation have reached a high level in the fields of parts production, which permits the efficient production of individual parts with a relatively low proportion of labor costs. In the field of assembly, by contrast, automation remains limited to large-volume production. In medium and short-run production, rationalization measures have been taken mainly in the area of work structuring and workstation design. For the following reasons, automation measures have scarely begun in assembly:

1. Assembly is identified by product-specific, and thus quite variable, work content (handling, joining, adjusting, testing activities). Once solutions have been found, however, they can be applied to other products or companies only with great difficulty, in contrast to parts manufacturing.
2. Assembly, as the final production stage, must cope extensively with continuously shifting market requirements in regard to timing, batch sizes, derivatives, and product structure.

Although the automation trend in the assembly field has been growing remarkably over the past years, many industrial sectors still have relatively unexploited potential for rationalization. Compared to other manufacturing branches, such as parts manufacturing, assembly is characterized by a relatively low degree of automation. It accounts for only 60–90% of the manufacturing sequences of parts manufacturing, spot welding, and press shop in the automotive industry, for example. Specifically in assembly, however, this percentage decreases dramatically to less than 15% because of the complexity of the assembly tasks. The vehicle assembly as a whole (inclusive of final assembly) is quite cost-intensive in terms of employed labor force, representing 30% portion of the whole vehicle cost of production. Because of the high assembly costs, the automation of further subassembly tasks is therefore considered a major objective towards which the most effort is being directed.

Over the past years, automation technologies have been evolving dramatically in terms of flexibility and user-friendly operation as for their use in assembly systems. Automated devices, especially industrial robots, have also become less and less cost-intensive. Prices have been reduced up to 50% in some cases. For this reason, the scope for rationalization and application in assembly is as promising as ever.

An automation study was conducted by the Fraunhofer IPA in order to ascertain which industrial branches are characterized by exceptional scope for rationalization in the assembly. The results of this investigation are shown in Figure 9.

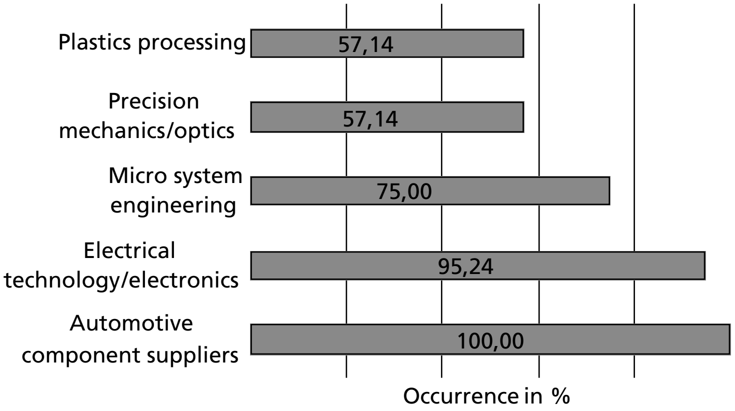


Figure 9 Industrial Branches Characterized by Exceptional Scope for Rationalization in Assembly.

The objectives achieved by the industrial application of automated solutions are shown in Figure 10. The investigation gave evidence that the cost-reduction objective, followed by quality improvement, was considered the first priority.

The most important preconditions to the realization of the still unexploited potential for rationalization in assembly are illustrated in Figure 11.

Despite the fact that assembly-oriented product design is perhaps the most important prerequisite for simple and automatic assembly, as has been well known for years, the enormous scope for rationalization of this field is very far from exhausted. Yet even though it is undeniable that the high manufacturing cost of a product becomes evident in assembly, the largest portion of it is due to construction. Investigations into different manufacturing areas have produced evidence that approximately 75% of the whole assembly costs originate in the early stages of product design and construction. The most important rules and methods of assembly-oriented product design are described in the following chapter.

The precondition to assembly systems covered next in the study is hardware and software modularization. The interconnecting implementation of standardized components, allowing assembly systems for the most varied assembly tasks to be designed simply and quickly, makes it possible to

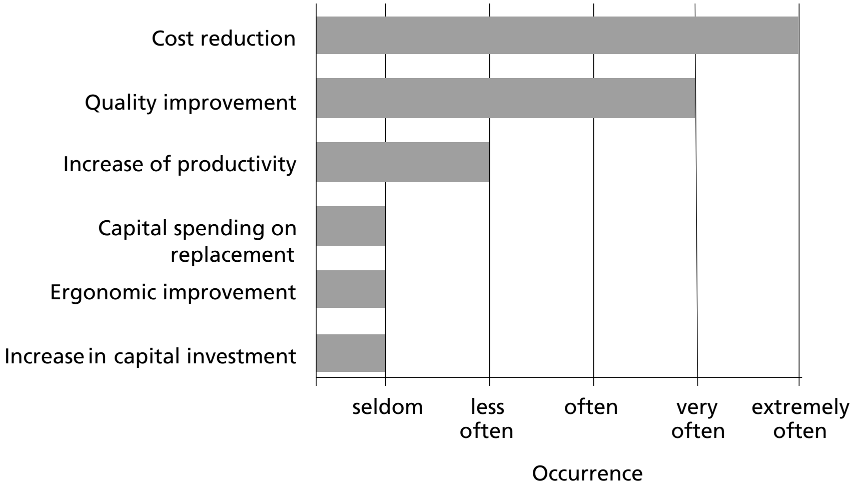


Figure 10 Objectives Pursued by Assembly Automation Projects.

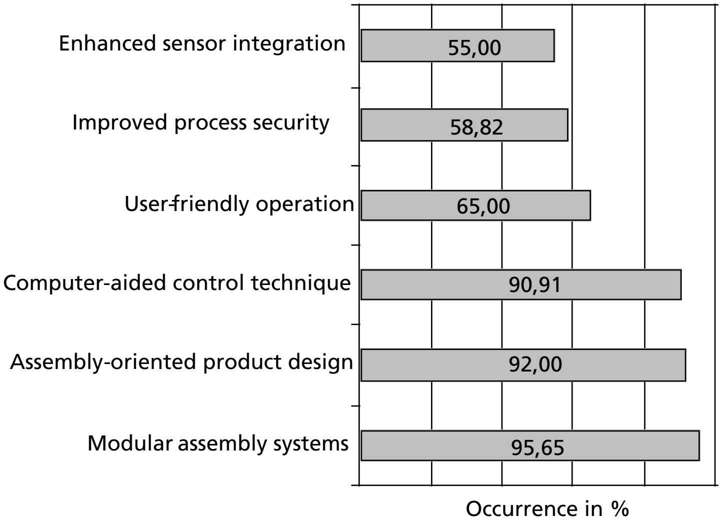


Figure 11 Preconditions to the Realization of the Assembly scope for Rationalization.

strive for the required flexibility. The investment risk decreases in proportion to the high recoverability of most components because the volume of investment would be well paid off in the case of product change if a large portion of the assembly components could be reused.

New requirements for the control technique are imposed by modularization, such as modular configuration and standardized interfaces. Conventional SPS control systems are largely being superseded by industrial PC. The result is decentralization of intelligence achieved by integration, for example, its transfer to the single components of the assembly systems. Assembly systems can thus be realized more quickly and cost-effectively. Further possibilities for rationalization could be exploited by simplification of the operation as well as programming of assembly systems, still characterized by a lack of user-friendliness. The service quality and operation of assembly systems could be greatly improved by the use of graphically assisted control panels allowing for better progress monitoring.

Industrial robots are being used more and more in flexible automation for carrying out assembly sequences. Flexibility is guaranteed by free programmability of the device as well as by the ever-increasing integration of intelligent sensory mechanisms into the systems. The application of image-processing systems directly coupled with robot control is expected to grow faster and faster in the next years. The increasing implementation of sensor technologies is due to the rapid development of computer hardware, allowing for better performance at lower prices.

Further on, the degree of automation in the assembly can be enhanced by improving the logistics and material flow around the workplaces as well as by reducing secondary assembly times. Logistical aspects can be improved by optimizing the layout of workplaces, paying special regard to the arm sweep spaces. Specific devices for local parts supply can help to reduce secondary assembly times consistently. They should be specially designed for feeding marshalled components correctly oriented as close as possible to the pick-up station to enable workers to grasp them ‘‘blindly’’ and quickly. Rigid workplace limitations can be relaxed by introducing the compact layout (see Figure 12), allowing workers to walk from one workstation to the next and carry out different assembling tasks corresponding to overlapping sequences. This allows the number of workers along the same assembly line to be varied flexibly (according to the task to perform) even though the whole assembly sequence is still divided proportionately among the workers employed.

As for the organizational and structural aspects in the assembly, recent research approaches and developments have revealed new prospects of better ‘‘hand-in-hand’’ cooperation between workers and robots along assembly lines or at workstations. Industrial robots, being intended to help workers interactively in carrying out assembly tasks, need not be isolated or locked any longer. For this purpose, man–machine interfaces should be improved and safety devices redeveloped. Further on, robots should learn to cope with partially undefined external conditions. This is still a vision at present, but it shows future opportunities for realizing the unexploited potential for rationalization in the assembly field.

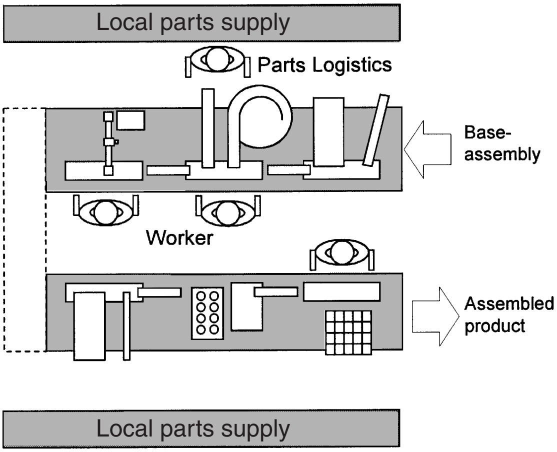


Figure 12 Layout of a Flexibly Varying Assembly System.

7. ASSEMBLY ENGINEERING: ESSENTIAL TECHNOLOGIES

The three essential technologies for assembly engineering explained in this section are product design for assembly, simultaneous engineering, and automation of connecting techniques.

7.1. Design for Assembly (DFA) and Assemblability Evaluation

7.1.1. Design for Assembly

When a product is designed, consideration is generally given to the ease of manufacture of its individual parts and the function and the appearance of the final product. One of the first steps in the introduction of automation in the assembly process is to reconsider the design of the product to make it simple enough to be performed by a machine.

The subject of design for automatic assembly can be conveniently divided into three sections: product design, design of subassemblies for ease of assembly, and design of parts for feeding and orienting.

The greatest scope for rationalization is in the measures affecting the whole product assembly system. However, these can only be executed for subsequent products, that is, in the long-term. Individual subassemblies can be made easier to assemble for existing products as long as this does not affect the function and the interfaces to the rest of the product remain the same. The success of these rationalization measures, however, is not as great as that concerning the whole assembly system. Measures affecting individual parts have the least potential for rationalization because only small modifications can be made. However, they can be realized quickly with existing products so that the positive effect of the measures can be perceived very quickly (Figure 13).

Various points concerning parts and product design for assembly are summarized below.

Assembly rules for product design and design of subassemblies:

1. Minimize the number of parts.
2. Ensure that the product has a suitable part on which to build the assembly, with accessibility from all sides and sufficient room for assembly.
3. Ensure that the base part has features that will enable it to be readily located in a stable position in the horizontal plane.
4. If possible, design the product so that it can be built up in layer fashion, each part being assembled from above and positively located so that there is no tendency for it to move under the action of horizontal forces during the machine index period.
5. Form subassemblies that can be assembled and inspected independently.
6. Try to facilitate assembly by providing chamfers or tapers that will help to guide and position.
7. Avoid time-consuming fastening operations (e.g., screwing, soldering) and separate joining elements and safety elements (e.g., rivets, safety panes).
8. Avoid loose, bendable parts (e.g., cable).

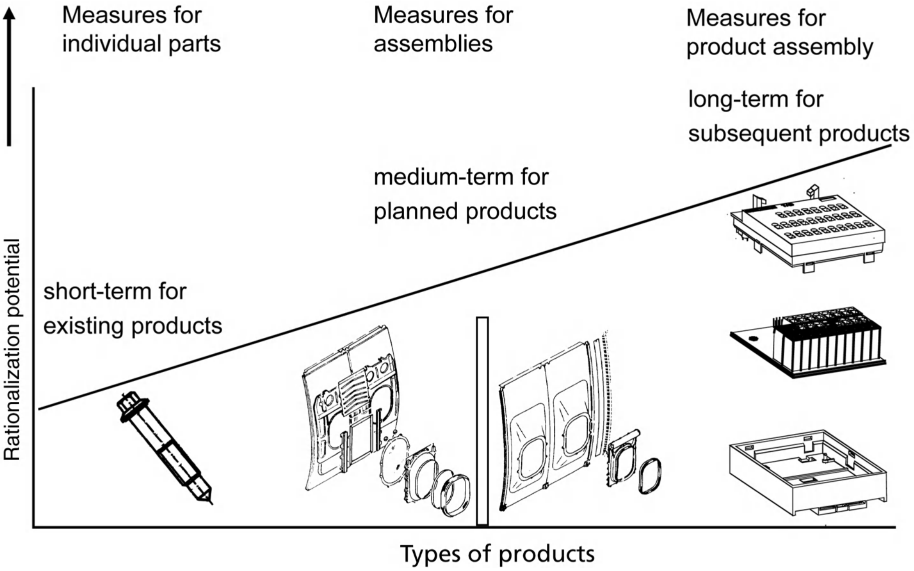

Measures for individual parts
Measures for assemblies
Measures for product assembly
long-term for subsequent products
medium-term for planned products
short-term for existing products
Rationalization potential
Types of products

Figure 13 Measures for Easy-to-Assemble Product Design.

Assembly rules for the design of parts:

1. Avoid projections, holes, or slots that will cause tangling with identical parts when placed in bulk in the feeder. This may be achieved by arranging that the holes or slots are smaller than the projections.
2. Attempt to make the parts symmetrical to avoid the need for extra orienting devices and the corresponding loss in feeder efficiency.
3. If symmetry cannot be achieved, exaggerate asymmetrical features to facilitate orienting or, alternatively, provide corresponding asymmetrical features that can be used to orient in the parts.

In addition to the assembly design rules, a variety of methods have been developed to analyze component tolerance for assembly and design for assembly by particular assembly equipment.

7.1.2. Assemblability Evaluation

The suitability of a product design for assembly influences its assembly cost and quality. Generally, product engineers attempt to reduce the assembly process cost based on plans drawn by designers. The recent trend has been toward product assemblability from the early design phase in order to respond to the need for reduction in time and production costs.

The method of assemblability evaluation is applied by product designers for quantitatively estimating the degree of difficulty as part of the whole product design process. In 1975, the Hitachi Corporation developed the pioneering method of assemblability evaluation called the Assemblability Evaluation Method (AEM). AEM analyzes the assembly structure using 17 symbols, thus aiming to give designers and production engineers an idea of how easily procedures can be assembled (see Figure 14). It points out weak areas of design from the assemblability viewpoint. The basic ideas of the AEM are:

1. Qualification of difficulty of assembly by means of a 100-point system of evaluation indexes
2. Easy analysis and easy calculation, making it possible for designers to evaluate the assemblability of the product in the early stage of design
3. Correlation of assemblability evaluation indexes to assembly cost

AEM focuses on the fundamental design phase. The approach is to limit the number of evaluation items so that designers can apply them in practice. The reason for concentrating on the early design phase was to achieve greater savings by considering assemblability issues as priorities from the very

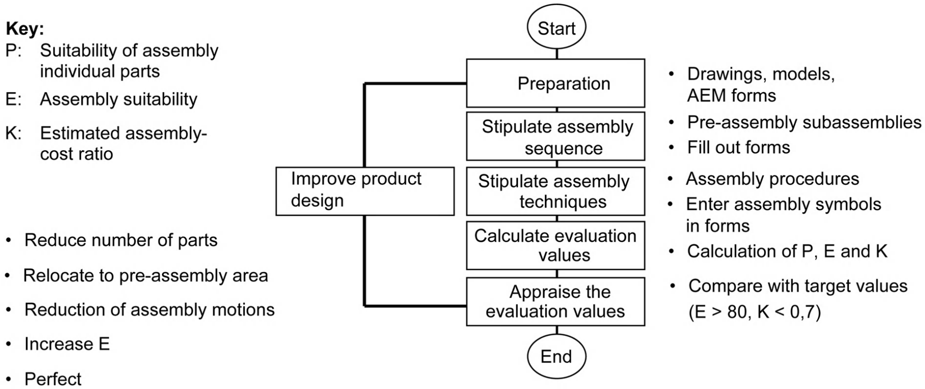

Key:
P: Suitability of assembly individual parts
E: Assembly suitability
K: Estimated assembly-cost ratio
• Reduce number of parts
• Relocate to pre-assembly area
• Reduction of assembly motions
• Increase E
• Perfect
Improve product design
Start
Preparation
Stipulate assembly sequence
Stipulate assembly techniques
Calculate evaluation values
Appraise the evaluation values
End
• Drawings, models, AEM forms
• Pre-assembly subassemblies
• Fill out forms
• Assembly procedures
• Enter assembly symbols in forms
• Calculation of P, E and K
• Compare with target values (E > 80, K < 0,7)

Figure 14 Evaluation of Assembly Suitability According to Hitachi (AEM).

beginning. As a result, however, the accuracy of this evaluation cannot be very high compared to more detailed analyses.

7.1.3. Boothroyd–Dewhurst DFA Method

The Boothroyd–Dewhurst method is widely used. The objective of this method is to minimize the amount of assembly required for existing products or product constructions. This should be achieved first and foremost by reducing to a minimum the number of individual parts that have to be assembled. The shape of the individual parts for handling and assembly is also optimized. The procedure can be divided into three steps:

1. Selection of an assembly principle
2. Analysis of the assembly task (Figure 15)
3. Improvement of design

In the Boothroyd–Dewhurst method, the assembly times and costs as well as the design efficiency (DE) play a decisive role in determining the assembly suitability. The DE is calculated by multiplying the number of individual parts by the calculated costs and comparing this figure with the theoretical minimum number of individual parts multiplied by the ideal costs for assembly. This means that the factor DE always has a value between 0 (very bad) and 1 (ideal assembly system for this product).

The theoretical minimum number of individual parts is determined by examining whether two assembly parts that are associated with another

- Must move toward each other relatively
- Must be made of different materials
- Must be separated from other assembly parts for assembly or dismantling

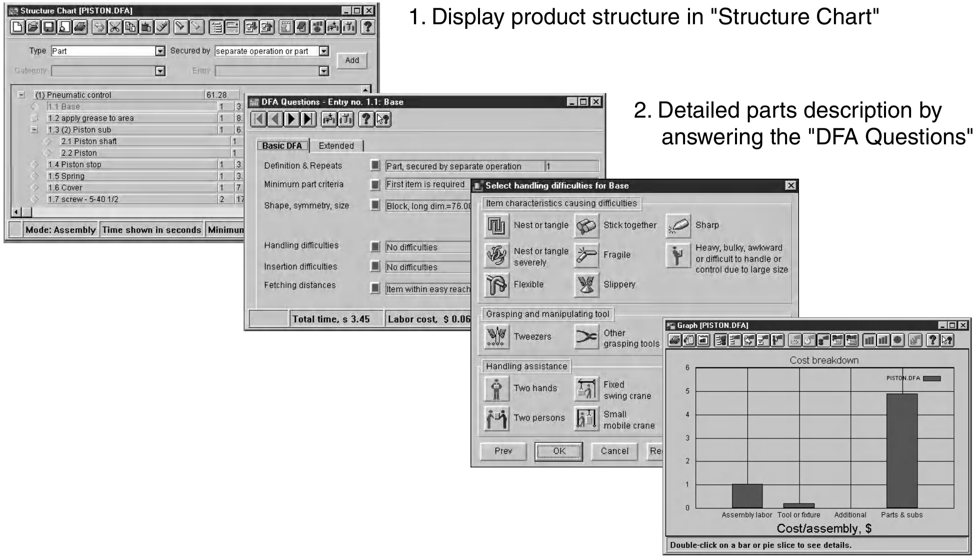
If none of these three requirements have to be satisfied, both assembly parts may be combined.

Based on the costs and the DE factor, a decision is made on whether reconstruction is possible. A reconstructed product is then reanalyzed in the same way. Figure 16 shows the original variants and a reconstructed valve. Originally, the lid had to be kept low against the screw compression spring so that both screws could be assembled. Designing the lid with snap connectors for attachment to the casing and the integration of guides to help the screw compression spring into the piston suggested that the number of parts could be reduced from seven to four and the expensive screwing assembly task could be eliminated.

The Boothroyd–Dewhurst method aims primarily at reducing assembly costs by means of an integrated construction system. Other measures, such as product structuring, sorting, the construction of subassemblies and the differential construction technique are not considered. The evaluation of the suitability of the assembly system requires that the product be a constructed one. This means that assembly suitability is taken into consideration only very late in the construction process. The Boothroyd–Dewhurst DFA method is also available for PCs.

7.2. Simultaneous Engineering

In the planning and realization of assembly systems, it is not only the selection of the correct operating materials which is important for the project success but also the optimal composition of the project



Structure Chart [PISTON.DFA]

Type: Part

Secured by: separate operation or part

(1) Pneumatic control 61.28
- 1.1 Base 1 3
- 1.2 apply grease to area 1 8
- 1.3 (2) Piston sub 1 6
  - 2.1 Piston shaft 1
  - 2.2 Piston 1
- 1.4 Piston stop 1 3
- 1.5 Spring 1 3
- 1.6 Cover 1 7
- 1.7 screw - 5-40 1/2 2 17

Mode: Assembly | Time shown in seconds | Minimum

DFA Questions - Entry no. 1.1: Base

Basic DFA | Extended

Definition & Repeats: Part, secured by separate operation 1

Minimum part criteria: First item is required

Shape, symmetry, size: Block, long dim.=76.00

Handling difficulties: No difficulties

Insertion difficulties: No difficulties

Fetching distances: Item within easy reach

Total time, s 3.45 | Labor cost, $ 0.06

Select handling difficulties for Base

Item characteristics causing difficulties
Nest or tangle
Stick together
Sharp
Nest or tangle severely
Fragile
Heavy, bulky, awkward or difficult to handle or control due to large size
Flexible
Slippery
Grasping and manipulating tool
Tweezers
Other grasping tools
Handling assistance
Two hands
Fixed swing crane
Two persons
Small mobile crane
Prev
OK
Cancel
Re

Graph [PISTON.DFA]

Cost breakdown
PISTON.DFA
6
5
4
3
2
1
0
Assembly labor
Tool or fixture
Additional
Parts & subs
Cost/assembly, $

Double-click on a bar or pie slice to see details.

1. Display product structure in "Structure Chart"

2. Detailed parts description by answering the "DFA Questions"

3. Presentation of the results

Figure 15 Analysis of the Assembly Tasks According to the Boothroyd–Dewhurst Method.

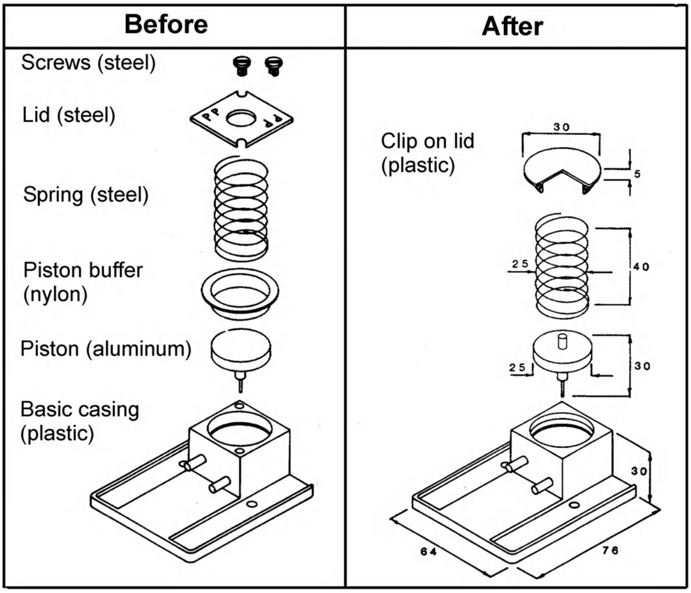


Figure 16 Easy-to-Assemble Product Design for a Pneumatic Piston.

groups. Teamwork, also called simultaneous engineering, is certainly the most decisive precondition for an optimal and mutually coordinated design of the product and assembly system. The team is made up of product developers, assembly system planners, quality management, purchasing, and sales and can vary depending on the project tasks and decisions to take. However, steps must be taken to ensure that all members of the team are aware of the rules of simultaneous engineering.

Another opportunity for interdisciplinary work arises when there is pressure to shorten product development times. By involving the assembly system planner in the construction process, it is possible to begin much earlier with the assembly system design. Some of the development and planning work may then be executed parallel to one another. The development and realization times are shortened and the costs for the whole project are usually reduced. Additionally, the products can be brought onto the market much earlier due to the shorter development times.

Figure 17 shows a few examples of how simultaneous engineering can help shorten the product development times for new products in comparison with preceding models.

Experience from numerous projects has also shown that, by means of simultaneous engineering, the number of product changes during the development phase can be significantly reduced.

Simultaneous engineering is not really a new way of thinking. However, in recent times it has been applied more consistently and systematically than ever. This is particularly true in Europe. Many Japanese companies have been practising simultaneous engineering successfully since the mid-1970s.

7.3. Connecting Technologies

The establishment of connections is one of the most common assembly tasks. The automation of this task is thus one of the central problems confronting assembly rationalization.

7.3.1. Screws

Among the most important assembly procedures is the making of screw connections. The connecting process ‘‘screwing’’ has therefore been examined in detail, and the use of flexible handling appliances such as screw assembly systems has further been tested in typical industrial tasks. The problems of positioning errors, controlled correction of the screwing tool, and diameter flexibility for handling different screw sizes without changing tools have been solved through the development of special hardware modules. The system offers the possibility, in assembling metric screws, of increasing the joining speed, handling screws with different heads (including Phillips), self-tapping in wood and/ or plastic, and mounting hexagon head cap screws of different diameters without changing tools.

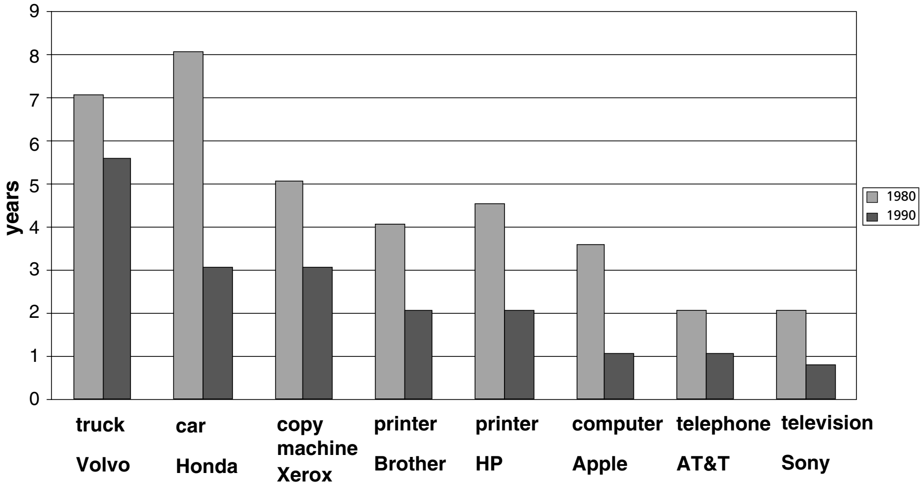


Figure 17 The Change in Product Development Times When Using Simultaneous Engineering. (*Source:* Prof. Dr. Gerpott)

7.3.2. Rivets

Like screwing, riveting is one of the classic connecting techniques. In recent times, rivets have in fact been gaining in popularity again through the development of new types and manipulation tools and the increasing replacement of steel and iron by other materials.

The greatest market potential among the different types of rivets is forecast for blind rivets. Access to the workpieces being joined is required from only one side, thereby offering the ideal conditions for automation. A blind riveter for industrial robots has been developed based on a standard tool that, apart from its size and weight, is also notable for its enhanced flexibility. With the aid of a changeover device, for example, different rivet diameters and types of blind rivet can be handled.

7.3.3. Self-Pierce Riveting

In contrast to traditional riveting, self-pierce riveting does not require pilot drilling of the steel plate at the joint. The most frequent technique is self-pierce riveting with semitubular rivets, characterized by the rivet punching out only the punch side of the sheet. Joining simply results from noncutting shaping of the die side and widening of the rivet base on the sheet die side. The necessary dies are to be matched with the rivet length and diameter, workpiece thickness, as well as number of interlayer connections, thickness variants, and materials. The workpiece punch side is flat when countersunk punching head rivets are used, whereas the die side is normally uneven after die sinking.

Punching rivet connections with semitubular rivets are gastight and waterproof. Punching rivets are mainly cold extruded pieces of tempering steel, available in different degrees of hardness and provided with different surfaces. Ideal materials for application are therefore light metals and plastics or combinations of both. Punching rivet connections with semitubular rivets show a far better tensile property under strain than comparable spot welding connections.

7.3.4. Press-fitting

Press-fitting as a joining process does not require any additional elements, making it particularly desirable for automation. At the same time, however, high seam joining and bearing pressures occur that may not be transferred to the handling appliance.

A press-fitting tool was therefore developed with allowance for this problem. It confines the bearing pressures within the tool with very little transfer to the handling appliance. Various process parameters, such as joining force, joining path, and joining speed, can be freely programmed so as to increase the flexibility of the tool. A tolerance compensation system is also integrated in the tool that compensates any positioning errors (axis and angle shift). For supply of the connecting elements the tool also has interfaces to interchangeable part-specific magazines (revolving and cartridge magazines) and the possibility of supply via formed hoses.

7.3.5. Clinching

Clinching is a way of joining sheet metal parts without connecting elements. A specially designed press-driven set of dies deforms the parts at the joint to provide a friction-locked and keyed connection.

Because no connecting elements are required, the problem of organizing, preparing, and feeding these parts does not apply. For this reason, this relatively recent connecting technique is being used for an increasing number of applications.

A robot-compatible, freely programmable tool has been developed for flexibly automated clinching. This tool is notable for its compact outer dimensions and low handling weight. With this robust but sensitive robot, tool joining forces up to 50 kN can be generated. Different sets of dies for the specific connecting task in question can also be used by means of the automated changeover system. The sensor system integrated in the tool allows precise and individual process control. Rapid working strokes and joining force can all be freely programmed. The quality and success of the join at each point are checked and documented.

8. INDUSTRIAL ROBOTS

8.1. Definitions

Industrial robots have become an important and indispensable component in today's world of flexible automation. After the initial technical problems and high financial risk that impeded the willingness to invest, they have become an important means of automation in recent years.

The large number of handling devices in use can be divided into parts handlers and industrial robots (programmable handling devices).

Parts handlers are used in a variety of industrial sectors. They are mainly equipped with handling devices with grippers that execute set motion processes in specified cycles. The motion processes are controlled by means of limit stops or control cams. These components present the biggest disadvantage of the parts handler because of the amount of effort required to retool the device if the process is changed.

Industrial robots are defined by ISO 8373 as automatically controlled, reprogrammable, multipurpose manipulators that are programmable in three or more axes and may either be fixed in place or mobile for use in industrial automation applications.

The number of installed industrial robots in all sectors of the most important industrial nations is shown in Figure 18.

The automotive, electrical, and electronic industries are the largest users of robots. The predominant applications are welding, assembling, painting, and general handling tasks. Flexibility, versatility, and the cost of robot technology have been driven strongly by the needs generated by these industries, which still account for more than 75% of the world's stock of robots. In their main application areas, robots have become a mature product exposed to enormous competition by international robot manufacturers, resulting in rapidly falling unit costs. However, the robot unit price accounts for less than 30% of the average work cell investment. A considerable share of total investment cost is attributed to engineering and programming the robot and putting it into operation.

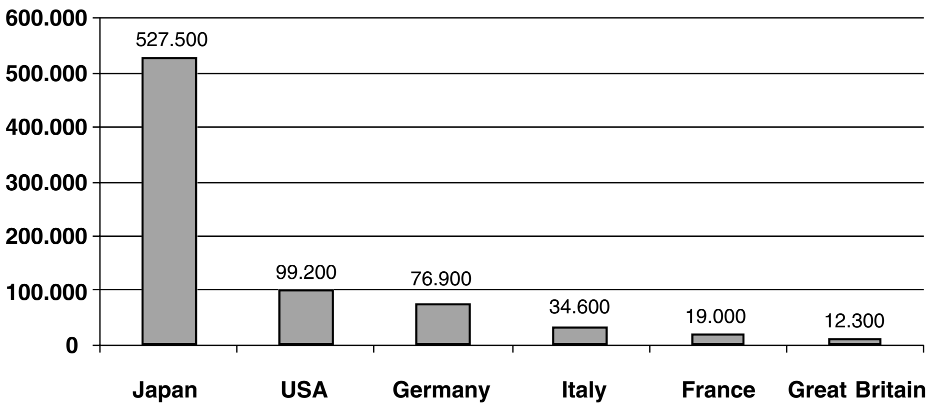


Figure 18 Estimate of the Number of Installed Industrial Robots in the Year 2000.

8.2. Classification and Types of Robots

Each of the industrial robot elements is linked to one another by linear guides and revolute joints in a kinematic chain, and the actuation of these elements forms the axes of the robot. Kinematics is the spatial allocation of the movement axes in order of sequence and construction.

Figure 19 illustrates the mechanical configurations of different types of robots.

The main axes help to position the final effector (tool or workpiece) spatially. The hand or adjacent axes are primarily responsible for the orientation of the tool and are therefore usually made of a series of revolute joints.

The following information may be useful for the selection of the industrial robot best suited to the respective application:

- The load-bearing capacity is the largest weight that has to be handled at the specified motion speed in relation to the robot flange. If the work speed or the reach is decreased, a larger weight can be moved.
- The mechanical structure refers to the kinematic chain as an ordered sequence and the type of motion axis on the robot.
- The number of axes is the sum of all the possible movements by the system. The higher the degree of freedom, the lower the system accuracy and the higher the costs. Therefore, it is advisable to limit the number of axes to the amount required.
- The work space is calculated from the structure of the kinematics and its dimensions. Its shape is also dependent on the movement areas of the individual axes and actuators.
- The positional accuracy determines the deviation during the run up to freely selected positions and orientations. The repetitive accuracy is the difference when the run-up to a spatial point is repeated.

Within a wide range of contructional types and variants, four versions have proved to be particularly suitable in practical operation:

1. Cartesian robots
2. Articulated robots
3. SCARA robots
4. Parallel robots

An extreme rigid structure is the first characteristic of Cartesian robots, whose single axes move only in the direction of the Cartesian space coordinates. For this reason, Cartesian robots are particularly suitable for operating in large-sized working areas. Major applications include workpiece palletizing and commissioning.

Standard articulated robots consist of six axes. They are available on the market in a large variety of types and variants. Characterized by a cylindrical working area occupying a relatively small volume, articulated robots easily allow failures to be repaired directly. Articulated robots are used primarily for spot welding, material handling, and painting as well as machining.

SCARA robots have a maximum of four degrees of freedom but can also be equipped with just three for very simple tasks. These robots are used for assembly tasks of all types, for simple loading and unloading tasks, and for fitting of electronic components to printed circuit boards.

The six linear driving axes of parallel robots are aligned between the base plate and the robot gripper so as to be parallel. Therefore, positioning accuracy and a high degree of rigidity are among the characteristics of parallel robots. The working area is relatively small. Parallel robots are particularly suitable for tasks requiring high accuracy and high-range forces, such as workpiece machining.

8.3. Major Robot Components

8.3.1. Power Supply

Three main power sources are used for industrial robot systems: pneumatic, hydraulic, and electric. Some robots are powered by a combination of electric and one other power source. Pneumatic power is inexpensive but is used mostly for simpler robots because of its inherent problems, such as noise, leakage, and compressibility. Hydraulic power is also noisy and subject to leakage but is relatively common in industry because of its high torque and power and its excellent ability to respond swiftly to motion commands. It is particularly suited for large and heavy part or tool handling, such as in welding and material handling, and for smooth, complex trajectories, such as in painting and finishing. Electric power provides the cleanest and most quiet actuation and is preferred because it is self-contained. On the other hand, it may present electric hazards in highly flammable or explosive environments.

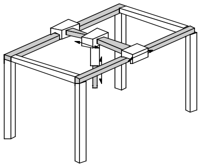
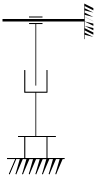
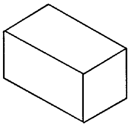
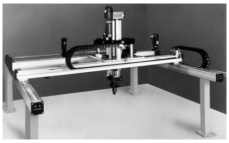
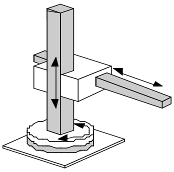
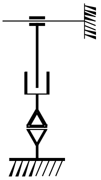
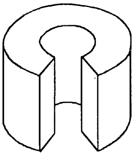
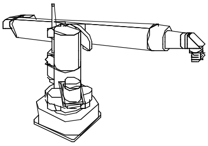
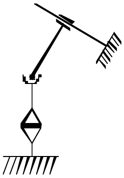
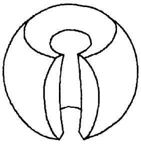
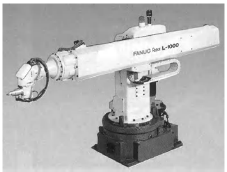
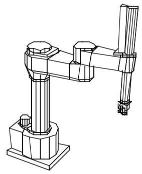
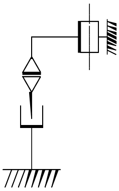
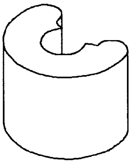
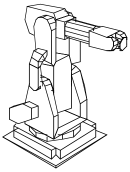
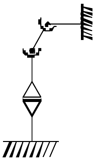
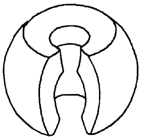
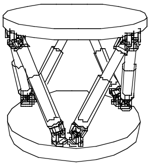
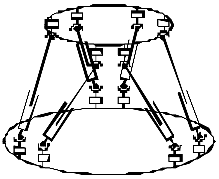
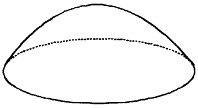
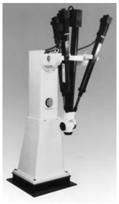
Robot	**Axes**		**Examples**
Principle	Kinematic Structure	Workspace	Photo
 Cartesian Robot			
 Cylindrical Robot			
 Spherical Robot			
 SCARA Robot			
 Articulated Robot			
 Parallel Robot			

Figure 19 Classification of Industrial Robots by Mechanical Structure.

The typical series of sequences to be carried out by the electrical drive of a robot axis is shown in Figure 20.

The reference variables, calculated in interpolation cycles by the motion-controlling device, are transmitted to the axle controller unit. The servodrive converts the electric command values into couple of forces to be transmitted to the robot axle through the linkage. The internal control circuit is closed by a speed indicator, the external one by a path-measuring system.

8.3.2. Measuring Equipment

Measuring devices able to provide the axis controller unit with the suitable input quantity are required for robot position and speed control. Speed measurement is carried out directly over tachometer generators or indirectly over differential signals of angular position measuring systems. Optoelectronical incremental transducers, optoelectronical absolute transducers, as well as resolvers are among the most common devices for robotic angular measuring. For the path measurement of linear motions, the measuring techniques used by incremental and absolute transducers are suitable as well.

8.3.3. Control System

The major task of the robot control system consists of piloting one or more handling devices according to the technologically conditioned handling or machining task. Motion sequences and operations are fixed by a user program and are carried out by the control system unit. The necessary process data are provided by sensors, which therefore make it possible for the unit to adapt to a certain extent the preset sequences, motions, and operations to changing or unknown environmental conditions.

Figure 21 shows the components of a robot controller system.

Data exchange with superset control system units is actuated through a communication module, such as for loading the user program in the robot control unit or exchanging status data. The process control device organizes the operational sequence of the user program which provides instructions for the motion, gripper, sensors, and program flow. The robot axes as well as the auxiliary axes involved in the task to perform are driven by the motion-controlling device.

There are three kinds of motion control:

1. Point-to-point positioning control
2. Multipoint control
3. Continuous path control

The axis controller carries out the task of driving the robot axes according to the reference variables. The sequential progression from one axis position to the next is monitored and readjusted comparatively to the actual positions. An appliance (teach pendent) allows the robot to be operated and programmed by hand. The operator can easily access all control functions in order to run the robot in manual mode or determine the program flow.

The control system also determines two major performance measures of the robot: its accuracy and repeatability. The first indicates the precision with which the robot can reach a programmed

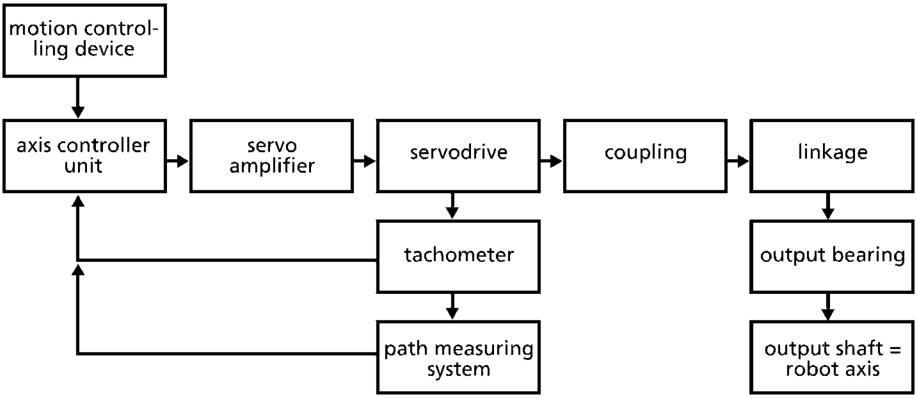


Figure 20 Electrical Drives of Robot Axis.

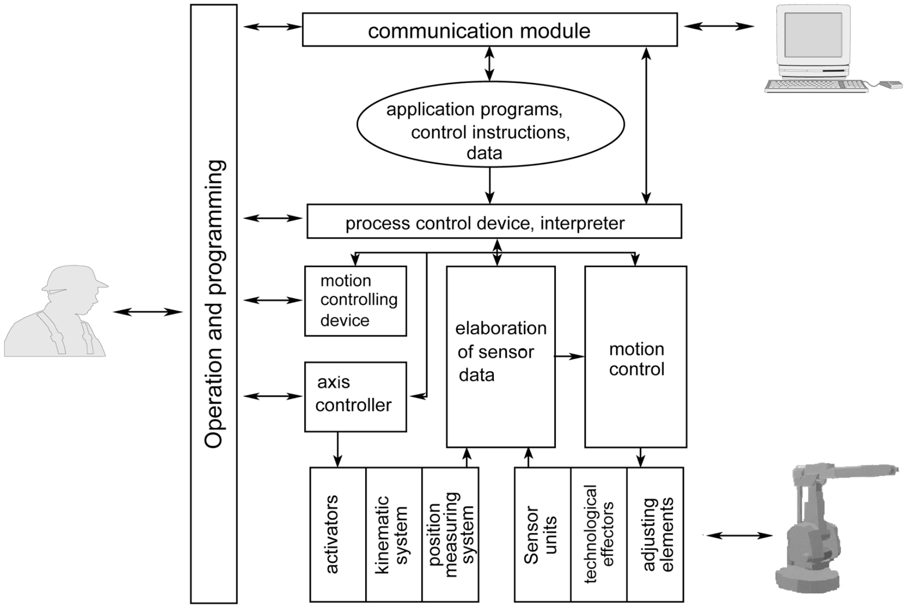

communication module
application programs,
control instructions,
data
Operation and programming
process control device, interpreter
motion controlling device
axis controller
elaboration of sensor data
motion control
activators
kinematic system
position measuring system
Sensor units
technological effectors
adjusting elements

Figure 21 Components of a Robot Controller System.

position or orientation. The second indicates the tolerance range over which a programmed motion can be repeated several or many times. While the accuracy of robots is difficult to establish in a standard way, repeatability has been standardized. Typically, industrial robots will have a repeatability better than ±0.01 in. and more precise robots will be within ±0.003 in.

8.3.4. Gripper

The gripper is the subsystem of an industrial robot able to transfer the power transmission from the workpiece to the industrial robot to make sure the robot has recorded the workpiece's exact position. A gripping system has to carry out the following tasks:

- Temporarily maintaining the allocation of the workpiece and the gripping device, defined on the basis of the gripping center line
- Recording the static forces and moments produced by the workpiece
- Recording the kinetic forces and coupling of forces produced by the acceleration during the operating sequence
- Recording process-bound forces (e.g., joining forces)

According to the complexity of the gripping task, the robot gripper can be equipped with several subsystems. A driving mechanism (electric, pneumatic, or hydraulic) is required if the gripping force is produced kinematically, which is often the case with mechanical grippers. A holding system allows the workpiece position to be fixed opposite the gripping kinematic device. The gripping area is among the major criteria for evaluating the flexibility of a gripping device.

The object (workpiece) to be manufactured, the space available in the building, and the required cycle time of handling and assembly systems play a major role in the decision which robotic gripper to use. If, for example, the standard gripper version proves to be inadequate for specific operations, it can be equipped with interchangeable gripping systems by the creation of standardized interfaces.

8.4. Programming and Robot Simulation

8.4.1. Programming

The robot control system contains a programming method by teaching, by using a teach pendent, by showing and actually leading the robot manipulator through the desired motions, or by programming.

A programming method is the planned procedure that is carried out in order to create programs. According to IRDATA, a program is defined as a sequence of instructions aimed at fulfilling a set of manufacturing tasks. Programming systems allow programs to be compiled and also provide the respective programming assistance (see Figure 22).

If direct methods (online programming methods) are used, programs are created by using the robot. The most common method is teach-in programming, in which the motion information is set by moving towards and accepting the required spatial points, assisted by the teach pendent. However, the robot cannot be used for production during programming.

One of the main features of indirect methods (offline programming methods) is that the programs are created on external computers independent of the robot control systems. The programs are generated in an offline programming system and then transferred into the robot's control system. The key advantage of this method is that the stoppage times for the robot systems can be reduced to a minimum by means of the configuration of the programs.

Hybrid methods are a combination of direct and indirect programming methods. The program sequence is stipulated by indirect methods. The motion part of the program can be defined by teach-in or play-back methods or by sensor guidance.

8.4.2. Robot Simulation

Computer-assisted simulation for the construction and planning of robot systems is becoming more and more popular. The simulation allows, without risk, examination and testing on the model of new experimental robot concepts, alternative installation layouts or modified process times within a robot system. "*Simulation* is the imitation of a system with all its dynamic processes in an experimental model which is used to establish new insights and to ascertain whether these insights can be transferred to real situations" (VDI Guideline 3633).

The starting point for the development of graphic 3D simulation systems was the problem of planning the use of robots and offline programming. Independent modules or simulation modules integrated into CAD system were created. To enable a simulation to be conducted, the planned or real robot system must first be generated and depicted as a model in the computer. The abstraction level of the simulation model created must be adjusted to the required imitation: "as detailed as necessary, as abstract as possible." Once the model has been completed, an infinite number of simulations can be carried out and modifications made. The objective is to improve the processes and eliminate the possibility of planning mistakes.

Figure 23 shows an example of the visualization of a robot welding station.

Simulation is now considered one of the key technologies. Modern simulation systems such as virtual reality are real-time oriented and allow interaction with the operator. The operator can change the visual point of view in the graphical, 3D picture by means of input devices (e.g., data gloves) and thus is able to take a look around in the "virtual world." The virtual objects can be interactively manipulated so that modifications can be executed more quickly, safely, and comfortably.

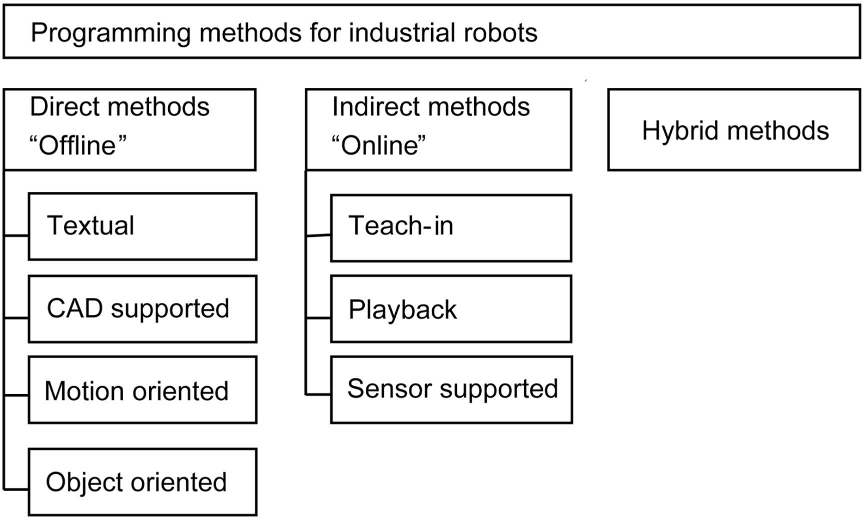


Figure 22 Overview of Programming Methods for Industrial Robots.

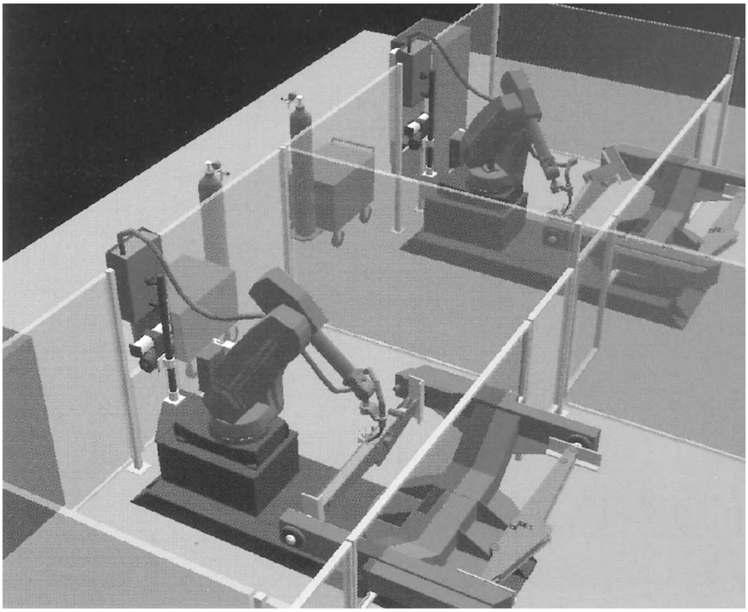


Figure 23 3D Simulation of a Robot Welding Station.

The main advantages and disadvantages of simulation can be summed up as follows:

- Dynamic analysis of modeled processes and verification of executed modifications
- Increased planning security due to the evaluation of the project planning before realization
- Reduction in planning time
- Reduction in development costs for the realization of robot systems
- Expenditure savings due to reduction of run-through times and stock levels and optimization of resource utilization

8.5. New Applications

In the 1980s, as industrial robots became a recognized indicator of modern production facilities, thought was also given to the use of robots outside the factory. The term *service robots* was coined for such new robot systems, which, however, still have no valid international definition today. The International Federation of Robotics (IFR) November 1997, Stockholm, Sweden suggested the following definition in 1997: “A service robot is a robot which operates semi or fully autonomously to perform services useful to the well-being of humans and equipment, excluding manufacturing operations.”

Because of the multitude of forms and structures as well as application areas, service robots are not easy to define. IFR has adopted a preliminary system for classifying service robots by application areas:

- Servicing humans (personal, safeguarding, entertainment, etc.)
- Servicing equipment (maintenance, repair, cleaning, etc.)
- Performing autonomous functions (surveillance, transport, data acquisition, etc.)

Following are examples of application fields for service robots.

8.5.1. Courier and Transportation Robots

The HelpMate, introduced in 1993, is a mobile robot for courier services in hospitals. It transports meals, pharmaceuticals, documents, and so on, along normal corridors on demand (see Figure 24). Clear and simple user interfaces, robust robot navigation, and the ability to open doors or operate

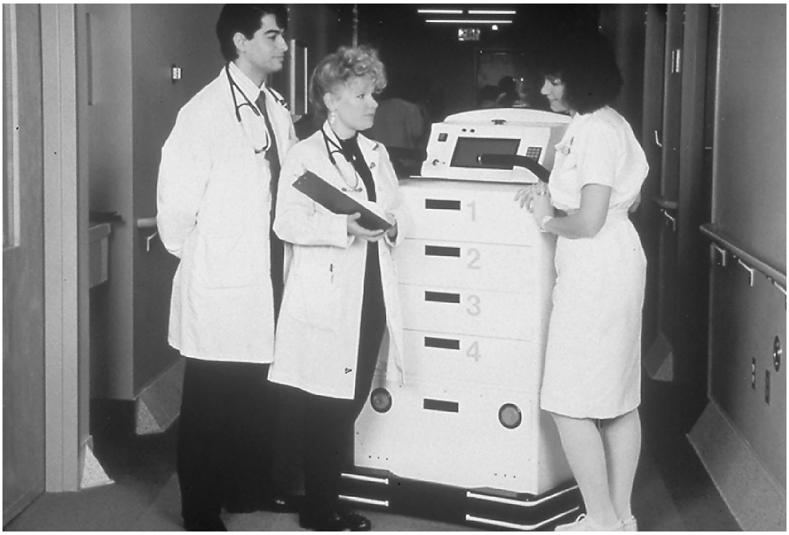


Figure 24 Mobile Robot for Courier Services in Hospitals. (*Source:* PYXIS)

elevators by remote control make this a pioneering system in terms of both technology and user benefit.

8.5.2. Cleaning Robots

A service robot climbs surfaces on suction cups for cleaning (see Figure 25), inspection, painting and assembly tasks. Tools can be mounted on the upper transversal axis. Navigation facilities allow accurate and controlled movement.

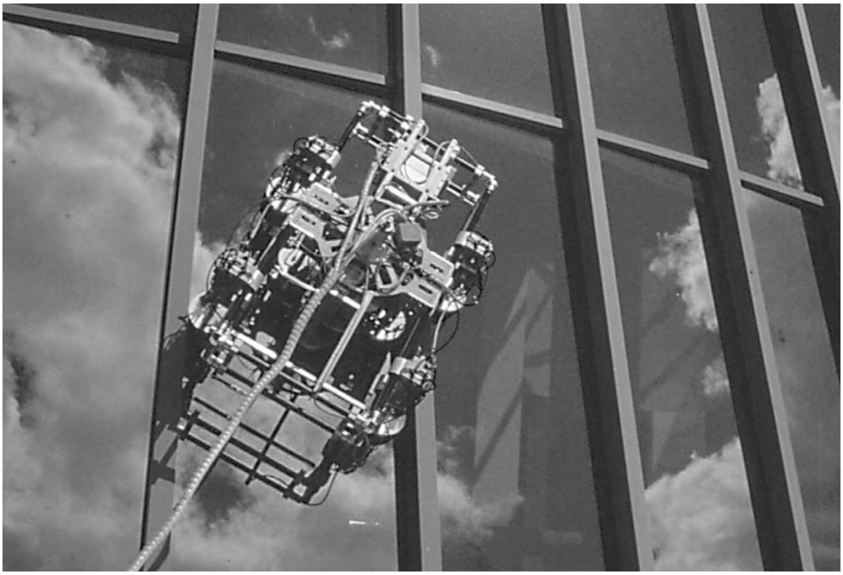


Figure 25 Cleaning Robot. (*Source:* IPA)

8.5.3. *Refueling by Robot*

Refueling a car by robot is as convenient and simple as entering a parking lot. Upon pulling up to the refueling station, the customer inserts a card and enters the PIN code and the refueling order. The robot locates the car, opens the tank flap, and docks onto the tank cap. Once the cap is open, the robot pumps the right grade and amount of fuel into the tank (see Figure 26).

8.5.4. *Medical Robot*

This mechatronic assistant system was developed by Siemens AG and Fraunhofer IPA. It is used for different operation tasks. The system consists of an operation robot (hexapod) whose kinematics concept is based on the Stewart platform. These kinematics allow extreme accuracy in micrometer ranges and a high level of stiffness despite the robot's relatively small size (see Figure 27).

The operator controls the operation robot from an ergonomic operation cockpit similar to a flight simulator. Tactile back-coupling of the motions helps the operator operate the system. The operation cockpit is also assembled on a hydraulic hexapod.

8.5.5. *Assistance and Entertainment*

This mobile robot has been created to communicate with and entertain visitors in a museum (see Figure 28). It approaches the visitors and welcomes them to the museum. Speech output is accompanied by movement of the robot head. The robot gives guided tours in the museum. Moving its head up and down symbolizes the robot looking at the object it is currently talking about. Explanations are further accompanied by pictures or video sequences shown on the screen of the robot.

9. TECHNOLOGIES FOR ASSEMBLY OPERATIONS

Several technologies highly relevant to assembly are discussed briefly below.

9.1. Feeding Systems

Sorting and feeding devices are an important component in the automated assembly of products. Their task is to sort out the often unsorted supplied workpieces and to deliver the right amount to the assembly station at the right time. The most commonly used feeding systems are vibratory bowl feeders, capable of sorting nearly 90% of all components automatically.

The unsorted workpieces that have been fed in are moved by triggered vibration energy to the edge of the bowl. They then run onto a sloped helical channel on the inside or outside wall of the bowl and are moved upwards over mechanical baffles. These baffles have the task of orienting the

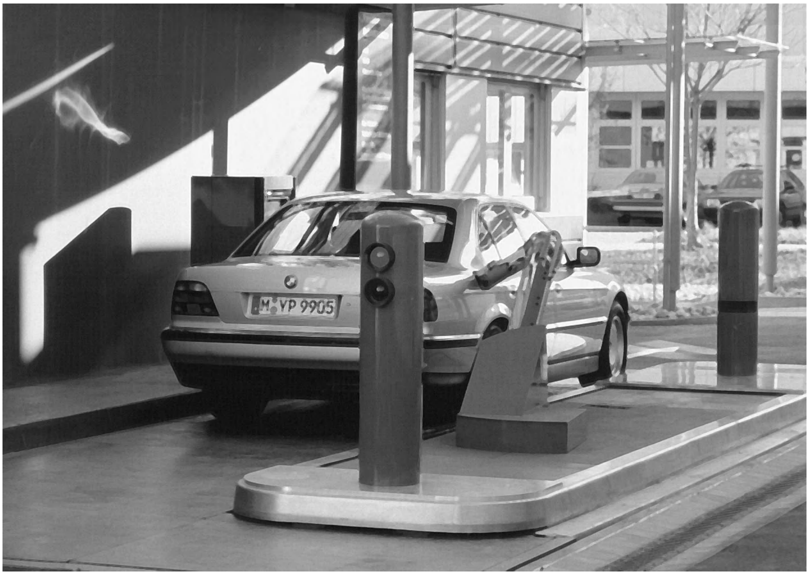


Figure 26 Refueling Station. (*Source:* IPA)

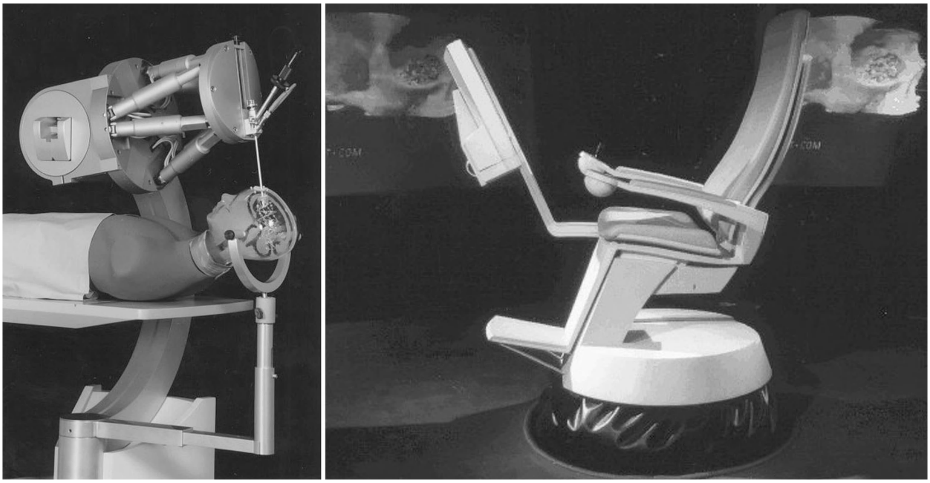


Figure 27 Operation Robot (Model number Vision OP 2015).

disoriented workpieces during the transfer movement by means of turning, twisting, tilting, or setting upright the workpieces and/or throwing back wrongly oriented workpieces into the bin. However, mechanical baffles are set up for only one type of workpiece and much retooling is required if the workpiece is changed. These pieces of equipment are therefore suitable only for large series assembly.

To increase flexibility, mechanical baffles are being replaced more and more with optical parts-recognition systems in the vibratory bowl feeder. The geometry of the various workpieces can be programmed, stored in, and retrieved from the control system if there is a change of product. This means that retooling can be kept to a minimum.

Image processing as a sensor for guiding robots is also being applied more and more in flexible workpiece-feeding systems. Isolation of unsorted workpieces is conducted on a circular system of servocontrolled conveyor belts. The correctly oriented workpieces on one of these conveyor bands

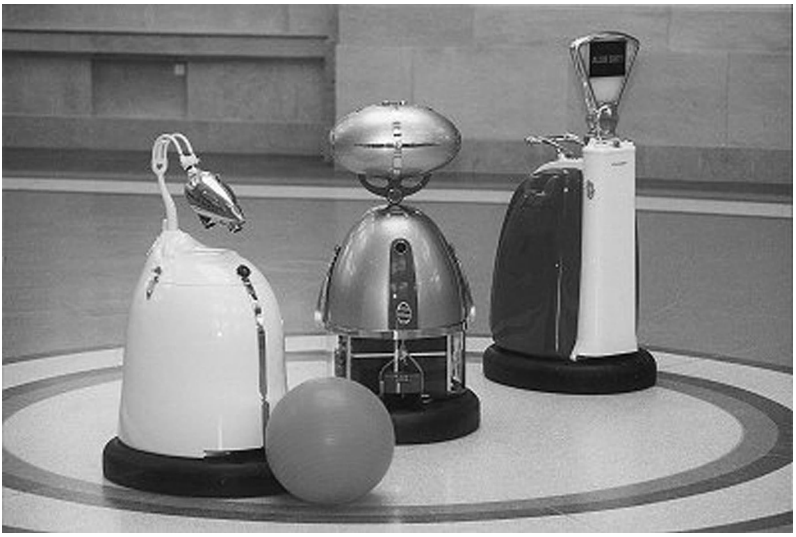


Figure 28 Entertainment Robots in the Museum of Communication, Berlin. (*Source:* IPA)

are recognized by a camera and the coordinates are determined. The robot can grab only recognized workpieces; the other workpieces remain in the circular system.

The various feeding and sorting systems are depicted in Figure 29.

9.2. Magazines

Parts that cannot be easily bowl-fed can be stored in an oriented fashion in a magazine. Magazining is the storage of workpieces in a special order for stock purposes before and after the manufacturing equipment. Magazines for the simple storage of workpieces are containers with grid elements, slot-in boards, and so on. When magazines for automatic handling are used, the magazine and forwarding functions usually blend into one function. The various types of magazine are depicted in Figure 30.

Static magazines (e.g., stack magazines) are widely used especially in component manufacturing. The magazines are simply constructed and make use of the falling or rolling motion of the workpieces. Pallet magazines are used for the storage of workpieces that are mainly in sheet arrangements and

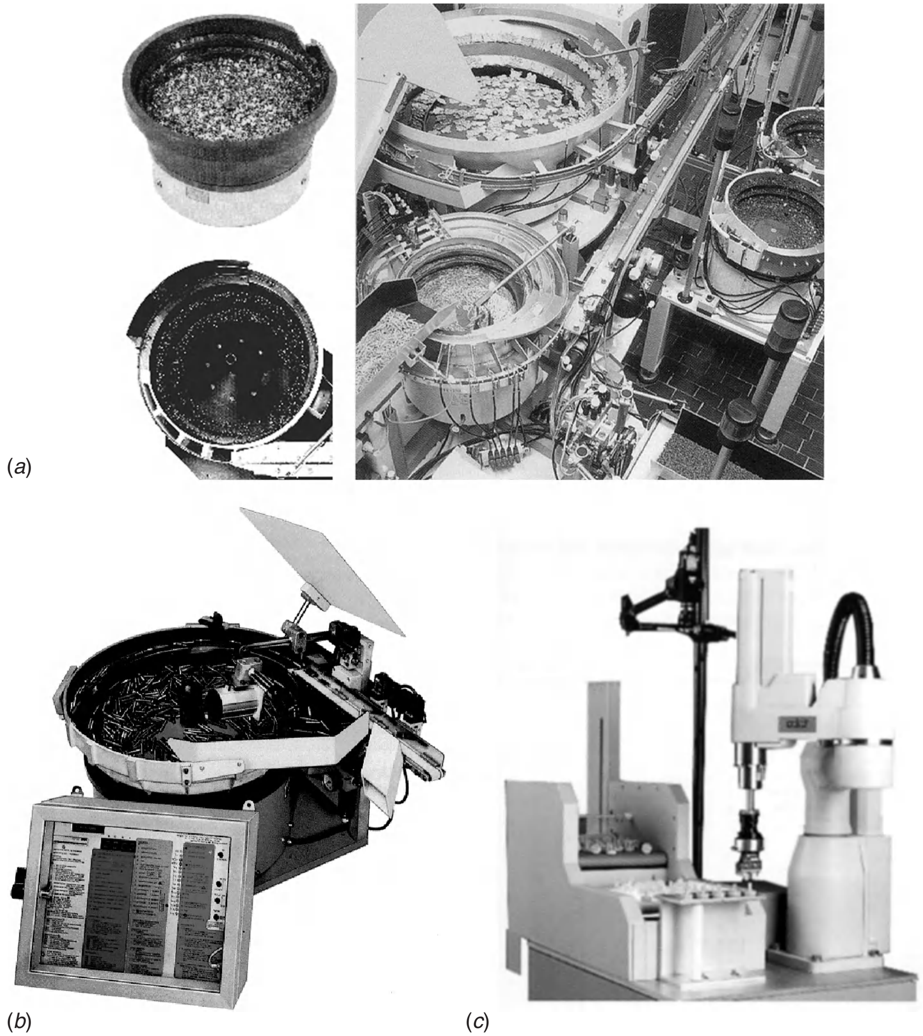


Figure 29 Examples of Standard and Flexible Part-feeding Devices. (*a*) Bowl feeders with mechanical baffles. (*Source:* RNA, MHK, Bihler) (*b*) Bowl feeder with parts-recognition system. (*Source:* MRW) (*c*) FlexFeeder. (*Source:* Adept)

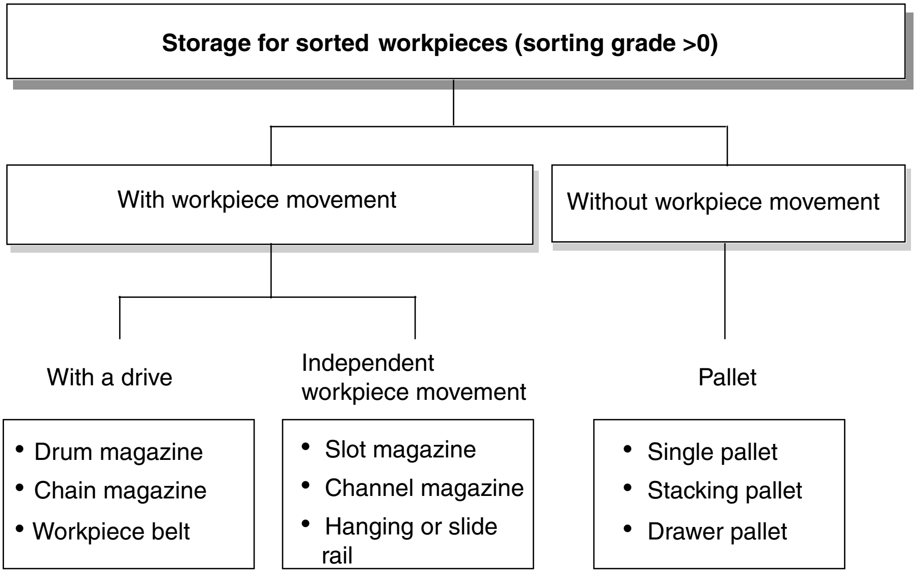

Storage for sorted workpieces (sorting grade >0)
With workpiece movement
Without workpiece movement
With a drive
Independent workpiece movement
Pallet
• Drum magazine
• Chain magazine
• Workpiece belt
• Slot magazine
• Channel magazine
• Hanging or slide rail
• Single pallet
• Stacking pallet
• Drawer pallet

Figure 30 Types of Magazines.

in a certain order. Workpiece positioning is conducted by means of a form closure. Pallet magazines are usually stackable. The pallets can be coded for controlling tasks in the material flow. They are usually used in flex-link assembly systems.

Driven and/or movable magazines are preferable when dealing with larger workpieces. These magazines are capable not only of storage but also of forwarding and/or transferring the workpieces.

9.3. Fixturing

Fixtures are the bases that securely position workpieces while the assembly system performs operations such as pick-and-place and fastening. They are generally heavy and sturdy to provide the precision and stability necessary for automated assembly. They carry a variety of substructures to hold the particular workpiece. While each fixture is different, depending on the workpiece it handles, several commonsense rules for their design are suggested:

1. Design for assembly (DFA) usually results in simpler fixture design. Simple vertical stacking of parts, for example, generally allows minimal fixtures.
2. Do not overspecify fixtures; do not demand fixture tolerances not required by product tolerances. Fixture cost is proportional to machining accuracy.
3. Fixture weight is limited by conveyor and drive capabilities. The transfer conveyor must be able to carry the fixtures. And because fixture bases get rough treatment, include replaceable wear strips where they contact the conveyor. Replacing wear strips is cheaper than replacing bases. In addition, plastic bumpers cut wear, noise, and shock.
4. Use standard toolroom components when possible. Drill bushings, pins, and clamps are available from many sources. Use shoulder pins in through-holes instead of dowel pins positioned by bottoming in the hole. Through-holes are easier to drill and do not collect dirt.
5. Make sure any fragile or vulnerable parts are easily replaceable.

9.4. Sensors and Vision Systems

Sensors provide information about the automated system environment. They are the link between the technical process and its control system and can be seen as the sense organ of the technical system. They are used for a number of tasks, such as material flow control, process monitoring and regulation, control of industrial robot movements, quality inspections and industrial metrology, and for security protection and as safeguards against collisions.

The variety of sensors can be divided into technology applications as depicted in Figure 31.

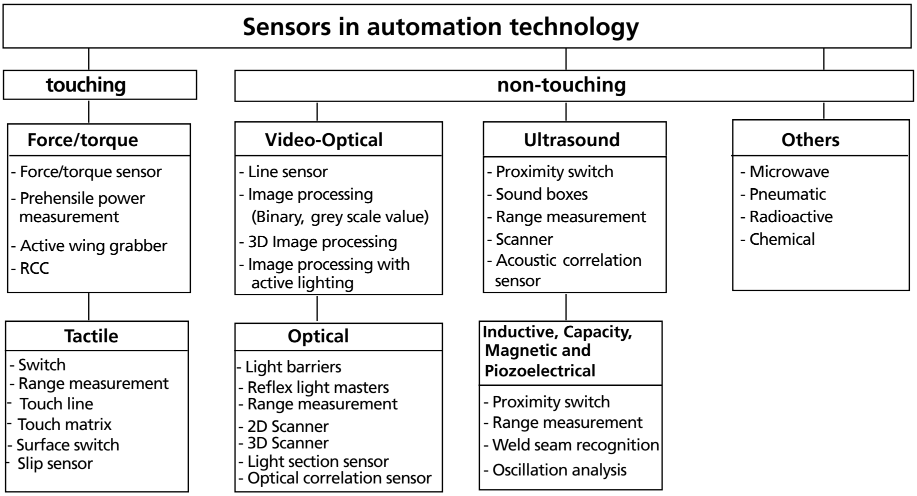

Sensors in automation technology
touching
non-touching
Force/torque
- Force/torque sensor
- Prehensile power measurement
- Active wing grabber
- RCC
Video-Optical
- Line sensor
- Image processing (Binary, grey scale value)
- 3D Image processing
- Image processing with active lighting
Ultrasound
- Proximity switch
- Sound boxes
- Range measurement
- Scanner
- Acoustic correlation sensor
Others
- Microwave
- Pneumatic
- Radioactive
- Chemical
Tactile
- Switch
- Range measurement
- Touch line
- Touch matrix
- Surface switch
- Slip sensor
Optical
- Light barriers
- Reflex light masters
- Range measurement
- 2D Scanner
- 3D Scanner
- Light section sensor
- Optical correlation sensor
Inductive, Capacity, Magnetic and Piozoelectrical
- Proximity switch
- Range measurement
- Weld seam recognition
- Oscillation analysis

Figure 31 Systemization of Sensors by Technologies.

9.4.1. Tactile Sensors

Tactile touching sensors allow simple information from the environment to be recorded by touching the objects directly. In the simplest case, the sensor is a switch that sends a binary signal when set or not set. A variety of miniaturized switches are combined with line and matrix arrangements with integrated sensor data preprocessing to enable the creation of a tactile pattern recognition similar to the human sense of touch.

9.4.2. Force/Torque Sensors

Force/torque sensors allow reaction forces and torques that occur during handling or processing procedures to be recorded. For example, when two parts with a low positional tolerance are joined in an assembly system, the respective motional corrections can be automatically executed due to reaction forces. A completed recording is usually conducted with rotational measurement strips and contains three forces and torques. Depending on the task, a smaller number may also be sufficient.

9.4.3. Video-optical Sensors

With video-optical systems, tapes of camera pictures and subsequent image processing on a computer allow statements to be made about the type, number, and position of objects at a workpiece scene. These objects are then taught to the image processing system in a specific learning process (see Figure 32).

The reflection of an illuminated workpiece passes through a lens onto the CCD chip in the camera. A 2D discrete position image is created. The camera electronically turns this image into a video signal. The image-processing computer uses a frame grabber to turn the video signal into digital images and store these in the memory. Image-processing procedures can access these image data. Despite the large amount of hardware required, the main component of image-processing systems is the image-processing software.

Image-processing systems are extremely sensitive to changes in lighting conditions. Lighting therefore plays a decisive part in the evaluation of the images. Different specific lighting procedures for gathering spatial information have been developed, such as:

- The transmitted light procedure
- The silhouette-projection procedure
- The light section procedure (triangulation)
- The structured light procedure
- The coded lighting procedure
- The Stereo image processing

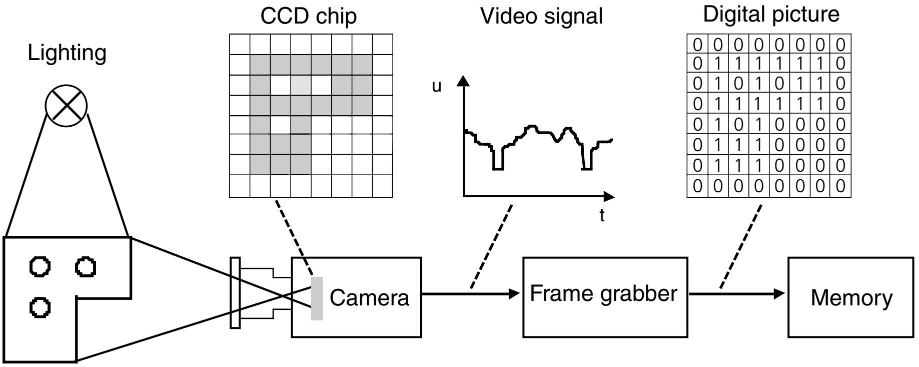


Figure 32 Basic Structure of an Image-processing System.

9.4.4. Ultrasound Sensors

An ultrasound sensor transmits a short high-frequency sound signal. If this signal is reflected by an object to the sensor and the echo is received by the sensor, it is possible to determine the distance to the object from the transfer time of the ultrasound impulse. The rate of measurement is physically limited due to the transfer time of the sound in the air. The range of sensors typically varies between 0.5 m and 6 m due to the attenuation of the sound in the air.

10. COMPUTER-AIDED METHODS FOR ASSEMBLY SYSTEMS

A large number of analytical models have been developed for planning manual and general assembly systems. Another objective has been to measure and evaluate the performance and productivity of assembly system design by applying simulation. Two essential technologies are explained in this section.

10.1. Layout Planning and Optimization

Nobody these days could imagine new products being designed and constructed without software tools. However, in the field of assembly line planning for new products, a lot of paperwork still has to be done. Most companies use computers only for documentation purposes and visualization of planning results. This means that many companies still use standard office tools for this field of activity and even execute the concept for the layout of assembly lines with a simple graphic tool. The high number of different tools and interfaces means that an actual data status for all project participants is impossible.

Meanwhile, some tools are available that are specially designed for the planning of assembly structures. The advantage of these tools is that they make continuous planning of assembly plants possible. All planning increments are entered into a central database that can show the actual status at any time.

The product structure is first defined in the tool, and then design proposals for the product are created by the software. The objective of these proposals is to create a consistent design for product assembly. MTM and UAS analysis are also supported by the software. The cycle time and the number of workstations are determined on the basis of the time data. The work steps for each station are also defined. After this, rough plans for each station can be made. This can be done using the database, which can be placed directly in the workstation. The database contains many common data such as workbenches, pressings, and robots. These are displayed as 3D graphics with additional parameters such as price and dimensions.

In addition, ergonomic investigations can be carried out in every workstation. The results of the time analyses and the appropriation of the parts can also be checked and optimized in the 3D model of the workstations (Figure 33).

The independent stations are finally united in a complete layout of the system, and the interlinking system and the bumpers are defined. In a final simulation of the material flow, the buffer sizes can be optimized and bottlenecks eliminated. The planning results are available as data and also as a 3D model of the assembly line. This provides all participants with enough data upon which to base their decisions (see Figure 34).

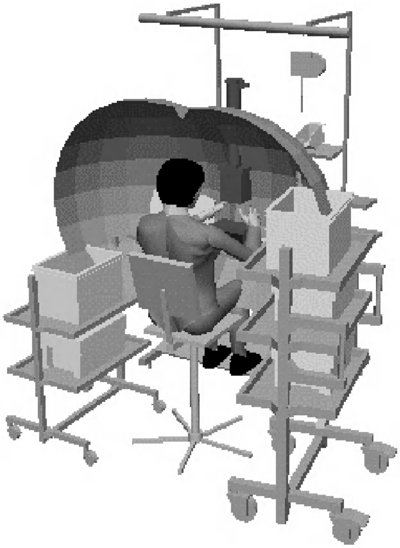


Figure 33 Software-Based Optimization of a Manual Assembly Station.

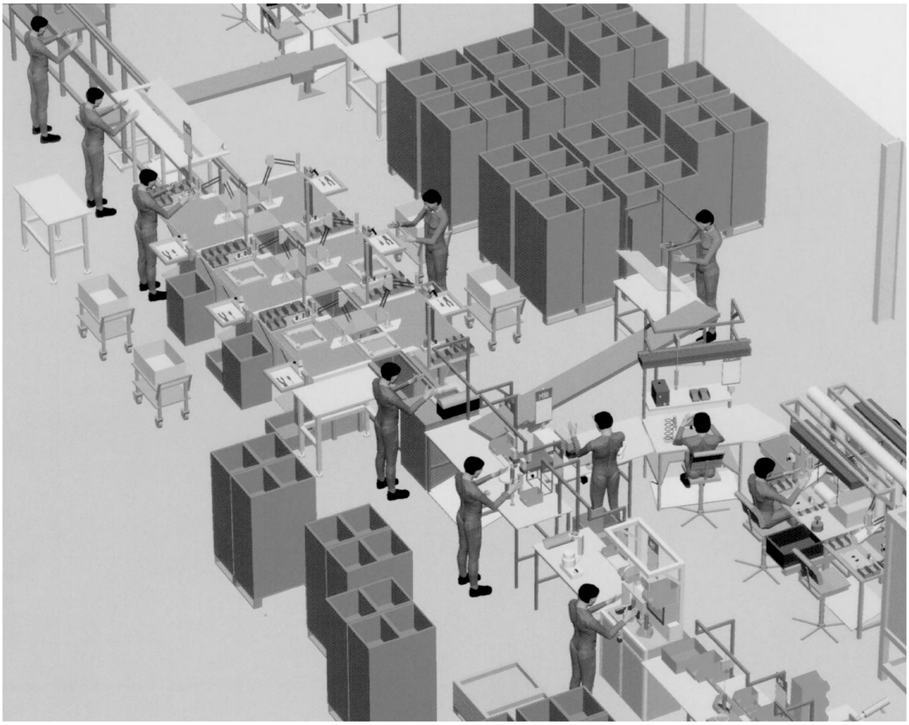


Figure 34 Software-Based Design of an Assembly Line. (*Source:* Vossloh Schwabe GmbH)

Cost analyses and efficiency calculations can also be executed with the software support. The use of software tools for the planning of assembly lines can reduce planning time and increase the quality of the results.

10.2. Simulation of Material Flow

The objective of material flow simulation is to predict the effect of actions before they are executed in order to be able to react if the results do not meet expectations. For making decisions especially in complex problems, the operations research procedure is also available. This procedure supposedly finds the most optimal solution on its own, which the simulation cannot accomplish. However, the immense number of model restrictions for complex problems with regard to model shaping and the great amount of calculation required make this practically unusable. In contrast, material flow simulation requires only a reasonable amount of modeling and calculation effort.

There are basically two application fields for material flow simulation: simulation for new plans and simulation aimed at optimizing existing systems. For new plans, the simulation aims to increase the planning security by determining initially whether the system works at all, whether the cycles of the individual stations are correct, where possible bottlenecks are in the system, and how malfunctions in individual stations affect the whole system. During the subsequent planning optimization process, the effects of malfunctions mainly determine the buffer size, which can vary considerably from the static buffer deviations. Furthermore, steps can be taken to review which measures will be needed to increase capacity by 10%, 20%, and more without the necessity of further shifts. This allows a change in the level of demand during the product's life cycle to be taken into consideration.

During operation after a product modification, new work tasks, or the assembly of a further variation on the existing system have been introduced, often a manufacturing system will no longer achieve the planned production levels and a clear deviation between target and actual production figures will become evident over a longer period of time. Due to the complexity of the manufacturing process, it is often not clear which actions would have the greatest effect. Simulation can determine, for example, what level of success an increase in the number of workpiece carriers or the introduction of an additional work cell would have. This helps in the selection of the optimal measures needed to increase productivity and the required levels of investment. Simulation of an existing system also has the major advantage that existing data pertaining to system behavior, such as actual availability, the average length of malfunctions, and the dynamic behavior of the system, can be used. These data can be determined and transposed to the model relatively easily, which makes the model very realistic.

For a simulation study to be executed, a simulation model must first be made that represents a simplified copy of the real situation. The following data are required: a scaled layout of the production, the cycle times of each of the processing steps, the logistic concept including the transportation facilities and the transportation speeds, the process descriptions, and data on the malfunction profiles, such as technical and organizational interruption times and the average length of these times. Once these data has been entered into the model, several simulations can be executed to review and, if necessary, optimize the behavior of the model.

The following example shows the possibilities of material flow simulation. In the model of an assembly system which is designed to produce 140 pieces/hr, the number of the workpiece carriers is increased. The result is a higher production level at some of the examined stations, but at some locations the higher number of workpiece carriers cause a blockage of stations that have successor stations with a higher cycle time. Thus, the increase of workpiece carriers has to be done in small steps with a separate simulation after each step to find the optimum. In this example, a 10% increase in the number of workpiece carriers and the installation of a second, parallel screwing station lead to a production increase from 120 to 150 pieces/hr.

The model reaches its optimum when each of the parameters is specifically modified and a combination of different measures is achieved. Aside from the creation of the model itself, this is the real challenge.

11. ASSEMBLY IN INDUSTRY: APPLICATIONS AND CASE STUDIES

Assembly is fundamental in almost all industrial activities. Assembly applications in several important industries are illustrated in the following case studies.

11.1. Automotive Assembly

More than half of the industrial robots throughout the world are used in the automotive industry. These robots are utilized in a number of different manufacturing sectors. The main areas of application are in car-body construction, mainly in spot welding and painting. Assembly lines are also being more and more robotized in the automotive industry. This is essential in the case of complex joining processes requiring a high standard of quality, which can be guaranteed only by robots.

But in order to promote automation in automotive assembly effectively, two conditions must be met first:

1. Assembly-oriented automotive design
2. Preassembly of as many subunits as possible

A few application areas of automatic assembly in the automotive industry (see Figures 35, 36, 37) are shown in the following examples.

11.2. Assembly of Large Steering Components

This non-synchronous pallet-based, power and free system manufactures a large steering component for the automotive industry (see Figure 38). Cycle time is approximately 20 seconds, three shifts per day. Unique pallet design helped this customer automate the assembly process while minimizing material handling of the final product. An important issue for the manufacturing process was the testing of each unit throughout the assembly sequence. Tests performed are leak, torque, function, and final performance. The automation system also featured manual stations for placing specific parts and for rework and evaluation.

11.3. Automatic Removal of Gearboxes from Crates

The combination of innovative sensory analysis with modern industrial robots is opening up more new fields of application for automation technology. Automatic unloading of workpieces from crates is a problem that automation technology researchers have been addressing for a long time. However, only recently have usable systems for the rough production environment been developed. The requirements for these unloading systems are:

- Localization accuracy better than 5 mm
- Cycle time maximum 24 sec (in accordance with subsequent manufacturing processes)
- No predefined lighting conditions for the recognition sensors
- Lowest possible tooling-up requirements for new types of workpiece
- If possible, no additional pieces of equipment for the recognition sensors in the work cell

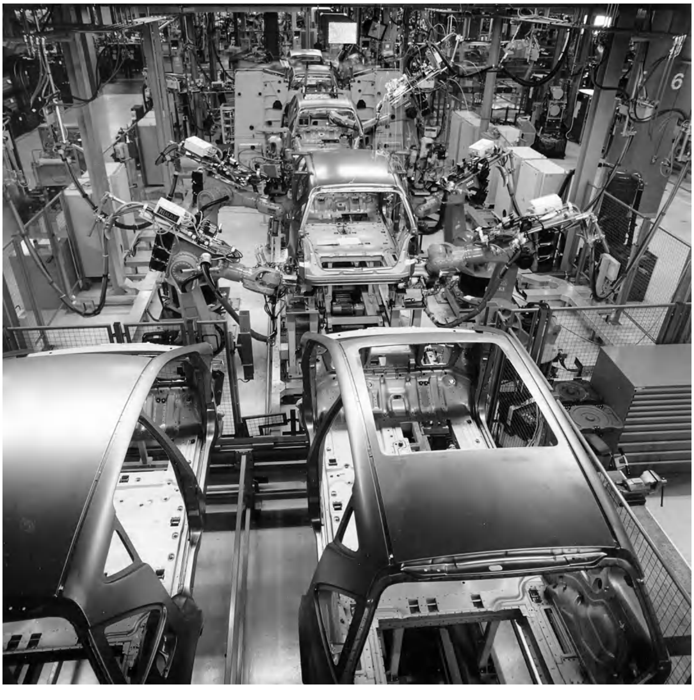


Figure 35 Robotized Spot Welding Line (*Source:* KUKA Roboter GmbH)

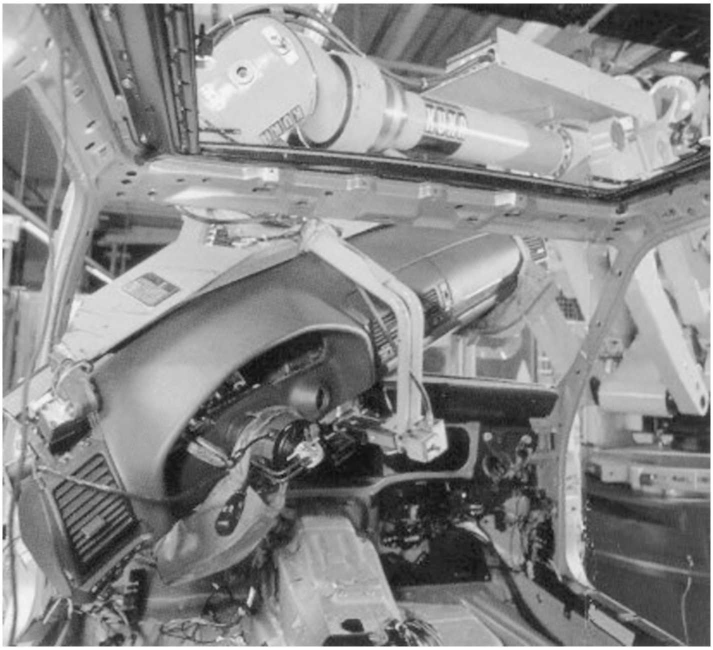


Figure 36 Assembly of Driver Control Instruments. (*Source:* KUKA Roboter GmbH)

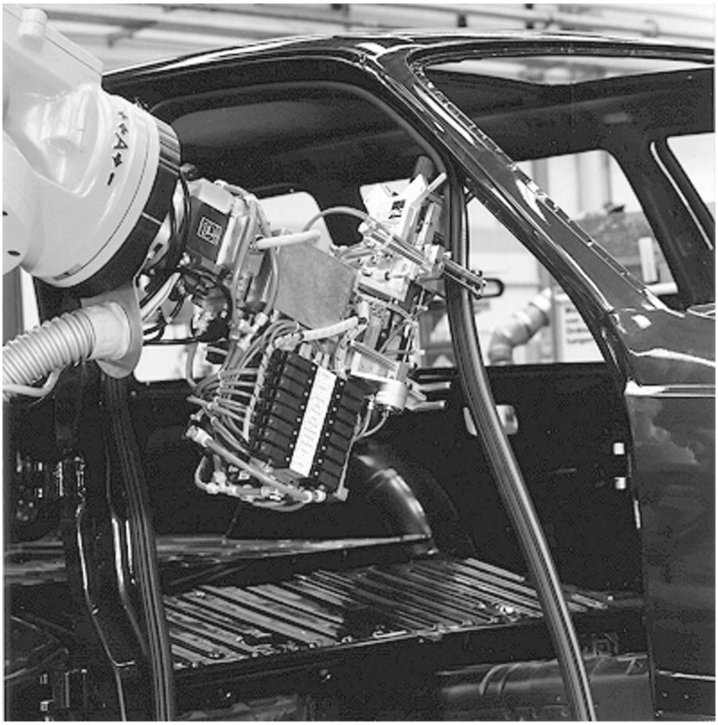


Figure 37 Assembly of Car Door Rubber Profiles (*Source:* IPA)

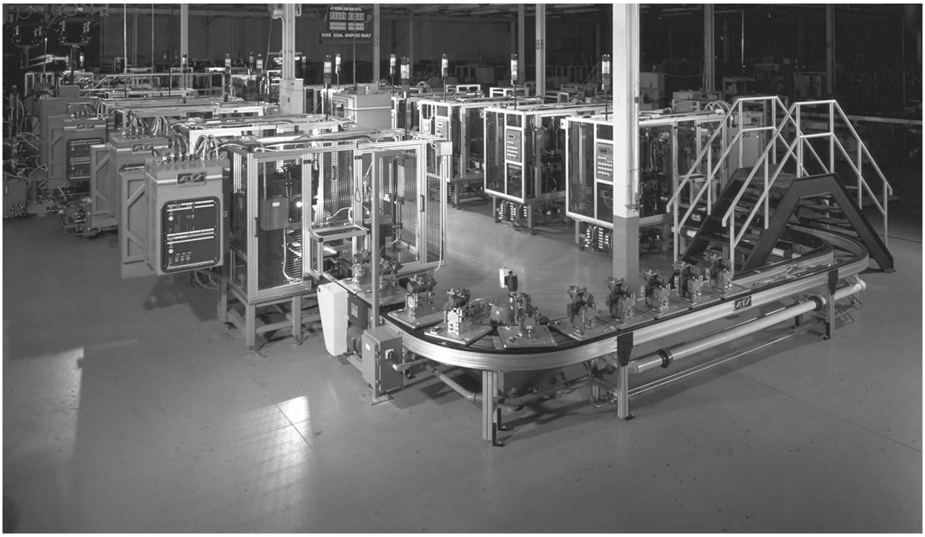


Figure 38 Assembly Line for Large Steering Components. (*Source:* ATS Inc.)

In the case outlined here, the processed and partly sorted cast parts are automatically unloaded from the crates. In order to avoid the cast parts damaging each other, they are stacked in the crates on cushioned layers (see Figure 39).

After the filled crate has been delivered to the work cell, the robot first measures the four upper edges with the help of a 2D triangulation sensor that is attached to the robot. This measurement is used to calculate the exact location and height of the crate.

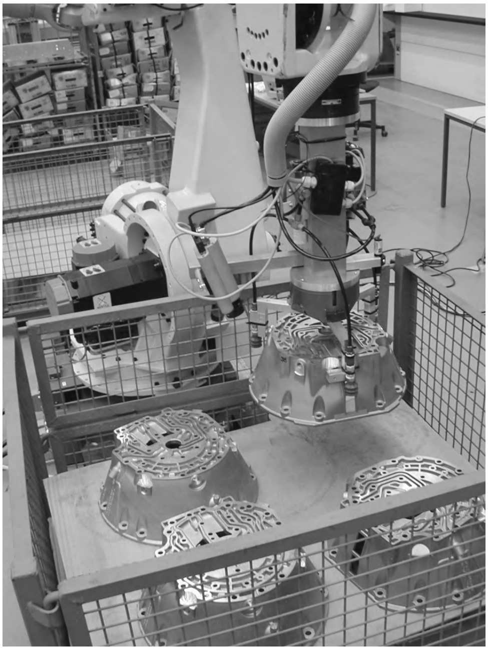


Figure 39 Robot Cell for Unloading Gearboxes from Crates. (*Source:* IPA)

The robot sensor is guided over the workpieces in the crate to determine the position of each of the workpieces. The workpiece surface that has been recognized is used to determine the three positions and orientations of the workpiece. The approach track with which the robot can grab and pick up the cast part is prepared from the workpiece position. The whole approach track is calculated in advance and reviewed for possible collisions. If the danger of a collision exists, alternative tracks are calculated and the next cast part is grabbed if necessary. In this way the system maintains a high level of operational security and has sound protection from damage.

11.4. Electronic Assembly

11.4.1. Assembly of an Overload Protector

This pallet-based assembly system produces a safety overload protection device and was designed with multiple zones (see Figure 40). The transport system is a flex-link conveyor with a dual-tooled pallet for cycle time consideration. The cycle time for the system is 1.2 sec. Tray handling and gantry robots were utilized for this application due to the high value of the component. The line features unique coil winding, wire-stripping technologies, and sophisticated DC welding for over 20 different final customer parts for the electronic assembly market.

11.4.2. Assembly of Measuring Instruments

The production of measuring instruments is often characterized by small batches, short delivery times, and short development times. Automated assembly of the measuring instrument cases enables products to be manufactured at lower costs. The cases are made of several frame parts. Each frame part is separated by a metal cord that provides screening against high-frequency radiation.

The different kinds of cases have various dimensions (six heights, three widths, and three depths). Altogether there are about 1,000 product variants with customer-specific fastening positions for the measuring instrument inserts. Therefore, the assembly line is divided into order-independent preassembly and order-specific final assembly (see Figure 41).

In the preassembly cell are automatic stations for pressing in different fastening nuts and screwing in several threaded bolts. An industrial robot handles the frame parts. After removing the frames from the supply pallets, the robot brings the frame into a mechanical centering device for fine

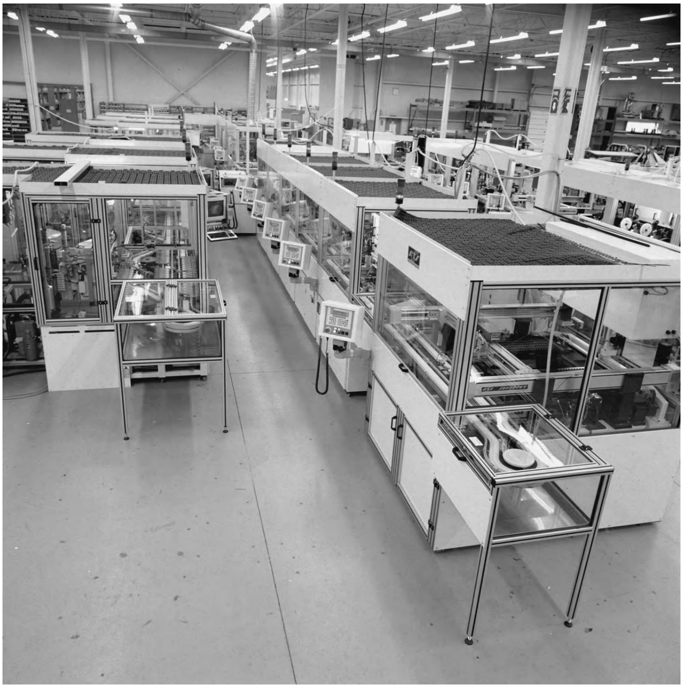


Figure 40 Assembly Line for Safety Overload Protection Device. (*Source:* ATS Inc.)

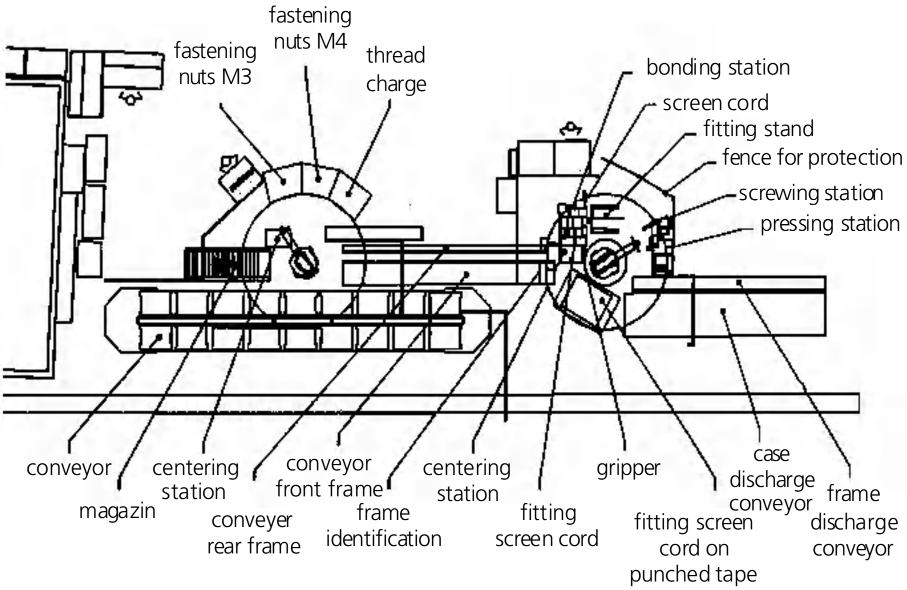

fastening nuts M3
fastening nuts M4
thread charge
bonding station
screen cord
fitting stand
fence for protection
screwing station
pressing station
conveyor
magazin
centering station
conveyer rear frame
conveyor front frame
frame identification
centering station
fitting screen cord
gripper
fitting screen cord on punched tape
case discharge conveyor
frame discharge conveyor

Figure 41 Layout of the Pre- and Final-Assembly Cell.

positioning and the elimination of tolerances in the pallets. In this way it is possible to achieve very exact positioning of the frame in the robot gripper.

During the pressing operation, the robot positions and fixes the frame in the press station (press force of about 3000 N). A compliance system is integrated in the gripper to eliminate tolerances in dimensions of the frame. The robot also positions and fixes the frames in the screwing station. There is a system to control the rotation angle, screwing torque, and screwing depth to reproduce a screwing depth of 0.1 mm.

After finishing preassembly, the robot places the frames on a conveyer system. The conveyer links the pre- and final-assembly cells. The final-assembly cell work steps are (see Figure 42):

- Fitting the metal cord for high-frequency screening into the frames
- Order-specific pressing of fastening elements
- Screwing together all frames to produce the finished case
- Lettering the finished case

The difficulty in assembling the metal cord is that these parts have no rigidity and can only transmit tractive forces. Two metal cords with different diameters (2.0 mm and 3.0 mm) have to be fitted into four fundamentally different running slots. To achieve the required stability of the cord in the rear-frame slot it is necessary to put in adhesive points. A cord-cutting system is integrated into the robot tool to achieve the right length of the cord (depending on the dimensions of the frames), and an adhesive dispensing system is integrated into the tool for placing the adhesive spots into the slots.

After the metal cord has been fitted into the different frames (front inside, front outside, rear, and side ledge), the fasteners for the inserts are pressed in order-specific positions into the side ledges. For this operation the robot takes the ledge with the required length out of a magazine and brings it into a press station. A guide rail defines the exact position. The fasteners are blown automatically from the feeder through a feed pipe to the press position.

The subsequent screwing of all frame components is also executed by the robot. The robot moves to the screwing position. The screws are blown automatically through the feed pipe from the feeder to the screwing tool. The rotation angle and screwing torque are monitored during the screwing operation to achieve a perfect result.

Depending on the construction of the case, it is important to have accessibility from four directions during the complete assembly process. Therefore, a clamping device that can turn the case in all

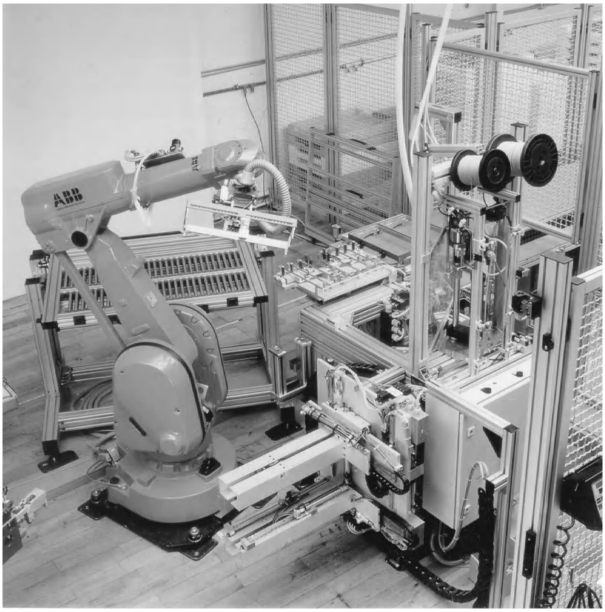


Figure 42 Final-Assembly Cell with Flexible Clamping System.

required positions is essential. The clamping device is made up of standard components that can clamp more than 25 cases with different dimensions. The robot arm has an automatic tool-changing system for changing tools quickly; each tool can be picked up within a few seconds.

Experience with the assembly cells outlined above has shown that even assembly tasks with extensive assembly steps producing less than 10,000 units per year can be automated in an efficient way.

11.4.3. Assembly of Luminaire Wiring

For the optimal economic assembly of luminaire wiring, a simultaneous development of new products and process techniques was necessary. A new method developed for the direct assembly of single leads has proved to be a great improvement over the traditional preassembly of wires and wiring sets. Single leads are no longer preassembled and placed in the end assembly. Instead, an endless lead is taken directly from the supply with the help of a newly developed fully automated system and built into the luminaire case. This greatly simplifies the logistics and material flow inside the company and reduces costs enormously.

The IDC technique has proved to be the optimal connection system for the wiring of luminaires. Time-consuming processes like wire stripping and plugging can now be replaced by a process combining pressing and cutting in the IDC connector. The introduction of this new connection technique for luminaire wiring requires modification of all component designs used in the luminaires.

Fully automatic wiring of luminaires is made possible by the integration of previously presented components into the whole system. Luminaire cases are supplied by a feeding system. The luminaire case type is selected by the vision system identifying the bar code on the luminaire case. Next, luminaire components are assembled in the luminaire case and then directed to the wiring station (Figure 43).

Before the robot starts to lay and contact the leads, a camera, integrated in the wiring tools, is positioned above the required position of the component designated for wiring. Each component is identified and its precise position is controlled by vision marks in the shape of three cylindrical cavities in the die casting of the IDC connector. Any deviation of the actual position from the target position is calculated and transmitted as compensation value to the robot's evaluation program.

Quality assurance for the processes, especially for pressing and contacting of the conductor in the IDC connector, is directly controlled during processing. When the wire is inserted into the connector, a significant and specific force gradient is shown within narrow tolerances. This measured gradient can be easily compared and evaluated immediately after pressing with the help of a reference

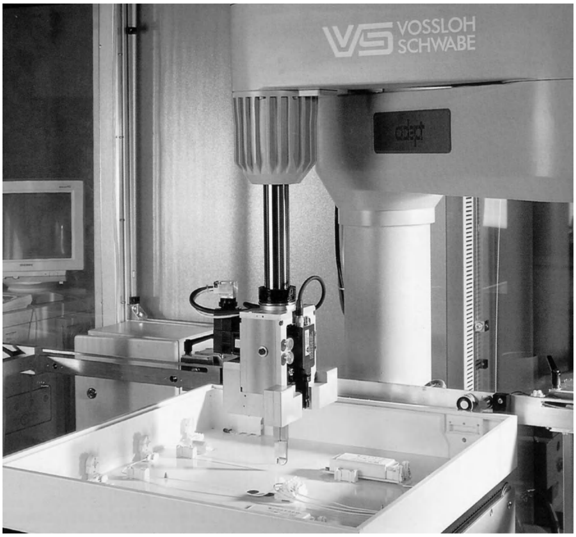


Figure 43 Assembly Cell for Automatic Wiring of Luminaires. (*Source:* Vossloh Schwabe GmbH)

graph. In the case of nonconformity, the luminaire is isolated, taken out of the material flow system, and directed to a manual work station for rework.

Depending on the geometry of the luminaires, it can be assumed that the application of this assembly system can reduce the share of wage costs for each luminaire by up to 65%.

11.4.4. Assembly of Fiberoptic Connectors

While communications and information technology are rapidly evolving, transmission media requirements are increasing more and more. The enormous increase in data rates, the increase in data transmission speeds, and the high level of reliability of EMI mean that copper-bounded transmission technology has reached its technical and economical limits because of its susceptibility to interference and limited transmission rates. Fiberoptic technology, by contrast, has already shown its potential and advantages in telecommunication and computer network (LAN) applications. Because manufactured units are increasing, automation of the fiberoptic connector assembly is becoming a more and more exciting field.

Fiberoptic cables are currently assembled almost completely manually. Semiautomatic devices are available on the market for the premanufacturing process. After this process is completed, the ceramic ferrule of the connector is filled with glue, then the glass fiber is introduced into the ferrule hole. This assembly sequence is executed only manually at present. However, because no measuring systems are available for this purpose, manual assembly cannot ensure that the critical joining force will not be exceeded. If this happens, the fiber will break off. The fiber will be damaged and the connector will become waste.

Figure 44 shows the construction of an ST connector and the simplex cable that is used.

An automated fiberoptic connector assembly, especially in the inserting process, is generally obstructed by following factors: glass fiber sensitivity to breaks, nonrigid cable, complex cable construction, inserting dimensions in the μm range, small backlash between fiber and ferrule (0–3 μm), varying glue viscosity, and maximal inserting forces in the mN range.

Figure 45 shows a prototype assembly cell where the above-mentioned subsystems are integrated.

11.5. Microassembly

Increasing miniaturization of mechanical components and rising demands in the precision mechanics and microtechnology industries are creating high demands on assembly automation. Today automated solutions only exist for mass series production systems (the watchmaking industry, the electronics industry). Automation of assembly tasks for small and medium-sized series is still rare. Flexible automated microassembly systems are required there due to the large number of different types and variants.

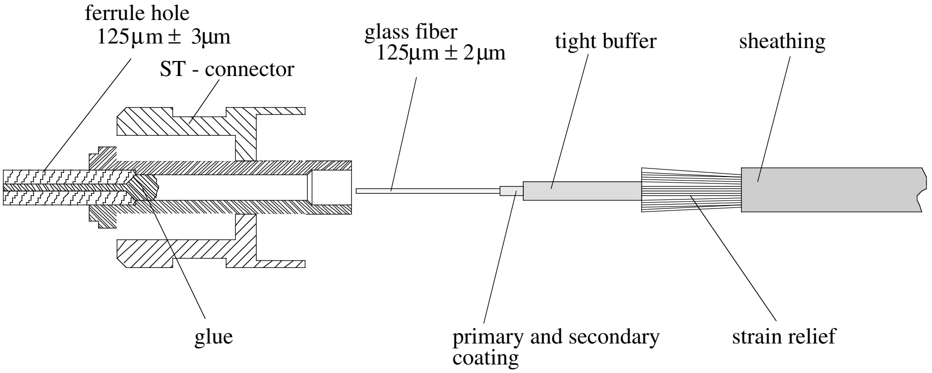


Figure 44 Construction of Connector and Cable.

The assembly of precision mechanics and technical microsystem components often requires joining work to be conducted in the micrometer range. Intelligent robot grabbing systems have been developed in order to achieve the join accuracy required. Motion-assisted systems are particularly suited for flexible offsetting of tolerances in short cycles.

Figure 46 shows an example of a oscillation-assisted joining robot tool for precision and microassembly of small planetary gears. Assisted by the adjustable oscillation in the grabber and the regulated tolerance offset, each of the gearwheels and the gearwheel housing can be process assembled very securely by means of the integrated fine positioning.

The special features of the joining tool are the joining accuracy of up to 2 μm that can be achieved, the process-integrated offsetting of orientation and position deviations (up to ±1 mm), and the high level of adjustment as well as flexibility essential for performing the different joining tasks. The robot tool is used especially for phaseless bolt-in-hole joining tasks with accuracy requirements in the micrometer range and for the assembly of complex toothed wheel pairs.

11.6. Food Industry

The number of convenience products sold annually in recent years has increased markedly. Particularly in everyday life and in cafeterias the trend is toward meals that can be prepared more quickly and conveniently. According to the experts at *Food Consulting,* the share of convenience products in

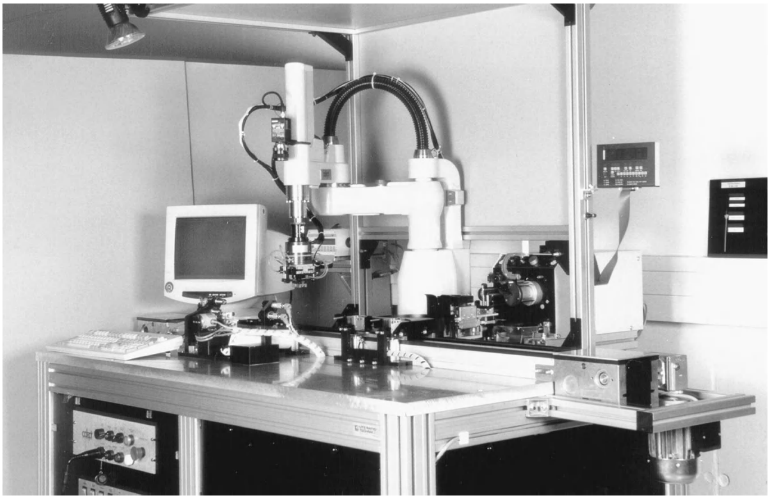
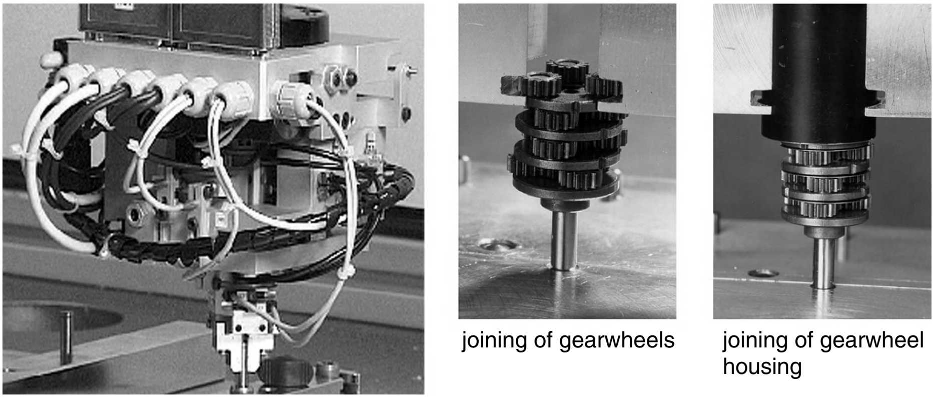


Figure 45 Prototype Assembly Cell for Manufacturing Fiberoptic connectors.



joining of gearwheels

joining of gearwheel housing

Figure 46 Oscillation-Assisted Joining Tool. (*Source:* IPA)

meat products will increase from 15% at present to above 30% in the next two years. The food industry is changing into a preparation industry and its vertical range of manufacture will increase dramatically. This will necessitate the introduction of automated manufacturing processes and thus enable the use of new automation and robot systems.

In addition to increasing productivity, automated systems allow improved levels of hygiene and thus extending the sell-by date of food products. They also allow ever-increasing consumer demands for more transparency regarding product origins to be met.

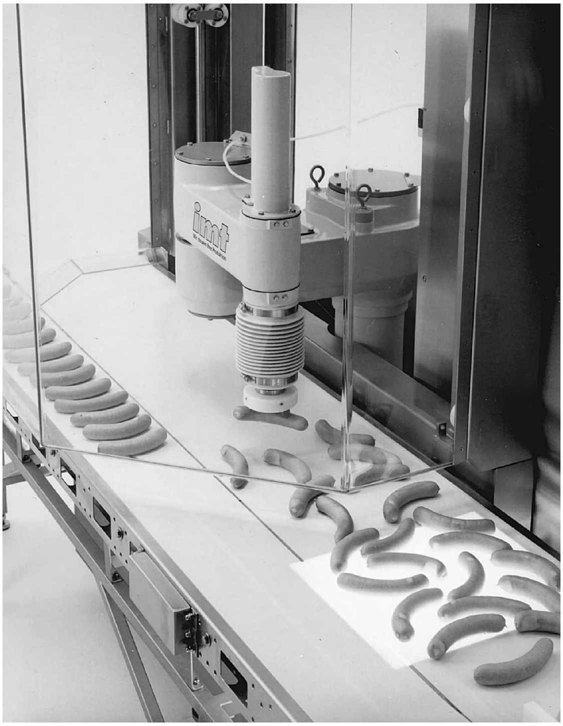


Figure 47 Sorting Sausages with Robots. (*Source:* imt)

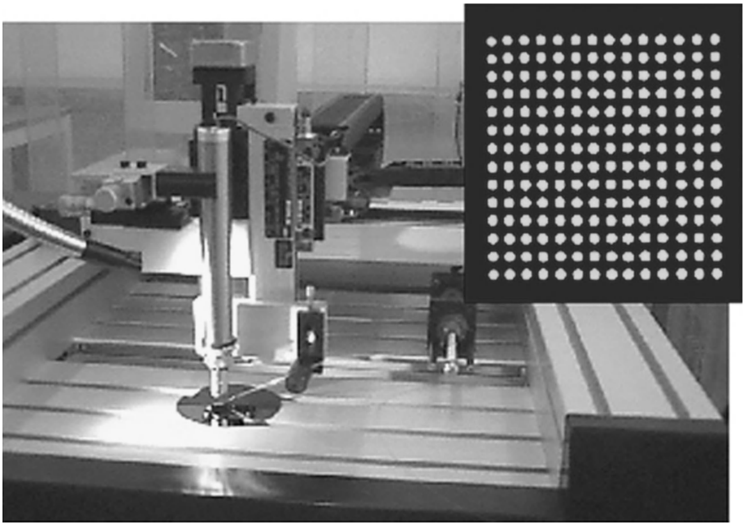


Figure 48 Automated Arrayer Based on a Gantry Robot. (*Source:* IMNT)

Figure 47 shows a robot in use in the food-packaging sector. At the end of a production line, the unsorted sausages that arrive on the conveyor belt are recognized by an image-processing system and automatically sorted according to the final packaging unit with the help of the robot.

11.7. Pharmaceutical and Biotechnological Industry

In the pharmaceutical industry, large sums of money are used for preclinical and clinical research. The development of a drug costs millions of dollars. Hundreds of tests have to be conducted. One of the innovative tools in the field of drug discovery is microchip technology. Microchips require less reagent volume, make analytical processes run faster because of their smaller size, and allow more sensitive detection methods to be implemented. In this way, they reduce costs, save time, and improve quality. Microchips come in two main categories: chips based on microfluidics, with components such as pumps, mixers, microinjectors, and microarray chips, which have numerous locations of samples, such as DNA samples, on their surface. Most of the interest now seems to be focused on this second category.

Figure 48 shows an arrayer for low-volume spots. With a capillary-based tip printing method, volumes in the lower pl-range can be produced. These very small drop volumes give the opportunity to use more sensitive methods of detection and produce high-density arrays. The automated arrayer is based on a gantry robot. Its three axes are driven by electronic step drives that have a positional accuracy of $\pm 1\ \mu$m and a speed up to 300 mm/sec. The work area is 300 × 300 mm and the total dimension is 500 × 500 mm. In order to have a controlled environment, the system is operated in a clean room.

So far, commercial available arraying systems are useful only for R&D applications and research labs. They have not achieved wide commercial availability.

REFERENCES

Turner, W. C. (1993), *Introduction to Industrial and Systems Engineering,* 3rd Ed., Prentice Hall, Englewood Cliffs, NJ.

ADDITIONAL READING

Armada, M., and de Santos, P. G., “Climbing, Walking and Intervention Robots,” *Industrial Robot,* Vol. 24, No. 2, pp. 158–163, 1997.

Barbey, J., Kahmeyer, M., and Willy, A., *Technikgestaltung in der flexibel automatisierten Serienmontage*, Fortschritt-Berichte VDI, Reihe 2, Vol. 262, VDI, Düsseldorf, 1992.

Bässler, R., "Integration der montagegerechten Produktgestaltung in den Konstruktionsprozess," Dissertation, Springer, Berlin, 1988.

Bito, J. F., and Rudas, I. J., "Industrial Robots: Automatic Handling and Assembly Equipment in Present and Future Manufacturing Systems," in *Proceedings, Workshop on Robotics* (Budapest, September 19–20, 1994).

Black, A., *An Industrial Strategy for the Motor Vehicle Assembly and Component Sectors,* UCT Press, Rondebosch, RSA.

Boothroyd, G., *Assembly Automation and Product Design,* Marcel Dekker, New York, 1992.

Boothroyd Dewhurst, Inc., *Proceedings of the International Forum on Design for Manufacture and Assembly* (Newport, RI, June 12–13, 1995), Boothroyd Dewhurst, Inc., Wakefield, RI.

Boothroyd Dewhurst, Inc., *Proceedings of the International Forum on Design for Manufacture and Assembly* (Newport, RI, June 8–10, 1998), Boothroyd Dewhurst, Inc., Wakefield, RI.

Boothroyd, G., and Dieter, G. E., *Manufacturing Engineering and Materials Processing,* 1998.

Cooke, D. S. Hewer, N. D., White, T. S., Galt, S. Luk, B. L., and Hammond, J., "Implementation of Modularity in Robug IV," in *Proceedings of First International Symposium on Climbing and Walking Robots* (Brussels, November 26–28, 1998).

Craig, J. J., *Introduction to Robotics*, 2nd Ed., Addison-Wesley, Reading, MA, 1989.

De Vries, W. R., and Yeong, M. Y., "A Methodology for Part Feeder Design," *Annals of the CIRP,* Vol. 43, No. 1, pp. 19–22, 1994.

EUREKA, Project EU 1002, Final Report, FAMOS-Flexfeed: Flexible Feeder Development, 1996.

Flowers, P., and Overbeck, J., "An Improved System for Making DNA Microarrays," in *2nd Workshop on Methods and Applications of DNA Microarray Technology*, Tucson, AZ, Genetic MicroSystems, 1998.

Göpel, W., Hesse, J., and Zemel, J. N., *Sensors: A Comprehensive Survey,* Vol. 1-9, VCH, Weinheim, 1996.

Hesse, S., *Vorrichtungen für die Montage: Praxisbeispiele für Planer, Konstrukteure und Betriebsingenieure,* Expert, Renningen, Malmsheim, 1997.

Jones, R., "Gene Chips for Every Laboratory," *International Biotechnology Laboratory,* No. 6, pp. 22–24, 1999.

Karlsson, J., in *World Robotics 1998,* International Federation of Robotics Staff, United Nations Publications, Geneva, pp. 13–32, 1998.

Katz, Z., "Advances in Manufacturing Technology: Focus on Assembly Systems," in *Proceedings of CIRP International Seminar on Manufacturing Systems* (Johannesburg, May 15–17, 1996).

Koller, S., "Direktmontage von Leitungen mit Industrierobotern," Dissertation, Springer, Berlin, 1994.

Lotter, B., *Manuelle Montage: Planung—Rationalisierung—Wirtschaftlichkeit,* VDI, Düsseldorf, 1994.

Müller, E., Spingler, J., and Schraft, R. D., "Flexible Montage feinwerk-und mikrotechnischer Bauteile," *Werkstattstechnik,* Vol. 89, No. 4, pp. 151–153, 1999.

Nof, S. Y., Wilhelm, W. E., and Warnecke, H.-J., *Industrial Assembly,*, 1st Ed., Chapman & Hall, London, 1997.

Rose, D., and Lemmo, T., "Challenges in Implementing High Density Formats for High Throughput Screening," *Laboratory Automation News,* No. 2, pp. 12–19, 1997.

Scholpp, C., "Automatisierte Montage von Glasfaser-Lichtwellenleiterkabeln in Steckverbinder," in *IPA-IAO Forschung und Praxis,* Jost-Jetter, Heimsheim, No. 308, 2000.

Scholpp, C., and Ankele, A., "Flexible Automation of Fiber Optical Connectors assembly," in *Proceedings of SPIE: Micro-Optics Integration and Asemblies* (San José, CA, January 29–30, 1998), pp. 103–113.

Schraft, R. D., Ed., *Proceedings of Institute for International Research Conference, Zukunft Montage: Wirtschaftlich, flexibel, kundenorientiert* (Böblingen, June 11–12, 1997), IIR, Sulzbach.

Schraft, R. D., and Kaun, R., *Automatisierung: Stand der Technik, Defizite und Trends,* Verlagsgruppe Handelsblatt GmbH, WirtschaftsWoche, Düsseldorf, 1999.

Schraft, R. D., Neugebauer, J., and Schmierer, G., "Service Robots: Products, demonstrators, visions," in *Proceedings of ICARCV '98, Fifth international conference on Control, Automation, Robotics and Vision* (Singapore, December 8–11, 1998).

Schraft, R. D., Spingler, J., and Wössner, J. F., "Automated High-Frequency Sealing in Measuring Instruments," *Robotic and Manufacturing Systems,* pp. 705–710, 1996.

Schraft, R. D., Wolf, A., and Schmierer, G., "A Modular Lightweight Construction System for Climbing Robots," in *Proceedings of First International Symposium on Climbing and Walking Robots* (Brussels, November 26–28 1998).

Schulte, J., "Automated Mounting of Connectors to Fiber Optic Cables," in *Proceedings of 40th International Wire & Cable Symposium* (St. Louis, November 18–21, 1991), pp. 303–308.

Takeda, H., *Automation ohne Verschwendung,* Verlag Moderne Industrie, Landsberg, 1996.

Takeda, H., *Das System der Mixed Production,* Verlag Moderne Industrie, Landsberg, 1996.

Taylor, R. H., Lavealle, S., Burdea, G. C., and Mosges, R., *Computer-Integrated Surgery: Technology and Clinical Applications,* MIT Press, Cambridge, MA, 1995.

Tsuda, M., Yuguchi, R., Isobe, T., Ito, K., Takeuchi, Y., Uchida, T. Suzuki, K., Masuoka, T., Higuchi, K., and Kinoshita, I., "Automated Connectorizing Line for Optical Cables," in *Proceedings of 43rd International Wire and Cable Symposium* (Atlanta, November 14–17, 1994), pp. 781–789.

Wapler, M., Weisener, T., and Hiller, A., "Hexapod-Robot System for Precision Surgery," in *Proceedings of 29th International Symposium on Robotics* (*ISR '98*) (Birmingham, UK, April 27–30, 1998), DMG Business Media Ltd., Redhill, UK.

Wapler, M., Neugebauer, J., Weisener, T., and Urban, V., "Robot-Assisted Surgery System with Kinesthetic Feedback," in *Proceedings of 29th International Symposium on Robotics* (*ISR '98*) (Birmingham, UK, April 27–30, 1998), DMG Business Media Ltd., Redhill, UK.

Warnecke, H.-J., et al., *Der Produktionsbetrieb 1: Organisation, Produkt, Planung,* Springer, Berlin, 1995.

Willemse, M. A., "Interactive Product Design Tools for Automated Assembly," Dissertation, Delft University of Technology, Fab's Omslag, Delft, Netherlands, 1997.

Xu, C., "Worldwide Fiberoptic Connector Market and Related Interconnect Hardware," in *Proceedings of Newport Conference on Fiberoptics Markets* (Newport, RI, October 20–2, 1997), p. 17.

Yeh, C., *Handbook of Fiber Optics: Theory and Applications*, Academic Press, San Diego, 1990.

CHAPTER 13

Assembly Process

K. FELDMANN
University of Erlangen-Nuremberg

1. CURRENT DEVELOPMENTS 402
  1.1. General Developments in Assembly 402
  1.2. Impact of Electronics on Assembly 404
2. ASSEMBLY TECHNOLOGIES AND SYSTEMS 407
  2.1. Basic Structures of Assembly Systems 407
  2.2. Joining Technologies 409
    2.2.1. Classification and Comparison of Joining Technologies 409
    2.2.2 Bolting 410
    2.2.3. Riveting/Clinching 411
    2.2.4. Sticking 412
    2.2.5. Welding 413
  2.3. Peripheral Functions 413
    2.3.1. Handling Devices 413
    2.3.2. Grippers 413
    2.3.3. Feeding Principles 415
    2.3.4. Linkage 415
  2.4. Manual Assembly Systems 416
    2.4.1. Description of Manual Assembly Systems and Their Components 416
    2.4.2. Criteria for the Design of Manual Assembly Systems 417
  2.5. Automated Assembly Systems 418
  2.6. Flexible Assembly Systems 419
    2.6.1. Assembly of Different Versions of a Product in Changing Amounts 419
    2.6.2. Flexible Handling Equipment 420
    2.6.3. CAD-CAM Process Chain 420
  2.7. Control and Diagnosis 422
3. ELECTRONIC PRODUCTION 423
  3.1. Process Chain in Electronic Production 423
  3.2. Electronic Components and Substrate Materials 423
  3.3. Application of Interconnection Materials 424
  3.4. Component Placement 425
    3.4.1. Kinematic Concepts 425
    3.4.2. Classification of Placement Systems 425
    3.4.3. Component and PCB Feeding 426
    3.4.4. Measures to Enhance Placement Accuracy 428
  3.5. Interconnection Technology 429
  3.6. Quality Assurance in Electronics Production 431
4. INTEGRATION OF MECHANICAL AND ELECTRONIC FUNCTIONS 432
  4.1. Structure of Molded Interconnect Devices 432
  4.2. Materials and Structuring 433
  4.3. Placement Systems for 3D PCBs 435
    4.3.1. Six-Axis Robot System for SMD Assembly onto MID 435
    4.3.2. Optimized MID Placement System 436
  4.4. Soldering Technology for 3D PCBs 438
5. DISASSEMBLY 439
  5.1. Challenge for Disassembly 439

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

5.2. Disassembly Processes and Tools 440

5.3. Applications 443

5.4. Entire Framework for Assembly and Disassembly 444

REFERENCES **445**

1. CURRENT DEVELOPMENTS

Today's competitive environment is characterized by intensified competition resulting from market saturation and increasing demands for a customer-oriented production. Technological innovations also have an influence on the competitive environment. These facts have dramatically altered the character of manufacturing. Meeting customers' demands requires a high degree of flexibility, low-cost/low-volume manufacturing skills, and short delivery times. Production and thereby manufacturing performance thus have gained increasing significance and are conceived as a strategic weapon for achieving and maintaining competitiveness (Verter and Dincer 1992). Especially in high-tech markets, where product technology is rapidly evolving, manufacturing process innovation is becoming an increasingly critical capability for product innovation. To meet the requirements of today's markets, new paths must be trodden both in organizational methods and in manufacturing and automation technology (Feldmann and Rottbauer 1999).

1.1. General Developments in Assembly

The great significance of assembly in a company's success is due to its function- and quality-determining influence on the product at the end of the direct production chain (Figure 1). Rationalization of assembly is still technologically impeded by high product variety and the various influences resulting from the manufacturing tolerances of the parts to be joined (Tönshoff et al. 1992). As a result, considerable disturbance rates are leading to reduced availability of assembly systems and delays in assembly operations. This complicates efficient automation and leads to consideration of displacing assembly plants into lower-cost regions. Assembly is also influenced by innovative developments in the manufacturing of parts, such as surface technology or connecting technology, which can have an important influence on assembly structures. The complex reflector assembly of a car headlight, for example, can be substituted for by surface coating technology.

Due to the rapidly changing market and production conditions, the role and the design of the individual functions within the entire value-adding chain are also changing. Other factors helping to bring this about include by the influence of microelectronics on product design and manufacturing structure and global communication possibilities. Facing the customer's demands for the after-sales or service functions, for example, is becoming of increasingly important to a company's success. More and more customers would like to purchase service along with the product contract.

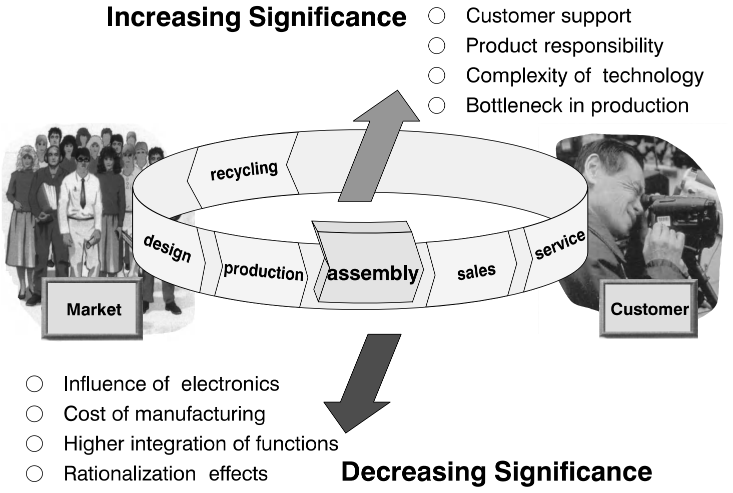

Increasing Significance
○ Customer support
○ Product responsibility
○ Complexity of technology
○ Bottleneck in production
recycling
design
production
assembly
sales
service
Market
Customer
○ Influence of electronics
○ Cost of manufacturing
○ Higher integration of functions
○ Rationalization effects
Decreasing Significance

Figure 1 Significance of Assembly in the Value-Adding Chain. (From Feldmann et al. 1996)

The growing complexity of the manufacturing process in industry has its origin in the globalization of the markets, customer demands for systems instead of single products, and the introduction of new materials and technologies. Designing a product for ease of assembly using design for manufacture and assembly (DFMA) methodology leads to a reduced number of variants and parts, higher quality, shorter time-to-market, lower inventory, and few suppliers and makes a significant contribution to the reduction of complexity in assembly (Boothroyd 1994).

The influence of the different branches and product structures is more evident in the range of assembly technology than in prefabrication. This also has a lasting impact on site selection because of the required workforce potential. Therefore, the global orientation of the assembly plants should be clarified on the basis of four major product areas (Figure 2).

Car assembly can be characterized as a serial assembly with trained workers. In contrast, the machine tool industry is characterized by a high degree of specializiation and small lot sizes, requiring highly skilled workers for assembly tasks. This makes the global distribution of assembly plants difficult, especially because the close interaction of development, manufacturing, and start-up still plays an important role. In contrast, the assembly of electronic components, inspired by the technological transition to surface mount technology (SMT), has been rapidly automated in recent years. In view of the small remaining share of personnel work, labor costs do not have a major influence on site selection in this product area any longer. In contrast to car assembly, the electronics industry is characterized by a more global distribution of production sites and comparatively smaller production units. However, this simplifies the regional or global distribution of assembly plants. Product size as well as logistical costs for the global distribution of products from central assembly sites are generally lower than in the automotive industry. In the white goods industry, a relatively small ratio of product value to product size is decisive. Serving global markets under minimized logistical costs requires corresponding global positioning of distributed assembly plants.

In general, for all industries, four fundamental solutions in assembly design can be distinguished (Figure 3). Manual assembly in small batch sizes is at one end and automated serial assembly at the other. Thus, the introduction of flexible assembly systems is reinforced. Again, these flexible automated assembly systems offer two alternatives. The integration of NC axes increases the flexibility of conventional machines, whereas the introduction of robot solutions is aimed at opening up further assembly tasks for efficient automation. There have been many technological responses to the global demand for large product variety coupled with short delivery times. In this context, the concept of human-integrated production systems is gaining ground. The intention here is to allow the human operator to be a vital participant in the future computer integrated manufacturing systems. This also has an impact on the design of assembly systems.

Changes in product structure and the influence of electronics are drivers of assembly rationalization. A common approach, from car assembly up to assembly in the electronics industry, is the

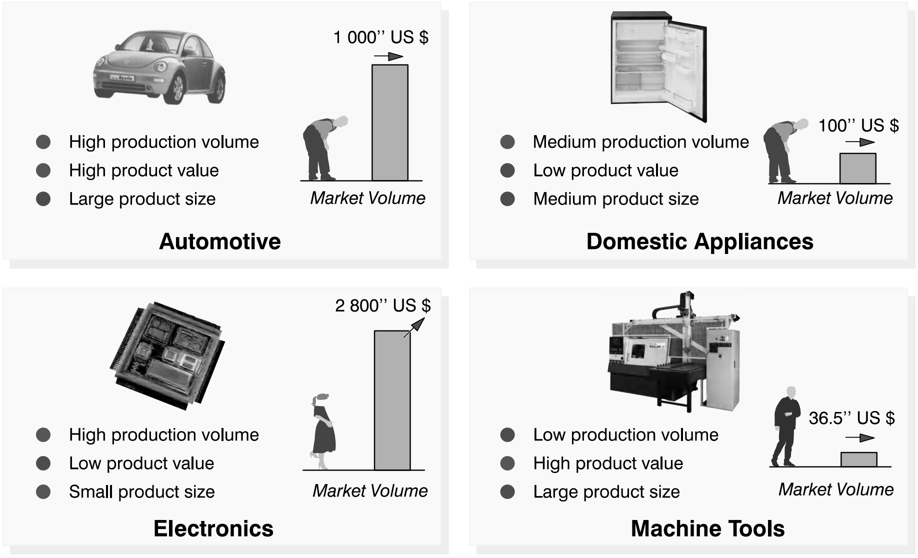

1 000” US $
High production volume
High product value
Large product size
Market Volume
Automotive
100” US $
Medium production volume
Low product value
Medium product size
Market Volume
Domestic Appliances
2 800” US $
High production volume
Low product value
Small product size
Market Volume
Electronics
36.5” US $
Low production volume
High product value
Large product size
Market Volume
Machine Tools

Figure 2 Major Product Areas with Different Conditions for Global Assembly.

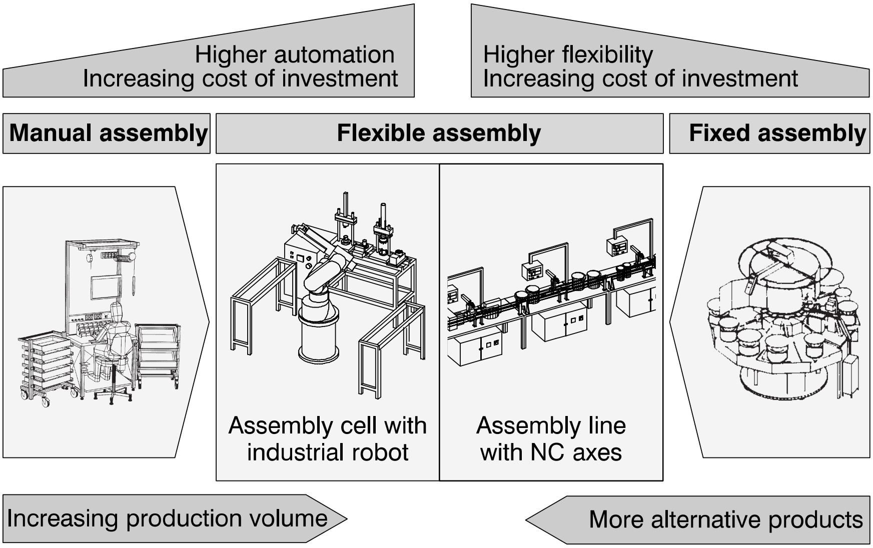

Higher automation
Increasing cost of investment
Higher flexibility
Increasing cost of investment
Manual assembly
Flexible assembly
Fixed assembly
Assembly cell with industrial robot
Assembly line with NC axes
Increasing production volume
More alternative products

Figure 3 The Basic Technological Alternatives in Assembly. (From Feldmann et al. 1996)

stronger concentration on higher functional density in subsystems (Figure 4). In car assembly, this means preassembling complex units like doors or cockpits; in electronics, this means circuit design with few but highly integrated circuits.

The new paradigm in manufacturing is a shift from Taylorism and standardization to small-lot, flexible production with emphasis on short lead-time and responsiveness to market. Organizational decentralization in autonomous business units, called fractals (Warnecke 1993), has also encouraged the distribution of assembly plants. The reduction of production complexity as a prerequisite for a faster, more flexible, and self-organizing adaptation of business units to changing market conditions is based on a redistribution of decision making responsibilities to the local or distributed unit.

The managing of complexity in manufacturing by outsourcing is rather common in the prefabrication field, whereas the reduction of production depth and the resulting cost advantages contribute to keep assembly sites in the countries of origin. In turn, the formation of decentralized business units results in more favorable conditions for the relocation of specific assembly tasks to other regions.

1.2. Impact of Electronics on Assembly

Within the framework of assembly rationalization, electronics almost has a double effect (Figure 5). In the first step, efficient assembly solutions can be built up by electronically controlled systems with programmable controllers and sensors. In the second step, the assembly task can be completely replaced by an electronically provided function. Examples are the replacement of electromechanical fluorescent lamp starters by an electronic solution and, on a long-term basis, the replacement of technically complex letter-sorting installations by purely electronic communication via global computer networks. Not only does the replacing electromechanical solutions with electronic functional carriers reduce assembly expenditure, but electronics production can be automated more efficiently. In many cases, the functionality and thus the customer benefit can be increased by the transition to entirely electronic solutions.

The further development of semiconductor technology plays an important role in electronics production. In addition to the direct consequences for the assembly of electronic components, further miniaturization, increasing performance, and the expected decline in prices have serious effects on the assembly of electronic devices and the further development of classical engineering solutions. Within a certain range, the degree of automation in mechanical assembly can be increased only up to a certain level. In contrast, in electronics assembly, there is greater higher potential for increasing the degree of automation. This also has consequences for product design. Figure 6 shows a comparison of the degree of automation in mechanical and electronics assembly.

Today's paradigms for manufacturing require a holistic view of the value-adding chain. The disadvantages of breaking up the value-adding chain and distributing the single functions globally can be compensated for by models and tools supporting integrated process optimization. Examples of this are multimedia applications based on new developments in information technology and the

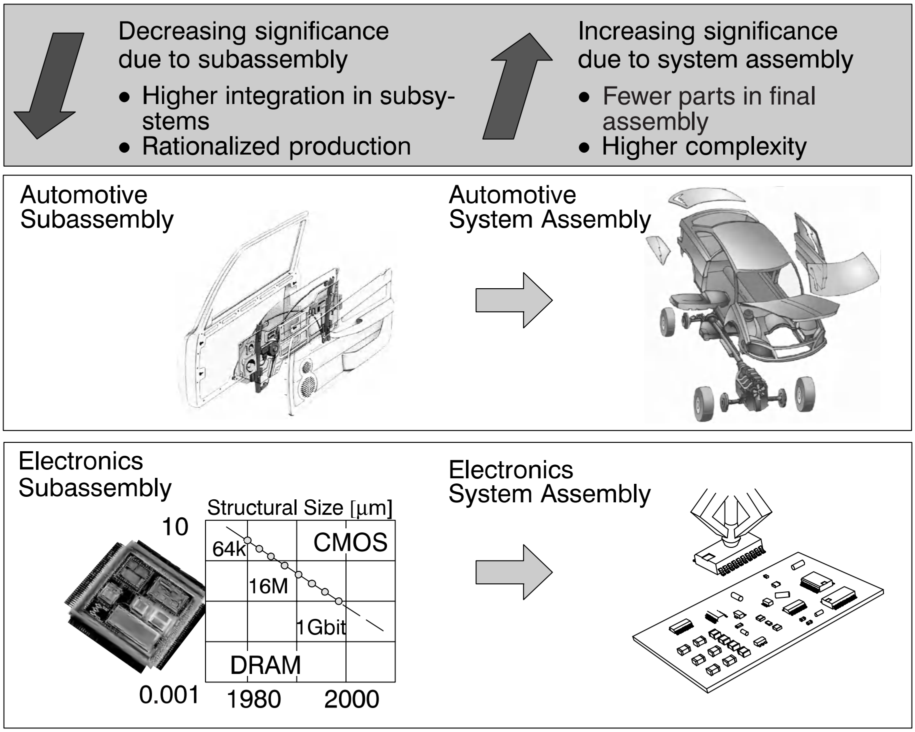

Decreasing significance due to subassembly
• Higher integration in subsystems
• Rationalized production
Increasing significance due to system assembly
• Fewer parts in final assembly
• Higher complexity
Automotive Subassembly
Automotive System Assembly
Electronics Subassembly
Structural Size [µm]
10
0.001
64k
16M
1Gbit
CMOS
DRAM
1980
2000
Electronics System Assembly

Figure 4 Trends toward New Assembly Structures. (From Feldmann et al. 1996)

concept of virtual manufacturing, which is based on simulation technology. The diffusion of systems such as electronic data interchange (EDI) and integrated service digital network (ISDN) allows a more efficient communication and information exchange (Figure 7).

Distributed and decentralized manufacturing involves the problem of locally optimized and isolated applications as well as incompatibilities of process and system. To ensure synergy potentials and stabilize the productivity of the distributed assembly plants, an intensive communication and data

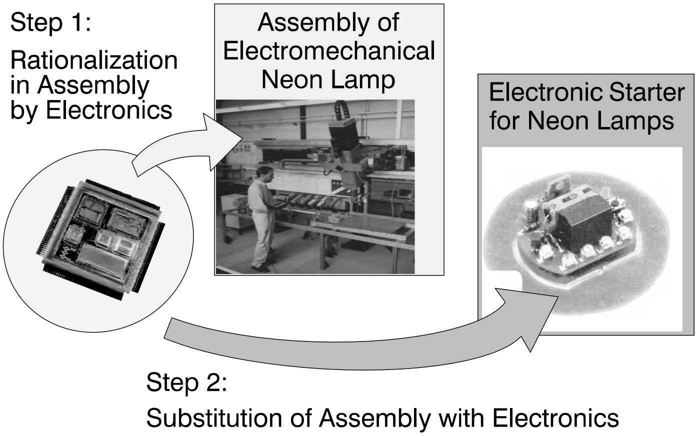

Step 1:
Rationalization in Assembly by Electronics
Assembly of Electromechanical Neon Lamp
Electronic Starter for Neon Lamps
Step 2:
Substitution of Assembly with Electronics

Figure 5 Impact of Electronics on Assembly Tasks. (From Feldmann et al. 1996)

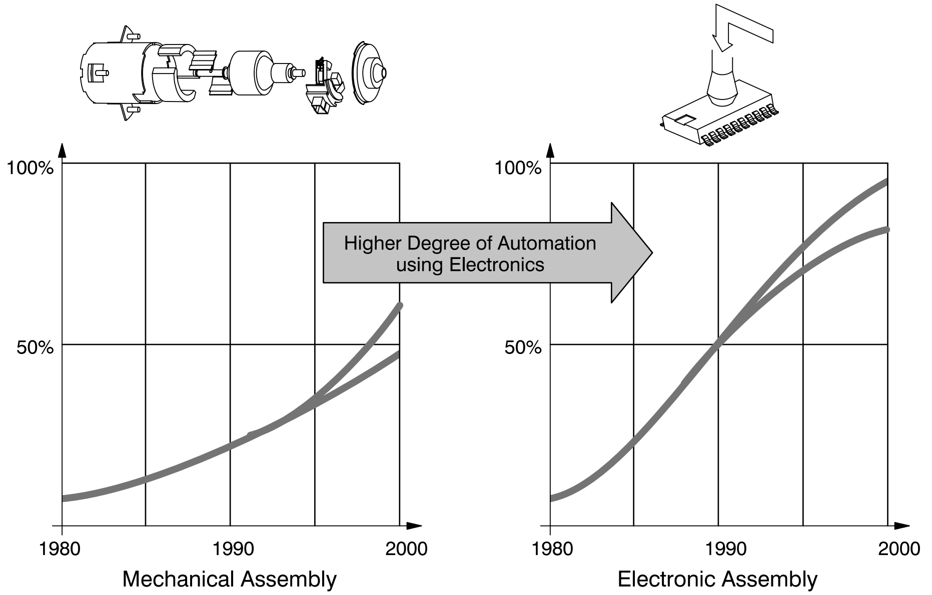


Figure 6 Comparison of the Degree of Automation in Mechanical and Electronics Assembly.

exchange within the network of business units is vital (Feldmann and Rottbauer 1999). New information technologies provide correct information and incentives required for the coordination of efficient global production networks. For instance, the diffusion of systems such as EDI and ISDN allows more efficient communication and information exchange among the productive plants linked in the same network. Furthermore, high-efficiency systems such as satellite information systems have contributed to perform operations more efficiently with regard to the critical success factors. For instance, tracking and expediting international shipments by means of preclearance operations at customs leads to a reduction in the delivery time. The coordination of a network of transplants dispersed throughout the world provides operating flexibility that adds value to the firm. From that

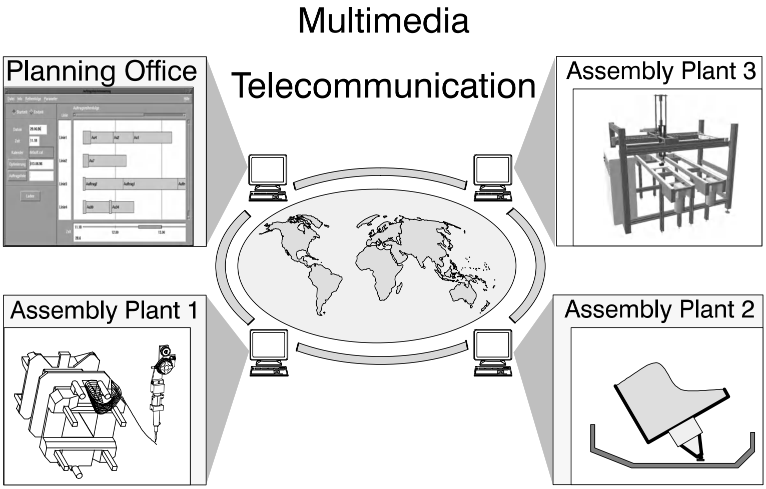


Figure 7 Global Engineering Network for Assembly Plants.

point of view, a decisive improvement in the conditions for global production networks has come about through the influence of microelectronics with the new possibilities in telecommunications.

As data definitions become more sophisticated under emerging standards such as STEP, corporate server networks can distribute a growing wealth of product and process information among different parts of the organization and its associates. Processing programs developed worldwide can be used in all production sites by means of direct computer guidance (Figure 7). The diagnosis of assembly systems is coordinated from a central control center. In this way, it becomes possible to transmit the management of a project in the global production network around the clock.

2. ASSEMBLY TECHNOLOGIES AND SYSTEMS

2.1. Basic Structures of Assembly Systems

Assembly is the sum of all processes needed to join together geometrically determined bodies. A dimensionless substance (e.g., lubricants, adhesives) can be applied in addition (VDI 1982).

In addition to joining in the manufacturing process, handling of components is the primary function of assembly. The assembly also contains secondary functions such as adjusting and inspecting as well as various special functions (see Figure 8).

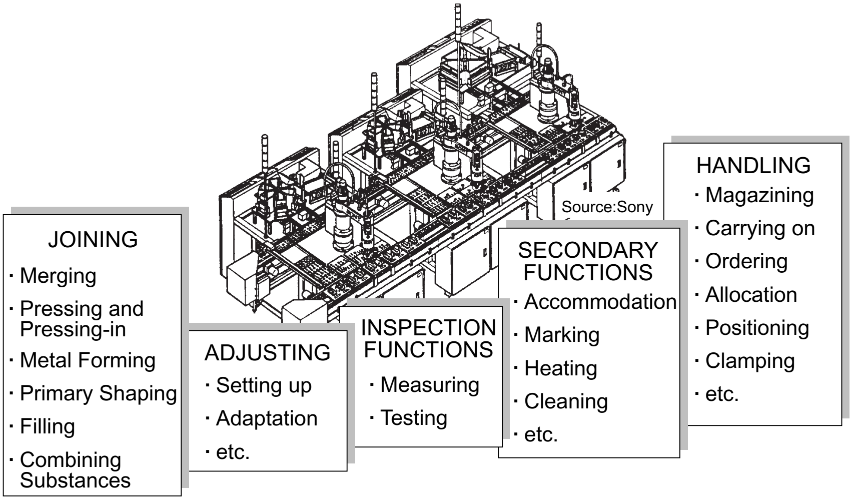
Joining is defined by DIN 8593 as a part of manufacturing processes. In this case, the production of a compound consisting of several parts can be achieved by merging, pressing, pressing in, metal forming, primary shaping, filling, or by combining substances.

Handling is defined in VDI Guideline 2860/1 as the creation, defined varying, or temporary maintaining of a prescribed 3D arrangement of geometrical defined solids in a reference coordinate system. For this, procedures such as ordering, magazining, carrying on, positioning, and clamping are important. It is simple for a human being to bring parts into correct position or move them from one place to another. However, a considerably larger expenditure is necessary to automate this task. An extensive sensory mechanism often must be used.

Manufacturing of components is subject to a great number of influences. As a result, deviations cannot be avoided during or after the assembling of products. These influences must be compensated for, and thus adjusting is a process that guarantees the required operating ability of products (Spur and Stöferle 1986).

Testing and measuring are contained in the inspection functions. Testing operations are necessary in all individual steps of assembly. Testing means the fulfillment of a given limiting condition. The result of the test operation is binary (true or false, good or bad). On the other hand, specifications are determined and controlled by given reference quantities while measuring. Secondary functions are activities, such as marking or cleaning operations, that can be assigned to none of the above functions but are nevertheless necessary for the assembly process.

Assembly systems are complex technical structures consisting of a great number of individual units and integrating different technologies.



Source:Sony

Figure 8 Functions in an Assembly System.

There are different possibilities for the spatial lineup of assembly systems. One possibility is a line structure, which is characterized by:

- Clear flow of materials
- Simple accessibility of the subsystems (e.g., for maintenance and retrofitting)
- Simple lineup of main and secondary lines
- Use mainly for mass production (the same work routine for a long time)

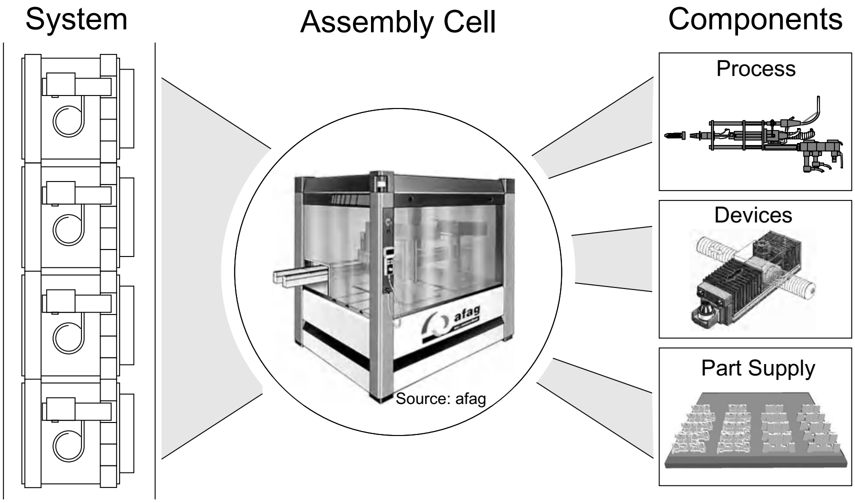
Alternatively, an assembly system can be arranged in a rectangular structure, which is characterized by:

- Very compact design
- High flexibility. The combination of opposing subsystems is easy to realize.
- Poor accessibility to the subsystems during maintenance and retrofitting
- Use mainly for small and middle lot sizes

In addition to different spatial lineups, other basic modifications of cell structure are possible for achieving the required efficiency (see Figure 10).

The number of work cycles that can be carried out on an assembly system depends on the size of the assembly system (number of cells) and the required productivity. The availability drops as the number of stations increases. Therefore, the distribution of work cycles onto several machines is necessary. In this case, the productivity increases with decreasing flexibility. The entire assembly system can be subdivided into different cells, which are connected by the flow of materials and information. The basic structure of a cell consists of a tabletop, a basic energy supply, the mounted handling devices, the internal conveyor system, and a safety device (protective covering, doors with electrical interlock). The individual cells should be built up modular and equipped with standardized interfaces (energy, pneumatic, and information technology). A strict module width for the spatial measurements is also useful. As a result, fast realization of the individual assembly cells, a high degree of reuse, and flexible use are guaranteed. The assembly cell itself consists of different components and units (Figure 9). The main components of an assembly cell are devices (integrated robots, modular system built of numerically controlled axes, pick-and-place devices, etc.), its mechanical construction, the design of the grippers, the actuating system, the control system, the method of programming, and the sensor technology used. Furthermore, the part supply is very important (ordered or disordered part supply; see also Section 2.3). Last but not least, the joining process (see also Section 2.2) must be adapted and optimized to meet all the necessary requirements.

The different alternatives for design of assembly systems are represented in Figure 10. The flexible cell is able to carry out all necessary assembly functions for a certain variant spectrum (three variants,



Source: afag

Figure 9 Structure and Components of Assembly Cells.

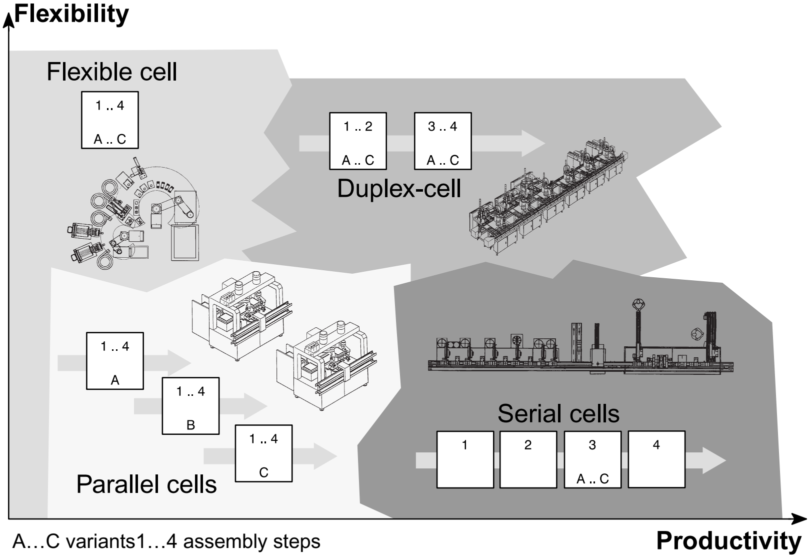


Figure 10 Alternative Cell Structures for Assembly Systems.

A, B, and C, are represented here). For this purpose, it has all the necessary main and secondary functions for the assembly process. High variant flexibility, sequential processing of the individual assembly tasks, as well as the resulting high nonproductive times, lead to the use of flexible cells, mostly for small and middle lot sizes.

If the duplex cell is used, the assembly tasks will be distributed over two handling units. These always have the full variant flexibility. With regard to function, the cells need not be completely flexible. They must be able to carry out only a part of the assembly tasks. The function size of the multiple grippers (in the case of unchangeable variation of the functions) can be increased by parallel assembly with two independent handling devices. Even shorter cycle times can be achieved than with the flexible cell.

Serial cells are flexible only with respect to the assembly of variants. However, the function size of an individual cell is highly limited. The spatial lineup of the cells is responsible for the fact that the cell with the longest cycle time determines the total cycle time of the assembly process. Considering the spatial extension of serial structures, it is obvious that the integration level of the system is smaller than that of flexible cells or duplex cells.

Parallel cells will be used if only one assembly cell is responsible for the complete assembly of a variant. Each individual cell has a high degree of functional flexibility, which unfortunately can only be used for one variant. Therefore, the potential of the resources cannot be exploited because identical assembly conditions must be used during the assembly of different variants. If the individual variants show only a small number of similar parts, which on top of that often have different gripping conditions, splitting up variants will be advantageous. As a result, a smaller variant flexibility will be necessary.

The serial structure will react extremely sensitively to fluctuation in the number of variants. At worst, restructuring or reconstruction of all cells will be required. In contrast, the rate of utilization of the system will remain almost stable if full variant flexible cells are used. Because the handling devices of parallel cells have full functional flexibility, it is not necessary to carry out an adaptation of the number of flexible systems. Due to changes in lot size, parallel cells designed for special variants will even have too much or too little capacity, which cannot be removed without adequate measures.

2.2. Joining Technologies

2.2.1. Classification and Comparison of Joining Technologies

Industrialized manufactured products predominantly consist of several parts that are usually manufactured at different times in different places. Assembly functions thus result from the demand for joining subsystems together into a product of higher complexity with given functions. According to German standard DIN 8580, joining is defined as bringing together two or more workpieces of geometrically defined form or such workpieces with amorphous material. The selection of a suitable joining technique by the technical designer is a complex function. In DIN 8593, manufacturing methods for joining (see Figure 11) are standardized.

In addition to the demands on the properties of the product, economic criteria have to be considered (e.g., mechanical strength, optics, repair possibilities). The ability to automate processes and design products with the manufacturing process view are important points to focus on. Therefore, examples of the joining techniques in automation presented are given here.

The joining processes can be divided into various classes on the basis of several criteria:

- Constructional criteria: strength, form, and material closure connections or combinations of these possibilities
- Disassembly criteria: in removable and nonremovable connections
- Use of auxiliary joining part, e.g., screws
- Influence of temperature on the parts

Some important joining techniques and applications are presented below. Pros and cons of the joining techniques will be discussed in regard to the features of function, application, and assembly.

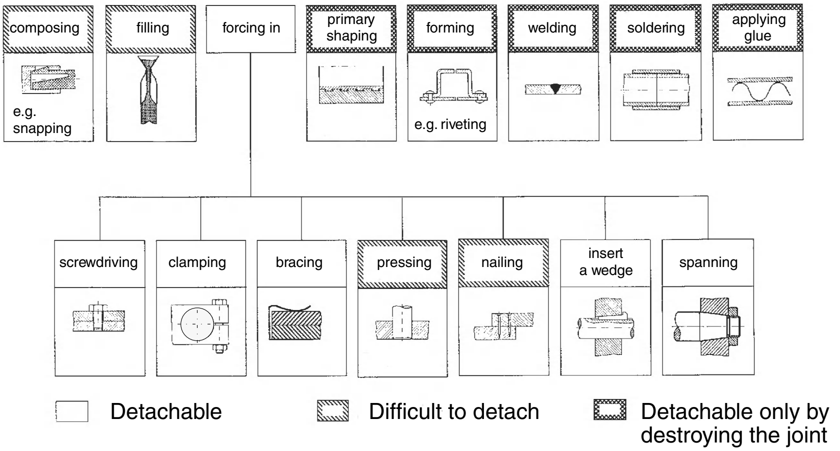
2.2.2. Bolting

Screwing is one of the most frequently used joining processes in assembly. Screwdriving belongs in the category of removable connections (ICS 1993). Joinings are removable if the assembly and separation process can be repeated several times without impairing the performance of the connection or modifying the components. The specific advantages of detachable connections are easy maintenance and repair, recycling, and allowing a broad spectrum of materials to be combined.

The most important component of the bolt connection is the screw as the carrier of substantial connecting functions and the link between the components. The screw's function in the product determines its geometrical shape and material. Important characteristics of a screw are the form of thread, head, shaft end, tensile strength, surface coating, tolerances, and quality of the screw lots.

Screwing can be manual, partly automated, or fully automated. The bolting tools used can be classified according to structural shape, control principle, or drive. Substantial differences occur in the drive assigned, which can be electric, pneumatic, or hydraulic (only for applying huge torques). Figure 12 shows the basic structure of a pneumatic screwdriving tool.

A basic approach is to increase the economy of automated screwing systems and reduce deadlock times. Process stabilization can be achieved by using fast control loops in process control and diagnostic support of the operators for error detection and recovery (see Figure 13) (Steber 1997). Further, position errors of the parts can be compensated automatically by adaptation of the coordinate



DIN 8593 Manufacturing process: joining

Figure 11 Survey of Joining Technologies.

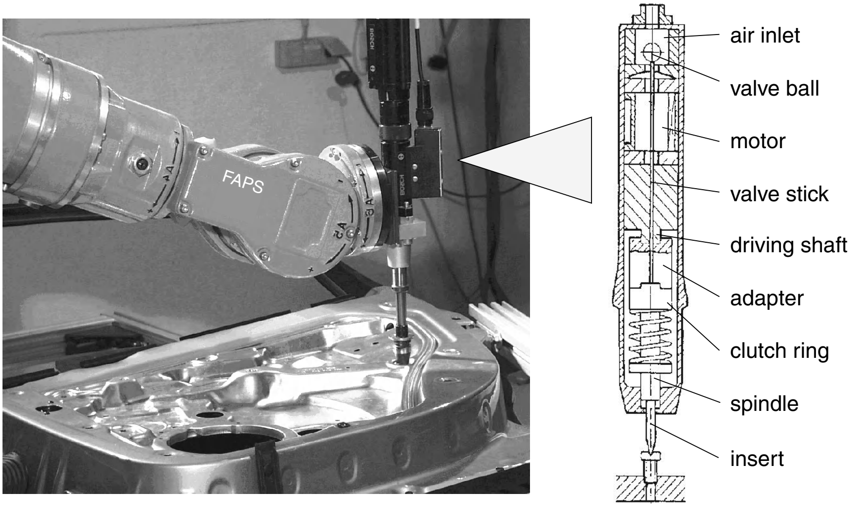

FAPS
air inlet
valve ball
motor
valve stick
driving shaft
adapter
clutch ring
spindle
insert

Figure 12 Structure of a Pneumatic Screwdriving System.

system in flexible handling systems, such as robots and image-processing systems. The automated screwing technique is becoming more and more important in mass or serial production. Apart from reduction unit costs, longer production times, and higher output, it offers the advantage of continuously high quality that can be documented.

2.2.3. Riveting/Clinching

Riveting is one of the classical joining processes. It is in particularly wide use in the aircraft industry. What all rivet methods have in common is that an auxiliary joining part, the rivet, must be provided. Blind rivets are often used in device assembly because accessibility from one side is sufficient.

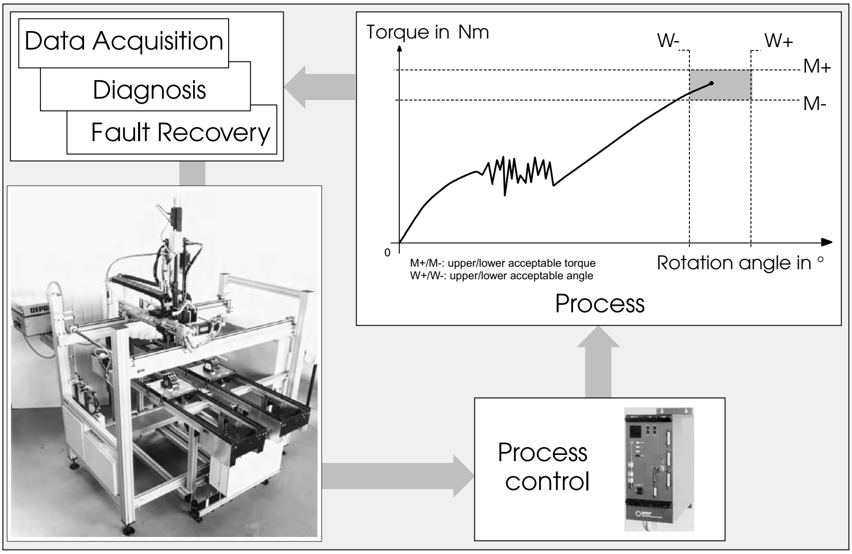

Data Acquisition
Diagnosis
Fault Recovery
Torque in Nm
W-
W+
M+
M-
0
M+/M-: upper/lower acceptable torque
W+/W-: upper/lower acceptable angle
Rotation angle in °
Process
Process control

Figure 13 Optimized Process Control of a Screwdriving Machine.

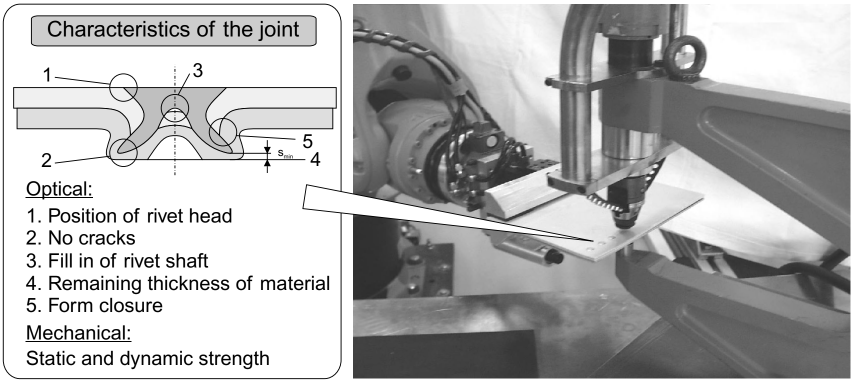

Characteristics of the joint
1
2
3
5
4
s_{min}
Optical:
1. Position of rivet head
2. No cracks
3. Fill in of rivet shaft
4. Remaining thickness of material
5. Form closure
Mechanical:
Static and dynamic strength

Figure 14 Characteristics of Punching Rivets.

Recently the punching rivet method has become increasingly important in the automobile industry. It is one of the joining processes with less heat influence on the material than with spot welding (Lappe 1997). The advantage of this technique is that no additional process step for punching the hole into the parts is necessary. The punching rivet itself punches the upper sheet metal, cuts through, and spreads in the lowest metal sheet. Figure 14 shows the profile of a punching rivet with typical characteristics for quality control. It is characterized by the fact that different materials, such as metal and aluminum, can be combined. The use of new lightweight design in automotive manufacturing means that materials such as aluminum, magnesium, plastics, and composites are becoming increasingly important. Audi, with the introduction of the aluminum space frame car body concept, is a pioneer in the application of punching rivets.

In clinching, the connection is realized through a specially designed press-driven set of dies, which deforms the parts at the join to provide a friction-locked connection. No auxiliary jointing part is necessary, which helps to save costs in supply and refill of jointing parts.

2.2.4. Sticking

Figure 15 shows a comparison of various destructive detachable joining technologies. The progress in plastics engineering has had positive effects on the engineering of new adhesives. Sticking is often

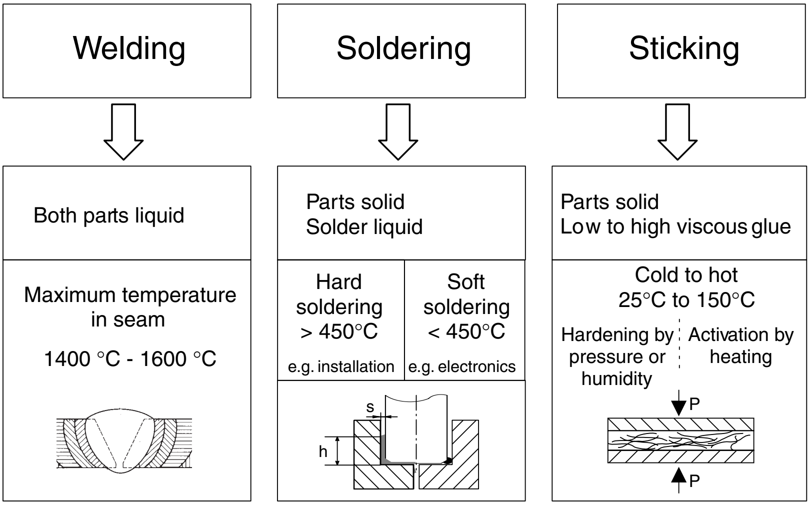

Welding
Soldering
Sticking
Both parts liquid
Parts solid
Solder liquid
Parts solid
Low to high viscous glue
Maximum temperature in seam
1400 °C - 1600 °C
Hard soldering > 450°C
e.g. installation
Soft soldering < 450°C
e.g. electronics
s
h
Cold to hot
25°C to 150°C
Hardening by pressure or humidity
Activation by heating
P
P

Figure 15 Comparison of Joining Technologies That Are Detachable Only by Destruction.

used in combination with other joining processes, such as spot welding and punching rivets. The connection is made by adhesion and cohesion (Spur and Stöferle 1986).

Adhesives can be divided into two groups. *Physically* hardening adhesives achieve adherence by two different mechanisms. The first is by cooling of the melted adhesive, and the second is by the evaporation of solvent or water (as the carrier) out of the adhesive. Because the adhesive does not interlace, it is less resistant to influences such as heating up, endurance stress, or interaction of solvent. *Chemically* hardening adhesives solidify themselves by a chemical reaction into a partially interlaced macromolecular substance characterized by high firmness and chemical stability. Adhesives can also be differentiated into aerobic and anaerobic adhesives.

The quality and firmness of an adhesive depends on the conditions at the part surface. The wettability and surface roughness of the parts, as well as contamination (e.g., by oil), play a substantial role. To ensure quality of sticking, therefore, often a special surface treatment of the jointing parts is necessary. This represents an additional process step, which can be automated too. Typically, car windows are automatically assembled into the car body in this way.

2.2.5. Welding

Welding methods can be subdivided into melt welding and press welding methods. In melt welding, such as arc welding, metal gas-shielded welding (e.g., MIG, MAG) or gas fusion welding, the connection is made by locally limited heating to just above the liquidus temperature of the materials. The parts that should be connected and the usually used additional welding materials flow together and solidify. In *pressure welding,* such as spot welding, the connection is realized by locally limited heating followed by pressing or hammers. Welding methods differ according to their capacity for automation. In the building of car bodies, fully automated production lines with robots are usually already in use. Fusion welding, such as gas-shielded arc welding, makes higher demands on automation. Therefore, robots are suitable, which should be also equipped with special sensors for seam tracking. Therefore, both tactile and contactless sensors (e.g., optical, capacitive, inductive) are used. With the increasing power density of diode lasers, laser beam welding with robots is becoming much studied.

2.3. Peripheral Functions

2.3.1. Handling Devices

For complex operations, industrial robots are normally used. If the handling task consists of only a simple pick-and-place operation, specially built devices with one or more linear axes are probably the better choice. These *handling devices* are classified by their degrees of freedom (DOF), which indicate the number of possible translational and rotational movements of a part. Therefore, six DOF—three translational and three rotational—are required to arrange an object in a defined way in a 3D room. Opening and closing of grippers is not counted as a degree of freedom, as this movement is used only for clamping and does not, strictly speaking, move the part.

Mechanically controlled inserting devices offer one to two DOF. They are suitable for simple handling tasks, such as inserting. Due to their strict mechanical setup, they are compact, robust, and very economical. Their kinematics allows them to reach different points with a predefined, mechanically determined motion sequence. This motion can be adapted to the different handling tasks by the use of different radial cams (control curves). Due to the sensor-less mechanical setup, it is an inherent disadvantage of the system that the precision of the movement is not very high. In an open control loop, only the end positions are detected by sensors.

If the handling task demands higher accuracy or flexibility, numerically controlled (NC) axes or industrial robots are recommended. Using two or three linear axes allows more complex and precise handling tasks to be performed than with mechanical handling devices.

Industrial robots are built in different setups. The most common are the SCARA (selective compliance assembly robot arm) robot and the six-DOF robot. The SCARA robot usually has four DOF, three translational and one rotational, whereas the z-stroke represents the actual working direction. The other axes are used for positioning. SCARA robots are often used for assembling small parts automatically with very short cycle times. Six-DOF robots are used for more complex tasks because they are more flexible in their movements so they can grip parts in any orientation.

2.3.2. Grippers

Grippers have to perform different tasks. Not only do they have to lock up the part to be handled (static force), but they also have to resist the dynamic forces resulting from the movement of the handling device. They also have to be made so that the parts cannot move within them. Additional characteristics such as low weight, fast reaction, and being fail-safe are required. There are different techniques for gripping parts. Grippers are typically classified into four groups by performing principle (Figure 17): mechanical, pneumatic, (electro-) magnetic, and other. With a market share of 66%,

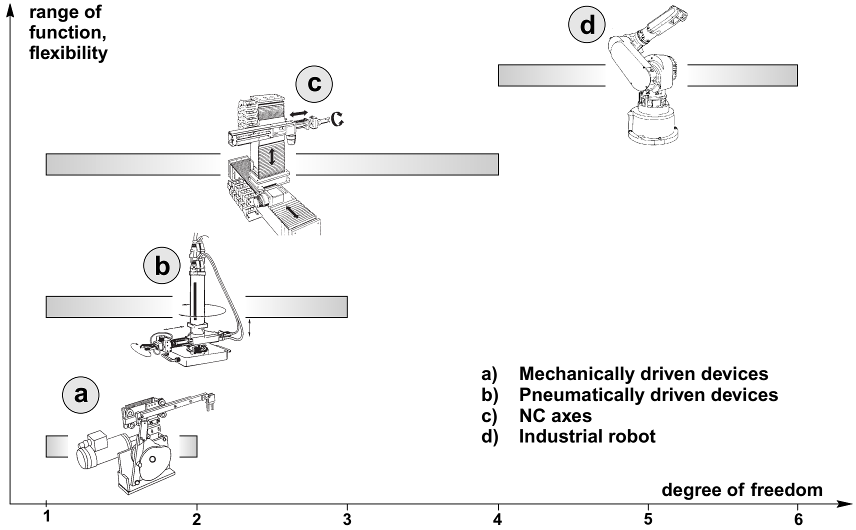

range of function, flexibility
d
c
b
a
a) Mechanically driven devices
b) Pneumatically driven devices
c) NC axes
d) Industrial robot
degree of freedom
1
2
3
4
5
6

Figure 16 Degrees of Freedom of Alternative Handling Devices.

mechanical grippers are the most commonly used. Mechanical grippers can be subdivided by their kind of closing motion and their number of fingers. Three-finger-centric grippers are typically used for gripping round and spherical parts. Two-finger-parallel grippers perform the gripping movement with a parallel motion of their fingers, guaranteeing secure gripping because only forces parallel to the gripping motion occur.

Because these gripper models can cause harm to the component surface, vacuum grippers are used for handling damageable parts. Two-dimensional parts such as sheet metal parts are also handled by vacuum grippers. Using the principle of the Venturi nozzle, an air jet builds up a vacuum in the suction cup that holds the parts. When the air jet is turned off, the parts are automatically released.

Heavy parts such as shafts are lifted not by mechanical grippers but with electromagnetic grippers. However, secure handling, not exact positioning, is needed when using these grippers.

To handle small and light parts, grippers using alternative physical principles, such as electrostatic and adherent grippers, are used because they do not exert any pressure that could cause damage to the part. Fields of application include microassembly and electronics production.

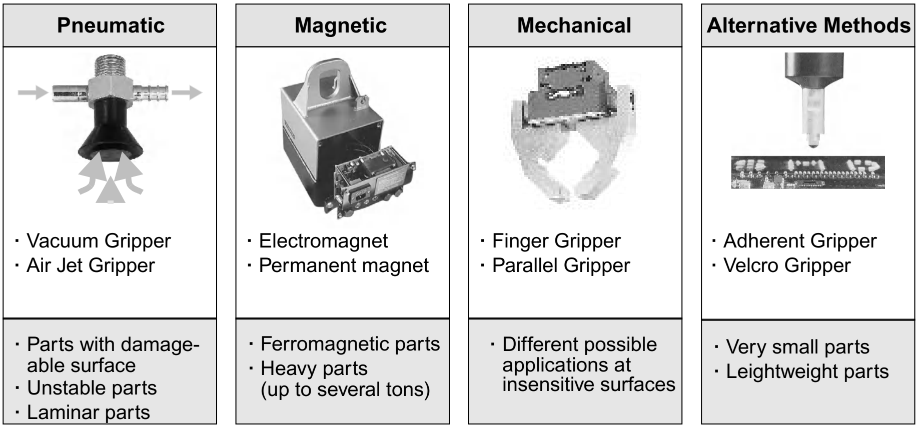
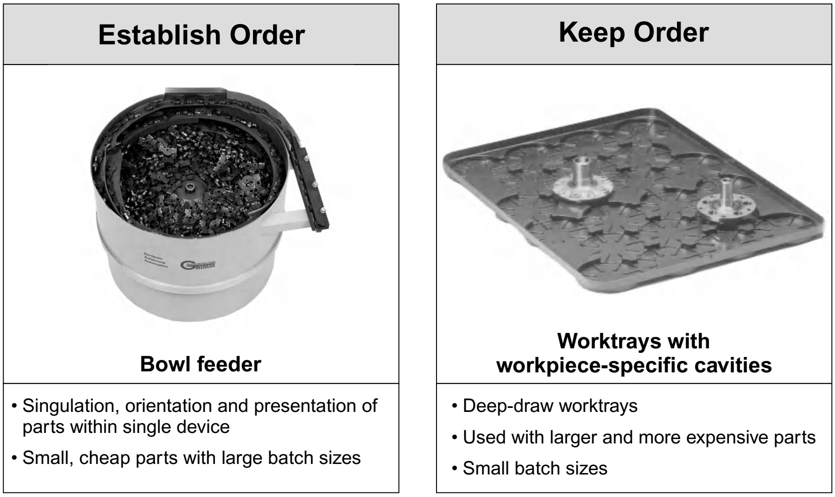

Pneumatic
Magnetic
Mechanical
Alternative Methods
· Vacuum Gripper
· Air Jet Gripper
· Electromagnet
· Permanent magnet
· Finger Gripper
· Parallel Gripper
· Adherent Gripper
· Velcro Gripper
· Parts with damageable surface
· Unstable parts
· Laminar parts
· Ferromagnetic parts
· Heavy parts (up to several tons)
· Different possible applications at insensitive surfaces
· Very small parts
· Leightweight parts

Figure 17 Different Kinds of Grippers. (Sources: Sommer, Truninger, Schunk, ipa)



- Singulation, orientation and presentation of parts within single device
- Small, cheap parts with large batch sizes

- Deep-draw worktrays
- Used with larger and more expensive parts
- Small batch sizes

Figure 18 Establish Order-Keep Order. (Source: Bowl feeder, Grimm Zufuhrtechnik GmbH & Co.)

Security of the gripped parts and the workers is an important consideration. Toggle lever grippers, for example, ensure that the part cannot get lost even if a power failure occurs. In contrast, parts handled by vacuum grippers fall off the gripper if the air supply fails.

2.3.3. Feeding Principles

To be gripped automatically, parts have to be presented to the handling device in a defined position and orientation irrespective of the above-mentioned gripping principles. For small parts, which are assembled in high volumes, the vibratory bowl feeder is commonly used. This feeder integrates three different tasks: singulation, orientation, and presentation.

The parts are stored in the bowl feeder as bulk material in the bunker, which is connected via a slanted feeding track to the pick-up point. Using suitable vibrations generated by lamellar springs through an electric motor, the parts move serially toward the pick-up point. The movement is a result of the superposition of micro-throws and a backwards slide.

A disadvantage of the bowl feeder is that the parts may be damaged by the micro-throws and the relative movement and contact between the parts and between the single part and the surface of the bowl feeder. Additionally, this feeding principle leads to the annoyance of noise. Another disadvantage is that the bowl feeder is extremely inflexible with regard to different parts because it is strictly constructed for the use of one special part. Also, it is not yet possible to automate the process of constructing a bowl feeder. Rather, whether the bowl feeder will meet the requirements is up to the experience and skill of the person constructing it. Nevertheless, this is and will continue to be one of the most important feeding technologies for small parts.

Larger and more expensive parts, or parts that must not be damaged on their surface can be presented to the handling device palletized by using deep-draw work trays.

Generally speaking, a degree of orientation, once reached, should never be allowed to be lost again, or higher costs will result.

2.3.4. Linkage

Normally a product is not assembled and produced by only one machine. Therefore, multiple handling devices, machines, and processes have to be arranged in a special way. The different ways of doing this are shown in Figure 19. The easiest, which has been widely used since the industrial revolution, is the loose linkage. This is characterized by the use of discrete ingoing and outgoing buffers for each machine or process. It is thus possible to achieve independence for the different cycle times. Disadvantages are that very high supplies are built up and the flow time is very high. An advantage is that the production will keep running for some time even if one machine fails, because the other machines have supplies left.

The stiff linkage works the opposite way. The workpieces are transported from station to station without a buffer in between. The corresponding cycle time is equal to the process time of the slowest machine. As a consequence, if one machine is out of order, the whole production line has to be

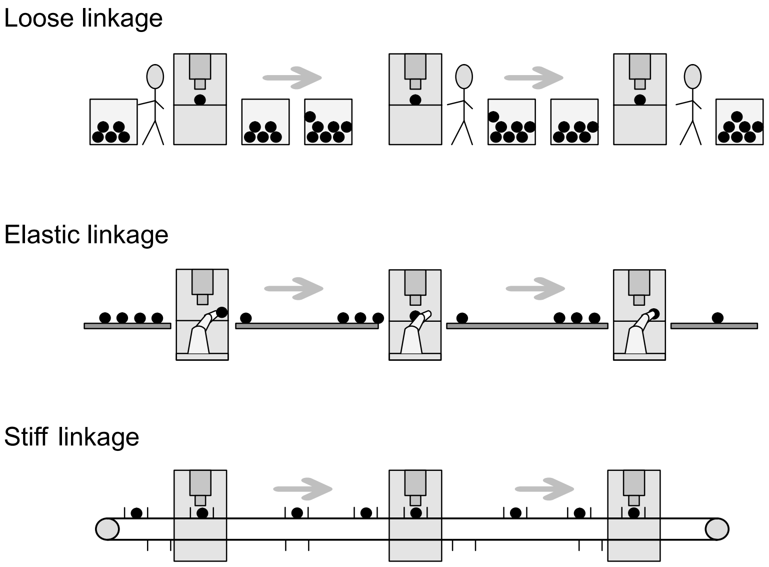


Figure 19 Different Possibilities for Linking Assembly Cells.

stopped. A well-known application of this principle can be found in automobile production, where, because there are no bypasses, the whole assembly line has to be stopped if problems occur at any working cell.

The elastic linkage can be seen as a combination of the other two. The stations are connected by different transports (belts, live roller conveyors, etc.). Because the single transports are not connected with each other, they can be used as additional buffers. The failure of one machine does not necessarily cause the whole production line to be stopped.

2.4. Manual Assembly Systems

2.4.1. Description of Manual Assembly Systems and Their Components

Manual assembly systems are often used within the area of fine mechanics and electrical engineering. They are suitable for the assembly of products with a large number of versions or products with high complexity. Human workers are located at the focal point of manual assembly systems. They execute assembly operations by using their manual dexterity, senses, and their intelligence. They are supported by many tools and devices (Lotter 1992). The choice of tools depends on the assembly problem and the specific assembly organization form. The most frequently used forms of organization are, on the one hand, assembly at one separate workstation, and, on the other, the flow assembly with chained assembly stations. Assembly at only one workstation is also an occasional form of assembly organization. The choice of the form of organization depends on the size of the product, the complexity of the product, the difficulty of assembly, and the number of units.

Workstations are used for small products or modules with limited complexity and a small number of units. High version and quantity flexibility are the most important advantages. Also, disturbances affect other workstations to only a small extent.

The components for the basic parts and the assembly parts are the substantial constituents of manual assembly systems. The assembly parts are often supplied in grab containers. The distances to be covered by the workers arms should be short and in the same direction. The intention is to shorten the cycle time and reduce the physical strain on the workers. This can be realized by arranging the grab containers in paternoster or on rotation plates. Further important criteria are glare-free lighting and adapted devices such as footrests or work chairs (Lotter and Schilling 1994).

When assembly at a workstation is impossible for technological or economical reasons, the assembly can be carried out with several chained manual assembly stations (Lotter 1992). Manual assembly systems consist of a multiplicity of components, as shown in Figure 20. The stations are chained by double-belt conveyors or transport rollers. The modules rest on carriers with adapted devices for fixing the modules. The carriers form a defined interface between the module and the

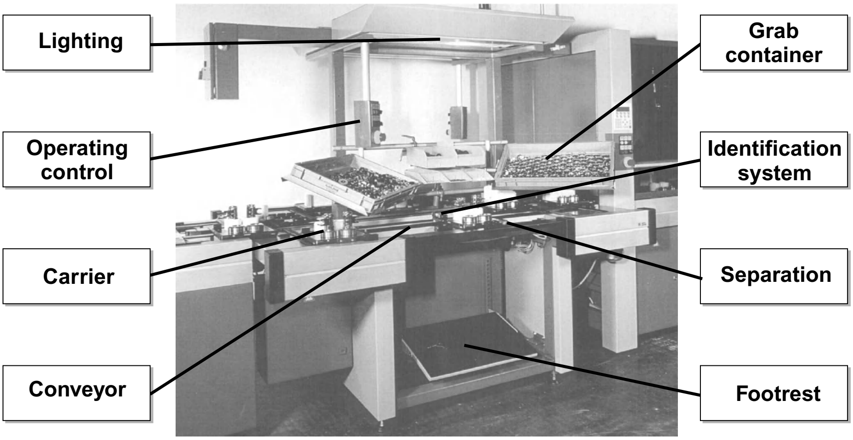

Lighting
Operating control
Carrier
Conveyor
Grab container
Identification system
Separation
Footrest

Figure 20 Components of Manual Assembly Systems. (Source: Teamtechnik)

superordinate flow of material. For identifying the different versions, the carriers can be characterized. Identification systems separate the different versions and help to transport them to the correct assembly stations.

2.4.2. Criteria for the Design of Manual Assembly Systems

There are many guidelines, tables, and computer-aided tools for the ergonomic design of workstations. The methodology of planning and realizing a manual workstation is shown in Figure 21.

The following criteria have to be considered in designing manual assembly systems.

- Design of the workstation (height of the workbench and work chair, dimensions of the footrest, etc.)
- Arrangement of the assembly parts, tools, and devices
- Physical strain on workers (forces, torque, noise)

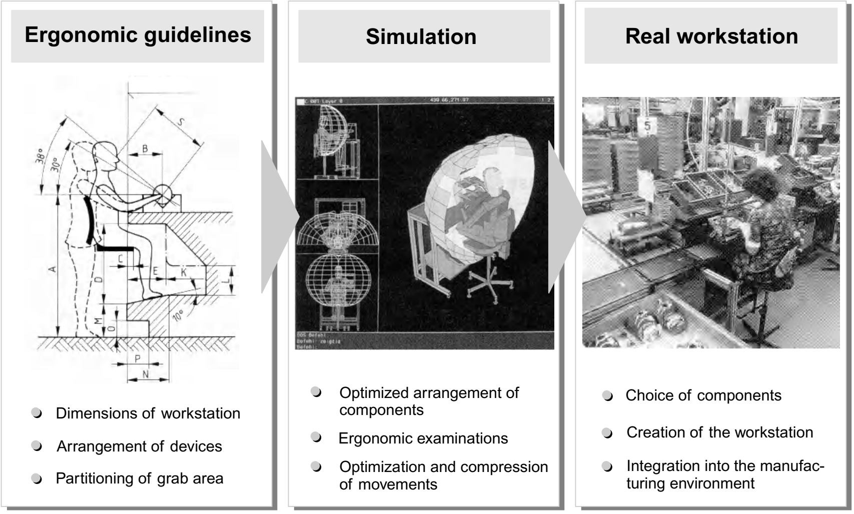

Ergonomic guidelines
Simulation
Real workstation
Dimensions of workstation
Arrangement of devices
Partitioning of grab area
Optimized arrangement of components
Ergonomic examinations
Optimization and compression of movements
Choice of components
Creation of the workstation
Integration into the manufacturing environment

Figure 21 Ergonomic Design of Manual Workstations. (Source: Bosch)

Ergonomic design requires the adjustment of the work height, chair height, grab distance, and so on to the human dimensions. It should be possible for workers to do the work by sitting or standing (Konold and Reger 1997). To ensure that most of the possible workers work under ergonomic conditions, the human dimensions between the 5th and 95th percentiles are considered in designing manual workstations.

One of the most important criteria is the organization of the assembly sequences. The assembly operation should be easy and require little force to be exerted by the worker. Devices like a screwdriver can reduce physical strain on the worker. An optimized arrangement of the assembly parts and components can rationalize the movements. Therefore, the grab room is partitioned into four areas (Lotter 1992):

1. The grab center
2. The extended grab center
3. The one-hand area
4. The extended one-hand area

Assembly should take place in the grab center because both hands can be seen. Part supply should be in the one-hand area.

The forces and torque also have to be considered. Overstressing the worker with too great or too extended physical strain is not permitted. The maximum limit values for static and dynamic load depend on the frequency of the operation, the hold time, and the body attitude. Further correction factors are sex, age, and constitution of the worker.

Today many possible forms of computer-aided tools have been developed for optimization of devices, minimization of forces, time registration, and simplification of movements. They also enable shorter planning time, minimized planning costs, and extended possibilities for optimization of movements (Konold and Reger 1997).

2.5. Automated Assembly Systems

Automated assembly systems are used mainly for the production of series and mass-produced articles. In the field of indexing machines, a distinction is made between rotary indexing turntables and rectilinear transfer machines. The essential difference between the two systems is the spatial arrangement of the individual workstations.

Rotary indexing turntables are characterized by short transport distances. Therefore, high clock speeds are possible. The disadvantage is the restricted number of assembly stations because of the limited place. Rectilinear transfer machines can be equipped with as many assembly stations as needed. However, the realizable cycle time deteriorates through the longer transport distances between the individual stations.

Indexing machines are characterized by a rigid chain of stations. The construction design depends mostly on the complexity of the product to be mounted. The main movements (drives for transfer systems) can be effected from an electrical motor via an adapted ratchet mechanism or cam and lever gears or can be implemented pneumatically and/or hydraulically. Secondary movements (clamping

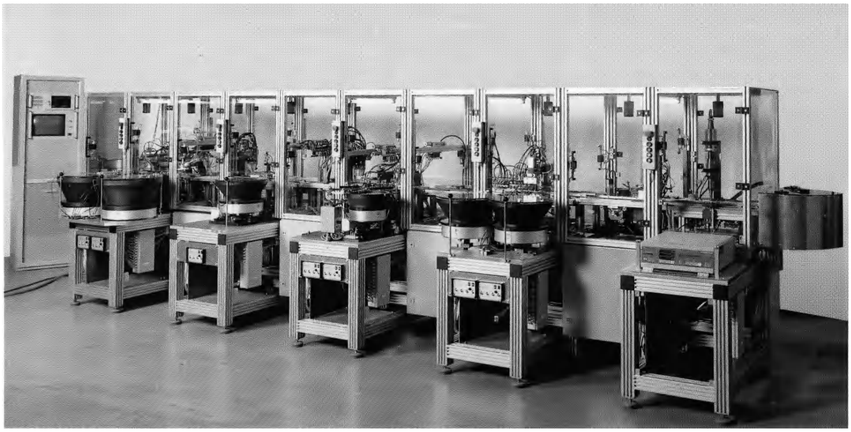


Figure 22 Rectilinear Transfer Machine with High Productivity. (Source: IMA Automation)

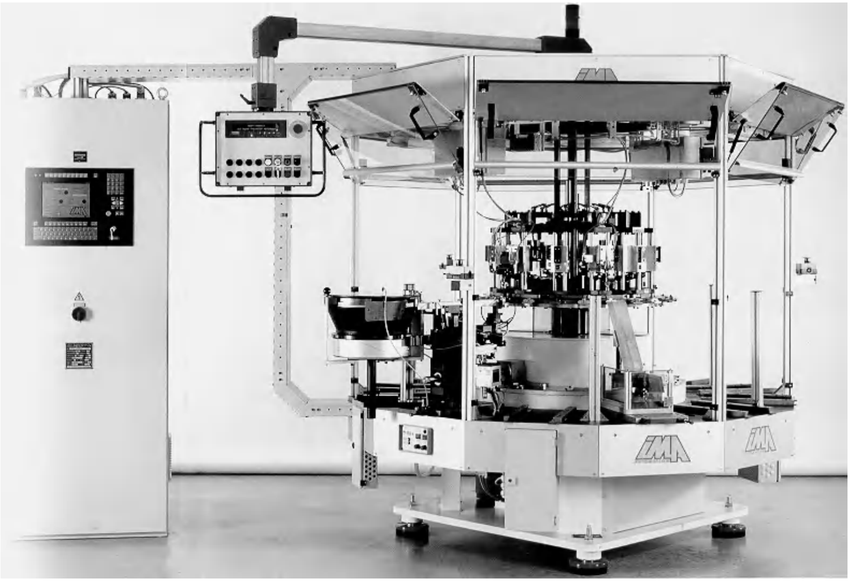


Figure 23 Rotary Indexing Turntable for an Efficient Assembly of Mass-Produced Articles. (Source: IMA Automation)

of parts, etc.) can be carried out mechanically (e.g., via cam and lever gears), electromechanically, or pneumatically. The handling and assembly stations are often driven synchronously over cam disks. If small products are assembled under optimal conditions, an output of up to 90 pieces/min will be possible. However, this presupposes a products design suitable for the automation, vertical joining direction, easy-to-handle components, processes in control, and a small number of product variants. The total availability of the assembly system is influenced by the availability of the individual feeding devices.

The number of stations needed depends on the extent of the single working cycles that have to be carried out (e.g., feeding, joining, processing, testing, adjusting). Therefore, the decision between rotary indexing and rectilinear transfer machines depends not only on the necessary number of workstations and the required space but also on the entire assembly task. On the one hand, short cycle times and high accuracy during the assembly of small products can be achieved by using typical indexing machines. On the other hand, only the assembly of mass-produced articles and small product variants is possible. Normally the reusability of the individual components is also very small. Consequently, these product-specific special-purpose machines can only be amortized heavily.

In order to balance this disadvantage, modern rotary indexing turntables are available on the market. These have higher flexibility, greater reusability, easier reequipping, and higher modularity. The product-independent components (basic unit with the drive, operating panel, protective device, control box, control system, transport unit) should be strictly separated from the product-dependent components (seat, handling and assembly modules). Defined and standardized interfaces are very useful for integrating the product-specific components into the basic system very quickly and with little effort. With modular construction, it is possible to combine the high capability of the indexing machines with economical use of the resources.

2.6. Flexible Assembly Systems

Different objectives can be pursued in order to increase the flexibility of assembly processes. One the one hand it may be necessary to handle different workpieces, on the other hand different versions of one product may have to be produced in changing amounts. Another contribution to flexibility is to use the advantages of off-line programming in order to optimize cycle time when the product manufactured is changed.

2.6.1. Assembly of Different Versions of a Product in Changing Amounts

For producing changing amounts of different versions of a product, an arrangement as shown in Figure 24, with a main assembly line to which individual modules from self-sufficient manufacturing

Figure 24 Flexible Manufacturing of Different Versions of a Product in an Assembly Line. (*Source:* Siemens)

cells are supplied, is very suitable. In order to decouple the module manufacturing from the main assembly line with respect to availability of the different modules, a storage unit is placed between module assembly and main assembly line to keep necessary numbers of presently needed versions of the modules permanently ready. Thus, even lot size 1 can be manufactured and at the same time the system remains ready for use even when one module of a module line fails. The individual manufacturing cells are connected by a uniform workpiece carrier-transport system, whereby the workpiece carriers—which are equipped with memory capacity—function as means of transport, assembly fixtures, and for product identification simultaneously. As the workpiece carriers are equipped with a mobile memory, all necessary manufacturing and test data are affiliated with the product through the whole process. Thus, the product data are preserved, even if the system controller breaks down.

2.6.2. Flexible Handling Equipment

Figure 25 shows an industrial robot with six DOF as an example of handling equipment with great flexibility of possible movement. Especially in the automobile industry, flexible robots are commonly used, and a further increase in investment in automation and process control is predicted. Robots are introduced in automobile assembly to reduce strenuous tasks for human workers, handle liquids, and precisely position complicated shapes. Typical applications include windshield-installation, battery mounting, gasoline pouring, rear seat setting, and tire mounting. Another important application where the flexibility of a robot is necessary is manufacturing automobile bodies with stop welding. In order to reach each point in the workspace with any desired orientation of the welding tongues in the tool center point, at least three main axes and three secondary axes are necessary. Because now each point of the workspace with each orientation can be reached, all kinds of workpieces can be handled and assembled.

2.6.3. CAD-CAM Process Chain

In traditional production processes such as lathing and milling, numerical control is widespread and process chains for CAD-based programming of the systems are commonly used. In contrast, winding

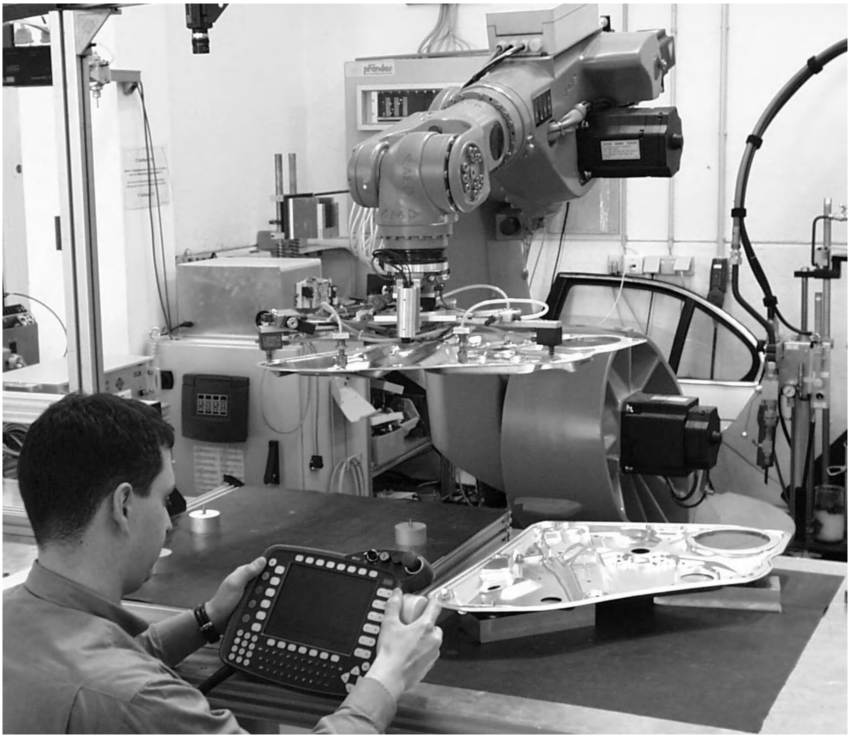


Figure 25 Most kinds of Workpieces can be assembled by an Industrial Robot Because of Its Great Flexibility of Movements.

systems are even now often programmed using the traditional teach-in method. Figure 26 shows a winding system for which a CAD-CAM process chain has been developed by the Institute for Manufacturing Automation and Production Systems in order to increase flexibility with respect to product changes. The disadvantages of on-line programming are obvious: because complex movements of the wire-guiding nozzle are necessary while the wire is fixed at the solder pins and guided from the pin to the winding space, considerable production downtimes during programming result. Furthermore collisions are always a danger when programming the machine with the traditional teach-in method. The tendency for lower production batches and shorter production cycles combined with high investment necessitates a flexible off-line programming system for assembly cells such as winding systems. Therefore a general CAD-CAM process chain has been developed. The process chain allows the CAD model of the bobbin to be created conveniently by adding or subtracting geometric primitives if necessary. Afterwards the path of the wire-guiding nozzle is planned at the CAD workstation, where the user is supported by algorithms in order to avoid collisions with the bobbin or the winding system. Simulating the wire nozzle movements at the CAD system allows visual and automated collision detection to be carried out. If a collision occurs, an automatic algorithm tries to find an alternative route. If no appropriate path can be calculated automatically, the wire nozzle movements can be edited by the user. To this point, the result of this CAD-based planning system is a data file that is independent from the winding machine used. A postprocessor translates this file into a machine-specific data file, which is necessary for the numerical control of the winding system. This fault-free NC data file can be sent from the CAD workstation to the machine control unit via network, and the manufacturing of the new product can start (Feldmann and Wolf, 1996, 1997; Wolf 1997).

2.7. Control and Diagnosis

Despite major improvements in equipment, advanced expertise on assembly processes, and computer-aided planning methods, assembly remains susceptible to failure because of the large variety of parts and shapes being assembled and their corresponding tolerances. Computer-aided diagnosis makes a substantial contribution to increasing the availability of assembly systems. Assuring fail-safe processes will minimize throughput times and enable quality targets to be achieved.

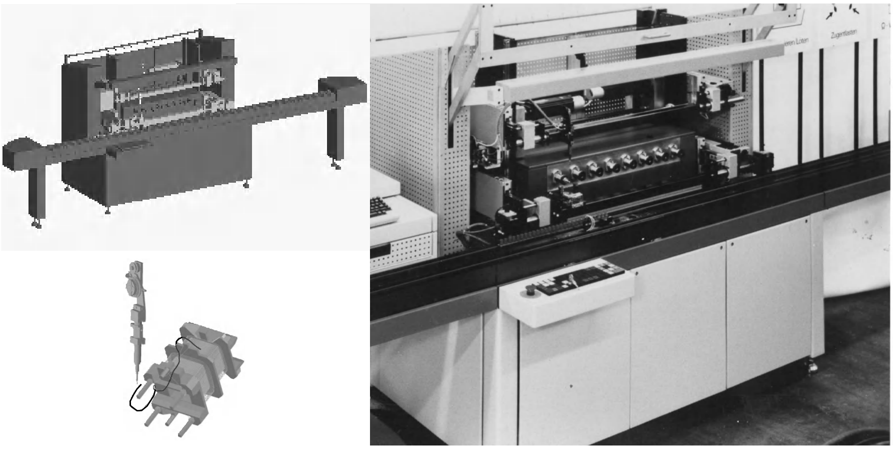


Figure 26 Flexible Automation in Winding Technology Using Off-Line Programming.

Diagnosis includes the entire process chain of failure detection, determination of the cause of failure, and proposal and execution of measures for fault recovery. Realizing and operating a high-performance diagnosis system requires comprehensive acquisition of the assembly data. For this purpose, three different sources of data entry in assembly systems can be used. First, machine data from controls can be automatically recorded. Second, for extended information, additional sensors based on various principles can be integrated into the assembly system. Third, the data can be input manually into the system by the operator.

Diagnosis systems can be divided into signal-based, model-based, and knowledge-based systems. A crucial advantage of knowledge-based systems is the simple extendability of the database, for example concerning new failures. Therefore, they are frequently used with assembly systems.

Dependent on the efficiency of the diagnosis system, different hierarchically graded control strategies are possible. In the simplest case, the diagnosis system only supplies information about the occurrence and number of a disturbance. The user must determine the cause of failure and execute

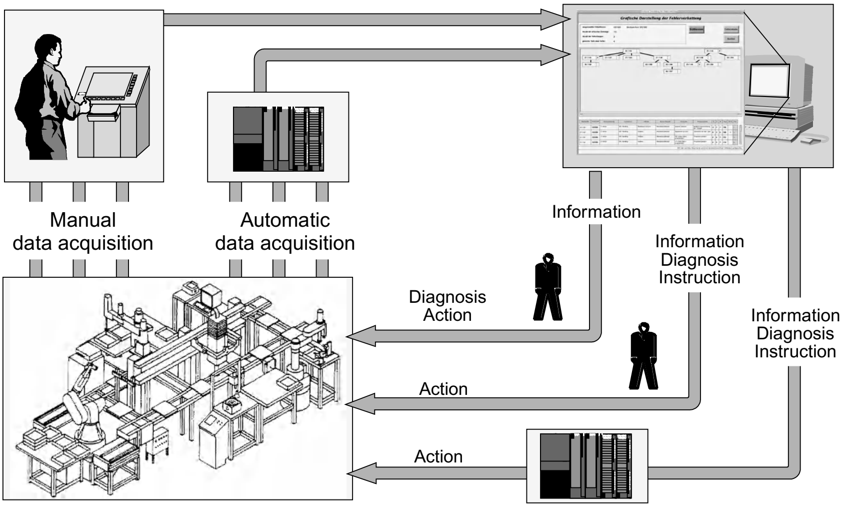

Manual
data acquisition
Automatic
data acquisition
Information
Information
Diagnosis
Instruction
Diagnosis
Action
Information
Diagnosis
Instruction
Action
Action

Figure 27 Computer-Aided Diagnosis with Hierarchically Graded Strategies for Fault Recovery.

the fault recovery on his own. Efficient diagnosis systems offer information to the user about the number and location of the failure. Further, appropriate diagnostic strategies are available that allow computer-aided detection of the cause of failure. The systems also suggest appropriate measures for fault recovery. In addition, diagnosis systems can independently initiate reaction strategies after the automatic determination of the cause of failure. These strategies directly affect the control of the assembly system or the assembly process. Due to the particularly high complexity of this procedure, it is applicable only to simple and frequently occurring errors.

3. ELECTRONIC PRODUCTION

3.1. Process Chain in Electronic Production

Production of electronic devices and systems has developed into a key technology that affects almost all areas of products. In the automotive sector, for example, the number of used electronic components has increased about 200,000% during the last 20 years, today accounting for more than 30% of the total value added.

The process chain of electronics production can be divided into three process steps: solder paste application, component placement, and soldering (Figure 28). Solder paste application can be realized by different technologies. The industry mainly uses stencil printing to apply the solder paste. Solder paste is pressed on the printed circuit board (PCB) by a squeegee through form openings on a metal stencil, allowing short cycle times. Several other technologies are on the market, such as single-point dispensing, which is very flexible to changes of layout of PCBs, but leads to long cycle times due to its sequential character.

In the next step the components of surface-mounted devices (SMDs) and/or through-hole devices (THDs) are placed on the PCB. Finally the PCB passes through a soldering process. Traditionally the soldering process can be classified into two groups: reflow and wave soldering. The purpose of reflow processing is to heat the assembly to specified temperatures for a definite period of time so that the solder paste melts and realizes a mechanical and electrical connection between the components and the PCB. Today three reflow methods are used in soldering: infrared (IR), convection, and condensation. A different method of molten soldering is wave or flow soldering, which brings the solder to the board. Liquid solder is then forced through a chimney into a nozzle arrangement and returned to the reservoir. For wave soldering, the SMDs must be glued to the board one step before. The quality of the electronic devices is checked at the end of the process chain by optical and electrical inspection systems (IPC 1996).

The special challenges in electronics production result from the comparatively rapid innovation of microelectronics. The continuing trend toward further integration at the component level leads to permanently decreasing structure sizes at the board level.

3.2. Electronic Components and Substrate Materials

Two main focuses can be distinguished concerning the development of electronic components. With the continuing trend toward function enhancement in electronic assembly, both miniaturization and integration are leading to new and innovative component and package forms. In the field of passive components, the variety of packages goes down to chip size 0201. This means that resistor or capacitor components with dimensions of 0.6 mm × 0.3 mm have to be processed. The advantages of

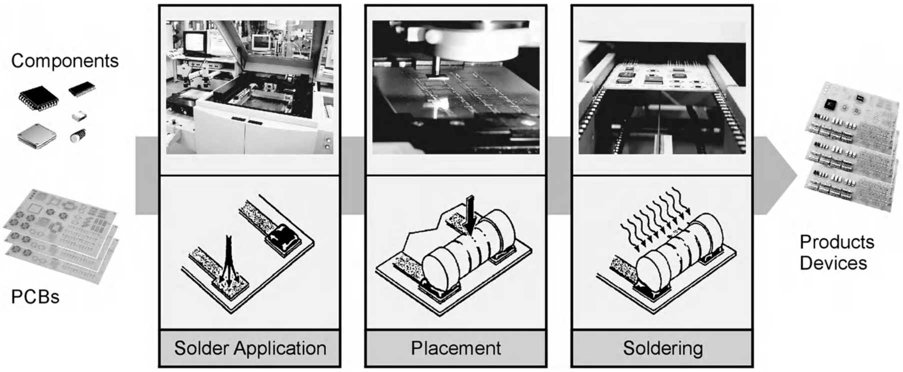


Figure 28 Process Chain in Electronics.

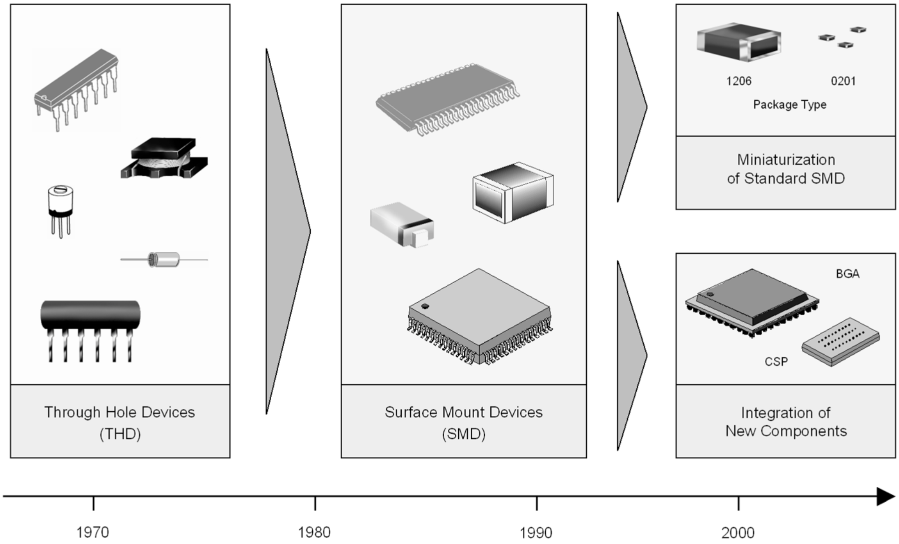

1206
0201
Package Type
Miniaturization of Standard SMD
BGA
CSP
Integration of New Components
Through Hole Devices (THD)
Surface Mount Devices (SMD)
1970
1980
1990
2000

Figure 29 Component Technology: Development through the Decades.

such components are obvious: smaller components lead to smaller and more concentrated assemblies. On the other hand, processing of such miniaturized parts is often a problem for small and medium-sized companies because special and cost-intensive equipment is needed (high-precision placement units with component specific feeders, nozzles, or vision systems).

Regarding highly integrated components two basic technologies should be mentioned. First, leaded components such as quad flat pack (QFP) are used within a wide lead pitch from 1.27 mm down to 300 μm. But these packages with a high pin count and very small pitch are difficult to process because of their damageable pins and the processing of very small amounts of solder. Therefore, area array packages have been developed. Instead of leaded pins at all four sides of the package area, array packages use solder balls under the whole component body. This easily allows more connections within the same area or a smaller package within the same or even a larger pitch. Ball grid arrays (BGAs) or their miniaturized version, chip scale packages (CSPs), are among the other packages used in electronics production (Lau 1993).

Further miniaturization is enabled by direct assembly of bare dies onto circuit carriers. This kind of component is electrically connected by wire bonding. Other methods for direct chip attachment are flip chip and tape automated bonding. All three methods require special equipment for processing and inspection.

In addition to planar substrates, many new circuit carrier materials are being investigated. Besides MID (see Section 4), flexible circuit technology has proved to be a market driver in the field of PCB production. Today, flexible circuits can be found in nearly every type of electronic product, from simple entertainment electronics right up to the highly sophisticated electronic equipment found in space hardware. With growth expected to continue at 10–15% per year, this is one of the fastest-growing interconnection sectors and is now at close to $2 billion in sales worldwide.

However, up to now, most flexible circuit boards have been based on either polyester or polyimide. While polyester (PET) is cheaper and offers lower thermal resistance (in most cases reflow soldering with standard alloys is not possible), polyimide (PI) is favored where assemblies have to be wave or reflow soldered (with standard alloys). On the other side, the relative costs for polyimide are 10 times higher than for polyester. Therefore, a wide gap between these two dominant materials has existed for a long time, prohibiting broad use of flexible circuits for extremely cost-sensitive, high-reliability applications like automotive electronics. Current developments in the field of flexible-base materials as well as the development of alternative solder alloys seem to offer a potential solution for this dilemma.

3.3. Application of Interconnection Materials

The mounting of SMD components onto PCBs or flexible substrates demands a certain amount of interconnection material (solder or conductive adhesive) to form a correct joint. In contrast to wave soldering, where heat and material (molten solder) are provided simultaneously during the soldering process, reflow soldering of SMDs necessitates the application of solder in the form of paste as the first process step in the assembly line. Also, for interconnections established by conductive adhesive, the application of interconnection material is necessary. This process is a decisive step in electronic device assembly because as failures caused here can cause difficulties during the following process steps, such as component placement and reflow soldering. In modern high-volume assembly lines, stencil printing is the predominant method for applying solder paste to the substrates. Other methods that are used industrially to a minor extent are automatic dispensing and screen printing. Each method has its specific advantages and drawbacks, depending on, for example, batch sizes or technological boundary conditions like component pitch.

The main advantage of stencil printing over screen printing occurs in applications where very small areas of paste have to be deposited. For components with pitches equal to or smaller than 0.65 mm, the stencil printing process is the only viable way for printing solder paste. Therefore, stencil printing has replaced screen printing in most cases.

Dispensing has the advantage over screen and stencil of being highly flexible. For small batches, dispensing may be an economical alternative to printing. On the other hand, dispensing is a sequential process and therefore quite slow. Additionally, paste dispensing for fine-pitch components is limited.

The principle of a stencil printer is shown in Figure 30(*a*). The major components of an automatic screen printer are squeegee blades (material stainless steel or polyurethane), the screen itself, the work nest (which holds the substrate during printing), an automatic vision system for PCB alignment, and sometimes a printing inspection system. During the printing process, solder paste is pressed by a squeegee through the apertures in a sheet of metal foil (attached in frame, stencil) onto the substrate.

Dispensing is an application of solder paste in which separate small dots of paste are deposited onto the substrate sequentially. Dispensing is a good way for applying solder paste when small batches have to be produced. The main advantages are short changeover times for new boards (due to loading only a new dispensing program) and low cost (no different screens are used). Several dispensing methods are realized in industrial dispensing systems. The two dominant principles are the time–pressure dispensing method and the rotary pump method (positive displacement). The principle of dispensing by the time–pressure method is shown in Figure 30(*b*). A certain amount of solder paste is dispensed by moving down the piston of a filled syringe by pressure air for a certain time. The paste flows from the syringe through the needle onto the PCB pad. The amount of solder can be varied by changing the dispensing time and the air pressure.

3.4. Component Placement

Traditional THD technology has nearly been replaced by SMD technology in recent years. Only SMD technology permits cost-efficient, high-volume, high-precision mounting of miniaturized and complex components. The components are picked up with a vacuum nozzle, checked, and placed in the correct position on the PCB.

3.4.1. Kinematic Concepts

Within the area of SMD technology, three different concepts have found a place in electronics production (Figure 31) (Siemens AG 1999). The starting point was the pick-and-place principle. The machine takes one component from a fixed feeder table at the pick-up position, identifies and controls the component, transports it to the placement position on the fixed PCB, and sets it down. New concepts have been developed because pick-and-place machines have not been able to keep pace with the increasing demands on placement rates accompanying higher unit quantities. The original pick-and-place principle process is now only used when high accuracy is required.

The first variant is the collect-and-place principle based on a revolver head (horizontal rotational axis) on a two-axis gantry system. Several components are collected within one placement circle and the positioning time per component is reduced. Additionally, operations such as component centering are carried out while the nozzle at the bottom (pick-up) position is collecting the next component, and the centering does not influence the placement time.

The highest placement rate per module can be obtained with the carousel principle. The most widely used version of this principle uses a movable changeover feeder table and a moving PCB. It is analogous to the revolver principle in that the testing and controlling stations are arranged around the carousel head and the cycle time primarily depends on the collecting and setting time of the components.

High-performance placement systems are available for the highest placement rates, especially for telecommunications products. These systems take advantage of several individual parallel pick-and-place systems. A placement rate of up to 140,000 components per hour can be attained.

3.4.2. Classification of Placement Systems

The classification and benchmarking of placement machines depend on different influence coefficients. In addition to accuracy and placement rate, placement rate per investment, placement rate per needed area, operation, maintenance, and flexibility have to be considered. In the future, flexible

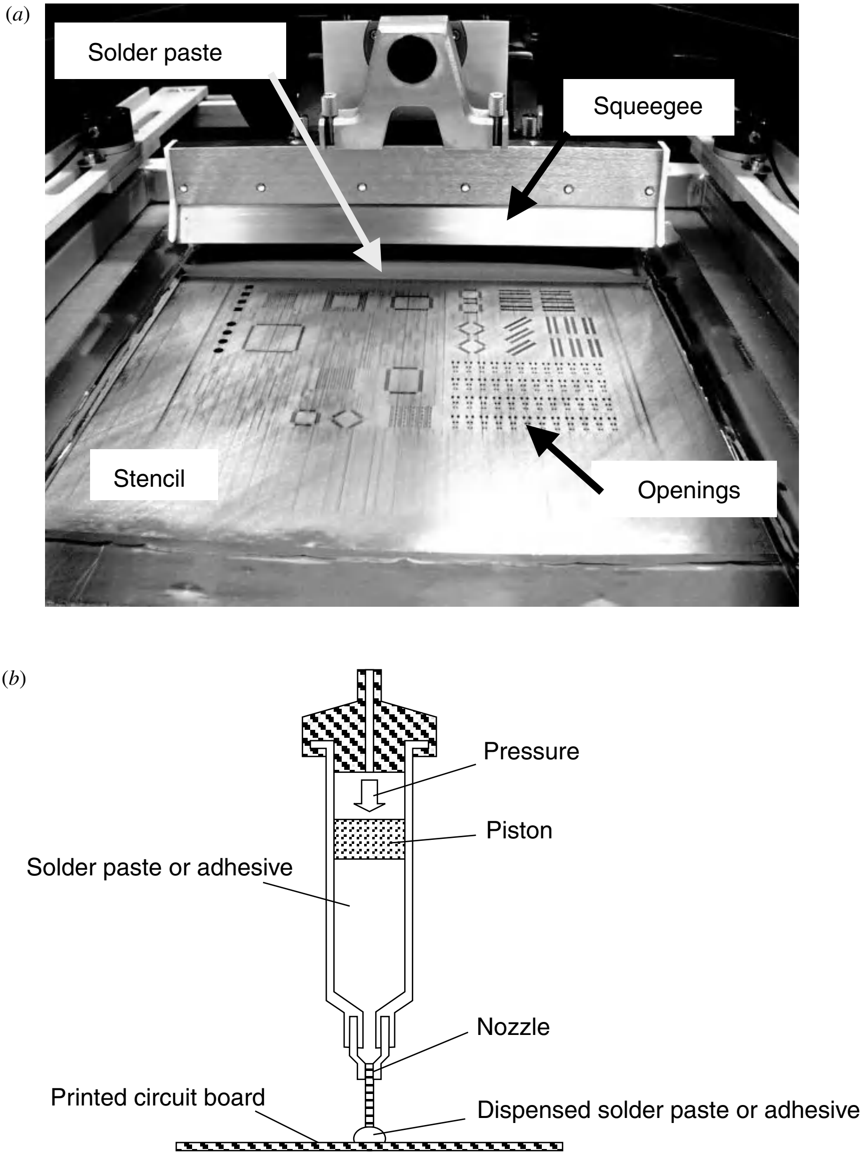

(a)
Solder paste
Squeegee
Stencil
Openings
(b)
Pressure
Piston
Solder paste or adhesive
Nozzle
Printed circuit board
Dispensed solder paste or adhesive

Figure 30 Alternative Principles for Application of Solder Paste. (*a*) Stencil printing. (*b*) Dispensing.

machines with changeable placement heads will be configurable to the current mounted component mix within a short time. The whole line can thus be optimized to the currently produced board in order to achieve higher throughput and reduce the placement costs.

3.4.3. Component and PCB Feeding

An important factor for the throughput and the availability of a placement machine is the component feeding. Depending on package forms, the user can choose between different types of packages (Figure 32). For packages with low or middle complexity, taped components are favored.

With the development of improved feeders, the use of components in bulk cases is becoming more important. Compared to tapes, bulk cases have reduced packaging volume (lower transport and

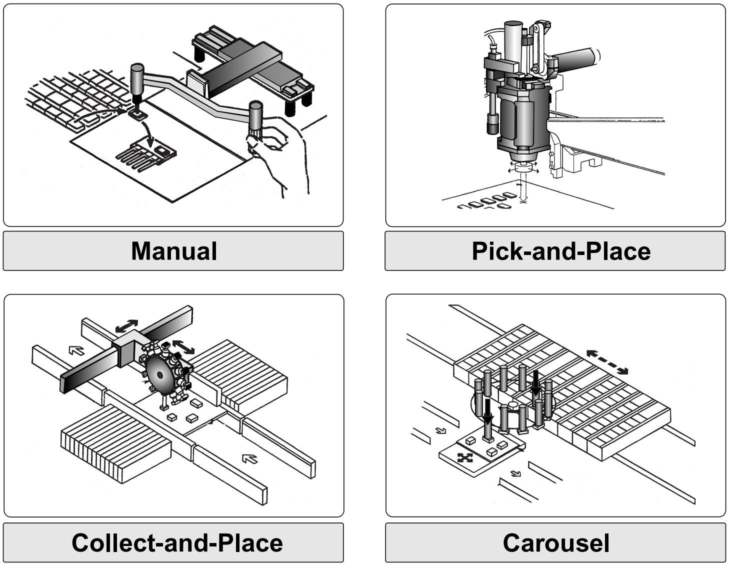

Manual
Pick-and-Place
Collect-and-Place
Carousel

Figure 31 Alternative Concepts for Component Placement.

stock costs) and higher component capacity (less replenishing) and are changeable during the placement process. The components can be disposed with higher accuracy, and the cassette can be reused and recycled. The disadvantages of the bulk cases are that they are unsuitable for directional components and the changeover to other components takes more time and requires greater accuracy.

For feeding complex components, often in low volumes, waffle pack trays are used. Automatic exchangers, which can take up to 30 waffle packs, reduce the space requirement.

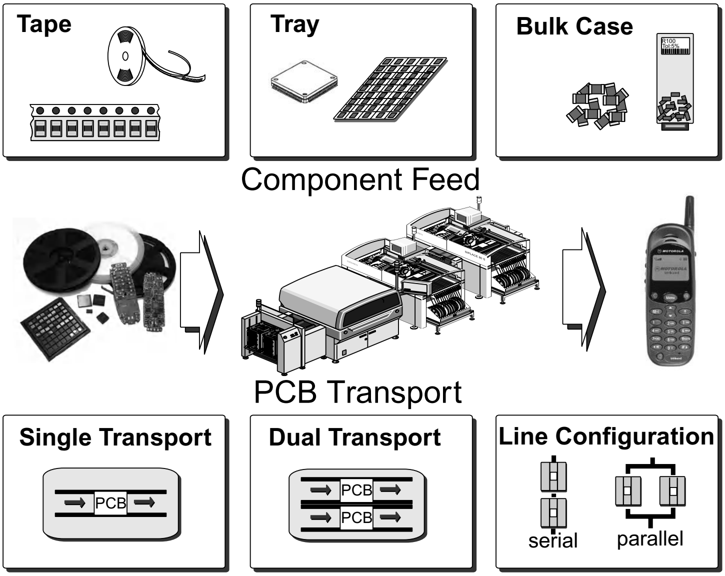

Tape
Tray
Bulk Case
R100
Tol:5%
Component Feed
PCB Transport
Single Transport
PCB
Dual Transport
PCB
PCB
Line Configuration
serial
parallel

Figure 32 Component and PCB Feeding.

The trend towards more highly integrated components is reducing placement loads for PCBs. As a direct result, the nonproductive idle time (PCB transport) has a greater influence on the processing time and leads to increased placement costs. With the use of a double transport (Figure 32) in asynchronous mode, the transport time can be eliminated. During component placement onto the first PCB, the next PCB is transported into the placing area.

For the optimal placement rate to be achieved, high-performance placement machines have to assemble more than 300 components per board. To reach this optimal operating point of the machine with an acceptable cycle time, the machines are connected in parallel instead of in a serial structure. Parallel machines place the same components. Depending on the product, the placement rate can be raised up to 30%.

3.4.4. Measures to Enhance Placement Accuracy

The trend in component packaging towards miniaturization of standard components and the development of new packages for components with high pin account (fine pitch, μBGA, flip-chip) are increasing requirements for placement accuracy.

For standard components, an accuracy of 100 μm is sufficient. Complex components must be placed with an accuracy better than 50 μm. For special applications, high-precision machines are used with an accuracy of 10 μm related to a normal distribution and at a standard deviation of 4σ. This means, for example, that only 60 of 1 million components will be outside a range of ± 10 μm. Due to the necessity for using a small amount of solder paste with fine-pitch components (down to 200 μm pitch), minimal vertical bending of the component leads (about 70 μm) causes faulty solder points and requires cost-intensive manual repairs.

Each lead is optically scanned by a laser beam during the coplanarity check after the pick-up of the component (Figure 33), and the measured data are compared with a default component-specific interval. Components outside the tolerance are automatically rejected.

Given the increasing requirements for placement accuracy, mechanical centering of PCBs in the placement process will not suffice. The accuracy needed is possible only if gray-scale CCD camera systems are used to locate the position of the PCB marks (fiducials) in order to get the position of the PCB proportional to the placement head and the twist of the PCB. Additionally, local fiducials are necessary for mounting fine-pitch components. Furthermore, vision systems are used to recognize the position of each component before placement (visual component centering) and to correct the positions. Instead of CCD cameras, some placement systems are fitted with a CCD line. The component is rotated in a laser beam and the CCD line detects the resulting shadow.

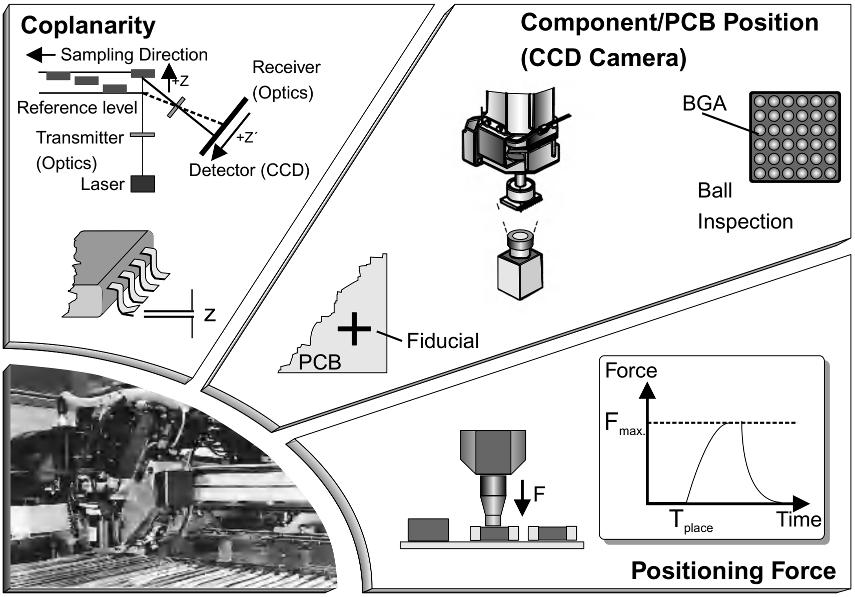


Figure 33 Measures to Enhance Placement Accuracy.

Component-specific illumination parameters are necessary for ball/bump inspection and centering of area array packages.

Direct optoelectronic scanning units are used for the positioning control of the axis to minimize positioning offsets. The glass scales have a resolution up to 1 increment per μm. The influence of the temperature drifts of, for example, the drives can be partly compensated for.

Despite the preventive measures and the high quality of the single systems in placement machines, reproducible errors caused by manufacturing, transport, or, for example, crashes of the placement head during the process must be compensated for. Several manufacturers offer different calibration tools that make inline mapping of the machines possible. Highly accurate glass components are positioned on a calibration board (also glass) with special position marks. The fiducial camera scans the position of the component proportional to the board (resolution about 2–3 μm). Extrapolating a correction value allows the placing accuracy of the machine to be improved.

Fine-pitch components must be placed within a close interval to guarantee sufficient contact with the solder paste but avoid deformations of the leads. Adapted driving profiles are necessary to reach the optimal positioning speed with accelerations up to 4 g, but the last millimeters of the placing process must be done under sensor control to take the positioning power down to a few newtons.

3.5. Interconnection Technology

In electronics production, two main principles of interconnection are used: soldering using metal-based alloys and adhesive bonding with electrically conductive adhesives (Rahn 1993).

Soldering is a process in which two metals, each having a relatively high melting point, are joined together by means of an alloy having a lower melting point. The molten solder material undergoes a chemical reaction with both base materials during the soldering process. To accomplish a proper solder connection, a certain temperature has to be achieved. Most solder connections in electronics are made with conventional mass soldering systems, in which many components are soldered simultaneously onto printed circuit boards. Two different mass soldering methods are used in today's electronics production. Wave or flow soldering is based on the principle of simultaneous supply of solder and soldering heat in one operation. The components on a PCB are moved through a wave of melted solder. In contrast, during reflow soldering process solder preforms, solid solder deposits or solder paste attached in a first operation are melted in a second step by transferred energy.

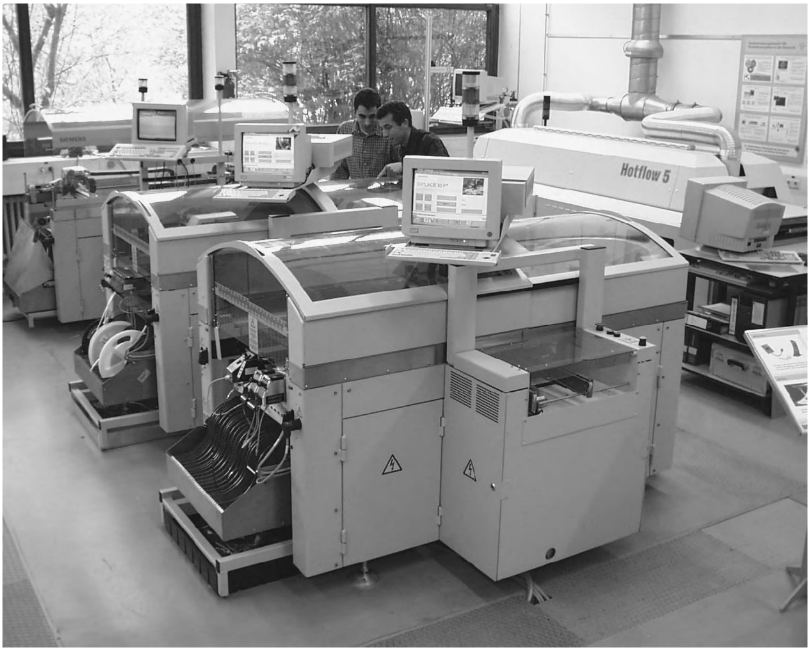


Figure 34 SMT Placement Line (Siemens Siplace).

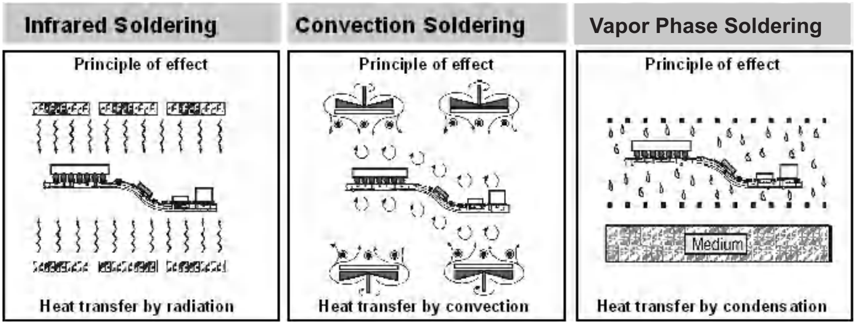

Infrared Soldering
Convection Soldering
Vapor Phase Soldering
Principle of effect
Principle of effect
Principle of effect
Medium
Heat transfer by radiation
Heat transfer by convection
Heat transfer by condensation

Figure 35 Reflow Soldering Methods.

Used in more than 80% of the electronic components processed worldwide, reflow soldering technology is the predominant joining technology in electronics production. Reflow soldering works satisfactorily with surface mount technology and can keep up with the constantly growing demands for productivity and quality. Current developments such as new packages or the issue of lead-free soldering, however, are creating great challenges for reflow soldering technology.

The reflow soldering methods of electronics production can be divided according to the type of energy transfer into infrared (IR), forced convection (FC), and condensation soldering. In contrast to the other versions of energy transfer, reflow soldering using radiant heating is not bound to a medium (gas, vapor), but is subject to electromagnetic waves. For IR reflow soldering, short- to long-wave infrared emitters are used as energy sources. Although heat transfer established with this method is very efficient, it is strongly influenced by the basic physical conditions. Depending on the absorptivity, reflectivity, and transmissivity of the individual surface, large temperature differences are induced

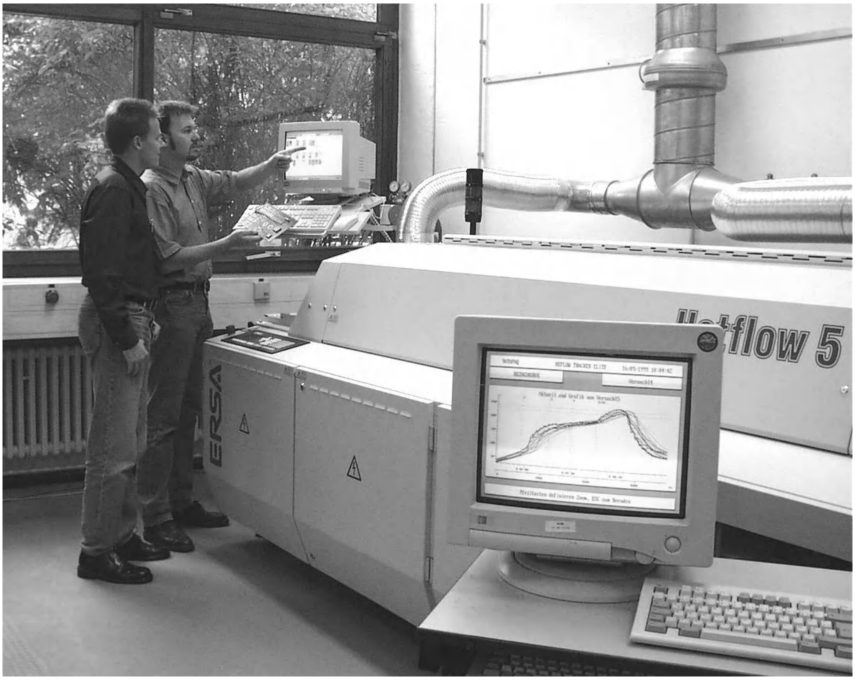

ERSA
Hotflow 5

Figure 36 Soldering Profile.

across the electronic modules. The problem of uniform heat distribution is increased by intricate geometries, low thermal conductivity, and variable specific heat and mass properties of the individual components. Therefore, this soldering method leads to very inhomogeneous temperature distributions on complex boards.

The transfer of energy by means of forced air convection induces a more uniform heat distribution than in IR soldering. The heating up of the workpiece is determined by, in addition to the gas temperature, the mass of gas transferred per time unit, the material data of the gas medium, and by the geometry and thermal capacity of the respective PCB and its components.

Reflow soldering in a saturated vapor (vapor phase or condensation soldering) utilizes the latent heat of a condensing, saturated vapor, whose temperature corresponds to the process temperature, to heat the workpiece. Due to the phase change from vapor to liquid during condensation, the heat transfer is very rapid, resulting in very uniform heating that is relatively independent of different specific heat and mass properties or geometric influences. An overheating of even complex surfaces is impossible due to the nature of vapor phase heating.

Despite the optimization of these mass soldering methods, an increasing number of solder joints are incompatible with conventional soldering methods, so microsoldering (selective soldering) methods have to be used. In the last few years there has been a steady flow of requests for selective soldering. Several systems have been designed and introduced into the market. Depending on the kind of heat transfer, different systems using radiation, convection, or conduction can be distinguished. The most popular selective soldering in modern assembly lines seems to be fountain soldering, a selective process using molten solder. The process is similar to the conventional wave soldering process, with the difference that only a few joints on a PCB are soldered simultaneously.

In contrast to soldering, conductive adhesives are used for special electronic applications. Conductive adhesives simultaneously establish mechanical and electrical joints between PCBs and components by means of a particle-filled resin matrix. Whereas the polymer matrix is responsible for the mechanical interconnection, the filling particles (silver, palladium, or gold particles) provide the electrical contact between PCB and component. Therefore, in contrast to solder joints, conductive adhesive joints have a heterogeneous structure.

3.6. Quality Assurance in Electronics Production

Quality assurance has become a major task in electronics production. Due to the complex process chains and the huge variety of used materials, components, and process steps, the quality of the final product has to be assured not only by tests of the finished subassemblies but also by integrated inspection steps during processing. It is common knowledge that the solder application process causes about two thirds of process-related failures but is responsible for only about 20% of process and inspection costs (Lau 1997). This means that the first step of processing leads to a huge amount of potential failures but is not or cannot be inspected sufficiently. Therefore, a combined strategy for the assembly of complex products such as automotive electronic or communication devices is useful in electronics production (Figure 38). Both capable processes and intelligent inspection tools are needed.

Several optical inspection systems are available that measure, for example, shape and height of the applied solder paste or placement positions of components. These systems, based on image

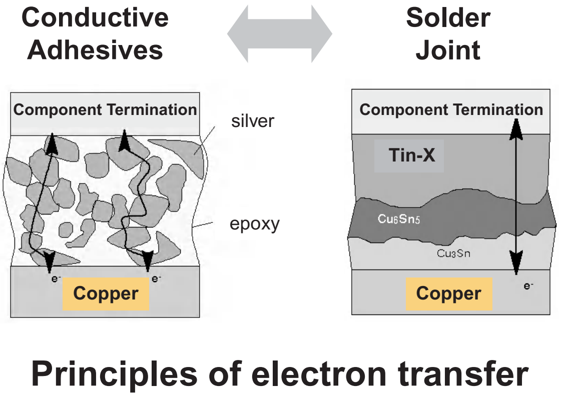


Figure 37 Comparison: Conductive Adhesive—Soldering Interconnection.

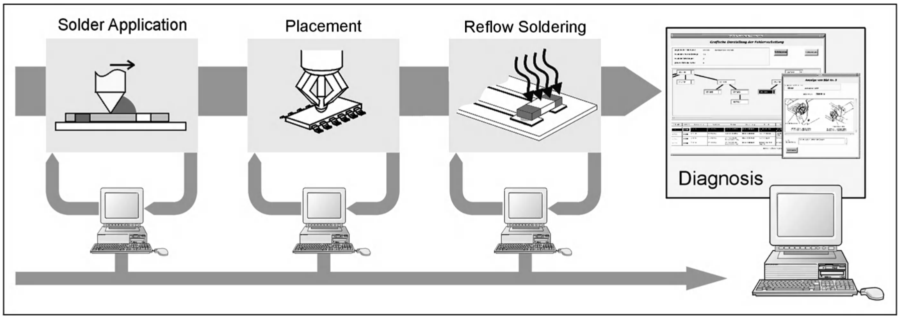


Figure 38 Basic Concept for Accompanying Quality Assurance.

processing with CCD cameras, capture 2D images for analysis. Advanced systems work with 3D images for volume analysis or with x-ray vision for visualization of hidden solder joints (as are typical for area arrays). The main problem with complex inspection strategies is inline capability and the long duration of the inspection itself. As a result, 3D inspection is used almost entirely for spot testing or in scientific labs. On the other hand, inline strategies are often combined with and linked to external systems for quality assurance. These systems are suitable for collecting, handling, and analyzing all relevant data during processing. They allow both short-term control loops within a single step and long-term coordination strategies for optimizing assembly yields. The vital advantage is the direct link between a real product and its database information.

Further tasks of quality assurance systems occur in diagnosis and quality cost analysis. Machine diagnosis is necessary to ensure and enhance machine or process capabilities. Especially, systematic errors such as constant placement offsets can only be detected and regulated by integrated diagnosis sensors. Further, telediagnosis and defect databases are possible. Both support quick and direct removal of problems with the assistance of the machine supplier, who remains distant. Another goal is to calculate process- and defect-related costs in electronics production. With an integrated quality cost-evaluation system, it should be possible to optimize not only quality, but costs as well within one evaluation step. This will finally lead to a process-accompanying system that focuses on technically and economically attractive electronics production.

4. INTEGRATION OF MECHANICAL AND ELECTRONIC FUNCTIONS

The use of thermoplastics and their selective metal plating opens a new dimension in circuit carrier design to the electronics industry: 3D molded interconnect devices (3D MIDs). MIDs are injection-molded thermoplastic parts with integrated circuit traces. They provide enormous technical and economic potential and offer remarkably improved ecological behavior in comparison to conventional printed circuit boards which they will, however, complement, not replace.

4.1. Structure of Molded Interconnect Devices

The advantages of MID technology are based on the broader freedom of shape design and environmental compatibility as well as on a high potential for rationalizing the process of manufacturing the final product. The enhanced design freedom and the integration of electrical and mechanical functions in a single injection-molded part allow a substantial miniaturization of modules to be obtained.

The rationalization potential comes from reduced part and component counts and shortened process chains. In several applications, it was possible to replace more than 10 parts of a conventional design solution with a single MID, thus reducing assembly steps and material use. Further potential lies in the increased reliability.

Additional advantages are obtainable regarding environmental compatibility and electronics waste-disposal regulations. MID technology allows the material mixture of a (conventional) combination of PCB and mechanical parts, which usually consist of a great number of materials, to be replaced by a metallized plastic part (MID). MIDs are made of thermoplastics, which can be recycled and are noncritical in disposal (Franke 1995).

The novel design and functional possibilities offered by MIDs and the rationalization potential of the respective production methods have inevitably led to a quantum leap in electronics production. The most important functions that can be integrated into an MID are depicted in Figure 39. So far, cost savings of up to 40% have been achieved by integration, depending, of course, on the specific product, lot sizes, and other boundary conditions.

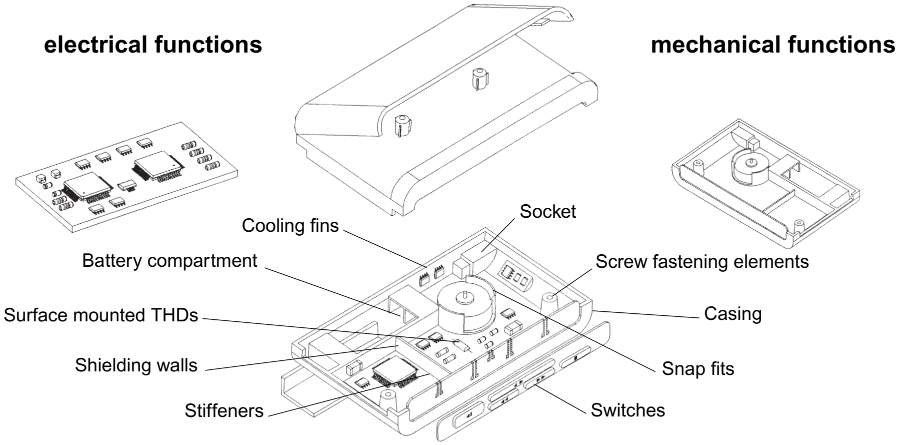


Figure 39 Advantages of MID Technology.

Key markets for MID technology are automotive electronics and telecommunication. MIDs are also suitable for use in computers, household appliances, and medical technology. The market currently shows an annual growth rate of about 25%.

Typically high geometrical and functional flexibility make MID suitable for applications in peripheral integrated electronics, such as small control units and portable devices. Past and present MID products bear this out. Standard electronics applications with high component density and multiple layers, such as motherboards in computers, will continue to be produced with conventional technologies.

MID product development shows a trend towards the implementation of MIDs in new fields of application. Examples include components in the telecommunications and computer industries, security-relevant parts in the automotive industry, and customized packages in components manufacturing (Figure 40) (Pöhlau 1999).

4.2. Materials and Structuring

The MID process chain can be divided in four main steps: circuit carrier production, metallization and structurization, component mounting, and finally joining of electronic components. A number of alternative technologies are available for each step.

Some of the production technologies are specific MID technologies, such as two-shot molding and hot embossing. Others are established processes in electronic production that have been altered and adapted to the requirements of MID. During the planning phase of a MID product it is necessary to choose a combination of the single process steps most suitable for the situation, given their respective benefits and limitations.

MIDs can be manufactured in a variety of ways, as shown in Figure 41.

Hot embossing is characterized by low investment for the embossing die or stamp and high efficiency. The layout is applied following injection molding. Because no wet chemistry is needed, hot-embossed MIDs are well suited for decorative surfaces.

The photoimaging process is perfectly suited for generating EMC screens (at the same time as the conductors) and is highly flexible as to the layer structure. The design freedom is good and few limitations exist. The use of 3D photomasks enables well-defined structures to be generated precisely where the mask contacts the MID directly (Franke 1995).

Laser imaging is highly flexible as to layer structure. It basically involves copperplating the entire workpiece surface first with a wet chemical and then with a galvanical treatment to the desired final coating thickness. Flexibility as to layout modification is very high and cost effective: all that is necessary is a rewriting of the laser-routing program.

Of all MID production processes, two-shot injection molding offers the greatest design freedom. Conductor geometries that raise major problems in the other processes, such as conductors on irregularly shaped surfaces in recesses as well as through-platings, are easily realized by means of the two-shot process. A platable and a nonplatable constituent material are injected on top of each other in two operations. In the second shot, the workpiece is supported inside the tool by the surfaces that were contoured in the first shot.

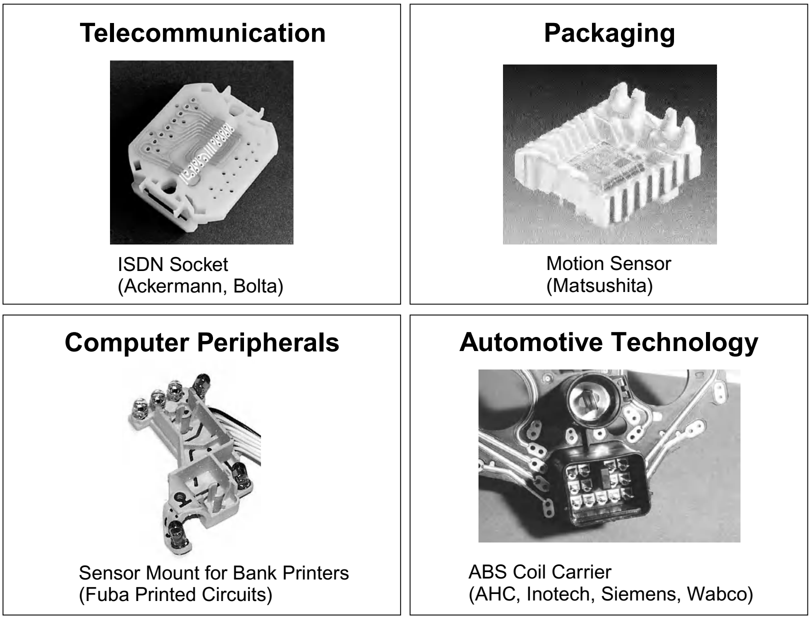

Telecommunication
ISDN Socket
(Ackermann, Bolta)
Packaging
Motion Sensor
(Matsushita)
Computer Peripherals
Sensor Mount for Bank Printers
(Fuba Printed Circuits)
Automotive Technology
ABS Coil Carrier
(AHC, Inotech, Siemens, Wabco)

Figure 40 Application Fields for 3D MID Technology.

The capture decal (film overmolding) technique features a short process chain. It is suitable for decorative surfaces unless the surfaces are finished later on. The conductor pattern can be generated by means of screen printing or conventional PWB technology; in the latter case, through-platings are also possible. The structuring takes place separately and before the injection-molding step.

MID suppliers use various materials to which they have adapted their manufacturing processes. Basically, high-temperature and engineering thermoplastics are plated with various surface materials. The key material properties to be considered are processing and usage temperatures, required flam-

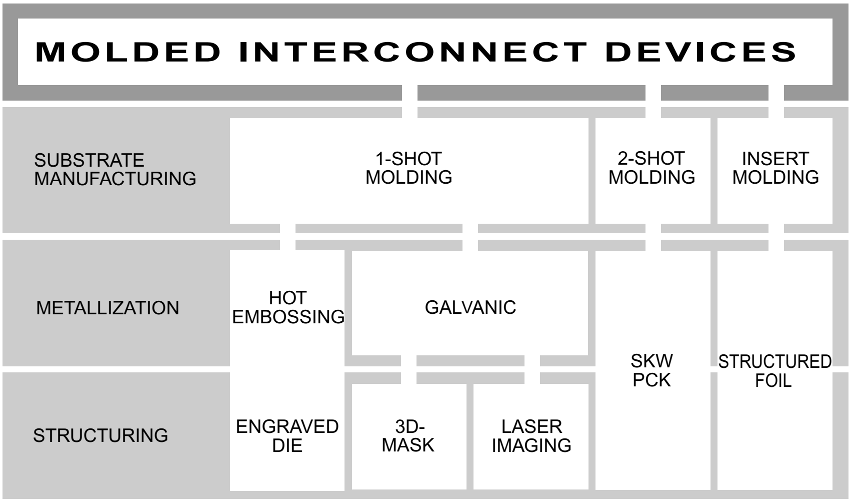

MOLDED INTERCONNECT DEVICES
SUBSTRATE MANUFACTURING
1-SHOT MOLDING
2-SHOT MOLDING
INSERT MOLDING
METALLIZATION
HOT EMBOSSING
GALVANIC
SKW PCK
STRUCTURED FOIL
STRUCTURING
ENGRAVED DIE
3D-MASK
LASER IMAGING

Figure 41 Manufacturing Methods for MIDs.

mability rating, mechanical and electrical properties, moldability and platability, and cost. As the price of thermoplastics generally increases with their thermal resistivity, low-temperature materials, such as ABS, PC, and PP, seem to be advantageous. These materials, however, exclude the use of conventional reflow soldering, thus necessitating an alternative joining method.

These problems can be avoided by the use of high-temperature resistance materials, such as PES, PEI, and LCP, which are also best for chemical plating.

For the completion of MID assemblies after substrate manufacturing and plating, several further steps are normally necessary.

In component placement and soldering several restrictions must be considered, as shown in Figure 42.

Conventional printing processes are less useful for applying conductive adhesives or solder paste. Dispensing of single dots in a complex geometric application is therefore necessary. Three-dimensional circuit carriers also decrease the freedom of component placement, which leads to restrictions for placement systems.

Conventional solder procedures require the use of high-temperature thermoplastics. Other plastics can be processed using low-melting-point solders, conductive adhesives, or selective soldering methods. Available assembly and interconnection technologies as well as processes for the production of the circuit carriers must be carefully considered, selected, and possibly modified to the specific MID materials.

4.3. Placement Systems for 3D PCBs

MIDs create new demands on assembly because of their complex geometric shape. The new requirements on assembly processes caused by different MID types in order to develop qualified production systems for MID assembly must be considered (Figure 43). The new task of mounting SMDs onto 3D circuit boards calls for new capabilities in dispensing and mounting systems. Both processes work sequentially and are realized with Cartesian handling systems for PCB production.

One way for SMD assembly onto MID is to use six-axis robots, which are designed as geometry-flexible handling systems for reaching each point in the workspace with every possible orientation of the tool. The available placement systems for SMD assembly onto PCBs are optimized for placement accuracy and speed. In order to work with these systems on MIDs, it is important to be able to move them during the process (Feldmann and Krimi 1998).

The first step was a systematic development of a MID placement system outgoing from an available four-axis system. This aim is realized by extending the machine by two DOF. First, an additional movement in the direction of *z*-axis allows high obstacles to be surmounted. Second, the MID can be inclined so that the process plane is oriented horizontally in the workspace. This concept is not suitable for MIDs with high geometrical complexity.

4.3.1. Six-Axis Robot System for SMD Assembly onto MID

Robot placement systems have till now been used primarily to place exotic THT components such as coils and plugs, whose assembly is not possible with pick-and-place machines. The application of

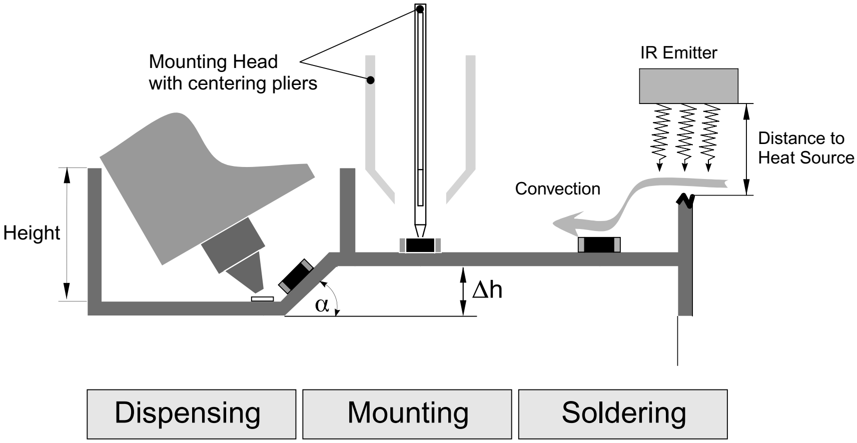


Figure 42 Definition of Design Rules for MID.

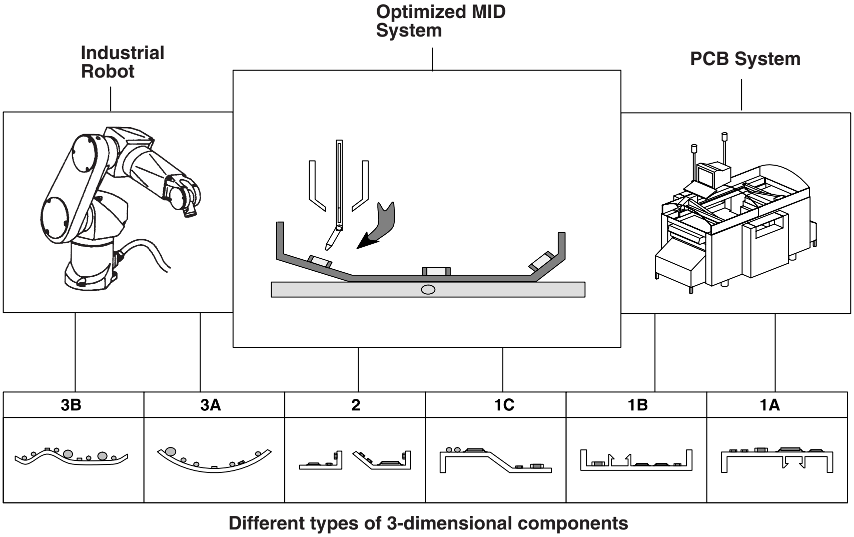


Figure 43 New Requirements for Assembly Process Caused by 3D Circuit Board.

MIDs has lead to a new task for robots in electronics assembly. Robots can be equipped with pipettes, which are built small enough to assemble SMDs with a small distance to obstacles. If the length of the free assembly perpendicular of a MID is limited, bent pipettes can be used without inclined feeders or a centering station.

Available tool-changing systems make it possible to integrate additional processes such as application of solder paste and soldering into one assembly cell. A critical point in surface-mounted technology is the necessary accuracy in centering the SMD to the pipette and the accurately placing components. This is realized by the use of two cameras, one mounted on joint five of the manipulator, pointing down toward a reference mark, and the other mounted under the stage, looking up for component registration.

The robot cell layout was designed to minimize the total distance to be covered by the robot in order to assemble onto MID. In addition, the feeder system is movable. As a result, the components can be placed at an optimized position. With this feeder moving system, the placement rate can reach 1800 components/hr (with a stationary feeder the placement rate is up to 1200 components/hr).

This geometric advantage and the flexibility of the robot system come at the expense of lower productivity and accuracy than with Cartesian systems. Therefore, Cartesian placement systems should be the basis for a specialized MID assembly system.

4.3.2. Optimized MID Placement System

The next step is to find out the possibilities for extending the detected limits of conventional SMD assembly systems. The insufficiency of Cartesian systems lies in the possible height and angle of inclination of the MID and the fact that an unlimited length of free assembly perpendicular is required. A concept has been developed to extend these limits. A module for handling the MID in the workspace of an SMD assembling system, as well as the use of a pipette with an additional axis, make 3D assembly of component onto MID possible.

For both these concepts to be realized, it is necessary to extend and work on hardware and control software of the PCB basic assembly system. These systems are realized with a widespread standard system at the FAPS Institute. Figure 45 shows the developed MID placement system with the different modules for the handling of SMD components and MID.

The way to enlarge the system kinematics of standard placement systems is to move the workpiece during the assembly process. Solutions for this problem are as follows.

First, an additional movement in the direction of the z-axis enlarges the possible height of obstacles that can be surmounted (six-axis pipette). Second, the MID can be inclined so that the process plane is oriented horizontally in the workspace (MID handling system). This allows the high accuracy and speed of a Cartesian system to be used. The tolerance chain is divided into the two parts: SMD handling system and MID moving system. Thus, it is possible to compensate for the possibly lower

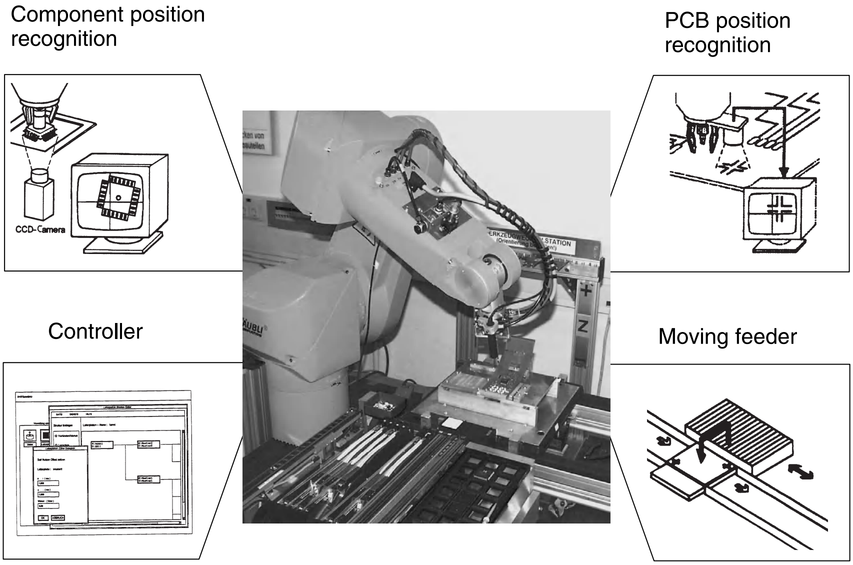

Component position recognition
CCD-Camera
PCB position recognition
Controller
Moving feeder

Figure 44 Flexible Robot Placement System for Assembly onto MID.

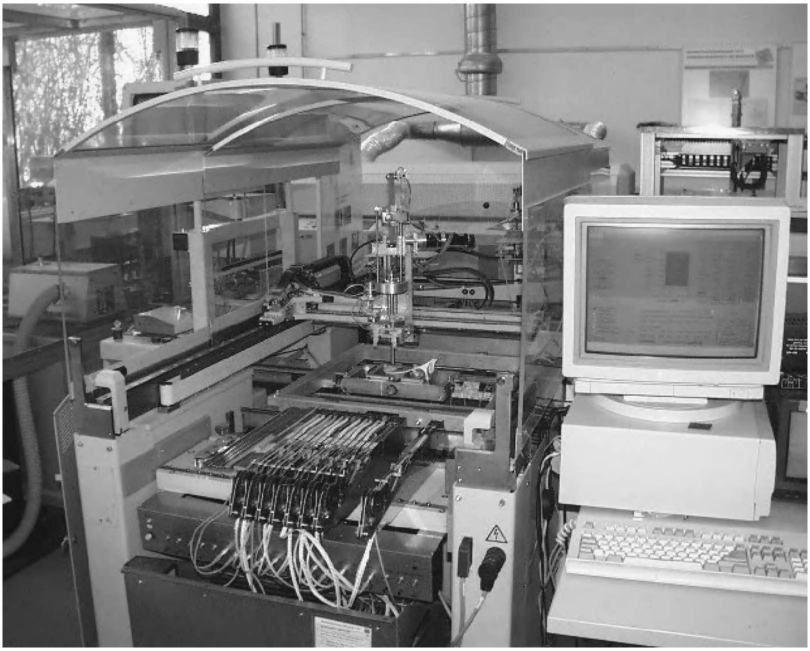


Figure 45 Prototypical Realization of MID Placement System for 3D assembly. (From Feldmann et al. 1998)

position accuracy of the MID moving system by correction of the handling system. The use of vision system is necessary here in order to detect the position of the inclined MID.

Inclination of the MID means that the process plane, which should be oriented horizontally, may be on another level than the usual. Its height must be compensated for by an additional *z*-axis. This means that both possibilities of enlargement of Cartesian assembly systems could be integrated into one lift-and-incline module. With this module, integrated into the Cartesian SMD assembly system, it is possible to attach SMD components onto MID circuit planes with an angle of inclination up to 70°. For this, the MID is fixed on a carrier system that can be transported on the double-belt conveyor. At the working position the carrier is locked automatically to the module, the connection to the conveyor is opened, and the MID can be inclined and lifted into the right position. It is thus possible to enlarge the capability of the Cartesian assembly system almost without reducing productivity.

4.4. Soldering Technology for 3D PCBs

Processing 3D molded interconnect devices in reflow soldering processes places stringent requirements on reflow technology due to the base materials used and the possible geometrical complexity of the circuit carriers. Most MID solutions, which are assembled with electronic components, consist of expensive high-temperature thermoplastics, such as PEI, PES, and LCP, which can resist high thermal stress during reflow soldering. However, in order for the total potential of the MID technique to be used, inexpensive technical thermoplastics like polyamide have been discussed as possible base materials, although the process window would clearly be limited regarding the tolerable maximum temperatures. As a consequence, the process parameters of the selected reflow soldering method have to be modified to reduce the thermal loading to a sufficient minimum value to avoid:

- Thermal damage to the base material
- Degradation of metallization adhesion
- Warping of the thermoplastic device

In addition, the highest solder joint quality has to be ensured for the best quality and reliability to be attained. In extensive investigations of the 3D MID soldering topic, the influence of the reflow soldering process on the points mentioned above has been examined. It can be stated that the different methods of heat transfer clearly influence the temperature distribution on the workpiece and therefore the process temperature level needed when processing complex MIDs. The example of a circuit carrier of MID type 3 (the highest level of geometrical complexity) demonstrates these facts. While during condensation or convection soldering very uniform temperature distributions, with temperature differences on the molded device of only up to a few degrees kelvin can be observed, for IR soldering, temperature differences of about 20–30°K on the surface of the 3D device can be shown by thermal measurement with an infrared scanner or attached thermocouples. Thus, particularly in the more heated upper areas of a 3D device the danger of local thermal damage to the base material exists due to the reduced distance to the infrared emitters.

If thermoplastic base materials are suspended above the usual admissible operating temperature, material damage can occur. For some hygroscopical thermoplastics (e.g., PA) the humidity stored in the material heats up, strongly expands with the transition to the vapor state, and causes a blistering on the surface of the already softened material (Gerhard 1998).

Besides thermal damage to base material caused by local overheating, the influences of thermal stress during soldering and of the soldering parameters (especially reflow peak temperature) on the geometrical shape of the molded part and the peeling strength of the metallization are important to the process quality of thermoplastic interconnection devices.

In summary, for processing 3D interconnect devices in reflow soldering processes, a minimum thermal load of the molded device has to be realized in order to eliminate unwanted thermal material damages as far as possible and achieve highest process yield. In the conflict field between solder joint quality and thermal load of the thermoplastic workpiece, fluid-based soldering methods such as convection and condensation reflow soldering represent the optimal solution. They guarantee homogeneous temperature distributions on low-temperature levels in connection with optimal soldered interconnection joints.

To reduce the thermal stress for components as well as for thermoplastic base substrates, conductive adhesives and selective soldering can be used. In working with adhesives, good interconnections with a maximum temperature of only about 100°C can be established. This temperature is necessary to harden the epoxy material. In the selective soldering technique, the heat is transferred only to places where interconnections have to be made between the component's termination and the pad on the substrate. This reduces the overall thermal stress situation for the whole 3D MID.

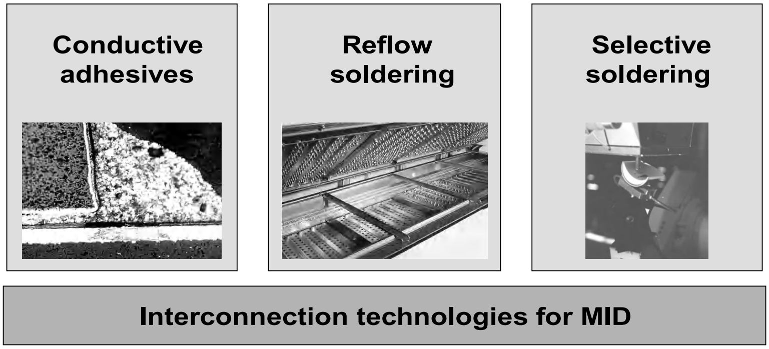

Conductive adhesives
Reflow soldering
Selective soldering
Interconnection technologies for MID

Figure 46 Interconnection Technologies for 3D MIDs.

5. DISASSEMBLY

5.1. Challenge for Disassembly

The development of disassembly strategies and technologies has formerly been based on the inversion of assembly processes. The characteristics of components and connections are like a systematic interface between both processes. In addition, assembly and disassembly can have similar requirements regarding kinematics and tools (Feldmann, et al. 1999). Nevertheless, the most obvious differences between assembly and disassembly are their goals and their position in the product's life cycle.

The objective of assembly is the joining of all components in order to assure the functionality of a product (Tritsch 1996). The goals of disassembly may be multifaceted and may have a major influence on the determination of dismantling technology. The economically and ecologically highest level of product treatment at its end of life is the reuse of whole products or components (Figure 47) because not only the material value is conserved but also the original geometry of components and parts, including its functionality. Thus, disassembly generally has to be done nondestructively in order to allow maintenance and repair of the components.

A further material loop can be closed by the disassembly and reuse of components in the production phase of a new product. In order to decrease disassembly time and costs, semidestructive dismantling technologies are applicable, such as to open a housing.

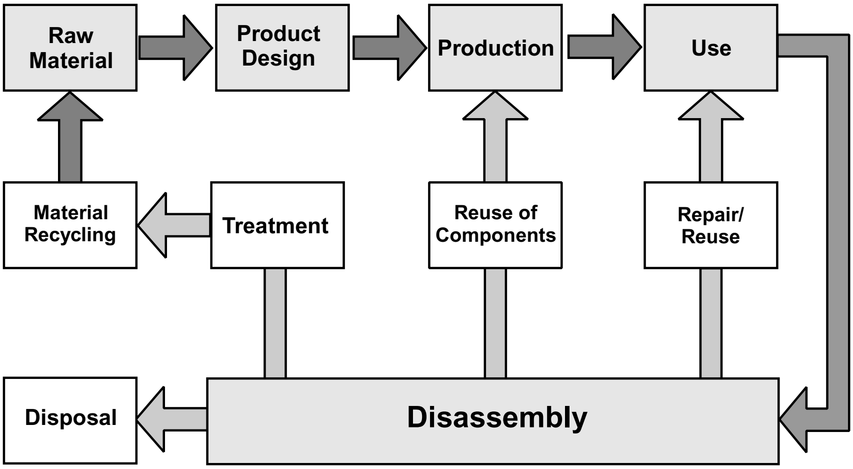

Raw Material
Product Design
Production
Use
Material Recycling
Treatment
Reuse of Components
Repair/ Reuse
Disposal
Disassembly

Figure 47 Role of Disassembly in the Product's Life Cycle.

In most cases, disassembly is done to remove hazardous or precious materials in order to allow ecologically sound disposal or high-quality-material recycling. All types of disassembly technology are used in order to decrease costs and increase efficiency (Feldmann et al. 1994).

In the determination of disassembly technologies, several frame conditions and motivations have to be taken into account (Figure 48). Because of rapidly increasing waste streams, the ecological problems of disposal processes, and the shortage of landfill capacities, the product's end of life has become a special focus of legislative regulation (Gungor and Gupta 1999; Moyer and Gupta 1997) that requires disassembly and recycling efforts.

Furthermore, the ecological consciousness of society (represented by the market and its respective requirements) and the shortage of resources, leading to increasing material prices and disposal costs, put pressure on the efficient dismantling of discarded products. The increasing costs and prices also provide economic reasons for dismantling because the benefits of material recycling increase as well. Finally, on the one hand, the technological boundary conditions indicate the necessity of disassembly because of increasing production and shorter life cycles. And on the other hand, the availability of innovative recycling technologies is leading to new opportunities and challenges for disassembly technology.

Due to influences in the use phase, unpredictable effects such as corrosion often affect dismantling. Also, the incalculable variety of products when they are received by the dismantling facility has a negative influence on the efficiency of disassembly. Together with the lack of information, these aspects lead to high dismantling costs because of the manpower needed (Meedt 1998).

On the other hand, some opportunities in comparison to the assembly, for example, can be exploited. Thus, with disassembly, only specific fractions have to be generated. This means that not all connections have to be released, and (semi)destructive technology is applicable for increasing efficiency.

5.2. Disassembly Processes and Tools

The advantages and disadvantages of manual and automated disassembly are compared in Figure 49. Because manual disassembly allows greater flexibility in the variety of products and types as well as the use of adequate disassembly tools, it is generally used in practice. Due to risk of injury, dirt, the knowledge needed for identification of materials, and wages to be paid, manual disassembly is not considered an optimal solution. On the other hand, problems of identification, the influence of the use phase (as mentioned above), damage to connections, and particularly the broad range of products, prevent the effective use of automated disassembly processes (Feldmann et al. 1994). Thus, for most disassembly problems partial automated systems are the best solution.

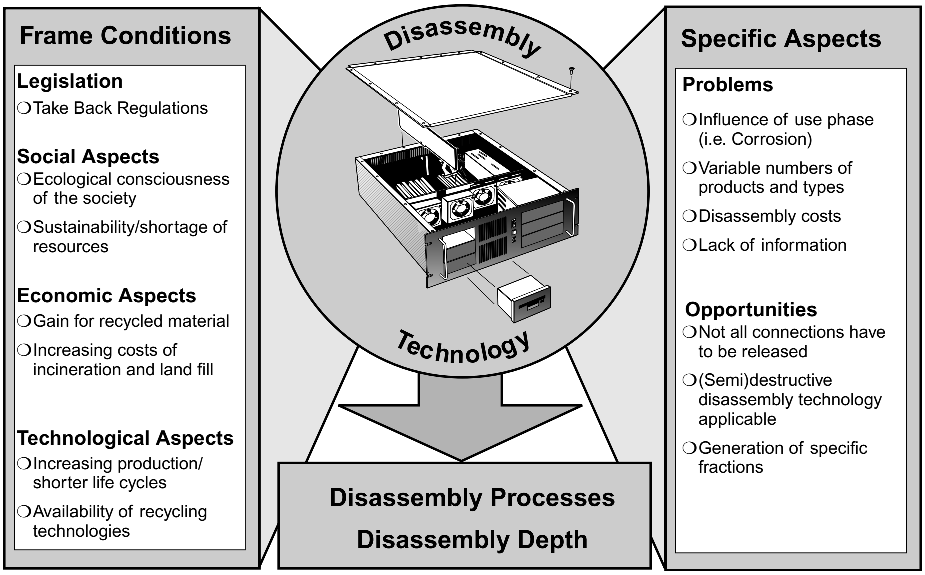

Frame Conditions
Legislation
❍Take Back Regulations
Social Aspects
❍Ecological consciousness of the society
❍Sustainability/shortage of resources
Economic Aspects
❍Gain for recycled material
❍Increasing costs of incineration and land fill
Technological Aspects
❍Increasing production/ shorter life cycles
❍Availability of recycling technologies
Disassembly
Technology
Disassembly Processes
Disassembly Depth
Specific Aspects
Problems
❍Influence of use phase (i.e. Corrosion)
❍Variable numbers of products and types
❍Disassembly costs
❍Lack of information
Opportunities
❍Not all connections have to be released
❍(Semi)destructive disassembly technology applicable
❍Generation of specific fractions

Figure 48 Frame Conditions and Specific Aspects of Disassembly.

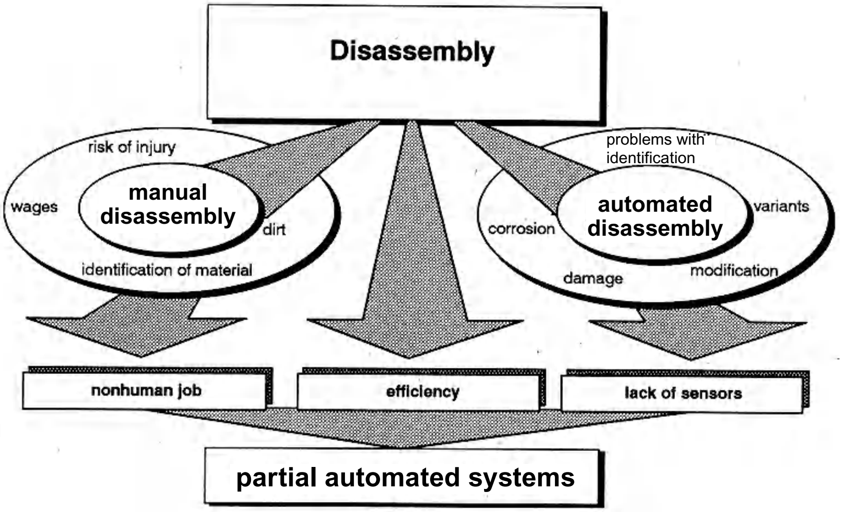


Figure 49 Features of Manual and Automated Disassembly Systems.

Along with the combination of manual and automated disassembly, other aspects, such as logistics, flow of information, and techniques for fractioning, storing, and processing residuals, have to be optimized for efficient dismantling.

In addition to the planning of the optimal disassembly strategy, the major area for improvement in the disassembly of existing products is in increasing the efficiency of disassembly processes. Highly flexible and efficient tools especially designed for disassembly are needed.

The function of tools is based on different types of disassembly: (1) destructive processes in order to bring about the destruction of components or joining; (2) destructive processes in order to generate working points for further destructive or nondestructive processes; (3) nondestructive processes such as the inversion of assembly processes (Feldmann et al. 1999).

Figure 50 shows an example of a flexible unscrewing tool developed at the Technical University of Berlin that generates working points in a destructive way. This tool consists of an impact mass that is speeded up in a first step by tangent pneumatic nozzles until a certain angular velocity is reached (Seliger and Wagner 1996).

In a second step, the rotating impact mass is accelerated towards a conical clutch and transmits the linear and rotational impulse to the end effector. Using W-shaped wedges of the end effector, a new acting surface is generated on the head of the screw, avoiding any reaction force for the worker.

Furthermore, the rotational impulse overcomes the breakaway torque and the linear impulse reduces pre-tension in order to ease loosening. The loosened screw is now unscrewed by a pneumatic drive and the impact mass is pushed to the starting position (Seliger and Wagner 1996). In case a screw cannot be released due to influences in the use phase (e.g., corrosion or dirt), the end effector can be used as a hollow drill with wide edges in order to remove the screw head.

A major problem in dismantling is in opening housings quickly and efficiently in order to yield optimal accessibility to hazardous or worthy materials. Especially with regard to small products, the effort in dismantling with conventional tools is very often too high compared to the achieved benefit. To solve this problem, a flexible tool, the splitting tool (Figure 51), has been developed (Feldmann et al. 1999).

The splitting tool has specific levers and joints that divide the entry strength or impulse (e.g., by a hammer or pneumatic hammer) into two orthogonal forces or impulses.

Through the first force component with the same direction as the original impact, the splitting elements are brought into action with the separating line of the housing in order to generate acting surfaces. The strength component set normally to the entry impact is used simultaneously as separation force. This way, screws are pulled out and snap fits are broken.

Thus, in general the housing parts are partially destroyed and torn apart without damaging the components inside. Using specially designed tools for disassembly allows unintentional destruction of other components, which is often an effect of dismantling with conventional tools such as hammers and crowbars, to be avoided.

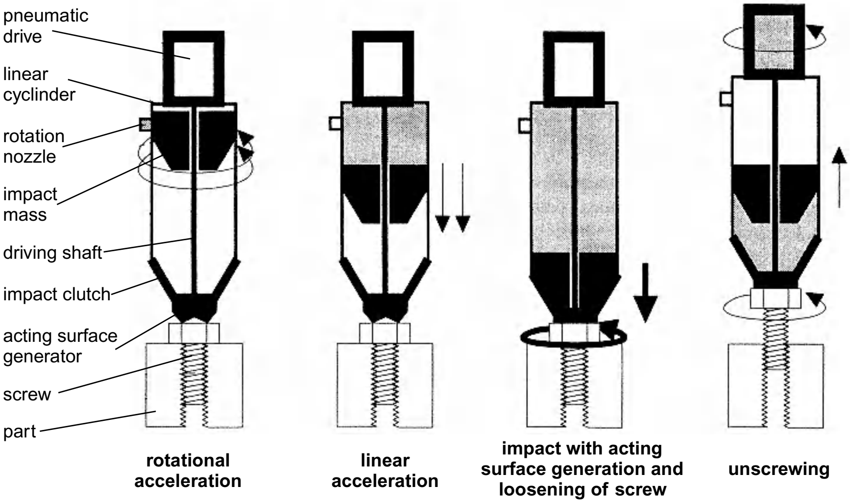

pneumatic drive
linear cyclinder
rotation nozzle
impact mass
driving shaft
impact clutch
acting surface generator
screw
part
rotational acceleration
linear acceleration
impact with acting surface generation and loosening of screw
unscrewing

Figure 50 Example of a Flexible Unscrewing Tool. (From Seliger and Wagner 1996)

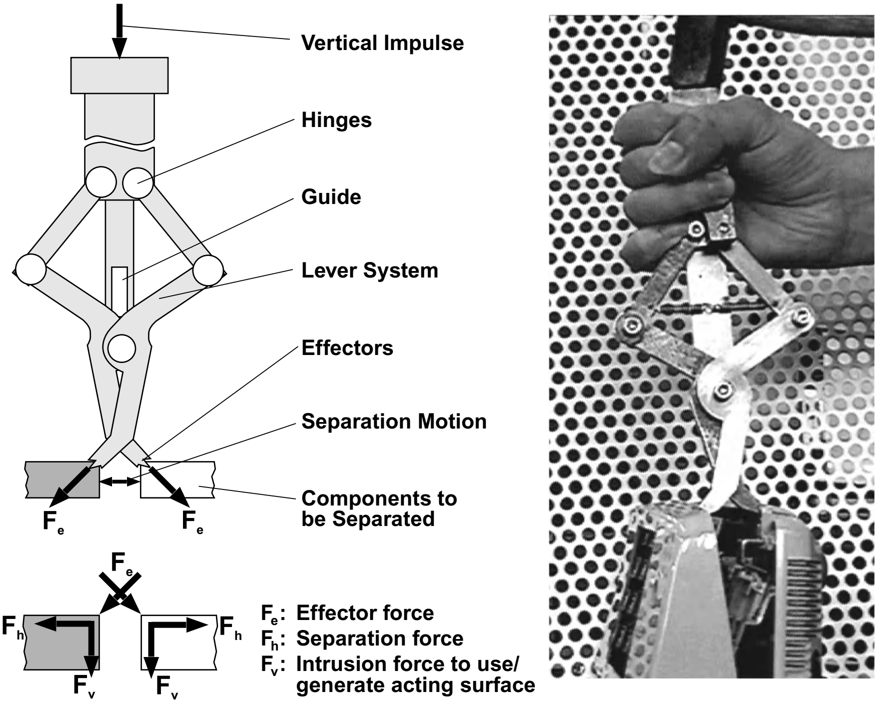

Vertical Impulse
Hinges
Guide
Lever System
Effectors
Separation Motion
Components to be Separated
F_e
F_e
F_e
F_h
F_h
F_v
F_v
F_e: Effector force
F_h: Separation force
F_v: Intrusion force to use/ generate acting surface

Figure 51 Example of a Flexible Splitting Tool.

The examples given show the basic requirements for disassembly tools: flexibility concerning products and variants, flexibility concerning different processes, and failure-tolerant systems.

5.3. Applications

Conventional tools such as mechanized screwdivers, forceps, and hammers are still the dismantling tools generally used. The use of tools, the choice of disassembly technology, and the determination of the disassembly strategy depend to a large degree on the knowledge and experience of the dismantler because detailed disassembly plans or instructions are usually not available for a product. This often leads to a more or less accidental disassembly result, where the optimum between disassembly effort and disposal costs/recycling gains is not met. Many approaches have been proposed for increasing the efficiency of disassembly by determinating an optimal dismantling strategy (Feldmann et al. 1999; Hesselbach and Herrmann 1999; Gungor and Gupta 1999).

Depending on the actual gains for material recycling and reusable components, the costs of disposal, and legislative regulations, products are divided into several fractions (Figure 52). First, reusable components (e.g., motors) and hazardous materials (e.g., batteries, capacitors) are removed. Out of the remaining product, in general the fractions ferrous and nonferrous metals and large plastic parts are dismantled so they can be recycled directly after the eventual cleaning and milling processes.

Removed metal–plastic mixes can be recycled after the separation in mass-flow procedures (e.g., shredding). Special components such as ray tubes are also removed in order to send them to further treatment. The remainder must be treated and/or disposed of.

As mentioned in Section 5.2, the economic margin of dismantling is very tight and the logistic boundary conditions—for example, regarding the sporadic number of similar units and the broad variety of product types—are rather unfavorable for automation. Depending on the flexibility of automated disassembly installations, three strategies can be used.

Some standardized products (e.g., videotapes, ray tubes of computer monitors) can be collected in large numbers so that automated disassembly cells especially designed for these products can work economically.

Greater flexibility from using PC-based control devices and specially designed disassembly tools, as well as the integration of the flexibility of human beings allow the automated dismantling of a certain range of discarded products in any sequence. Figure 53 shows an example of a hybrid disassembly cell that was realized on a laboratory scale. One of the integrated robots, used for unscrewing actions, is equipped with a flexible unscrewing tool that can release various types of screws. The second robot, using a special multifunctional disassembly and gripping device, is able to split components, cut cables, and remove parts without changing the device.

Another strategy for automated disassembly is the enhancement of flexibility using sensors that detect automatically connection locations and types. By an evaluation of the sensor data, the re-

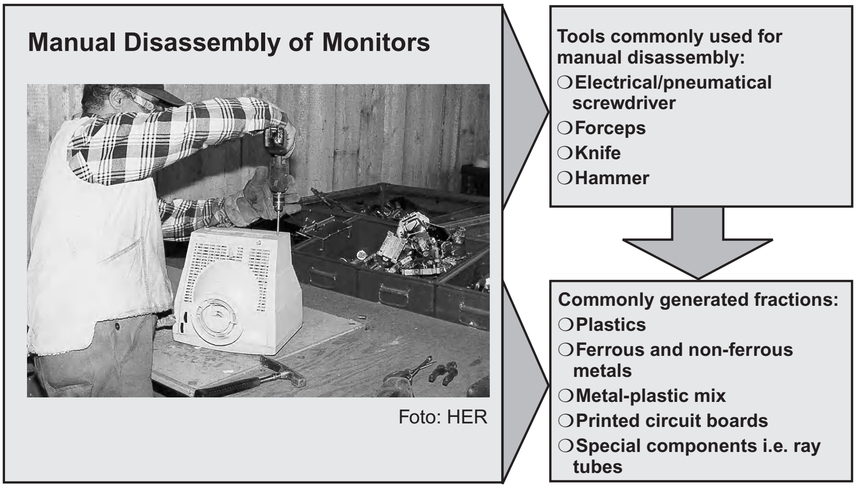

Manual Disassembly of Monitors
Foto: HER
Tools commonly used for manual disassembly:
○ Electrical/pneumatical screwdriver
○ Forceps
○ Knife
○ Hammer
Commonly generated fractions:
○ Plastics
○ Ferrous and non-ferrous metals
○ Metal-plastic mix
○ Printed circuit boards
○ Special components i.e. ray tubes

Figure 52 Manual Disassembly of Monitors.

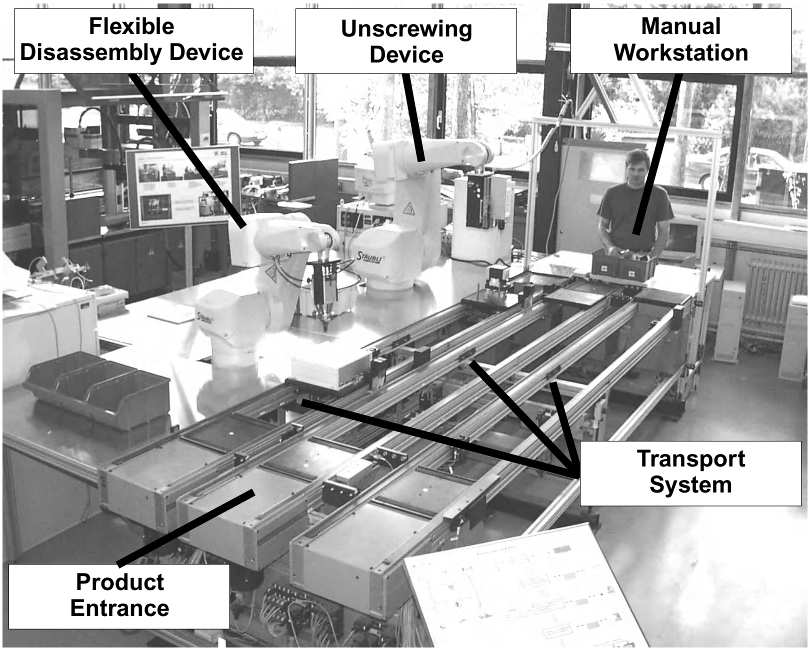

Flexible Disassembly Device
Unscrewing Device
Manual Workstation
Transport System
Product Entrance

Figure 53 Layout of a Hybrid Disassembly Cell.

spective tools are determined and controlled (Weigl 1997). This strategy in general leads to high investments that have to be compensated for by disassembly efficiency.

5.4. Entire Framework for Assembly and Disassembly

Mostly in regard to maintenance and reuse of components, but also in regard to optimal disassembly planning, an integrated assembly and disassembly approach should be considered—just with the design of a product (DFG 1999). Five stages of integration can be distinguished (Figure 54).

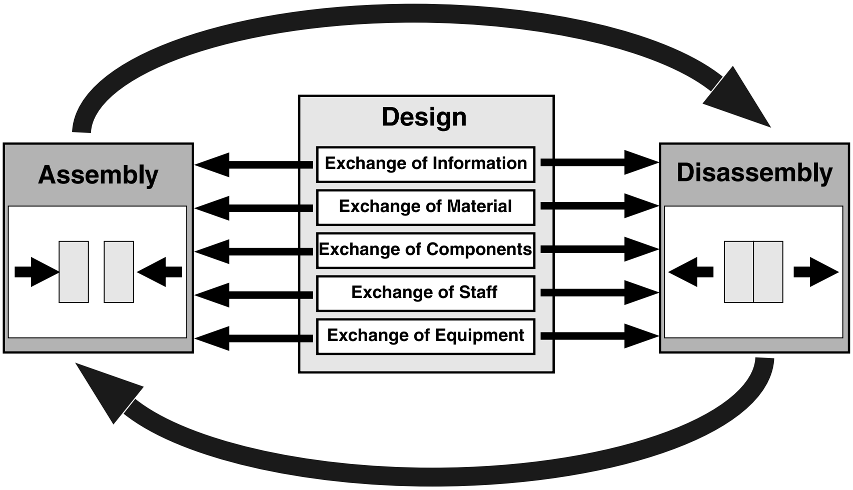

Design
Exchange of Information
Exchange of Material
Exchange of Components
Exchange of Staff
Exchange of Equipment
Assembly
Disassembly

Figure 54 Stages of the Integration of Assembly and Disassembly.

1. Information is exchanged between assembly and disassembly. Data on, for example, connection locations and types ease substantially the disassembly of discarded products.
2. Recycled material, such as auxiliary materials, can be delivered from disassembly plants to the production plants.
3. The reuse and reassembly of components, which first requires disassembly, is already practiced with many products. Nevertheless, the integrated consideration of both assembly and disassembly is important for efficiency.
4. The exchange of staff requires the proximity of assembly and disassembly facilities. At this stage, the knowledge and experience of the workers can be used for both processes and thus provide major benefits in efficiency.
5. The highest stage of integration is the use of a device for both assembly and disassembly. Examples are repairing devices for printed circuit boards.

The possible level of integration is determined mainly by the design of a product. Thus, an overall approach that looks at all phases of the product's life cycle is required.

REFERENCES

Boothroyd, G. (1994), "Product Design for Manufacture and Assembly," *Computer-Aided Design,* Vol. 26, pp. 505–520.

Brand, A. (1997), "Prozesse und Systeme zur Bestückung räumlicher elektronischer Baugruppen," Dissertation, Meisenbach, Bamberg.

DFG (1999), *Integration der Montage- und Demontageprozessgestaltung in einem produktneutralen Ansatz,* Project Report, DFG, Erlangen.

Feldmann, K., and Krimi, S. (1998), "Alternative Placement Systems for Three-Dimensional Circuit Board," *Annals of the CIRP,* Vol. 47, No. 1, pp. 23–26.

Feldmann, K., and Rottbauer, H. (1999), "Electronically Networked Assembly Systems for Global Manufacturing," in *Proceedings of the 15th International Conference on Computer-Aided Production Engineering* (Durham), pp. 551–556.

Feldmann, K., and Wolf, K. U. (1996), "Computer Based Planning of Coil Winding Processes for Improvements in Efficency and Quality," in *Proceedings of the Electrical Manufacturing and Coil Winding Conference* (Chicago), pp. 299–305, EMCWA.

Feldmann, K., and Wolf, K. U. (1997), "Improved Winding Quality and Production Efficiency with the help of Computer Based Planning and Programming Systems," in *Proceedings of the Coil Winding, Insulation and Electrical Manufacturing Conference* (Berlin).

Feldmann, K., Meedt, O., and Scheller, H. (1994), "Life Cycle Engineering—Challenge in the Scope of Technology, Economy and General Regulations," in *Proceedings of the 2nd International Seminar on Life Cycle Engineering* (Erlangen, October 10–11), pp. 1–17.

Feldmann, K., Rottbauer, H., and Roth, N. (1996), "Relevance of Assembly in Global Manufacturing," *Annals of the CIRP,* Vol. 45, No. 2, pp. 545–552.

Feldmann, K., Luchs, R., and Pöhlau, F. (1998), "Computer Aided Planning and Process Control with Regard to New Challenges Arising through Three-Dimensional Electronics," in *Proceedings of the 31st CIRP International Seminar on Manufacturing Systems* (Berkeley, CA, May 26–28), pp. 246–251.

Feldmann, K., Trautner, S., and Meedt, O. (1999), "Innovative Disassembly Strategies Based on Flexible Partial Destructive Tools," *Annual Reviews in Control,* Vol. 23, pp. 159–164.

Franke, J. (1995), "Integrierte Entwicklung neuer Produkt- und Produktionstechnologien für räumliche spritzgegossene Schaltungsträger (3D MID)," Dissertation, Carl Hanser, Munich.

Gerhard, M. (1998), "Qualitätssteigerung in der Elektronikproduktion durch Optimierung der Prozessführung beim Löten komplexer Baugruppen," Dissertation, Meisenbach, Bamberg.

Gungor A., and Gupta, S. M. (1999), "Issues in Environmentally Conscious Manufacturing and Product Recovery: A Survey," *Computers and Industrial Engineering,* Vol. 36, No. 4, pp. 811–853.

Hesselbach, J., and Herrmann, C. (1999), "Recycling Oriented Design Weak-Point Identification and Product Improvement," *Proceedings of the International Seminar on Sustainable Development* (Shanghai, November 16–17).

Informationscentrum des Deutschen Schraubenverbandes (ICS) (1993), *ICS-Handbuch Automatische Schraubmontage,* Mönnig, Iserlohn.

Institute for Interconnecting and Packaging Electronic Circuits (IPC) (1996), "Acceptability of Electronic Assemblies," ANSI/IPC-A-610, IPC, Chicago.

Klein Wassink, R. J., and Verguld, M. M. F. (1995), *Manufacturing Techniques for Surface Mounted Assemblies,* Electrochemical Publications, Port Erin, Isle of Man.

Konold, P., and Reger, H. (1997), *Angewandte Montagetechnik,* F. Vieweg & Sohn, Brunswick.

Lappe, W. (1997), *Aufbau eines Systems zur Prozessüberwachung bei Stanznieten mit Halbhohlniet,* Shaker, Aachen.

Lau, J. H. (1993), *Ball Grid Array Technology,* McGraw-Hill, New York.

Lau, J. H. (1997), *Solder Joint Reliability of BGA, CSP, Flip Chip, and Fine Pitch SMT Assemblies,* McGraw-Hill, New York.

Lotter, B. (1992), *Wirtschaftliche Montage,* VDI, Düsseldorf.

Lotter, B., and Schilling, W. (1994), *Manuelle Montage,* VDI, Düsseldorf.

Meedt, O. (1998), "Effizienzsteigerung in Demontage und Recycling durch optimierte Produktgestaltung und flexible Demontagewerkzeuge," Dissertation, Meisenbach, Bamberg.

Moyer, L. K., and Gupta, S. M. (1997), "Environmental Concerns and Recycling/Disassembly Efforts in the Electronics Industry," *Journal of Electronics Manufacturing,* Vol. 7, No. 1, pp. 1–22.

Pöhlau, F. (1999), "Entscheidungsgrundlagen zur Einführung räumlicher spritzgegossener Schaltungsträger (3-D MID)," Meisenbach, Bamberg.

Rahn, A. (1993), *The Basics of Soldering,* John Wiley & Sons, New York.

Seliger, G., and Wagner, M. (1996), "Modeling of Geometry-Independent End-effectors for Flexible Disassembly Tools," in *Proceedings of the CIRP 3rd International Seminar on Life Cycle Engineering: ECO-Performance '96* (March 18–29), ETH Zürich.

Siemens AG (1999), *The World of Surface Mount Technology,* Automation Technology Division, Siemens AG, Munich.

Spur, G., and Stöferle, T. (1986), *Fügen, Handhaben, Montieren,* Vol. 5 of *Handbuch der Fertigungstechnik,* Carl Hanser, Munich.

Steber, M. (1997), *Prozeßoptimierter Betrieb flexibler Schraubstationen in der automatisierten Montage,* Meisenbach, Bamberg.

Tönshoff, H. K., Metzel, E., and Park, H. S. (1992), "A Knowledge-Based System for Automated Assembly Planning," *Annals of the CIRP,* Vol. 41, No. 1, pp. 19–24.

Tritsch, C. (1996), *Flexible Demontage technischer Gerbauchsgüter,* Forschungsberichte wbk, University of Karlsruhe.

Verter, V., and Dincer, M. (1992), "An Integrated Evaluation of Facility Location, Capacity Aquisition and Technology Selection for Designing Global Manufacturing Strategies," *European Journal of Operational Research,* Vol. 60, pp. 1–18.

Verein Deutscher Ingenieure (VDI) (1982), Richtlinie 2860, Bl. 1, Entwurf, *Montage- und Handhabungstechnik: Handhabungsfunktionen, Handhabungseinrichtungen, Begriffe, Definitionen, Symbole,* VDI, Düsseldorf.

Warnecke, H. J. (1993), *Revolution der Unternehmenskultur,* Springer, Berlin.

Weigl, A., (1997), *Exemplarische Untersuchungen zur flexiblen automatisierten Demontage elektronischer Geräte mit Industrierobotern,* Berichte aus der Automatisierungstechnik, Technical University of Darmstadt, Shaker, Aachen.

Wolf, K. U. (1997), "Verbesserte Prozessführung und Prozessplanung zur Leistungs- und Qualitätssteigerung beim Spulenwickeln," Dissertation, Meisenbach, Bamberg.

CHAPTER 14

Manufacturing Process Planning and Design

TIEN-CHIEN CHANG
Purdue University

YUAN-SHIN LEE
North Carolina State University

1. INTRODUCTION 448
- 1.1. The Product-Realization Process 448
- 1.2. From Design to Process Planning to Production 449
  - 1.2.1. Selection of Materials 449
  - 1.2.2. Geometry Creation 449
  - 1.2.3. Function Analyses 450
  - 1.2.4. Design Evaluation 450
  - 1.2.5. Process Planning 450
  - 1.2.6. Production Planning and Scheduling 451
  - 1.2.7. Consideration of Production Quantity in Process Planning 452

2. PROCESS PLANNING 452
- 2.1. Geometry Analysis 452
- 2.2. Stock Selection 452
- 2.3. Gross Process Determination 453
  - 2.3.1. Casting, Machining, and Joining 453
  - 2.3.2. Product Strength, Cost, etc. 454
- 2.4. Setup and Fixture Planning and Design 455
- 2.5. Process Selection 456
- 2.6. Process Detailing 457
  - 2.6.1. Tool Selection 457
  - 2.6.2. Process Parameters Determination 458
  - 2.6.3. Process Optimization 458
- 2.7. Plan Analysis and Evaluation 458
  - 2.7.1. Machining Time and Cost Estimation 459
  - 2.7.2. Estimated Product Quality 460

3. TOOLS FOR PROCESS PLANNING 460
- 3.1. Group Technology 461
  - 3.1.1. How GT Is Used in Process Planning 461
  - 3.1.2. Coding and Classification 461
  - 3.1.3. Family Formation 462
  - 3.1.4. Composite Component Concept 462
- 3.2. Process Mapping 463
  - 3.2.1. Process for Features Mapping 463
  - 3.2.2. Relative-Cost Table for Manufacturing Processes 465
- 3.3. Process Capability Analysis 465
- 3.4. Cost Model 465
- 3.5. Tolerance Charting 472

4. COMPUTER-AIDED PROCESS PLANNING 473
- 4.1. Variant Approach 475
- 4.2. Generative Approach 477
  - 4.2.1. Part-Description Methods for Generative Process-Planning Systems 478

5. COMPUTER-AIDED PROCESS-PLANNING SYSTEMS SELECTION CRITERIA 478

6. CONCLUSIONS 482

REFERENCES 482

ADDITIONAL READING 483

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

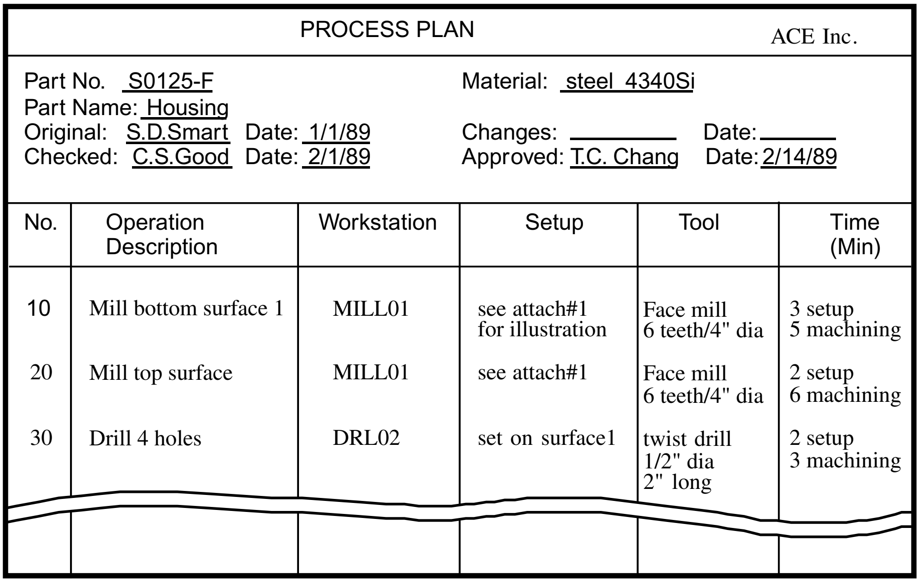
Manufacturing process planning is an important step in the product-realization process. It can be defined as "the function within a manufacturing facility that establishes which processes and parameters are to be used (as well as those machines capable of performing these processes) to convert a part from its initial form to a final form predetermined (usually by a design engineer) in an engineering drawing" (Chang et al. 1998, p. 515). Alternatively, it can be defined as the act of preparing detailed work instructions to produce a part. The result of process planning is a process plan. A process plan is a document used by the schedulers to schedule the production and by the machinist /NC part programmers to control/program the machine tools. Figure 1 shows a process plan for a part. The process plan is sometimes called an operation sheet or a route sheet. Depending on where they are used, some process plans are more detailed than others. As a rule, the more automated a manufacturing shop is, the more detailed the process plan has to be.

To differentiate the assembly planning for an assembled product, process planning focuses the planning on the production of a single part. In this chapter, when a product is mentioned, it refers to a discrete part as the final product. One important step in process planning is process selection, which is the selection of appropriate manufacturing processes for producing a part. When none of the existing processes can produce the part, a process may have to be designed for this purpose. Process design can also be interpreted as determining the parameters of a process for the manufacture of a part. In this case, process design is the detailing of the selected processes. Thus, process planning and process design are used for the same purpose—determining the methods of how to produce a part.

In this chapter, process planning and design are discussed. Techniques employed for process planning and process design are also introduced. Due to the vast number of manufacturing processes, it would be impossible to cover them all in this chapter. Only machining processes are focused upon here. However, early in the chapter, casting, forming, and welding examples are used to illustrate alternative production methods.

1.1. The Product-Realization Process

Manufacturing is an activity for producing a part from raw material. In discrete product manufacturing, the objective is to change the material geometry and properties. A sequence of manufacturing processes is used to create the desired shape. The product-realization process begins with product design. From the requirements, an engineering design specification is prepared. Through the design process (the details of which are omitted here), a design model is prepared. Traditionally, the design model is an engineering drawing (drafting) either prepared manually or on a CAD system. Since the



PROCESS PLAN ACE Inc.

Part No. S0125-F Material: steel 4340Si
Part Name: Housing
Original: S.D.Smart Date: 1/1/89 Changes: ________ Date: ______
Checked: C.S.Good Date: 2/1/89 Approved: T.C. Chang Date: 2/14/89

No.	Operation Description	Workstation	Setup	Tool	Time (Min)
10	Mill bottom surface 1	MILL01	see attach#1 for illustration	Face mill 6 teeth/4" dia	3 setup 5 machining
20	Mill top surface	MILL01	see attach#1	Face mill 6 teeth/4" dia	2 setup 6 machining
30	Drill 4 holes	DRL02	set on surface1	twist drill 1/2" dia 2" long	2 setup 3 machining

Figure 1 Process-Plan.

1990s, solid model for engineering design has gained popularity for representing design models. A design model must contain the complete geometry of the designed part, the dimensions and tolerances, the surface finish, the material, and the finished material properties. The design model is a document, or contract, between the designer and the manufacturing facility. The finished product is checked against the design model. Only when all the specifications on the design model are satisfied is the product accepted.

There are several steps in the product-realization process (Figure 2): design, process planning, manufacturing, and inspection. Process planning is a function linking the design and the manufacturing activities. The objective of manufacturing is to produce the product at an acceptable quality (instead of the best quality), in a desired time frame (not necessarily the shortest time), and at the lowest cost (lowest cost is always desirable). Because manufacturing follows the process plan, the quality of the process plan is critical to the success of manufacturing and thus product realization. In the following section, the more detailed steps of product realization are discussed.

1.2. From Design to Process Planning to Production

Before a product is materialized, it has to be designed and manufactured. Following are the major steps in this product realization process.

1.2.1. Selection of Materials

Materials are selected based on the functionalities of the part being made. Most parts are made from a single material. Material selection may not be the first step in design. However, it is an important decision to be made. Often, several materials all satisfy the functional requirements of the part. For example, one may choose steel, aluminum, of composite material for the part. Although the physical, mechanical, and electrical properties all satisfy the design requirements, the material and processing costs might be very different. The material cost is easily estimated (Table 1), but estimating the processing cost is more involved. For example, steel-part manufacturing is very different from composite-part manufacturing. Totally different machines and material-handling methods are needed. A good designer will take manufacturing issues into consideration. Design for manufacturing should begin with the proper selection of materials for manufacturing. In some cases, due to the material property, the geometry of the part will need to be changed.

1.2.2. Geometry Creation

The shape of a product can be determined by functional or aesthetic considerations. Individual parts in an assembly must fit together to form the assembly. They use the geometric shape to carry out a specific function. For example, an angle bracket is used for mounting a machine, a hole is used to

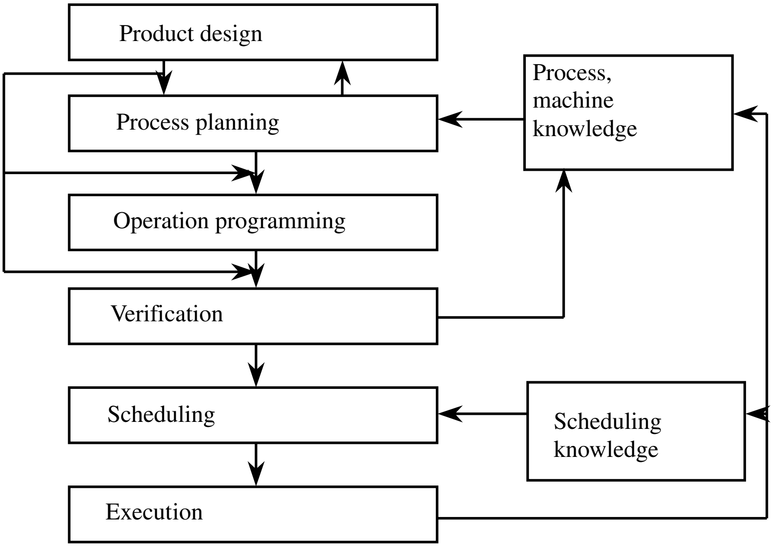


Figure 2 Product-Realization Process.

TABLE 1 Approximate Cost of Raw Materials as a Function of their Condition, Shape, and Size

Material	Cost ($)	Material	Cost ($)
Carbon-steel plate and sheet		Aluminum plate	
Hot rolled	60–70	2024 T351	530–590
Cold rolled	75–90	6061 T651	330–350
Carbon-steel bars		7075 T651	560–620
Hot rolled, round	55–80	Aluminum sheet	
Cold finished, round	60–200	2024 T3	610–650
Cold finished, square	90–170	3003 H14	275–300
Stainless steel sheet		6061 T6	360–400
304	230	Aluminum bars	
316	300–340	Round	275–510
410	375	Square	575–700
Stainless steel bars		Rectangular	550–1000
304 round	310–730	Aluminum extrusions	260–310
303 square	560–000		

From S. Kalpakjian, *Manufacturing Engineering and Technology,* 3d Ed., © 1995. Reprinted by permission of Prentice-Hall, Inc., Upper Saddle River, NJ.

fit an axle, and a T-slot is used to hold a bolt. The designer must create the appropriate geometry for the part in order to satisfy the functional requirements. The ultimate objective of the designer is to create a functionally sound geometry. However, if this becomes a single-minded mission, the designed part may not be economically competitive in the marketplace. Cost must also be considered.

Manufacturing processes are employed to shape the material into the designed geometry. A large number of unrelated geometries will require many different processes and/or tools to create. The use of standard geometry can save money by limiting the number of machines and tools needed. For example, a standard-size hole means fewer drill bits are needed. Design for manufacturing also means imposing manufacturing constraints in designing the part geometry and dimension.

The designed geometry is modeled on a CAD system, either a drawing or a solid model (see Figure 3). More and more designs are modeled using 3D solid modelers, which not only provide excellent visualization of the part and assembly but also support the downstream applications, such as functional analysis, manufacturing planning, and part programming. The key is to capture the entire design geometry and design intents in the same model.

1.2.3. Function Analyses

Because the designed part must satisfy certain functional requirements, it is necessary to verify the suitability of the design before it is finalized. Engineering analyses such as kinematic analysis and heat transfer are carried out from the design. Finite element methods can be used, often directly from a design model. The more critical a product or part is, the more detailed an analysis needs to be conducted.

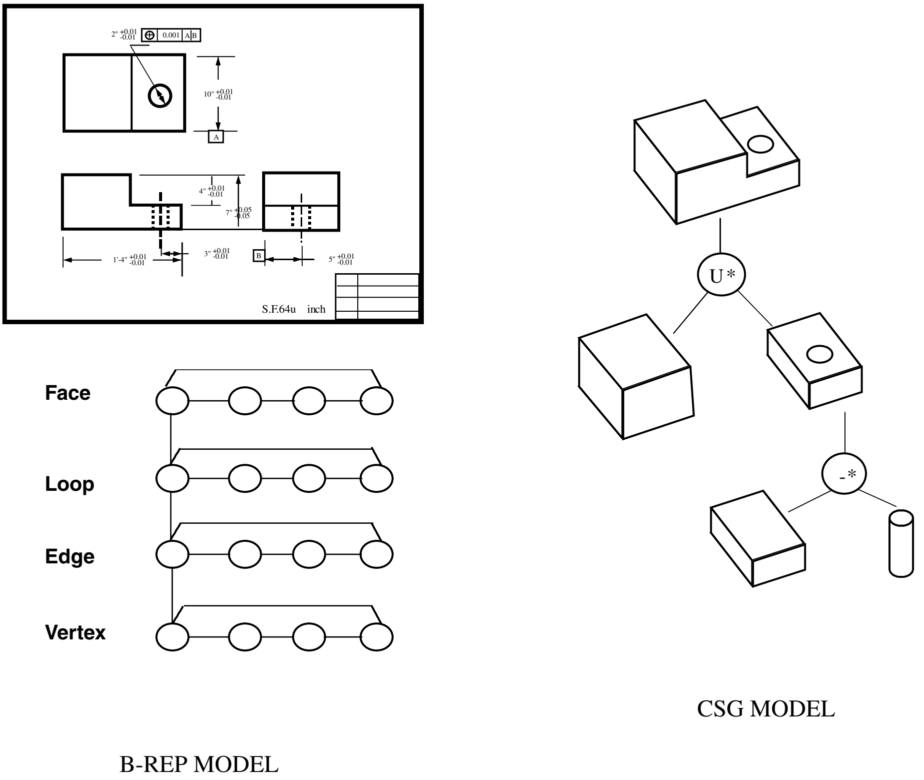
1.2.4. Design Evaluation

The task of design evaluation is to separate several design alternatives for the final selection of the design. Cost analysis, functionality comparison, and reliability analysis are all considerations. Based on the predefined criteria, an alternative is selected. At this point the design is ready for production.

1.2.5. Process Planning

Production begins with an assembly/part design, production quantity, and due date. However, before a production order can be executed, one must decide which machines, tools, and fixtures to use as well as how much time each production step will take. Production planning and scheduling are based on this information. As noted earlier, process planning is used to come up with this information. How to produce a part depends on many factors. Which process to use depends on the geometry and the material of the part. Production quality and urgency (due date) also play important roles. A very different production method will definitely be appropriate for producing a handful of parts than for a million of the same part. In the first case, machining may be used as much as possible. However, in the second case, some kind of casting or forming process will be preferable.

When the due date is very close, existing processes and machines must be used. The processes may not be optimal for the part, but the part can be produced in time to meet the due date. The cost



CSG MODEL

B-REP MODEL

Figure 3 Design Representations.

will be high—one pays for urgent orders. On the other hand, when there is plenty of lead time, one should try to optimize the process. When the production quantity justifies the cost, it might be necessary to design new processes or machines for the part. One good example is the use of a transfer line for engine block production. Machines (stations) in a transfer line are specially designed (or configured) for a part (e.g., an engine block). Production is optimized. Lower cost and higher quality can be expected. Table 2 shows the recommended production systems for different production quantities and production lead times.

1.2.6. Production Planning and Scheduling

After process planning is complete, production is ready to begin. Production planning and scheduling are important functions in operating the manufacturing facility. Because multiple products or parts are being produced in the same manufacturing facility, the resource allocation must be done appropriately in order to maximize the production output. When a transfer line or production line (assembly line) is the choice, the line is balanced (equal numbers of tasks are allocated to each machine station)

TABLE 2 Recommended Production Methods

	Long Lead Time	Medium Lead Time	Short Lead Time
Mass production	Transfer-line	Product line	Job shop
Medium batch	Manufacturing cell Job shop	Job shop Manufacturing cell	Job shop
Small volume	Job shop	Job shop	Job shop

and designed. After a line is installed, the operation of the line is simple. Upon the workpiece being launched at one end of the line, the product is built sequentially through the line. However, in a shop environment, production scheduling is much more complex. For each planning horizon (day or week), what each machine should process and in which sequence must be decided. Predefined objectives such as short processing time, maximum throughput, and so on are achieved through proper scheduling. Scheduling uses information provided by the process plan. A poorly prepared process plan guarantees poor results in production.

1.2.7. Consideration of Production Quantity in Process Planning

As noted above, production quantity affects the manufacturing processes selected for a part. If the production quantity is not considered in process planning, the result may be an expensive part or a prolonged production delay. Large quantities make more specialized tools and machines feasible. Small-quantity production must use general-purpose machines. Although the total quantity may be high, in order not to build up inventory and incur high inventory cost, parts are usually manufactured in small batches of 50 parts or less. Therefore, production batch size is also a factor to be considered. There is decision making loop. The process plan is the input to production planning; production planning determines the most economical batch size; batch size, in turn, changes the process plan. A decision can be made iteratively.

2. PROCESS PLANNING

As noted above, process planning is a function that prepares the detailed work instructions for a part. The input to process planning is the design model and the outputs include processes, machines, tools, fixtures, and process sequence. In this section, the process planning steps are discussed. Please note that these steps are not strictly sequential. In general, they follow the order in which they are introduced. However, the planning process is an iterative process. For example, geometry analysis is the first step. Without knowing the geometry of the part, one cannot begin the planning process. However, when one selects processes or tools, geometric reasoning is needed to refine the understanding of the geometry. Another example is the iterative nature of setup planning and process selection. The result of one affects the other, and vice versa.

2.1. Geometry Analysis

The first step in process planning is geometry analysis. Because the selection of manufacturing processes is geometry related, the machining geometries on the part, called manufacturing features, need to be extracted. An experienced process planner can quickly and correctly identify all the pertinent manufacturing features on the part and relate them to the manufacturing processes. For example, the manufacturing features of holes, slots, steps, grooves, chamfers, and pockets are related to drilling, boring, reaming, and milling processes. The process planner also needs to note the access directions for each feature. Also called approach directions, these are the unobscured directions in which the feature can be approached by a tool. When features are related to other features, such as containment, intersection, and related in a pattern, these relationships must be captured. Feature relations are critical in deciding operation (process) sequence. Figure 4 shows a few features and their approach directions. The pocket at the center and steps around the center protrusion are approachable from the top. The hole may be approached from the top or from the bottom.

In computer-aided process planning, the geometry analysis (or geometric reasoning) is done by computer algorithms. The design model in the form of a solid model is analyzed. Based on the local geometry and topology, regions are extracted as features. To be significant to manufacturing, these features must be manufacturing features (Figure 5). Again, manufacturing features are geometric entities that can be created by a single manufacturing process or tool. Like manual geometry analysis, geometric reasoning must find feature access directions (approach directions) and feature relations. Due to the vague definitions of features, the large number of features, and the complexity in feature matching (matching a feature template with the geometric entities on the solid model), geometric reasoning is a very difficult problem to solve. This is one of the reasons why a practical, fully automatic process planner is still not available. However, a trained human planner can do geometry analysis relatively easily. More details on geometric reasoning can be found in Chang (1990).

2.2. Stock Selection

Manufacturing is a shape-transformation process. Beginning with a stock material, a sequence of manufacturing processes is applied to transform the stock into the final shape. When the transformation is done using machining, minimizing the volume of materials removed is desirable. Less material removal means less time spent in machining and less tool wear. The stock shape should be as close to the finished part geometry as possible. However, in addition to the minimum material removal rule, one also has to consider the difficulty of work holding. The stock material must be

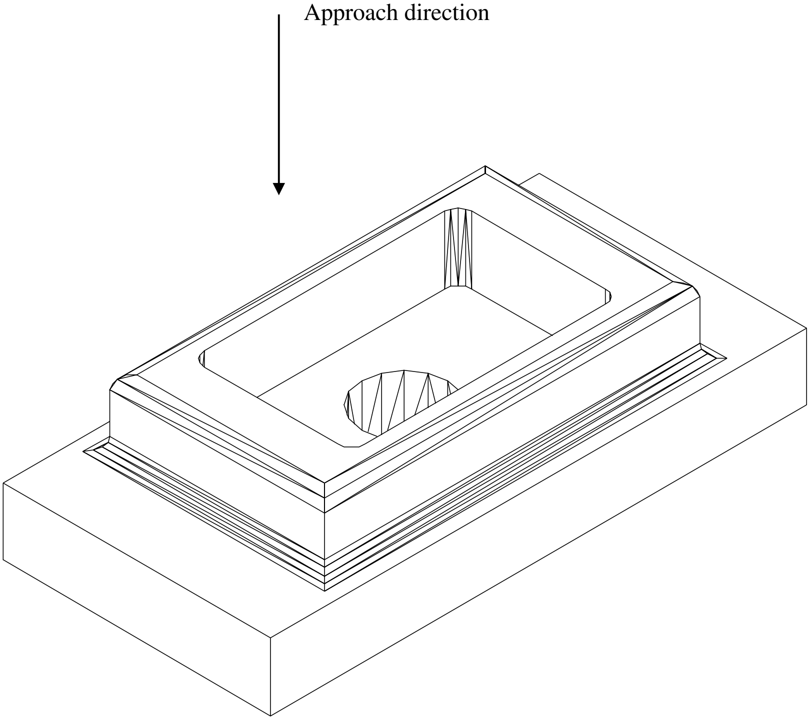


Figure 4 Protrusion, Pocket, and Hole.

clamped or chucked on the machine tool before cutting can be performed. This fixturing consideration has practical importance. However, raw materials can be purchased only in limited shapes and dimensions. For example, steel is supplied in sheets of different thickness, length, and width, strips, wires, bars, and so on. The general shape has to be determined, and then stock preparation (cutting) has to be done to produce the stock. This is also true for the forming processes.

After the stock material is selected, one can compare the stock with the finished part and decide the volume of material to be removed for each process. In automated process planning, often the difference between the stock and the part, called the delta volume, is calculated. Geometric reasoning is performed on the delta volume. Without first selecting the stock, one cannot be certain exactly how to proceed with the manufacturing processes.

In some cases, especially for mass production, minimizing metal removal means preparing a casting as the stock material. Machining is used to improve critical or hard-to-cast surfaces. In this case, the casting has to be designed based on the part. Using casting as the stock minimizes the machining yet requires a high initial investment (casting design, mold making, etc.).

2.3. Gross Process Determination

Process planning can be separated into two stages: gross planning and detailed planning. Often only detailed planning is discussed in the process planning literature. Actually, the gross planning is even more critical than the detailed planning. Gross planning is used to determine the general approach to produce a part. For example, the same part geometry may be created through casting, machining, 3D fabrication, or welding of sheet metal. Only when a general approach is determined may one proceed with the detailed planning.

2.3.1. Casting, Machining, and Joining

One of the first decisions a process planner needs to make is whether to cast the part or machine it. Rough casting, such as sand casting, requires a good amount of machining. However, precision casting, such as die casting and investment casting, can produce almost net shape part (finished part). The decision is based on both economics and material properties; this issue will be addressed below.

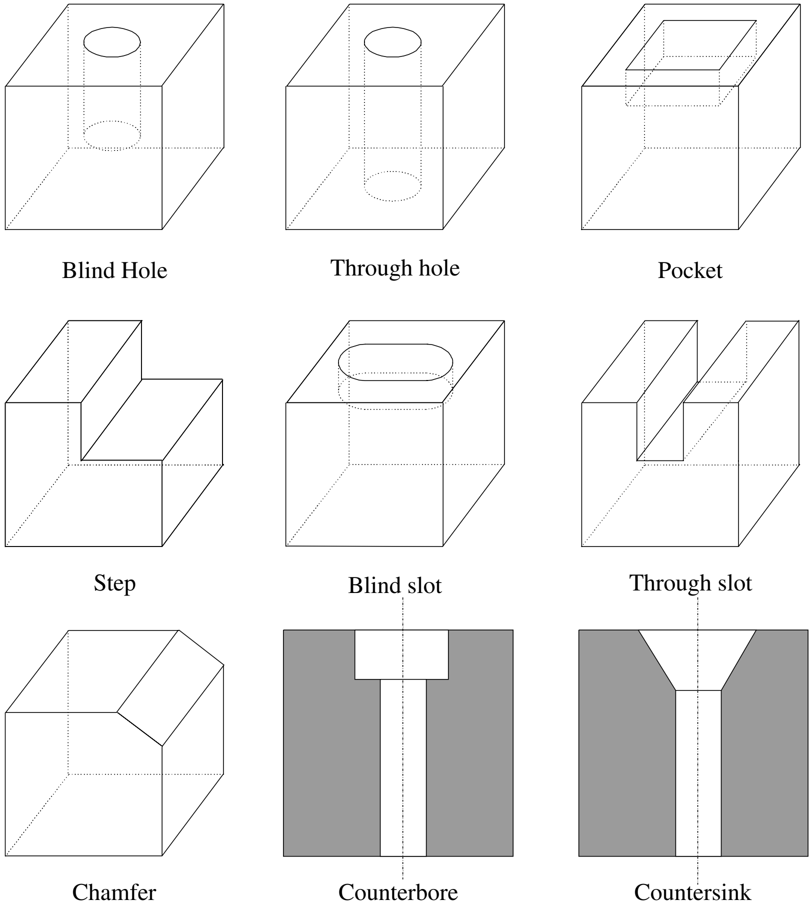


Figure 5 Some manufacturing Features.

In the initial evaluation, the technical feasibility and relative costs are taken into consideration. Process capabilities and relative cost for each casting and machining process are used in the evaluation. Based on the geometry to be created, material used, and production volume, one can propose several alternative processes. The process capability includes shape capabilities and accuracy. Casting processes, such as die casting and investment casting, can create intricate geometry with relative high precision (0.01 in.). However, sand casting is much worse. For tighter dimensional control, secondary machining is a must.

Certain parts can be built using the welding process as well. This is especially true for structural parts. Large structural parts such as ship hulls are always built using welding. Medium-sized structural parts such as airplane structure parts may be joined together from machined parts. Sectional structures for jet fighters are always machined from a solid piece of metal in order to obtain the maximum strength. For small structures, all manufacturing methods are possible. When there are several alternatives, an initial selection needs to be made.

2.3.2. Product Strength, Cost, etc.

Parts made using different stocks and processes exhibit different strength. As mentioned above, sectional structures for jet fighters are always machined from a solid piece of alloy steel. The raw material is homogeneous and rolled into the shape in the steel mill. It has high strength and consistent properties. When casting is used to form a part, a certain number of defects can be expected. The cooling and thus solidification of the material are not uniform. The surface area always solidifies first and at a much higher rate than the interior. A hard shell forms around the part. Depending on the complexity of the part, the type of mold used (sand, metal, etc.) and the mold design, voids, hot tear, cold shut, or other problems can happen. The part is not as strong as those produced using machining. In the case of welded parts, the welded join may not be as strong as the rest of the part.

As for the manufacturing cost, there is an even greater difference. For machining, the initial cost is low but the incremental cost is higher. The opposite is true for casting. Figure 6 shows the comparison. The initial cost for casting includes the cost of designing and building the mold and is relatively high. The slope of the incremental cost is the casting cost per piece. For machining, the initial cost is relatively much lower. On a manually controlled machine, only tools and fixtures need to be purchased. When a CNC machine is used, the programming cost has to be added to the fixed cost. However, the machining cost per piece will be lowered.

There are always alternative ways to make a part. Unless the way is specified by the designer, a good process planner always considers all possible alternatives and evaluate them. This evaluation need not always be carried out formally and precisely. Using past experience and with rough estimates, one can quickly eliminate most alternatives. The most promising alternatives have to be explored further before they are accepted or discarded.

2.4. Setup and Fixture Planning and Design

Let us assume that machining is the best alternative for the production of a part. Using the result of geometry analysis, we can group machining features based on their feasible approach directions. Most machining processes require the workpiece be positioned so that the tool orientation matches with the feature approach direction. For example, the part in Figure 7 consists of four holes. All holes have the same approach direction. Therefore, it is best to set up the workpiece with the holes aligned with the machine spindle. The position of the workpiece is called the setup position. When there are multiple approach directions, multiple setups may be needed to finish the machining operations.

A fixture is used to hold the workpiece at the desired setup position. For simple workpieces, a vise is sufficient to do the job. However, for more complex workpieces, locators and clamps are needed to position and clamp the workpiece on the machine table. After each setup the workpiece geometry is transformed (Figure 8). The finished part is obtained after the last setup is done. After a fixture is designed or configured (using modular figures), tool interference has to be checked. Obviously, fixture elements should not interfere with the tool motion. If the current fixture does interfere with the tool, either a new fixture needs to be designed or the machining feature that causes interference will not be cut during this setup. Again, this illustrates the iterative nature of process planning steps.

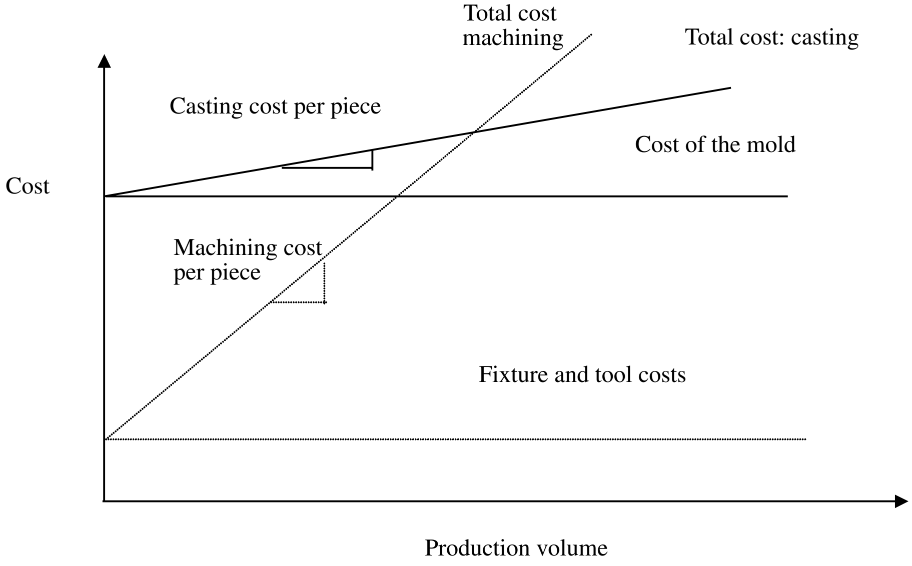


Figure 6 Costs vs. Production Volumes for Different Production Methods.

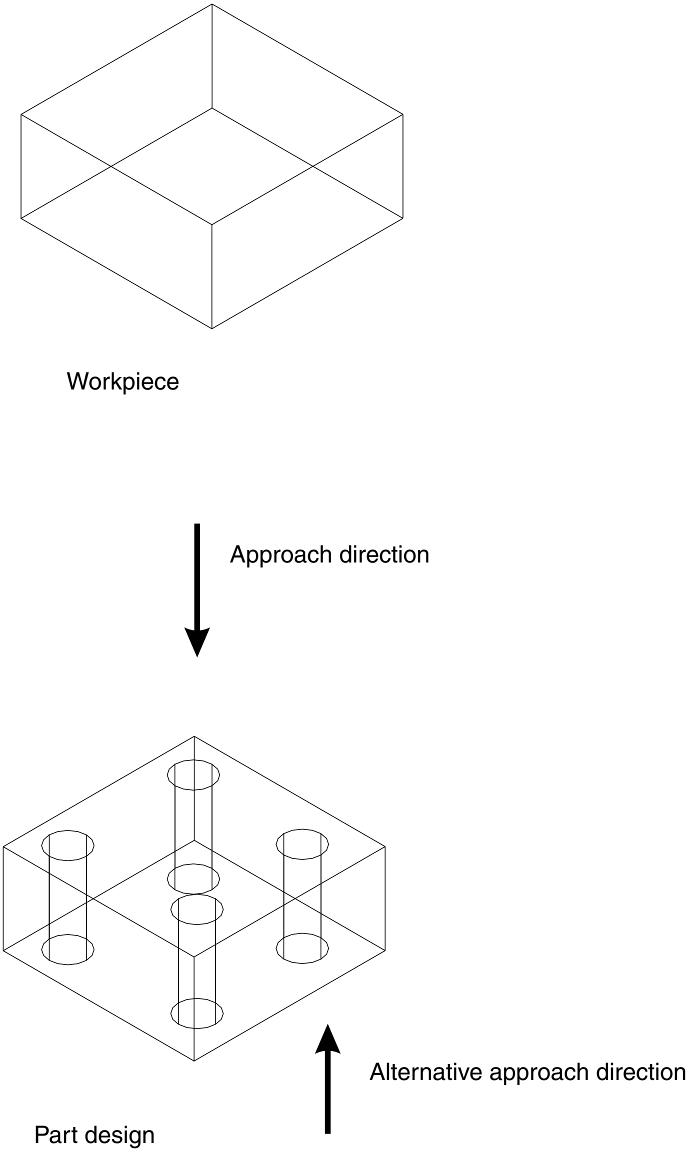


Figure 7 Workpiece and Features on a Part.

2.5. Process Selection

Process is defined as a specific type of manufacturing operation. Manufacturing processes can be classified as casting, forming, material-removal, and joining processes. Under casting are sand casting, investment casting, die casting, vacuum casting, centrifugal casting, inject molding, and so on. Forming includes rolling, forging, extrusion, drawing, powder metallurgy, thermoforming, spinning, and so on. Material removal includes milling, turning, drilling, broaching, sawing, filing, grinding, electrochemical machining (ECM), electrical-discharge machining (EDM), laser-beam machining, water-jet machining, ultrasonic machining, and so on. Joining processes include arc welding, electron-beam welding, ultrasonic welding, soldering, brazing, and so on. In this chapter only material-removal processes examples are used.

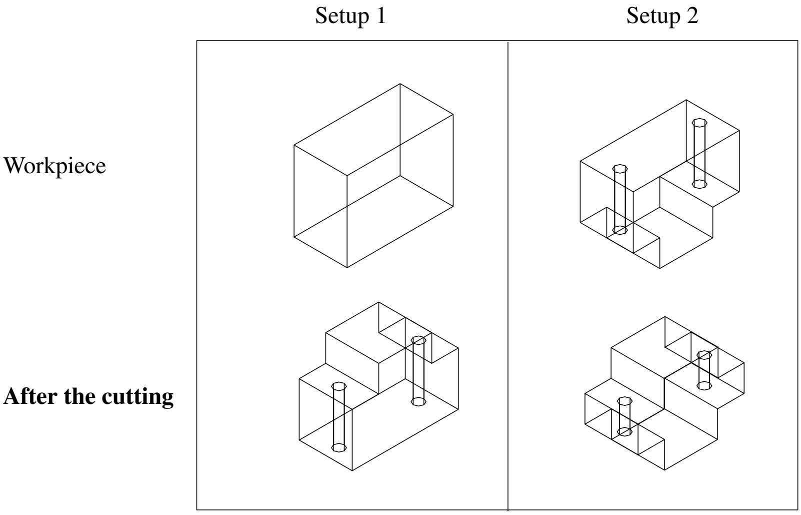


Figure 8 Setup of a Part.

Each process has its own geometric and technological capabilities. Geometric capability means the shapes a process can create and the geometric constraints it might have. For example, the drilling process usually creates round holes and due to the drill nose cone the hole bottom has the cone shape. Technological capability includes tolerances (both dimensional and geometrical), surface finish, and surface integrity. Processes are selected based on the machining features in a setup. In the previous example (Figure 7), a setup consists of four holes. The hole geometry matches the drilling process capability. Therefore, drilling is selected for the holes. In this example, the drilling sequence has no effect on the final product or the manufacturing cost. However, in many other cases the process sequence does matter. For example, in Figure 9 there are four holes and a step. It makes more sense to mill the steps before drilling the holes than the other way around. When the two left-hand holes are drilled first, much of the drilling is done in vain. The milling process will remove the top half of the holes drilled. The process sequence is determined based on the relationship between features and the technological constraints of the processes. A good process planner takes all these into consideration when selecting the processes.

2.6. Process Detailing

Process detailing involves filling the details for the process selected. It includes determining the tool for the process, tool parameters (total length, diameter, cutting length, etc.), and process parameters (speed, feed, depth of cut, etc.).

2.6.1. Tool Selection

In order to carry out the process selected for a feature, a cutting tool is needed. Many different cutting tools can be used for the same process. In drilling, for example, there are different types of drill bites, such as twist drill, spade drill, and gun drill. Each drill also comes with different diameters, cutting length, total length, nose angle, and tool material. For drilling the holes in the example (Figure 9), two different drill lengths with the same diameter are needed. Of course, the longer drill can be used to drill shorter holes, too. However, if the diameters are slightly different, separate drills need to be specified.

Figure 10 shows different kinds of drills and turn tools. The selection of a tool depends on the feature geometry and geometric constraints. In considering milling, there are even more tool parameters to consider. In addition to tool diameter, cutting depth, there are also such factors as number of cutting teeth, insert material, and rake angle. For end mills, there are also bottom-cutting and non-bottom-cutting types. Faced with this vast amount of choices, one must often rely on past experience and, for unfamiliar tools, handbooks.

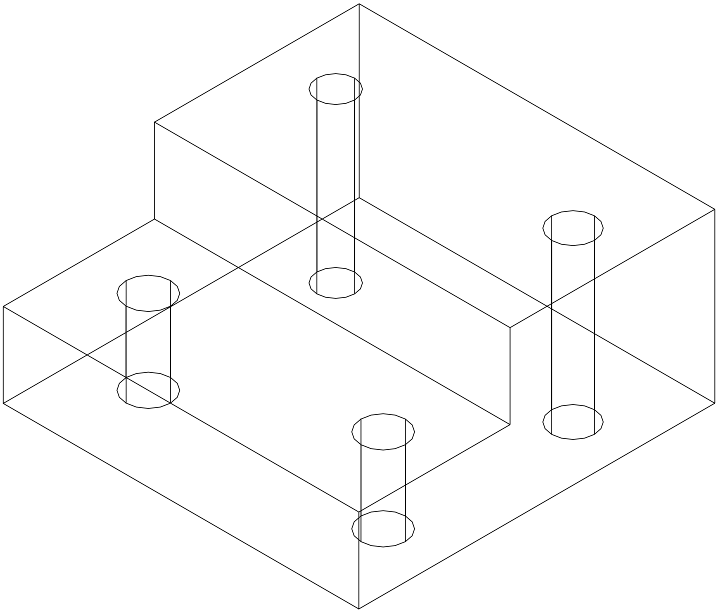


Figure 9 Sample Part with Holes and Step.

2.6.2. Process Parameters Determination

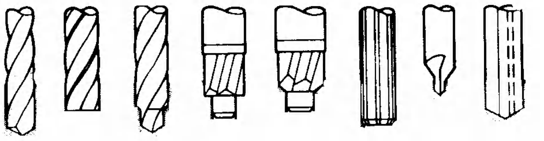
Process parameters include speed, feed, and depth of cut. They are critical to the cutting process. Higher parameter values generate higher material-removal rate, but they also reduce the tool life. High feed also means rougher surface finish. In drilling and turning, feed is measured as how much the tool advances for each rotation of the tool. In milling, it is the individual tooth advancement for each tool rotation. In turning, for example, smaller feed means closely spaced tool paths on the part surface. The finish will be better in this case. Higher feed separates the tool paths and in the worst case creates uncut spacing between two passes. Types of process, the tool and workpiece materials, and hardness of the workpiece material affect process parameters. The parameter values are determined through cutting experiments. They can be found in the tool vendor's data sheets and in the *Machining Data Handbook* (Metcut 1980). These data are usually based on the constant tool life value, often 60 minutes of cutting time. When the required surface finish is high, one must modify the recommended parameters from the handbook. For new materials not included in any of the cutting parameter handbooks, one must conduct one's own cutting experiments.

2.6.3. Process Optimization

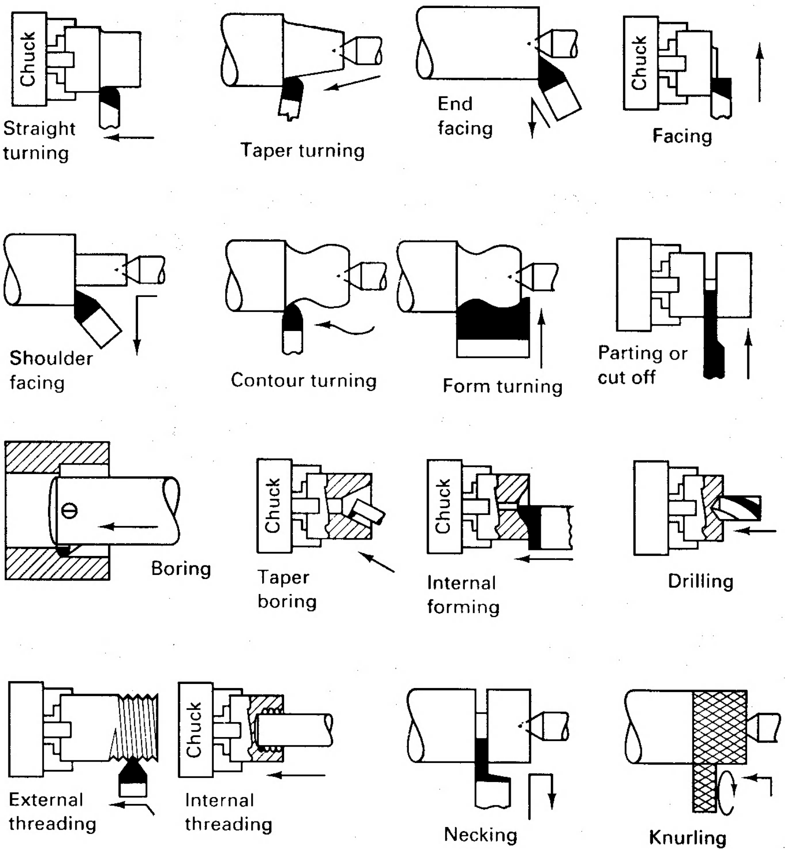
It is always desirable to optimize the production. While global optimization is almost impossible to achieve, optimization on the process level is worth trying. The total production time is the sum of the cutting time, the material handling time, and the tool-change time. Shorter cutting time means faster speed and feed and thus shorter tool life. With a shorter tool life, the tool needs to be changed more frequently and the total time is thus increased. Because there is a trade-off between cutting time and tool life, one may find the optimal cutting parameters for minimum production time or cost. The techniques of process optimization are based on an objective function (time, cost, or another criterion) and a set of constraints (power, cutting force, surface finish, etc). The process-optimization models will be discussed later in the chapter. Finding optimal cutting parameters is fairly complex and requires precise data on both tool life model and machining model. Except in mass production, process optimization is generally not considered.

2.7. Plan Analysis and Evaluation

In on an old study conducted in industry, when several process planners were given the same part design, they all came up with quite different process plans. When adopted for the product, each



Drills


Chuck
Straight turning
Taper turning
End facing
Chuck
Facing
Shoulder facing
Contour turning
Form turning
Parting or cut off
Boring
Chuck
Taper boring
Chuck
Internal forming
Drilling
External threading
Chuck
Internal threading
Necking
Knurling

Turn tools

Figure 10 Different Cutting Tools.

process plan resulted in different costs and part quality. If multiple process plans can be prepared for the same part, each plan must be analyzed and the best selected based on some preset criteria. If only one plan is prepared, it must be evaluated to ensure that the final result is acceptable. The final result here means the product quality.

2.7.1. Machining Time and Cost Estimation

Machining time can be calculated based on the cutting parameters and the feature geometry and dimension. It is used to estimate the production time and cost. It is also used in scheduling for determining the machine time. For turning and drilling, machining time can be calculated by the following formula:

$$T_m = \frac{L}{V_f}$$

$$V_f = fn$$

$$n = \frac{V}{\pi D}$$

where T_m = machining time, min
L = length of cut, in.
V_f = feed rate, ipm
f = feed, ipr (in. per revolution)
n = tool rpm
D = tool diameter, in.

For complex features, it is harder to estimate the length of the tool path. A rough estimation may be used. The material-removal rate (MRR) of the tool is calculated. The machining time is therefore the feature volume divided by the MRR. MRR for turning and drilling is:

$$\text{MRR} = V_f A$$

where A = the cross-sectional area of the cutting

For hole drilling,

$$A = \frac{\pi D^2}{4}$$

For turning,

$$A = 2\pi r^2 a_p$$

where r = cutting radius
a_p = the depth of cut

Machining cost can be calculated by the machining time times a machine and operator overhead rate.

2.7.2. Estimated Product Quality

The commonly considered technological capabilities of a process include tolerances and surface finish. Surface finish is determined by the process and process parameters. It is affected not by the order in which processes are applied but only by the last process operated upon the feature. However, tolerances are results of a sequence of processes. Operation sequences will affect the final tolerance. Using a simple 2D part, Figure 10 shows the result of different process sequences. The arrow lines are dimension and tolerance lines. The drawing shows that the designer had specified dimensions and tolerances between features *AB*, *BC*, and *CD*. Notice that in this case features are vertical surfaces. If one uses feature *A* as the setup reference for machine *B*, *B* for *C*, and *C* for *D*, the produce part tolerance will be the same as the process tolerance. However, it would be tedious to do it this way. One may use *A* as the reference for cutting *B*, *C*, and *D*. In this case, tolerance on *AB* is the result of the process that cut *B* (from *A* to *B*). However, the tolerance on *BC* is the result of processes that cut feature *B* (from *A* to *B*) and feature *C* (from *A* to *C*). The finished tolerance on *BC* is twice the process tolerance and twice that of *AB*. The same can be said for *CD*. If we choose *D* as the reference, of course, the tolerance on *CD* is smaller than that for *AB* and *BC*. So we may conclude that process sequence does affect the quality of the part produced. The question is whether the current process sequence satisfies the designed tolerance requirements. Often this question is answered with the tolerance charting method. Tolerance charting will be introduced in the next section.

3. TOOLS FOR PROCESS PLANNING

Process-planning steps were introduced in the previous section. This section discusses the tools used to carry out these steps. Manual process planning, which relies on human experience, will not be discussed. The tools discussed in this section are those used in computer-aided process-planning systems. They are used to assist the human planner in developing process plans. Most of these tools have been used in practical process-planning systems. Methodologies or algorithms used only in advanced research will not be introduced here.

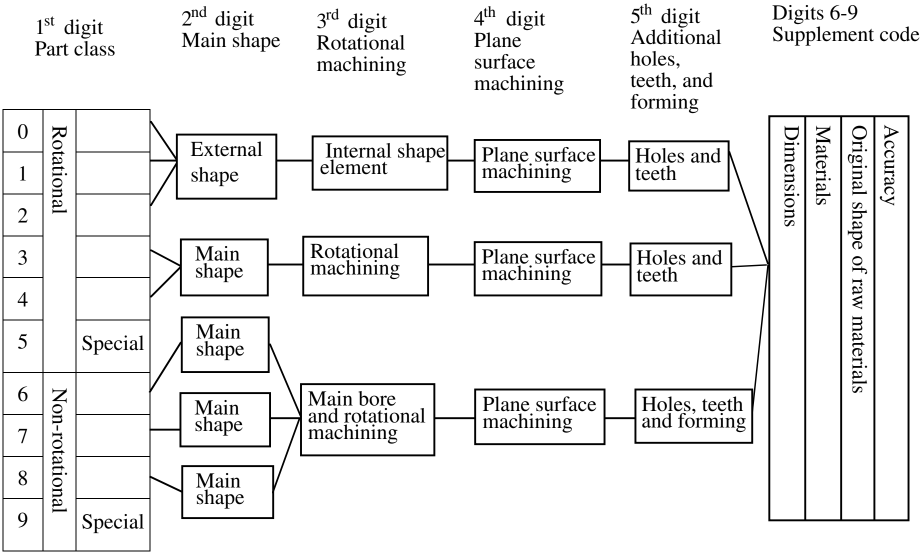

1st digit
Part class
2nd digit
Main shape
3rd digit
Rotational machining
4th digit
Plane surface machining
5th digit
Additional holes, teeth, and forming
Digits 6-9
Supplement code
0
1
2
3
4
5
6
7
8
9
Rotational
Non-rotational
Special
Special
External shape
Internal shape element
Plane surface machining
Holes and teeth
Main shape
Rotational machining
Plane surface machining
Holes and teeth
Main shape
Main shape
Main bore and rotational machining
Plane surface machining
Holes, teeth and forming
Main shape
Dimensions
Materials
Original shape of raw materials
Accuracy

Figure 11 Opitz Code.

3.1. Group Technology

Group technology is a methodology of using the similarity among parts to simplify the production. The key is to identify the similarity. In a machine shop, the thousands of different parts produced may share a small number of geometric characteristics. Parts of similar shape may be produced on the same set of machine tools. The material-handling requirements may also be the same. Manufacturing cells can be designed to produce similar parts. This simplifies the material handling, reduces the complexity of scheduling, and increases the slope of the learning curve. A complex problem can be decomposed into smaller and simpler problems. This set of similar parts is called a part family. Identifying part families is an important step in applying group technology to different manufacturing problems. Group technology has been used in manufacturing system design, scheduling, product retrieval, fixture design, and process planning.

3.1.1. How GT Is Used in Process Planning

As mentioned in the introduction, similar parts can be produced on the same set of machine tools. Most likely they will follow the same process sequence. Therefore, if one collects all the existing process plans used in the shop and groups them based on the part family, one will find that the process plans for the family members are similar. Summarizing the process plans for the family allows a standard process plan to be defined. All parts in the family may share this standard process plan. When a new part is to be planned, one can find the part family of this part, based on the geometric characteristics. The standard process plan for the family is then modified for this new part. Using group technology for manufacturing is like using a library to find references for writing a paper. Without the library database, locating appropriate references will be much more difficult and thus so will the writing.

3.1.2. Coding and Classification

Group technology is based on the concept of similarity among parts. Classification or taxonomy is used for this purpose. The dictionary definition of taxonomy is "orderly classification of plants and animals according to their presumed natural relationships." Here, taxonomy is used to classify parts in manufacturing. There are many methods of part classification. To name just a few: visual observation, manual sorting of the parts, sorting photographs of the parts, and sorting engineering drawings. Because keeping the physical parts or drawing in the sorted order is tedious or sometimes impossible, it is necessary to create a convenient representation, called a coding system.

A coding system uses a few digits or alphanumeric codes to represent a group (family) of similar parts. The classification system is embedded into the coding system. For example, one can easily see that parts can be classified as rotational and nonrotational. A crude coding system can have "0" representing any rotational parts and "1" representing any nonrotational parts. Rotational parts can further be classified as rods, cylinders, and disks, based on the length-to-diameter ratio. The code can be refined to have "0" represent rod, "1" cylinder, "2" disk, and "3" nonrotational parts. Further, the external and internal shapes of the part can be classified as smooth, step, screw thread, and so on. Additional digits may be used to represent the external and internal shapes. Each digit refines the classification or adds additional characteristics.

There are many public domain and proprietary coding systems. Opitz (Opitz 1970), Dclass (Allen 1994), MICLASS (OIR 1983), KK3 (Chang et al. 1998) are but a few popular ones. Opitz code (Figure 12), developed by Professor Opitz of Aachen University in the 1960s, uses five digits to represent the geometry of the part and four supplemental digits to represent part dimension, material, raw material shape, and accuracy. If only the geometry is of concern, the supplemental digits need not be coded. Extensive code tables and illustrations are given to guide the user in coding parts. Given a five-digit code, one can have a rough idea of the shape of the part. The code can also be used for searching the part database to find similar parts. Other coding systems may be more detailed or cover a different part domain, but they all serve the same purpose.

3.1.3. Family Formation

To take advantage of the similarity among parts, one must group parts into families. If the geometry is used in defining the part family, the coding and classification system discussed in the previous subsection can be used. Such families are called design families because they are design geometry based. However, often one would like to form part families based on the production methods used. In this case, the part family is called a production family. Because the manufacturing methods are geometry related, members of a production family share many similar feature geometries as well.

Families may be formed using the visual observation method, as mentioned above, or using the sorting approach. In forming design families the coding system can be used as well. Parts with the same codes always belong to the same family. To enlarge the family, several codes may be included in the same family. The determination as which codes should be included is based purely on the applications considered.

To form a production family, one has to consider the manufacturing method. The first well-known production family formation method was called production flow analysis (Burbidge 1975). First, production flows, or process sequences, for all parts are collected. An incidence matrix with columns representing each part and rows representing each process, machine, or operation code (a certain process performed on a machine) is prepared based on the production flows. By sorting the rows and columns of the matrix, one may move the entries along the diagonal of the matrix (see Figure 13). In the matrix, one may conclude that parts 8, 5, 7, 2 belong to one family and 4, 1, 3, 6, 9, 10 belong to another family. Family one needs processes P1, P5, P6, and P3. Family two needs processes P3, P2, and P4. This approach can be tedious and requires human judgment on separating families. As can be seen, P3 is needed for both families. It is not uncommon to have overlapping entries.

Over the years, many mathematics-based sorting methods, called clustering analysis, have been developed. For more details see Kusiak (1990) and Chang et al. (1998).

3.1.4. Composite Component Concept

Given a family of parts, one can identify several features. When one takes all the features and merges them into an imaginary part, this imaginary part, called a *composite component,* is a superset containing the features of the family members. The composite component concept was developed before World War II in Russia, where it was used to design flexible fixtures for a part family. By adjustment of the fixture, it can accommodate all the parts belonging to a family. This concept can also be used in design. For example, parametric design uses a parameterized design model for a narrowly defined

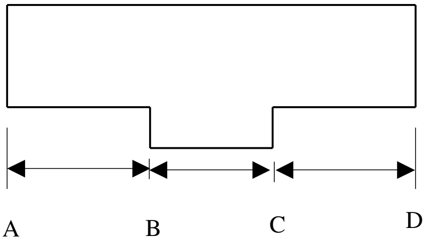


Figure 12 Tolerances.

Parts

Processes

	8	5	7	2	4	1	3	6	9	10
P1	1	1	1	1						
P5	1	1	1	1						
P6		1	1	1						
P3				1	1	1	1	1	1	
P2					1	1	1	1	1	1
P4					1	1	1	1	1	1

Figure 13 Production Flow Matrix.

family. One is the model for spur gears. Assigning parameters, such as pitch, number of teeth, outer diameter, pressure angle, teeth face width, hub diameter and length, hole diameter, and keyway and set screw, allows a drawing of the gear to be generated.

The same concept can be applied to process planning. When a process model (processes, tools, and cutting parameters) is developed for each of the features belonging to the composite component, a process plan can be generated using the parameters specified by the user. This concept has been used in developing several experimental and commercial process planners, such as CPPP (Dunn and Mann 1978). Because the same set of parameter data can be used to generate a process plan or a design drawing, CAD/CAM integration can be done. The limitation of this approach is that the family members must have very similar geometry. Not only do features have to be shared, but the relative position of features on family members have to be maintained. Otherwise, not only can the drawing not be done correctly, but the process plan generated will not be usable. The composite component concept is a tool for part families with minor differences among family members.

3.2. Process Mapping

Why a process can generate certain shapes depends on the geometry generation process of the tool. For example, a drill bit has two cutting edges (lips) (Figure 14). The drilling process requires the drill bit to rotate along its axis, then move the cutting edges downward. The rotating cutting edge creates a cone. Sweeping down the cone will remove a cylindrical volume of materials. Thus, the holes created always have a cone-shaped bottom. The turn tool has a single cutting edge. A layer of the material on a rotating workpiece is shaved off by the cutting edge. This layer is actually a tube-like volume. Therefore, the turn tool can reduce the diameter of a rotational workpiece.

A human process planner can use his or her experience and imagination to envision the shape a process/tool can create. However, in trying to automate process planning, it is essential to define the geometric capabilities of manufacturing processes explicitly. During process planning, for a given feature an inverse search is conducted to find the candidate process(es) for the feature. The relationship between features and processes is defined in a mapping between the two. In an automated process planning system, rules or algorithms are written based on this mapping.

3.2.1. Process for Features Mapping

The geometric capability of a process is summarized in Table 3. As can be seen, milling can create many different features (Volume Capabilities column). Drilling, reaming, and boring primarily create holes. Turning can create different axial symmetric parts. Other processes are also listed in the table.

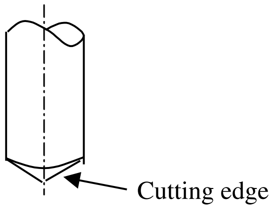


Figure 14 Drill Bit.

TABLE 3 Geometric Capabilities

Process	Subprocess	Cutters	Volume Capabilities
Milling	Face milling	Plain	flat bootom volume
		Inserted-tooth	
	Peripheral milling	Plain	flat bottom volume
		Slitting Saw	slot
		Form	formed volume
		Inserted-tooth	
		Staggered-tooth	
		Angle	
		T-slot cutter	T-slot
		Woodruff keyseat cutter	Internal groove
	End milling	Plain	pocket, slot, flat
		Shell end	
		Hollow end	
		Ball end	sculptured surface, flat
Drilling		Twist drill	round hole
		Spade drill	round hole
		Deep-hole drill	deep round hole
		Gun drill	deep round hole
		Trepanning cutter	large round hole
		Center drill	shallow round hole
		Combination drill	multiple diameter round hole
		Countersink	countersink hole
		Counterbore	counterbore hole
Reaming		Shell reamer	thin wall of round hole
		Expansion reamer	thin wall of round hole
		Adjustable reamer	thin wall of round hole
		Taper reamer	thin wall of round hole
Boring		Adjustable boring bar	thin wall of round hole
		Simple boring bar	thin wall of round hole
Turning	Turning	Plain	?
	Facing	Inserted	disk
	Parting		disk
	Knurling	Knurling tool	?
	Boring	Boring bars	thin wall of round hole
	Drilling	Drills	round hole
	Reaming	Reamers	thin wall of round hole
Broaching			flat bottom volume
			slot
			step
		Form tool	polyhedral through hole
			formed through volume
Sawing		Hacksaw	?
		Bandsaw	
		Circular saw	
Shaping		Form tool	flat bottom volume, slot
Planing		Inserted tool	flat bottom volume
Grinding	Cylindrical grinding	Grinding wheels	?
	Centerless grinding	Points	Internal wall of round hole
	Internal grinding		flat bottom volume
	External grinding		
	Surface grinding		
Honing		Honing stone	?
Lapping		Lap	most surfaces
Tapping		Tap	threaded wall of hole

One can easily find the entries for each process and determine the geometric capabilities of the process. Process planning is an inverse mapping. Given a feature, a process planner tries to find all processes that can create that feature.

The table alone is not sufficient. To select the correct process based on the geometry, one needs to look into the geometric constraints as well. Figure 15 provides a small sample of process constraints based on geometry. For example, on the upper-right corner is the ''large hole through a small slot'' constraint. One should not try to drill such a hole after the slot has been cut. The drill center will be in the air and not cutting any material. It will tend to slip and thus produce an inaccurate hole.

3.2.2. Relative-Cost Table for Manufacturing Processes

When conducting a feature-to-process mapping, one may find several candidate processes for the feature. Which process to choose also depends on the cost of the process. The process cost equation consists of a few terms: the tool and machine costs, the material removal rate, and the energy consumption. The relative cost of a process is the cost of removing a unit volume of material. Since the machining time is the inverse of the material removal rate (for a given machining volume), the cost is:

$$C = \frac{\text{tool and machine rates} + \text{energy cost}}{\text{MRR}}$$

where tool and machine rates are overhead cost of using the tool and the machine and energy cost is the energy cost per unit time.

Processes such as drilling, milling, and turning have higher material-removal rates and thus can finish a job faster at a lower cost. Finishing processes such as grinding, boring, and polishing have very low material-removal rates, and also consume more energy for the same amount of material removed. The relative cost is higher. Table 4 gives the price of machine costs, which are one of the factors in the relative cost equation. The energy consumption, for example, for cutting cast iron is 0.5–1.2 $\text{hp} \cdot \text{min/in}^3$. When grinding is used, the energy consumption is 4.5–22 $\text{hp} \cdot \text{min/in}^3$. Nontraditional processes such as a laser process consume much more energy.

3.3. Process Capability Analysis

Table 5 shows the technological capabilities of 13 processes. Because each shop may use machines and tools of different precision, the data are for reference only. Please note that all dimensions and tolerances are in inches and all surface finish values are in microinches. The process capability values can be used to decide whether a process can satisfy the design specifications of a feature. They can also be used to determine the need of a secondary process (finishing process). For example, a flat surface has a specified surface finish of 20 μin. Using Table 3, we chose flat end mill to cut the surface. From Table 5, we find that finish cut of end mill can create a surface finish of 50 μin. This is definitely not sufficient. Yet finish grinding can create a surface finish of 2 μin. Therefore, finish grinding will be used for finishing and milling for roughing. Milling is chosen for roughing because grinding has a very low material-removal rate. It is not economical to remove all the feature volume using grinding.

Process capability is shop specific. Each shop needs its own process capability database of its own before process planning can be done automatically. Collecting and analyzing capability data can be tedious. Some of these data can be collected through inspection, such as from the control charts. Others require experiments on the machines. Most of the processes that remove material quickly, such as milling, drilling, and turning, create poorer surface finish and accuracy.

3.4. Cost Model

Another extremely important factor is process economics. We are always interested in finding the most economical solution. Often it means the survival of the company. Process economics means the cost efficiency of the processes. For mass production, a very detailed economic analysis is necessary before a specific processing method can be selected. However, for the usual small to medium batch production, it is not practical to conduct a very detailed study. The savings cannot justify the amount of effort spent. Some rough estimation or just common sense should be used to select a better process. Whenever there are more than two candidate processes, both technologically suitable for the task, it is time to compare their relative costs. A process cost model can be stated as:

$$C = \text{labor cost} + \text{machine overhead} + \text{tool change cost} + \text{tool cost}$$

$$C = C_m(T_m + T_h) + (C_t + C_m\, t_t)\,\frac{T_m}{T_l}$$

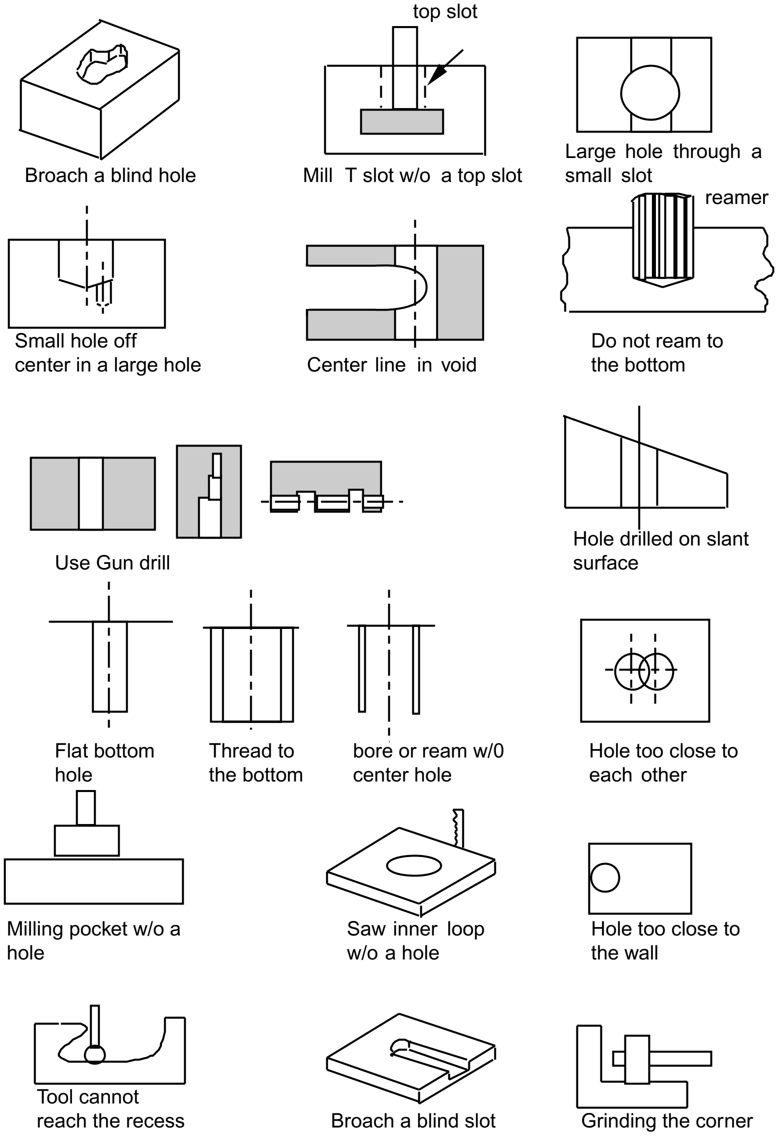

top slot
Broach a blind hole
Mill T slot w/o a top slot
Large hole through a small slot
reamer
Small hole off center in a large hole
Center line in void
Do not ream to the bottom
Use Gun drill
Hole drilled on slant surface
Flat bottom hole
Thread to the bottom
bore or ream w/0 center hole
Hole too close to each other
Milling pocket w/o a hole
Saw inner loop w/o a hole
Hole too close to the wall
Tool cannot reach the recess
Broach a blind slot
Grinding the corner

Figure 15 Process Constraints.

where C = total cost for the operation ($)
C_m = operator rate and machine overhead ($/hr)
C_t = cost of tool ($)
T_m = processing time (hr)
T_h = material-handling time, if any (hr)
T_t = tool change time (hr)
T_l = tool life (hr)

In the equation, T_m/T_l is the number of tool changes for the operation. It is determined by the tool life and the processing time. Processing time can be calculated by the necessary tool travel

TABLE 4 Cost of Machinery

Type of Machinery	Price range ($000)
Broaching	10–300
Drilling	10–100
Electrical discharge	30–150
Electromagnetic and electrohydraulic	50–150
Gear shaping	100–200
Grinding	
Cylindrical	40–150
Surface	20–100
Headers	100–150
Injection molding	30–150
Boring	
Jig	50–150
Horizontal boring mill	100–400
Flexible manufacturing system	>1000
Lathe	10–100
Single- and multi-spindle automatic	30–250
Vertical turret	100–400
Machining center	50–1000
Mechanical press	20–250
Milling	10–250
Robots	20–200
Roll forming	5–100
Rubber forming	50–500

From S. Kalpakjian, *Manufacturing Engineering and Technology,* 3d Ed., © 1995. Reprinted by permission of Prentice-Hall, Inc., Upper Saddle River, NJ.

divided by the feed speed. For example, for drilling an *x*-in. deep hole using a feed speed of *a* ipm, $T_m = x/a/60$.

The tool life equation can be expressed as (for milling):

$$T_l = \frac{C}{V^{\alpha} f^{\beta} a_p{}^{\gamma}}$$

where C = a constant determined by the tool geometry, tool material, and workpiece material
V = the cutting speed, fpm
f = the feed, ipr
a_p = the depth of cut, in.
$^{\alpha}, {}^{\beta}, {}^{\gamma}$ = coefficients

Unfortunately, several difficulties prohibit us from using this model to predict the operation cost. First, the cutter path is not known at the process-selection time. Generating a cutter path for each possible process to be machined would be very time consuming. The second problem is the availability of coefficients for each combination of tool-material type and workpiece-material type. There is little published data for tool life equations. Most of the tool life and machinability data are published in terms of recommended feed and speed. With these two major problems, this approach will probably not work for real-world problems. A quick and dirty way must be found to estimate the cost.

Since we are dealing with the machining of a single feature, it is reasonable to assume that the material-handling time is negligible. The chance of changing a tool during the operation is also minimal. Also, the feed and speed recommended by the *Machining Data Handbook* (Metcut 1980) usually make the tool life to be about 60 minutes. Since the recommended feed and speed are what are used in most machining operations, it is reasonable to assume that T_l = 1 hr. Therefore, the cost function can be simplified to:

TABLE 5 Process Technological Capabilities

Process	Subprocess	Cutters	Tolerances, Surface Finish, etc. Capabilities				
Milling	Face milling	Plain Inserted-tooth		roughing	finishing		
			tol	0.002	0.001		
			flatness	0.001	0.001		
			angularity	0.001	0.001		
			parallelism	0.001	0.001		
			surface finish	50	30		
	Peripheral milling	Plain Slitting saw Form Inserted-tooth Staggered-tooth Angle T-slot cutter Woodruff keyseat cutter Form milling cutter		roughing	finishing		
			tol	0.002	0.001		
			flatness	0.001	0.001		
			surface finish	50	30		
	End milling	Plain Shell end Hollow end Ball end		roughing	finishing		
			tol	0.004	0.004		
			parallelism	0.0015	0.0015		
			surface finish	60	50		
Drilling		Twist drill Spade drill Trepanning cutter Center drill Combination drill Countersink Counterbore	length/diam = 3 usual = 8 maximum mtl < Rc 30 usual < Rc 50 maximum				
			Dia	Tolerance		usual	best
			0–1/8	+0.003–0.001	true position	0.008	0.0004
			1/8–1/4	+0.004–0.001	roundness	0.004	
			1/4–1/2	+0.006–0.001	surface finish	100	
			1/2–1	+0.008–0.002			
			1–2	+0.010–0.003			
			2–4	+0.012–0.004			
		Deep-hole drill Gun drill	Dia	Tolerance			
			<5/8	0.0015	surface finish >100		
			>5/8	0.002	straightness 0.005 in 6 inch		

Units: Tolerances in inches; Surface finish in μ inches; Diameter and length in inches

		Shell reamer	Dia	Tolerance			roughing	finishing
		Expansion reamer	0–1/2	0.0005 to 0.001		roundness	0.0005	0.0005
Reaming		Adjustable reamer	1/2–1	0.001		true position	0.01	0.01
		Taper reamer	1–2	0.002		surface finish	125	50
			2–4	0.003				
			length/dia	5 to 8				
						straightness	0.002	
			Dia	Tolerance		roundness	0.003	
				roughing	finishing	true position	0.0001	
Boring		Adjustable boring bar				surface finish	8	
		Simple boring bar	0–3/4	0.001	0.0002			
			3/4–1	0.0015	0.0002			
			1–2	0.002	0.0004			
			2–4	0.003	0.0008			
			4–6	0.004	0.001			
			6–12	0.005	0.002			
	Turning		diameter	tolerance		surface finish	250 to 16	
	Facing		to 1.0	0.001				
	Parting		1–2	0.002				
		Plain	2–4	0.003				
Turning	Knurling	Inserted						
		Knurling tool						
	Boring	Boring bars						
	Drilling	Drills						
	Reaming	Reamers						
Broaching		Form tool	tolerance 0.001 surface finish 125 to 32					

TABLE 5 (*Continued*)

Process	Subprocess	Cutters	Tolerances, Surface Finish, etc. Capabilities				
			length to	squareness	surface finish	cutting rate	material
Sawing		Hacksaw	0.01	0.2	200–300	3–6 sq in./min	to Rc45
		Bandsaw	0.01	0.2	200–300	4–30 sq in./min	to Rc45
		Circular saw	0.008	0.2	125	7–36 sq in./min	to Rc45
Shaping		Form tool		roughing	finishing		
Planing		Inserted tool	location tol	0.005	0.001		
			flatness	0.001	0.0005		
			surface finish	60	32 (cast iron)		
			surface finish	125	32 (steel)		
Grinding	Internal grinding	Internal	Dia	Tolerance			
				roughing	finishing		
			0–1	0.00015	0.00005		
			1–2	0.0002	0.00005		
			2–4	0.0003	0.0001		
			4–8	0.0005	0.00013		
			8–10	0.0008	0.0002		
	Cylindrical grinding	Center ground and centerless		roughing	finishing		
	Centerless grinding		tolerance	0.0005	0.0001		
	External grinding		parallelism	0.0005	0.0002		
	Surface grinding		roundness	0.0005	0.0001		
			surface fin	8	2		
		flat		roughing	finishing		
			tolerance	0.001	0.0001		
			parallelism	0.001	0.0001		
			surface fin	32	2		

		Dia	Tolerance			
			roughing	finishing	surface finish	4
Honing	Honing stone				roundness	0.0005
		1	+0.0005–0.0	+0.0001–0.0		
		2	+0.0008–0.0	+0.0005–0.0		
		4	+0.0010–0.0	+0.0008–0.0		
			roughing	finishing		
Lapping	Lap	tolerance	0.000025	0.000015		
		flatness	0.000025	0.000012		
		surface fin	4–6	1–4		
		tolerance	0.003			
Tapping	Tap	roundness	0.003			
		surface fin	75			

$$C = \frac{(C_m + C_t)T_m}{60}$$

The machining time can be estimated by the material removal rate (MRR) and the volume (V) to be removed (current feature volume):

$$T_m = \frac{V}{\text{MRR} \cdot 60}$$

Therefore, the cost function can be rewritten as:

$$C = \frac{(C_m + C_t)V}{\text{MRR} \cdot 60}$$

The maximum material-removal rate data can be estimated using some simple equations (Chang 1990). First the tool size, feed, and speed need to be found. The volume to be removed can be estimated much more easily than the length of cutter path. Cost data can also be obtained from the accounting department. With these data the processing cost can be calculated. This cost information can be used to select the most economical process to use for machining a volumetric feature. Because in the processing cost function two variables, C_t and MRR, are related to the process, other capabilities of interest are tool cost and material–removal rate. These capabilities should be used for selecting economical machining processes.

Example. The hole to be drilled is 3 in. deep. The machine and operator rate is $40/hr. The tool cost is $10 each. What is the production cost of the hole?

$$V = \frac{\pi\, 1^2}{4}\, 3 = 2.356 \text{ in.}^3$$

$$C = (40 + 10)\,\frac{2.356}{6.93 \cdot 60} = \$0.283$$

The above model does not consider the fixed cost of tooling. The tool cost used in the model is the incremental tool cost. If special tools are needed, a fixed cost may be incurred. In that case, the fixed cost must be evenly distributed to the entire batch of parts made.

3.5. Tolerance Charting

Tolerance charting is a method for checking the proper in-process dimensions and tolerances from a process plan. It is used to verify whether the process sequence will yield the designed dimensions and tolerances. In most of the literature, *process* is replaced by *operation*. In this section we will use the term *process*. Tolerance charting begins with a part drawing and the process plan. On the process plan are processes and machines. Consequences of processes in terms of resultant dimensions and tolerances are marked on the chart. The processes that were used to produce the dimension and the tolerance are labeled for trace. This is done step by step following the sequence of processes. Finally, the specified dimensions and tolerances are compared to the resultant dimensions and tolerances. If any of the dimensions and tolerances are not satisfied, one can trace back to the sources. Then either a different process/machine is used to reduce the process tolerance or the process sequence is changed.

Figure 16 illustrates how a simplified tolerance chart works. The example part is a cylindrical part to be turned. The calculation section of the chart is omitted. On the top of the chart is the design. Note that the tolerance chart can handle one dimension at a time. The drawing is 2D and features are vertical lines (representing surfaces). The solid line shows the boundary of a cylinder with greater diameter at the center. The dashed line shows the stock boundary that encloses the part boundary. The designer has specified dimensions and tolerances between three feature sets. The dimension values are omitted in this example. The next section is the dimension and tolerance section. The thick horizontal lines show where the dimension and tolerance are specified. For example, the overall dimension is 3 and tolerance is 0.01. The following section is the process plan section. Four cuts are shown in the process plan. The first two cuts (10 and 12) use the left-hand side of the stock as the reference. They create two surfaces: surface *C* and surface *D* (at the same time the diameters are turned). The next two cuts (20 and 22) create the dimensions between *BD* and *AD*. Dimension *AB* is the result of cuts 20 and 22. Therefore, the tolerance for *AB* equals the sum of process tolerances for 20 and 22. In this case both are the same. To achieve the designed tolerance of 0.01, the process

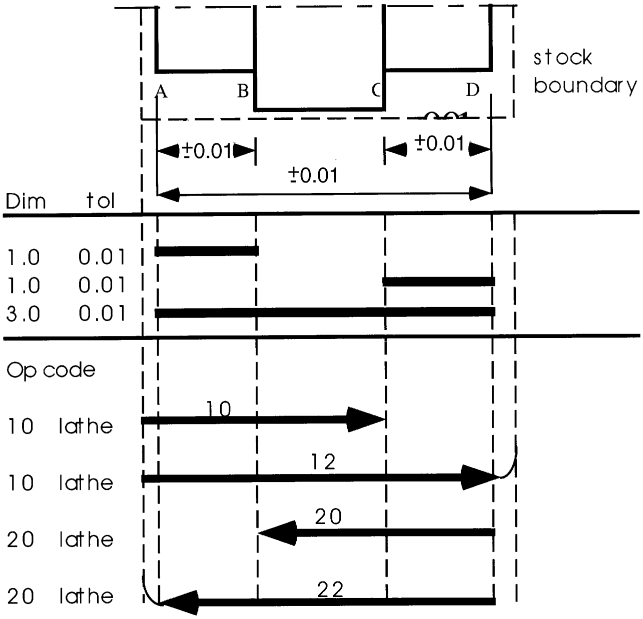


Figure 16 Tolerance Chart.

tolerance must be less than or equal to 0.01/2 = 0.005. The same can be found for *CD*, which is the result of processes 10 and 12. However, *AD* is the result of only one cut and thus the tolerance is better.

More details on this subject can be found in Curtis (1988).

4. COMPUTER-AIDED PROCESS PLANNING

Process planning has traditionally been experience based and performed manually. A problem facing modern industry is the lack of a skilled labor force to produce machined parts as in the past. Manual process planning also has other problems. Variability among the planners' judgment and experience can lead to differences in the perception of what constitutes the optimal or best method of production. This manifests itself in the fact that most industries have several different process plans for the same part, which leads to inconsistent plans and an additional amount of paperwork. To alleviate this problem, a computer-aided approach can be taken. Development in computer-aided process planning attempts to free the process planner from the planning process. Computer-aided process planning can eliminate many of the decisions required during planning. It has the following advantages:

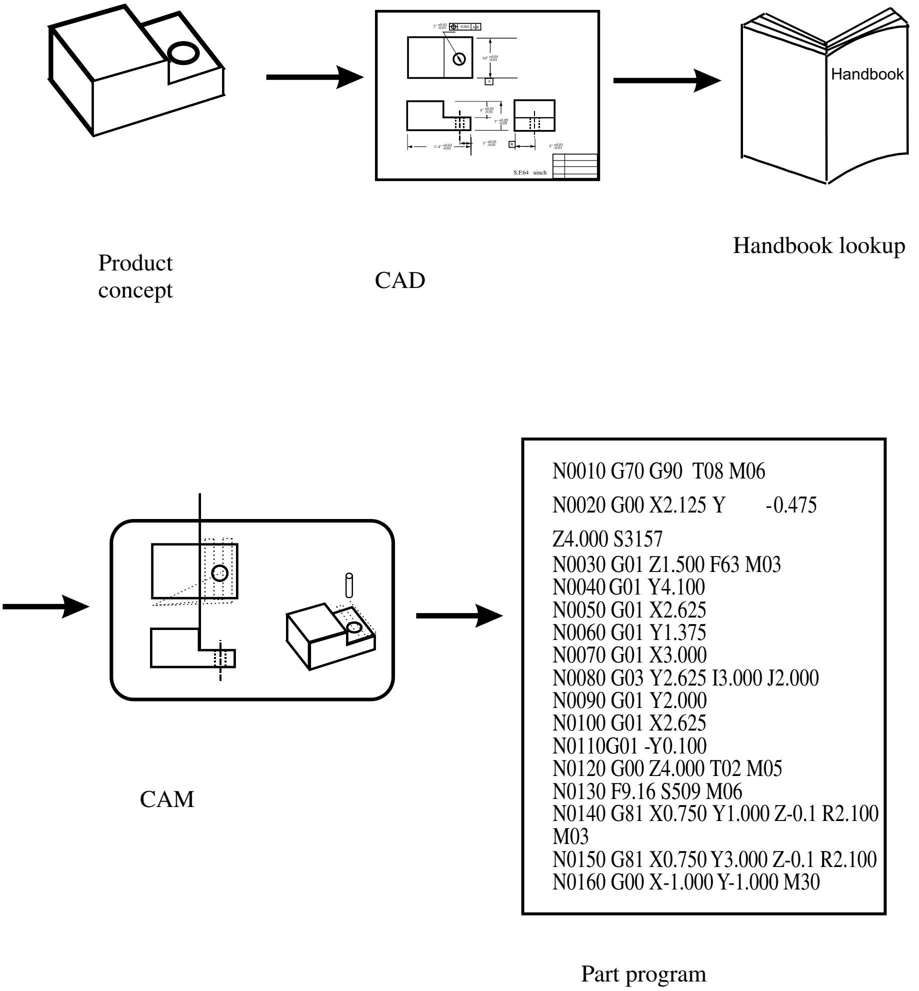
- It reduces the demand on the skilled planner.
- It reduces the process-planning time.
- It reduces both process-planning and manufacturing cost.
- It creates consistent plans.
- It produces accurate plans.
- It increases productivity.

The benefits of computer-aided process planning systems have been documented in several industries. Such systems can reduce planning time from days to hours and result in large cost savings.

The idea of using computers in the process planning activity was discussed by Niebel (1965). Other early investigations on the feasibility of automated process planning can be found in Scheck (1966) and Berra and Barash (1968). Many industries also started research efforts in this direction in the late 1960s and early 1970s. Early attempts to automate process planning consisted primarily of building computer-assisted systems for report generation, storage, and retrieval of plans. A database system with a standard form editor is what many early systems encompassed. Formatting of plans was performed automatically by the system. Process planners simply filled in the details. The storage and retrieval of plans are based on part number, part name, or project ID. When used effectively, these systems can save up to 40% of a process planner's time. A typical example can be found in Lockheed's CAP system (1981). An example of a modern version is Pro/Process for Manufacturing (launched in 1996 and since discontinued). Such a system can by no means perform the process-planning tasks; rather, it helps reduce the clerical work required of the process planner.

The typical organization of using a process-planning system is shown in Figure 17. A human planner interprets an engineering drawing and translates it into the input data format for a process-planning system. Either interactively or automatically, a process plan is produced. The plan is then used by production planners for scheduling of production and used by industrial engineers to lay out the manufacturing cell and calculate production cost and time. A part programmer follows the instructions on the process plan and the engineering drawing to prepare NC (numerical control) part programs. The same organization applies to all kinds of process planning systems.

Perhaps the best-known automated process planning system is the CAM-I automated process planning system (CAPP) (Link 1976). (CAM-I stands for ComputerAided Manufacturing International, a nonprofit industrial research organization.) In CAPP, previously prepared process plans are stored in a database. When a new component is planned, a process plan for a similar component is retrieved and subsequently modified by a process planner to satisfy special requirements. The tech-



```
N0010 G70 G90  T08 M06
N0020 G00 X2.125 Y      -0.475
Z4.000 S3157
N0030 G01 Z1.500 F63 M03
N0040 G01 Y4.100
N0050 G01 X2.625
N0060 G01 Y1.375
N0070 G01 X3.000
N0080 G03 Y2.625 I3.000 J2.000
N0090 G01 Y2.000
N0100 G01 X2.625
N0110G01 -Y0.100
N0120 G00 Z4.000 T02 M05
N0130 F9.16 S509 M06
N0140 G81 X0.750 Y1.000 Z-0.1 R2.100
M03
N0150 G81 X0.750 Y3.000 Z-0.1 R2.100
N0160 G00 X-1.000 Y-1.000 M30
```

Figure 17 From Design to Part Program.

nique involved is called group technology (GT)-based variant planning (Burbidge 1975). Variant planning will be discussed in more detail in the next section.

Figure 18 represents the structure of a complete computer-aided process-planning system. Although no existing turnkey system integrates all of the functions shown in the figure (or even a goodly portion of them), it illustrates the functional dependencies of a complete process-planning system. It also helps to illustrate some of the constraints imposed on a process-planning system (e.g., available machines, tooling, and jigs).

In Figure 18, the modules are not necessarily arranged based on importance or decision sequence. The system monitor controls the execution sequence of the individual modules. Each module may require execution several times in order to obtain an optimum process plan. Iterations are required to reach feasibility as well as good economic balance.

The input to the system will most probably be a 3D model from a CAD database. The model contains not only the shape and dimensioning information, but also the tolerances and special features. The process plan can be routed directly to the production-planning system and production-control system. Time estimates and resource requirements can be sent to the production-planning system for scheduling. The part program, cutter location (CL) file, and material-handling control program can also be sent to the control system.

Process planning is the critical bridge between design and manufacturing. Design information can be translated into manufacturing language only through process planning. Today, both automated design (CAD) and manufacturing (CAM) have been implemented. Integrating, or bridging, these functions requires automated process planning as the key component.

There are two basic approaches to computer-aided process planning: variant and generative. The variant approach is used by the computer to retrieve plans for similar components using table look-up procedures. The process planner then edits the plan to create a variant to suit the specific requirements of the component being planned. Creation and modification of standard plans are the process planner's responsibility. The generative approach is based on generating a plan for each component without referring to existing plans. Generative systems are systems that perform many of the functions in a generative manner. The remaining functions are performed with the use of humans in the planning loop.

4.1. Variant Approach

The variant approach to process planning was the first approach used to computerize planning techniques. It is based on the idea that similar parts will have similar process plans. The computer can

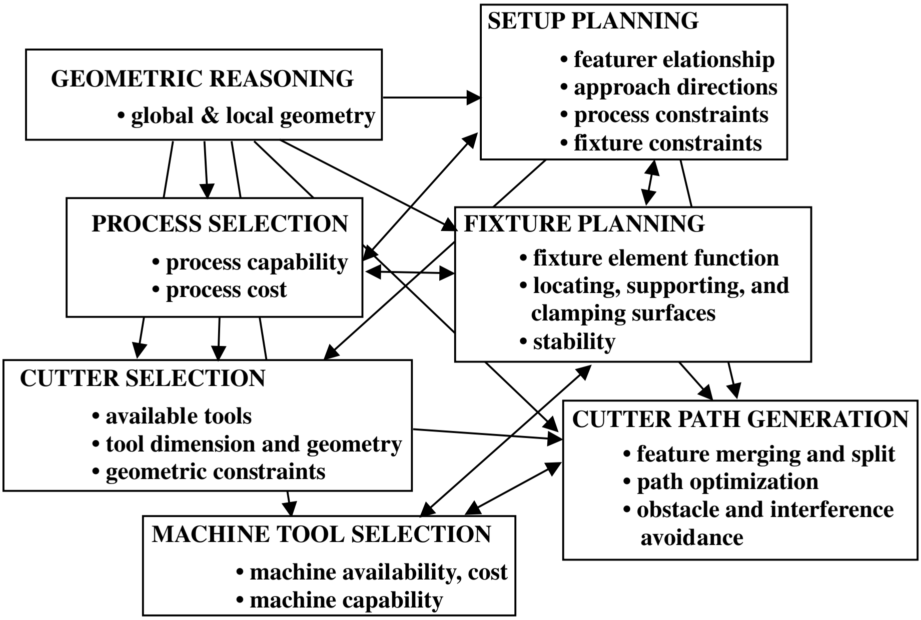

SETUP PLANNING
• featurer elationship
• approach directions
• process constraints
• fixture constraints
GEOMETRIC REASONING
• global & local geometry
PROCESS SELECTION
• process capability
• process cost
FIXTURE PLANNING
• fixture element function
• locating, supporting, and clamping surfaces
• stability
CUTTER SELECTION
• available tools
• tool dimension and geometry
• geometric constraints
CUTTER PATH GENERATION
• feature merging and split
• path optimization
• obstacle and interference avoidance
MACHINE TOOL SELECTION
• machine availability, cost
• machine capability

Figure 18 Process-Planning Functions.

be used as a tool to assist in the identification of similar plans, retrieving them and editing the plans to suit the requirements for specific parts.

A variant process planning system includes the following functions:

- Family formation
- Standard plan preparation
- Plan editing
- Databases

In order to implement such a concept, GT-based part coding and classification are used as a foundation. Individual parts are coded based upon several characteristics and attributes. Part families are created of ''like'' parts having sufficiently common attributes to group them into a family. This family formation is determined by analyzing the codes of the part spectrum. A ''standard'' plan consisting of a process plan to manufacture the entire family is created and stored for each part family. The development of a variant process-planning system has two stages: the preparatory stage and the production stage.

During the preparatory stage, existing components are coded, classified, and later grouped into families (Figure 19). The part family formation can be performed in several ways. Families can be formed based on geometric shapes or process similarities. Several methods can be used to form these groupings. A simple approach would be to compare the similarity of the part's code with other part codes. Since similar parts will have similar code characteristics, a logic that compares part of the code or the entire code can be used to determine similarity between parts.

Families can often be described by a set of family matrices. Each family has a binary matrix with a column for each digit in the code and a row for each value a code digit can have. A nonzero entry in the matrix indicates that the particular digit can have the value of that row. For example, entry (3,2) equals one implies that a code x3xxx can be a member of the family. Since the processes of all family members are similar, a standard plan can be assigned to the family.

The standard plan is structured and stored in a coded manner using operation codes (OP codes). An operation code represents a series of operations on one machine/workstation. For example, an OP code DRL10 may represent the sequence center drill, change drill, drill hole, change to reamer, and ream hole. A series of OP codes constitutes the representation of the standard process plan.

Before the system can be of any use, coding, classification, family formation, and standard plan preparation must be completed. The effectiveness and performance of the variant process-planning system depends to a very large extent on the effort put forth at this stage. The preparatory stage is a very time-consuming process.

The production stage occurs when the system is ready for production. New components can be planned in this stage. An incoming component is first coded. The code is then sent to a part family search routine to find the family to which it belongs. Because the standard plan is indexed by the family number, the standard plan can be easily retrieved from the database. Because the standard

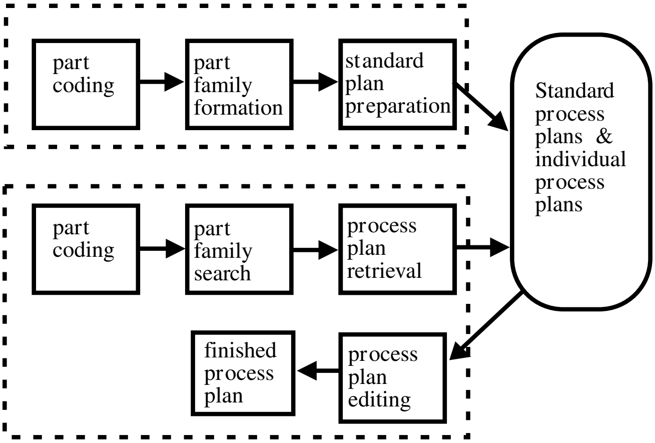


Figure 19 Variant Process Planning.

plan is designed for the entire family rather than for a specific component, editing the plan is unavoidable.

Variant process-planning systems are relatively easy to build. However, several problems are associated with them:

- The components to be planned are limited to similar components previously planned.
- Experienced process planners are still required to modify the standard plan for the specific component.
- Details of the plan cannot be generated.
- Variant planning cannot be used in an entirely automated manufacturing system, without additional process planning.

Despite these problems, the variant approach is still an effective method, especially when the primary objective is to improve the current practice of process planning. In most batch manufacturing industries, where similar components are produced repetitively, a variant system can improve the planning efficiency dramatically. Some other advantages of variant process planning are:

- Once a standard plan has been written, a variety of components can be planned.
- Comparatively simple programming and installation (compared with generative systems) is required to implement a planning system.
- The system is understandable and the planner has control of the final plan.
- It is easy to learn and easy to use.

The variant approach is the most popular approach in industry today. Most working systems are of this type, such as CAPP of CAM-I (Link 1976) and Multiplan of OIR (OIR 1983).

4.2. Generative Approach

Generative process planning is the second type of computer-aided process planning. It can be concisely defined as a system that automatically synthesizes a process plan for a new component. The generative approach envisions the creation of a process plan from information available in a manufacturing database without human intervention. Upon receiving the design model, the system is able to generate the required operations and operation sequence for the component.

A generative process-planning system consists of the following important functions:

- Design representation
- Feature recognition
- Knowledge representation
- System structures

Knowledge of manufacturing has to be captured and encoded into computer programs. A process planner's decision-making process can be imitated by applying decision logic. Other planning functions, such as machine selection, tool selection, and process optimization, can also be automated using generative planning techniques.

A generative process-planning system contains three main components:

- Part description
- Manufacturing databases
- Decision-making logic and algorithms

The definition of generative process planning used in industry today is somewhat relaxed. Thus, systems that contain some decision-making capability in process selection are called generative systems. Some of the so-called generative systems use a decision tree to retrieve a standard plan. Generative process planning is regarded as more advanced than variant process planning. Ideally, a generative process-planning system is a turnkey system with all the decision logic built in. Since this is still far from being realized, generative systems currently developed provide a wide range of capabilities and can at best be described as only semigenerative.

The generative process-planning approach has the following advantages:

- It generates consistent process plans rapidly.
- New components can be planned as easily as existing components.

- It has potential for integrating with an automated manufacturing facility to provide detailed control information.

Successful implementation of this approach requires the following key developments:

- The logic of process planning must be identified and captured.
- The part to be produced must be clearly and precisely defined in a computer-compatible format.
- The captured logic of process planning and the part-description data must be incorporated into a unified manufacturing database.

4.2.1. Part-Description Methods for Generative Process-Planning Systems

Part description forms a major part of the information needed for process planning. The way in which the part description is input into the process-planning system has a direct effect on the degree of automation that can be achieved. Since the aim is to automate the system, the part description should be in a computer-readable format. Traditionally, engineering drawings have been used to convey part descriptions and communicate between design and manufacturing. Understanding the engineering drawing was a task suited for well-trained human beings and initially not suitable for direct input for process planning. The requirements of the part-description methods include:

- Geometrical information
- Part shape
- Design features
- Technological information
- Tolerances
- Surface quality (surface finish, surface integrity)
- Special manufacturing notes
- Feature information
- Manufacturing features (e.g., slots, holes, pockets)

Before the representation method is decided, the following factors have to be determined:

- Data format required
- Ease of use for planning
- Interface with other functions, such as part programming and design
- Easy recognition of manufacturing features
- Easy extraction of planning information from the representation

Some of the representations used in a generative process-planning system include: GT code, line drawing, special language, symbolic representation, solid model, CSG, B-Rep, feature-based model. Extract and decompose features from a geometric model.

- Syntactic pattern recognition
- State transition diagram and automata
- Decomposition
- Logic
- Graph matching
- Face growing

5. COMPUTER-AIDED PROCESS PLANNING SYSTEMS SELECTION CRITERIA

Selecting a process-planning system for a company is not an easy task. Each machine shop has its own unique characteristics. Decision tables have been developed for selecting computer-aided process-planning systems. For example, Steudel and Tollers (1985) present a decision table for selecting CAPP solutions (the tables are also reprinted in Chang et al. [1991, chap. 13]). This approach assumes that certain decision rules can be written for each selection decision. However, often such rules are not available. Another approach is to compare the characteristics of the competing systems based on weights. Table 6 shows a comparison table for two process-planning systems. In the table are 12 major categories, most of which are divided into several subcategories. Weights can be assigned

TABLE 6 System Comparison Table

		Systems	
	Weight	System 1	System 2
1. Input data			
1.1. CAD file format		1	4
2. Workpiece understanding	**3**		
2.1. Shape analysis		0	5
2.2. Material analysis		1	1
2.3. Specification analysis		1	1
2.4. Tolerance analysis		2	3
Average		1	3
3. Process selection	**2**		
3.1. Feature hierarchy		3	5
3.2. Tolerance analysis		0	5
3.3. Process capability model		0	5
3.4. Specification analysis		0	2
3.5. Cost analysis		0	1
3.6. Capacity analysis		0	2
Average		**1**	**3**
4. Machine tool management	**2**		
4.1. Machine tool selection		0	5
4.2. Machine capability model		0	5
4.3. Maintenance planning		0	0
4.4. Environmental analysis		0	0
4.5. Flow analysis		0	0
4.6. Life-cycle analysis		0	0
4.7. Supplier development		0	0
4.8. Control specification		0	0
4.9. Facility planning		0	0
Average		**0**	**5**
5. Quality management	**2**		
5.1. Process-control planning		0	0
5.2. Gage planning		1	0
5.3. Scrap and rework management		0	0
5.4. Quality assurance		0	0
Average		**1**	**0**
6. Process design	**3**		
6.1. Tool selection		1	5
6.2. Fixture design		1	4
6.3. Gage design		1	0
6.4. Tool path generation		2	5
6.5. Operation sequencing		1	5
6.6. Process parameters		1	5
6.7. Feature accessibility		0	5
6.8. Tolerance analysis		0	4
Average	**1**	**4**	
7. Evaluation and validation	**3**		
7.1. Target cost		0	5
7.2. Tool path verification		5	4
7.3. Workpiece quality		0	0
7.4. Process reliability		0	0
7.5. Production times		1	5
Average		**2**	**3**
8. Document generation	**3**		
8.1. Tool/gage orders		2	4
8.2. Equipment orders		0	5
8.3. Operation sheets		2	4
8.4. Process routing		2	4
8.5. Part programs		5	5
8.6. Setup drawings		3	5
Average		**2**	**5**
9. Machine domain	**3**	**4**	**5**
10. Part domain	**3**	**4**	**4**
11. Platform	**1**	**5**	**3**
12. Technical support	**3**	**5**	**4**
Total		**65**	**101**

to either each major category or subcategory. The total weights may be used to determine the final selection. The comparison is not limited to the pair comparison as it is in the example; more systems can be added to the table.

The meanings of the categories in Table 6 are defined below:

- Input data: The process-planning system must interface with the existing CAD design system. The system can either read the native CAD data or can input through an data-exchange format such as IGES or STEP.
  - CAD file format: the design data file acceptable, e.g., Pro/E, CADDS, CATIA, IGES, STEP.
- Workpiece understanding:
  - Shape analysis: feature recognition; converting design data into manufacturing feature data.
  - Material analysis: identification of raw material shape, sizes, type (cast, bar, etc.), and weight.
  - Specification analysis: extracting other manufacturing specifications from the design (e.g., paint, hardness, surface treatment).
  - Tolerance analysis: extracting tolerance data from the design.
- Process selection:
  - Feature hierarchy: process selection based on hierarchically arranged manufacturing features.
  - Tolerance analysis: process selection based on the tolerance capabilities of processes vs. the design specifications.
  - Process capability model: a computer model that captures the capability of a process. Process models are used in process selection. They may be customized.
  - Specification analysis: understanding other manufacturing specifications and using them as the basis for selecting processes.
  - Cost analysis: estimating and analyzing manufacturing cost and using it as the basis for selecting processes.
  - Capacity analysis: machine selection with the consideration of the throughput of each machine.
- Machine tool management:
  - Machine tool selection: selecting appropriate machine tools for the job.
  - Machine capability model: a computer model that captures the process capability of a machine tool.
  - Maintenance planning: maintenance schedule for machine tools.
  - Environmental analysis: assessment of environmental requirements (temperature, pressure, humidity) necessary to achieve quality targets.
  - Flow analysis: production flow analysis (i.e., throughput).
  - Life-cycle analysis: the economic analysis, from cradle to grave, of asset costs to derive metrics like ROA.
  - Supplier development: identifying supplier for the machine tool.
  - Control specification: managing NC controller specifications for postprocessors.
  - Facility planning: integration with the facility planning functions.
- Quality management
  - Process control planning: generating control charts for process control.
  - Gage planning: CMM programming, gage calibration, certification planning.
  - Scrap and rework management: planning for management of the disposition of parts for scrap and rework
  - Quality assurance: generating quality assurance plan for certification. Process plans may be used for this purpose.
- Process design
  - Tool selection: selecting tools to be used for the machining of a part.
  - Fixture design: designing the fixture necessary for the workpiece holding under a given setup.
  - Gage design: designing the gage for inspecting the workpiece.
  - Tool path generation: generating the NC tool path for the part.
  - Operation sequencing: sequencing the NC tool path for the part
  - Process parameters: selecting process parameters, e.g., feed and speed, for each tool/cut.
  - Feature accessibility: analyzing the tool accessibility of each feature to be machined.

TABLE 7 Summary of System Functionalities

	System	
	System 1	System 2
1. Input data		
1.1. CAD file format	IGES	PDES, DXF
2. Workpiece understanding		
2.1. Shape analysis	N/A	Feature recognition
2.2. Material analysis	limited	limited
2.3. Specification analysis	Manual	Manual
2.4. Tolerance analysis	Manual	Manual input
3. Process selection		
3.1. Feature hierarchy	Pro/E	Yes
3.2. Tolerance analysis	Manual	Rules
3.3. Process capability model	N/A	Method hierarchy/DB
3.4. Specification analysis	Manual	limited
3.5. Cost analysis	N/A	implicit
3.6. Capacity analysis	N/A	implicit
4. Machine tool management		
4.1. Machine tool selection	Manual	Automatic
4.2. Machine capability model	N/A	Method hierarchy/DB
4.3. Maintenance planning	N/A	N/A
4.4. Environmental analysis	N/A	N/A
4.5. Flow analysis	N/A	N/A
4.6. Life-cycle analysis	N/A	N/A
4.7. Supplier development	N/A	N/A
4.8. Control specification	N/A	N/A
4.9. Facility planning	N/A	N/A
5. Quality management		
5.1. Process control planning	N/A	N/A
5.2. Gage planning	CMM/interactive	N/A
5.3. Scrap and Rework Management	N/A	N/A
5.4. Quality assurance	N/A	N/A
6. Process design		
6.1. Tool selection	Manual	Automatic
6.2. Fixture design	Manual	Automatic
6.3. Gage design	Manual	N/A
6.4. Tool path generation	Interactive	Automatic
6.5. Operation sequencing	Manual	Automatic
6.6. Process parameters	User-supplied DB	User-supplied RDB
6.7. Feature accessibility	N/A	Yes
6.8. Tolerance analysis	N/A	Yes
7. Evaluation and validation		
7.1. Target cost	N/A	Yes
7.2. Tool path verification	Pro/NC-Check	Yes
7.3. Workpiece quality	N/A	N/A
7.4. Process reliability	N/A	N/A
7.5. Production times	N/A	Yes
8. Document generation		
8.1. Tool/gage orders	User-defined template	Tool list
8.2. Equipment orders	N/A	Machine list
8.3. Operation sheets	User-defined template	Yes
8.4. Process routing	User-defined template	Yes
8.5. Part programs	Pro/E	APT
8.6. Setup drawings	Manual	Yes
9. Machine domain	Mill, lathe, etc.	Machining center
10. Part domain	Prismatic, turned	Prismatic 2 1/2D, turned
11. Platform	Workstations and PCs	SGI/HP, PC
12. Technical support	Local	Headquarters

  - Tolerance analysis: comparing the tolerance specification against the tolerance capabilities of the selected tools and machine tools; stacking up tolerance based on the setup and machine accuracy and setup error.
- Evaluation and validation:
  - Target cost: estimating the manufacturing cost of that part.
  - Tool path verification: graphically simulating the tool motion to ensure that the tool path has no collision with the fixture and the machine and produces correct part geometry.
  - Workpiece quality: confirming that quality indices are met.
  - Process reliability: reliability indexes for the processes selected for the part.
  - Production times: estimating the time it takes to produce the part.
- Document generation:
  - Tool/gage orders: job orders for tools and gages needed.
  - Equipment orders: job orders for equipment used.
  - Operation sheets: detailed description of each operation and operation parameters including gaging detail/instructions.
  - Process routing: process routing sheet. Lists operation sequence and processes.
  - Part programs: NC part program and part program format, e.g., RS 274, CL, BCL, APT source, COMPACT II source, postprocessed N-G code.
  - Setup drawings: drawings of the workpiece with fixture for each setup.
- Machine domain: machine tools that can be modeled and planned by the system.
- Part domain: types of part that can be planned by the system (e.g., turned, prismatic, casting).
- Platform: computer platform on which the software can be used.
- Technical support: technical support provided by the vendor.
- Total: total composite score based on the criteria above.

A composite score may not be sufficient for making the final decision. It is necessary to record the specifications in each subcategory. Table 7 illustrates such a table for the same comparison as in Figure 6. With these two tables, an appropriate computer-aided process planning system can be selected.

6. CONCLUSIONS

In this chapter, we have provided a general overview of manufacturing process planning and design. We have also discussed the tools used in process planning. As mentioned earlier in the chapter, automated process planning is especially important for small batch production. At a time when competition is keen and product changeover rapid, shortening the production lead time and production cost are critical to the survival of modern manufacturing industry. Process planning plays an important role in the product-realization process. In order to achieve the goals stated above, it is essential to put more focus on this planning function.

Process planning is a function that requires much human judgment. Over the past three decades, much effort has been put into automating process planning. Although progress has been made on geometric modeling and reasoning, knowledge-based systems, and computational geometry for toolpath generation, no robust integrated process-planning system has been developed. Developing an automated process-planning system is still a challenge to be fulfilled. In the meantime, many of these technologies have been integrated into CAD/CAM systems. They provide human planners with powerful tools to use in developing process plans. We can expect an increasing amount of planning automation be made available to manufacturing engineers. This chapter provides a general background to this important topic. Additional readings are included for those interested in a more in-depth investigation of the subject.

REFERENCES

Allen, D. K. (1994), “Group Technology,” *Journal of Applied Manufacturing Systems,* Vol. 6, No. 2, pp. 37–46.

Berra, P. B., and Barash, M. M. (1968), “Investigation of Automated Process Planning and Optimization of Metal Working Processes,” Report 14, Purdue Laboratory for Applied Industrial Control, West Lafayette, IN, July.

Burbidge, J. L. (1975), *The Introduction of Group Technology,* John Wiley & Sons, New York.

Chang, T. C. (1990), *Expert Process Planning for Manufacturing,* Addison-Wesley, Reading, MA, 1990.

Chang, T.-C., Wysk, R. A., and Wang, H. P. (1991), *Computer-Aided Manufacturing,* 1st Ed., Prentice Hall, Englewood Cliffs, NJ.

Chang, T.-C., Wysk, R. A., and Wang, H. P. (1998), *Computer-Aided Manufacturing,* 2nd Ed., Prentice Hall, Upper Saddle River, NJ.

Curtis, M. A. (1988), *Process Planning,* John Wiley & Sons.

Dunn, M. S., and Mann W. S. (1978), ‘‘Computerized Production Process Planning,’’ in *Proceedings of 15th Numerical Control Society Annual Meeting and Technical Conference* (Chicago, April).

Kalpakjian, S. (1995), *Manufacturing Engineering and Technology,* 3rd Ed., Addison-Wesley, Reading, MA.

Kusiak, A. (1990), *Intelligent Manufacturing Systems,* Prentice Hall, Englewood Cliffs, NJ.

Link, C. H. (1976), ‘‘CAPP-CAM-I Automated Process Planning System,’’ in *Proceedings of 13th Numerical Control Society Annual Meeting and Technical Conference* (Cincinnati, March).

Metcut Research Associates, Inc. (1980), *Machining Data Handbook,* 3d Ed., Machinability Data Center, Metcut Research Associates, Inc., Cincinnati.

Niebel, B. W. (1965), ‘‘Mechanized Process Selection for Planning New Designs,’’ ASTME Paper 737.

Organization for Industrial Research, Inc. (OIR) (1983), MULTIPLAN, OIR, Waltham, MA.

Opitz, H. (1970), *A Classification System to Describe Workpieces,* Pergamon Press, Elmsford, NV.

Scheck, D. E. (1966), ‘‘Feasibility of Automated Process Planning,’’ Ph.D. Thesis, Purdue University, West Lafayette, IN.

Steudel, H. J, and Tollers, G. V. (1985), ‘‘A Decision Table Based Guide for Evaluating Computer-Aided Process Planning Systems,’’ in *Proceedings of 1985 ASME Winter Annual Meeting* (Miami Beach), pp. 109–119.

ADDITIONAL READING

Allen, D., and Smith, P., *Part Classification and Coding,* Monograph No. 3., CAM Software Laboratory, Brigham Young University, Provo, UT, 1988.

Chang, T. C., *An Introduction to Automated Process Planning Systems,* Prentice Hall, Englewood Cliffs, NJ, 1985.

Drozda, T. J., and Wick, C., *Machining,* Vol. 1 of *Tool and Manufacturing Engineers Handbook,* Society of Manufacturing Engineers, 1983.

Halevi, G., and Weill, R. D., *Principles of Process Planning: A Logical Approach,* Chapman & Hall, London, 1995.

Houtzeel, A., and Schilperoort, B. A., ‘‘A Chain Structured Part Classification System (MICLASS) and Group Technology,’’ in *Proceedings of the 13th Annual Meeting and Technical Conference* (Pittsburgh), pp. 383–400.

Japan Society for the Promotion of Machine Industry, ‘‘Group Technology,’’ March 1980.

Jones, S. W., *Product Design and Process Selection,* Butterworths, London, 1973.

Nolen, J., *Computer Aided Process Planning for World-Class Manufacturing,* Marcel Dekker, New York, 1989.

Swift, K. G., and Booker, J. D., *Process Selection: From Design to Manufacture,* Arnold, London, and John Wiley & Sons, New York, 1997.

Tulkoff, J., ‘‘Lockheed’s GENPLAN,’’ in *Proceedings of 18th Numerical Control Society Annual Meeting and Technical Conference* (Dallas, May 1981), pp. 417–421.

Wang, H.-P., Computer-Aided Process Planning, Elsevier, Amsterdam, 1991.

CHAPTER 15

COMPUTER INTEGRATED MANUFACTURING

CHENG WU
FAN YUSHUN
XIAO DEYUN
Tsinghua University

1. **INTRODUCTION** **485**
2. **CIM DEFINITIONS AND CONCEPTS** **485**
    - 2.1. Manufacturing Environment 485
    - 2.2. Features of a General Manufacturing System 486
    - 2.3. CIM Definitions 487
    - 2.4. Integration: The Core of CIM 489
3. **CIMS STRUCTURE AND FUNCTIONS** **491**
    - 3.1. CIMS Structure 491
    - 3.2. Components of CIMS 491
        - 3.2.1. Management Information System 491
        - 3.2.2. CAD/CAPP/CAM System 494
        - 3.2.3. Manufacturing Automation System 496
        - 3.2.4. Computer-Aided Quality Management System 497
        - 3.2.5. Computer Network and Database Management Systems 498
4. **FLEXIBLE MANUFACTURING SYSTEMS** **499**
    - 4.1. Flexibility and Components of FMS 499
        - 4.1.1. Flexibility of Manufacturing System 499
        - 4.1.2. FMS Definition and Components 499
    - 4.2. General FMS Considerations 500
        - 4.2.1. FMS Design 500
        - 4.2.2. FMS Planning, Scheduling, and Control 501
        - 4.2.3. FMS Modeling and Simulation 503
    - 4.3. Benefits and Limitations of FMS 506
5. **CIM ARCHITECTURE AND ENTERPRISE MODELING** **507**
    - 5.1. Views of the Enterprise Model 507
        - 5.1.1. Process View 507
        - 5.1.2. Function View 508
        - 5.1.3. Information View 509
        - 5.1.4. Organization View 510
        - 5.1.5. Resource View 510
    - 5.2. Enterprise Modeling Methods 510
        - 5.2.1. CIMOSA 510
        - 5.2.2. ARIS 512
        - 5.2.3. GIM 512
6. **CIM IMPLEMENTATION** **514**
    - 6.1. General Steps for CIM Implementation 514
        - 6.1.1. Feasibility Study 514
        - 6.1.2. Overall System Design 514
        - 6.1.3. Detailed System Design 515
        - 6.1.4. Implementation and Operation 516
    - 6.2. Integration Platform Technology 516
        - 6.2.1. Requirements for Integration Platform 516

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

6.2.2. The Evolution of Integration Platform Technology 516

6.2.3. MACIP System Architecture 517

7. CIMS IN PROCESS INDUSTRY 518

7.1. Introduction 518

7.1.1. Definitions 518

7.1.2. Key Technologies 519

7.2. Reference Architecture of CIMS in Process Industry 519

7.2.1. Architecture Structure Model 520

7.2.2. Hierarchical Structure Model 521

7.3. Approach to Information Integration for CIMS in Process Industry 522

7.3.1. Production Process Information Integration 522

7.3.2. Model-Driven Approach to Information Integration 522

7.4. An Application Example 523

7.4.1. Refinery Planning Process 523

7.4.2. Integrated Information Architecture 523

7.4.3. Advanced Computing Environment 524

7.5 Conclusions 525

8. BENEFITS OF CIMS 525

8.1. Technical Benefits 525

8.2. Management Benefits 525

8.3. Human Resource Quality 525

9. FUTURE TRENDS OF CIM 527

9.1. Agile Manufacturing 527

9.2. Green Manufacturing 527

9.3. Virtual Manufacturing and Other Trends 527

REFERENCES 528

1. INTRODUCTION

Joseph Harrington introduced the concept of computer integrated manufacturing (CIM) in 1979 (Harrington 1979). Not until about 1984 did the potential benefits the concept promised begin to be appreciated. Since 1984, thousands of articles have been published on the subject. Thanks to the contributions of researchers and practitioners from industries, CIM has become a very challenging and fruitful research area. Researchers from different disciplines have contributed their own perspectives on CIM. They have used their knowledge to solve different problems in industry practice and contributed to the development of CIM methodologies and theories.

2. CIM DEFINITIONS AND CONCEPTS

2.1. Manufacturing Environment

From the name ‘‘computer integrated manufacturing,’’ it can be seen that the application area of CIM is manufacturing. Manufacturing companies today face intense market competition, and are experiencing major changes with respect to resources, markets, manufacturing processes, and product strategies. Manufacturing companies must respond to the rapidly changing market and the new technologies being implemented by their competitors. Furthermore, manufacturing, which has been treated as an outcast by corporate planning and strategy, must become directly involved in these critical long-range decisions. Manufacturing can indeed be a formidable competitive weapon, but only if we plan for it and provide the necessary tools and technologies (Buffa 1984).

Besides the traditional competitive requirements of low cost and high quality, competitive pressure for today’s manufacturing companies means more complex products, shorter product life cycles, shorter delivery time, more customized products, and fewer skilled workers. The importance of these elements varies among industries and even among companies in the same industry, depending on each company’s strategy.

Today’s products are becoming much more complex and difficult to design and manufacture. One example is the automobile, which is becoming more complex, with computer-controlled ignition, braking, and maintenance systems. To avoid long design times for the more complex products, companies should develop tools and use new technologies, such as concurrent engineering, and at the same time improve their design and manufacturing processes.

Higher quality is the basic demand of customers, who want their money’s worth for the products they buy. This applies to both consumers and industrial customers. Improved quality can be achieved through better design and better quality control in the manufacturing operation. Besides demanding higher quality, customers are not satisfied with the basic products with no options. There is a competitive advantage in having a broad product line with many versions, or a few basic models that can be customized. A brand new concept in manufacturing is to involve users in the product design. With the aid of design tools or a modeling box, the company allows the users to design the products in their own favor.

In the past, once a product was designed, it had a long life over which to recover its development costs. Today many products, especially high-technology products, have a relatively short life cycle. This change has two implications. First, companies must design products and get them to the market faster. Second, a shorter product life provides less time over which to recover the development costs. Companies should therefore use new technologies to reduce both time and cost in product design. Concurrent engineering is one method for improving product design efficiency and reducing product costs. Another new paradigm is represented by agile manufacturing, in which the cost and risks of new product development are distributed to partners and benefits are shared among the partners. This requires changes to or reengineering of traditional organization structures.

Several demographic trends are seriously affecting manufacturing employment. The education level and expectations of people are changing. Fewer new workers are interested in manufacturing jobs, especially the unskilled and semiskilled ones. The lack of new employees for the skilled jobs that are essential for a factory is even more critical. On the other hand, many people may not have sufficient education to be qualified for these jobs (Bray 1988).

To win in the global market, manufacturing companies should improve their competitive ability. Key elements include creative new products, higher quality, better service, greater agility, and low environmental pollution. Creative new products are of vital importance to companies in the current "knowledge economy."

Figure 1 presents a market change graph. From this figure, it can be seen that the numbers for lot size and repetitive order are decreasing, product life cycle is shortening, and product variety is increasing rapidly.

End users or customers always need new products with advancements in function, operation, and energy consumption. The company can receive greater benefits through new products. A manufacturing company without new products has little chance of surviving in the future market. Better services are needed by any kind of company. However, for manufacturing companies, better service means delivering products fast, making products easy to use, and satisfying customer needs with low prices and rapid response to customer maintenance requests.

2.2. Features of a General Manufacturing System

The manufacturing company is a complex, dynamic, and stochastic entity consisting of a number of semi-independent subsystems interacting and intercommunicating in an attempt to make the overall system function profitably. The complexity comes from the heterogeneous environment (both hardware and software), huge quantity of data, and the uncertain external environment. The complex structure of the system and the complex relationships between the interacting semi-autonomous subsystems are also factors making the system more complicated.

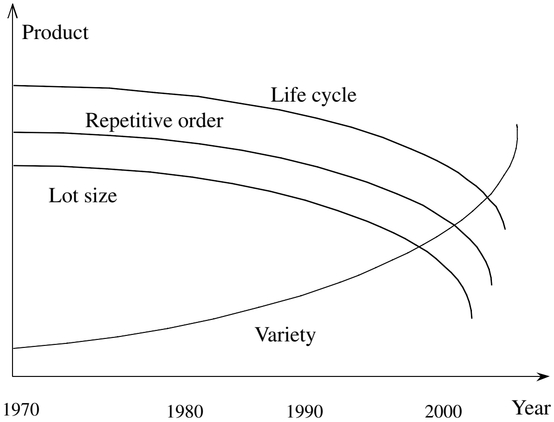


Figure 1 Market Change.

A simple model of a manufacturing system can be a black box that takes as input materials, energy, and information and gives as output products. The internal details of the manufacturing system depend on the particular industry involved, but the key feature common to all manufacturing organizations is that the system processes both materials and information. General manufacturing systems can be decomposed into seven levels of decision hierarchies (Rogers et al. 1992) (Figure 2). Decisions at the upper levels are made at less frequent intervals (but have implications for longer periods into the future) and are made on the basis of more abstract (and slower to change) information on the state of the system. Decisions at the lower levels are made more frequently using much more detailed information on the state of the system.

Three kinds of decisions should be made for any manufacturing company: (1) what kinds of products to make, (2) what resources will be needed to make the products, and (3) how to control the manufacturing systems. These decisions cannot be made separately. If the company wishes to make a decision at a certain level, such as at the business level, it should also get access to the information at other levels. In the whole process of decision making, the core concept is integration. This is the fundamental requirement for the research and development of computer integrated manufacturing.

2.3. CIM Definitions

There are many definitions for CIM, emphasizing different aspects of it as a philosophy, a strategic tool, a process, an organizational structure, a network of computer systems, or a stepwise integration of subsystems. These different definitions have been proposed by researchers working in different areas at different times from different viewpoints. Since the concept of CIM was put forward in 1973, it has been enriched by the contributions of many researchers and practitioners. One earlier definition of CIM is "the concept of a totally automated factory in which all manufacturing processes are integrated and controlled by a CAD/CAM system. CIM enables production planners and schedules, shopfloor foremen, and accountants to use the same database as product designers and engineers" (Kochan and Cowan 1986). This definition does not put much emphasis on the role of information.

Another definition is given by Digital Equipment Corporation (DEC): "CIM is the application of computer science technology to the enterprise of manufacturing in order to provide the right information to the right place at the right time, which enables the achievement of its product, process and business goals" (Ayres 1991). This definition points out the importance of information in manufacturing enterprise, but unfortunately it does not give much emphasis to the very important concept of integration.

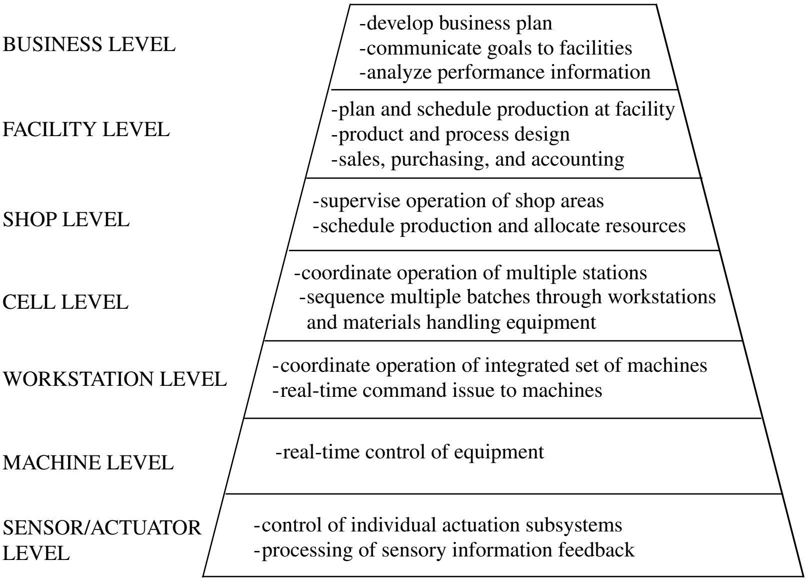

BUSINESS LEVEL
-develop business plan
-communicate goals to facilities
-analyze performance information
FACILITY LEVEL
-plan and schedule production at facility
-product and process design
-sales, purchasing, and accounting
SHOP LEVEL
-supervise operation of shop areas
-schedule production and allocate resources
CELL LEVEL
-coordinate operation of multiple stations
-sequence multiple batches through workstations and materials handling equipment
WORKSTATION LEVEL
-coordinate operation of integrated set of machines
-real-time command issue to machines
MACHINE LEVEL
-real-time control of equipment
SENSOR/ACTUATOR LEVEL
-control of individual actuation subsystems
-processing of sensory information feedback

Figure 2 A Seven-Level Manufacturing Hierarchy. (From Rogers et al. 1992)

Other definitions have pointed out that CIM is a philosophy in operating a manufacturing company. For example: "CIM is an operating philosophy aiming at greater efficiency across the whole cycle of product design, manufacturing, and marketing, thereby improving quality, productivity, and competitiveness" (Greenwood 1989).

To stress the importance of integration, the Computer and Automation Systems Association of the Society of Manufacturing Engineers gives the following definition: "CIM is the integration of the total manufacturing enterprise through the use of integrated systems and data communications coupled with new managerial philosophies that improve organizational and personnel efficiency" (Singh 1996).

CIM does not mean replacing people with machines or computers so as to create a totally automatic business and manufacturing processes. It is not necessary to build a fully automatic factory in order to implement a CIM system. It is especially unwise to put a huge investment into purchasing highly automation-flexible manufacturing systems to improve manufacturing standards if the bottleneck in the company's competitiveness is not in this area. In the current situation, the design standards for creative and customized products are more important than production ability in winning the market competition.

The importance of human factors should be emphasized. Humans play a very important role in CIM design, implementation, and operation. Although computer applications and artificial intelligence technologies have made much progress, even in the future, computers will not replace people. To stress the importance of the role of humans, the idea of human-centered CIM has been proposed.

Two views of CIM can be drawn from these definitions: the system view and the information view. The system view looks at all the activities of a company. The different functions and activities cannot be analyzed and improved separately. The company can operate in an efficient and profitable way only if these different functions and activities are running in an integrated and coordinated environment and are optimized in a global system range. The SME CIM wheel (Figure 3) provides a clear portrayal of relationships among all parts of an enterprise. It illustrates a three-layered integration structure of an enterprise.

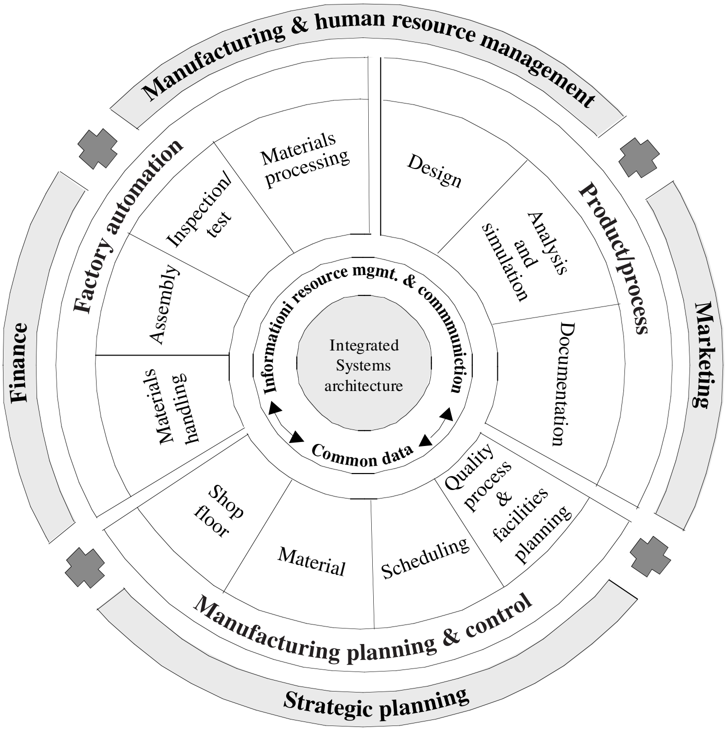

Manufacturing & human resource management
Finance
Marketing
Strategic planning
Factory automation
Product/process
Manufacturing planning & control
Materials processing
Inspection/ test
Assembly
Materials handling
Shop floor
Material
Scheduling
Quality process & facilities planning
Documentation
Analysis and simulation
Design
Information resource mgmt. & communication
Common data
Integrated Systems architecture

Figure 3 The SME CIM Wheel. (From *Systems Approach to Computer-Integrated Design and Manufacturing,* by N. Singh, copyright 1996 by John Wiley and Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)

The outer layer represents general management and human resources management. The middle layer has three process segments: product and process definition, manufacturing planning and control, and factory automation. These segments represent all the activities in the design and manufacturing phases of a product life cycle, from concept to assembly. The center of the wheel represents the third layer, which includes information resources management and the common database.

The other view of CIM is the information view. As stated in the definition given by Digital Equipment Corporation, the objective of CIM implementation is to enable the right information to be sent to the right person at the right time. The information system plays a vital role in the operation of CIM. Although many kinds of activities are involved in managing a manufacturing company, each activity has a different function in business management and production control. The associated function unit for the information system of CIM normally can be classified into three kinds of tasks: information collection, information processing, and information transfer.

Information collection is the basic function of an information system. The information collected forms the basis of decision making at different levels from business management to device control. There are many methods of information collection, depending on the information sources and technologies used. Device sensors may provide data regarding device status; barcode scanners may provide data about the production status of online products; and form scanners and database table view interfaces may provide data about order, raw material purchasing, and user requirements. Some data may also come from e-mail systems. The data collected can be stored in different data formats and different repositories.

Information processing is closely related to the business functions of a company. Business functions range from strategy planning, process planning, product design, warehouse management, and material supply to production management and control. The upper-stream process data are processed by algorithms or human intervention and the instructions produced are used for the downstream process. In data processing, different decisions will be made. The decisions can be used to optimize the production processes or satisfy user requirements such as delivery time and quality requirements.

Information transfer between different function units has three main functions: data output from application software in a certain data format to a certain kind of data repository; data format transformation, and data transfer from one application to another application within the same computer or in a network environment.

2.4. Integration: The Core of CIM

The core of CIM is usually seen to be integration. In our opinion, computer technology is the basis of CIM, manufacturing is the aim, and integration is the key technology. Why should integration be considered the core of CIM? This can be seen from different aspects. The system view of CIM was described above. *System* means the whole company, including people, business, and technology. In order to form a coordinated system, these elements must be integrated. Material flow, information flow, and capital flow must also be integrated. Although those aims seem clear, the technology for realizing this integration is far from mature.

CIMOSA (Esprit Consortium AMICE 1993) identifies enterprise integration as an ongoing process. Enterprises will evolve over time according to both internal needs and external challenges and opportunities. The level of integration should remain a managerial decision and should be open to change over a period of time. Hence, one may find a set of tightly coupled systems in one part of a CIM enterprise and in another a set of loosely coupled systems according to choices made by the enterprise. The implementation of multivendor systems in terms of both hardware and software and easy reconfiguration requires the provision of standard interfaces. To solve the many problems of the industry, integration has to proceed on more than one operational aspect. The AMICE (European Computer Integrated Manufacturing Architecture) project identifies three levels of integration covering physical systems, application and business integration (see Figure 4).

Business integration is concerned with the integration of those functions that manage, control, and monitor business processes. It provides supervisory control of the operational processes and coordinates the day-to-day execution of activities at the application level.

Application integration is concerned with the control and integration of applications. Integration at this level means providing a sufficient information technology infrastructure to permit the system wide access to all relevant information regardless of where the data reside.

Physical system integration is concerned with the interconnection of manufacturing automation and data-processing facilities to permit interchange of information between the so-called islands of automation (intersystem communications). The interconnection of physical systems was the first integration requirement to be recognized and fulfilled.

Even when business integration has been achieved at a given time, business opportunities, new technologies, and modified legislation will make integration a vision rather than an achievable goal. However, this vision will drive the management of the required changes in the enterprise operation.

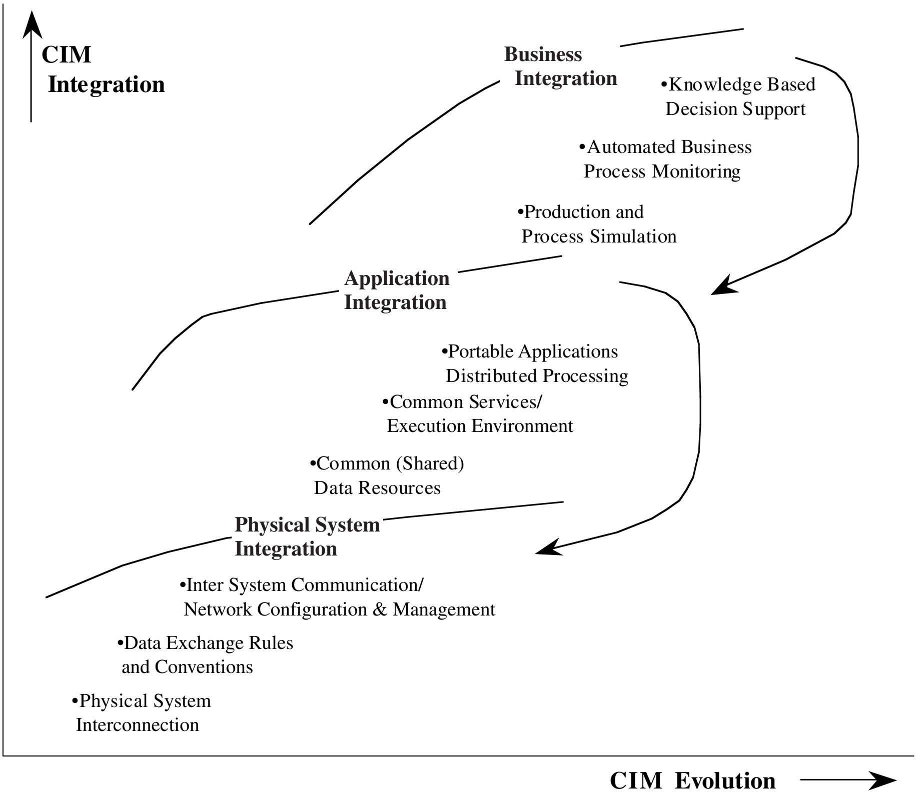

CIM Integration
Business Integration
•Knowledge Based Decision Support
•Automated Business Process Monitoring
•Production and Process Simulation
Application Integration
•Portable Applications Distributed Processing
•Common Services/ Execution Environment
•Common (Shared) Data Resources
Physical System Integration
•Inter System Communication/ Network Configuration & Management
•Data Exchange Rules and Conventions
•Physical System Interconnection
CIM Evolution

Figure 4 Three Levels of Integration. (From Esprit Consortium AMICE 1993. Reprinted by permission of Springer-Verlag.)

The classification of integration can also be given in another method, which is different from that given by CIMOSA. Regarding integration objectives and methods, integration can be classified as information integration, process integration, and enterprise-wide integration.

Information integration enables data to be shared between different applications. Transparent data access and data consistency maintenance under heterogeneous computing environments is the aim of information integration. Information integration needs the support of communication systems, data-representation standards, and data-transfer interfaces. Communication systems provide a data transfer mechanism and channel between applications located at different computer nodes. Data-representation standards serve as common structures for data used by different applications. Data-transfer interfaces are used to transfer data from one application to another. They fulfill two kinds of functions: data format transfer (from application-specified data structure to common structure and vice versa) and data transfer from application to interface module and vice versa. The traditional information-integration methods include database data integration, file integration, and compound data integration. The most efficient support tool for information integration is the integration platform (Fan and Wu 1997).

Process integration is concerned with the collaboration between different applications in order to fulfill business functions, such as product design and process control. The need to implement process integration arises from companies' pursuit of shorter product design time, higher product quality, shorter delivery time, and high business process efficiency. Business process reengineering (BPR) (Jacobson 1995) and concurrent engineering (CE) (Prasad 1996) have promoted the research and application of process integration. Business process modeling, business process simulation, and business process execution are three important research topics related to process integration.

A number of methods can be used in modeling business processes: CIMOSA business process modeling, IDEF3 (Mayer et al. 1992), Petri nets (Zhou 1995), event driven process chain (Keller 1995), and workflow (Georgakopoulos et al. 1995). The modeling objective is to define the activities within a business process and the relationships between these activities. The activity is a basic function unit within a business process. The control and data flow between these activities form the business process, which fulfils the business task of a company. Optimizing the flow path and shortening the flow time can help the company increase its working efficiency and reduce cost.

The third type of integration is called enterprise-wide integration. With the advent of agile manufacturing, virtual organization is ever more important than before. In agile manufacturing mode, a number of companies collaborate in a virtual company to gain new opportunities in the market. Enterprise-wide integration is required to enhance the exchange of information between the companies. The success of virtual organizations is predicated on the empowerment of people within the enterprise with the aid of computer technology including communication networks, database management systems, and groupware. These allow team members in the virtual organization to make more effective and faster group decisions. Such interaction lays the foundation for enterprise-wide integration, encompassing various plants and offices of an enterprise, possibly located in different cities, as well as customers and suppliers worldwide. Therefore, enterprise-wide integration is much broader than factory automation integration. It is the integration of people, technology, and the business processes throughout the enterprise.

Enterprise-wide integration is required to ensure that all the technical and administrative units can work in unison. This requires a great deal of information about a large number of activities, from product conception through manufacturing, customer delivery, and in-field support. All these life-cycle steps require a large volume of data. The transformation process from one stage to another yields volumes of new data. Furthermore, many of these design, manufacturing, distribution, and service activities responsible for generating and using volumes of data are scattered across a wide spectrum of physical locations. The information is generated by a diverse set of highly specialized software tools on heterogeneous computing hardware systems. Often, incompatible storage media with divergent data structures and formats are used for data storage. This is due to the peculiarities of the tools and systems that generate data without any regard to the needs of the tools or systems that will eventually use the data.

The main idea of enterprise-wide integration is the integration of all the processes necessary for meeting the enterprise goals. Three major tools for integration that are required for overcoming the local and structural peculiarities of an enterprise's data processing applications are network communications, database management systems, and groupware. A number of methods for enterprise-wide integration have been proposed; supply chain management, global manufacturing, and a virtual information system supporting dynamic collaboration of companies. The Web and CORBA (Otte et al. 1996) technologies are playing important roles in the realization of enterprise-wide integration.

3. CIMS STRUCTURE AND FUNCTIONS

3.1. CIMS Structure

The components of CIMS include both hardware and software. The hardware includes computer hardware, network, manufacturing devices, and peripherals. The software includes operating systems, communication software, database management systems, manufacturing planning and control software, management information software, design software, office automation software, and decision support software. These different hardware and software systems have different functions and work together to fulfill the company's business goals. To make it easier to understand, CIMS is normally decomposed into a number of subsystems interacting with each other. Unfortunately, no unique and standard decomposition method exists. Every company can define a method according to its specific situation and requirements. One decomposition method is shown in Figure 5.

From Figure 5, it can be seen that CIMS consists of four functional subsystems and two support subsystems. The four functional subsystems are management information, CAD/CAPP/CAM, manufacturing automation, and computer-aided quality management. These functional subsystems cover the business processes of a company. The two support subsystems are computer network and database management. They are the basis that allows the functional subsystems to fulfill their tasks. The arcs denote the interfaces between different subsystems. Through these interfaces, shared data are exchanged between different subsystems.

3.2. Components of CIMS

This section briefly describes the components of CIMS.

3.2.1. Management Information System

Management information system (MIS) plays an important role in the company's information system. It manages business processes and information based on market strategy, sales predictions, business decisions, order processing, material supply, finance management, inventory management, human resource management, company production plan, and so on. The aims of MIS are to shorten delivery time, reduce cost, and help the company to make rapid decision to react to market change.

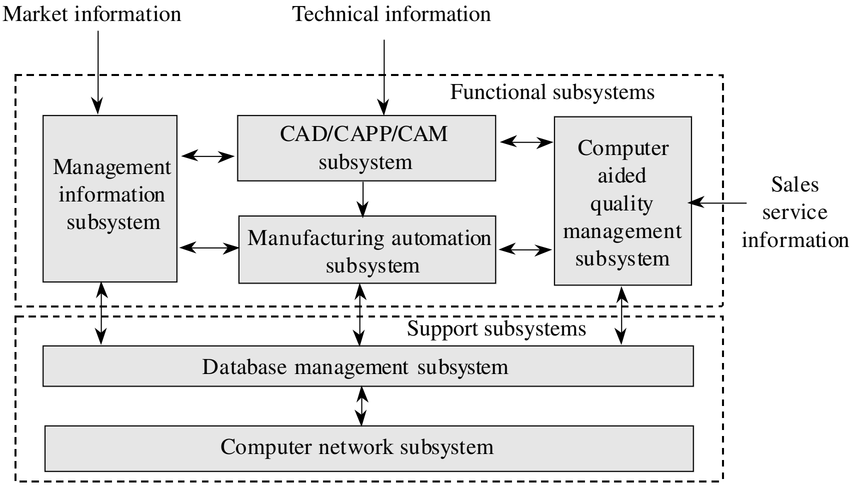

Market information
Technical information
Functional subsystems
Management information subsystem
CAD/CAPP/CAM subsystem
Manufacturing automation subsystem
Computer aided quality management subsystem
Sales service information
Support subsystems
Database management subsystem
Computer network subsystem

Figure 5 Decomposition of CIMS.

Currently, Enterprise Resource Planning (ERP) software is normally used as the key application software in MIS. Many commercial ERP software products are on the market, such as SAP R/3, developed by SAP, and BaanERP, developed by Baan.

3.2.1.1. Basic Concept of ERP In balancing manufacturing, distribution, financial, and other business functions to optimize company productivity, ERP systems are considered to be the backbone of corporate infrastructure. The ERP concept is derived from and extends the functions of MRPII (Manufacturing Resources Planning) system (Wright 1992). Besides the traditional functions of MRPII in manufacturing management, material supply management, production planning, finance, and sales management, ERP introduces new functions, such as transportation management, supply chain management, corporate strategy planning, workflow management, and electronic data exchange, into the system. The ERP system thus provides more flexibility and ability to the company in business process reengineering, integration with customers, and integration with material suppliers as well as product dispatchers.

3.2.1.2. Manufacturing Resource Planning The basis of MRPII is MRP (material requirements planning), which dates back to the 1940s. MRPII uses computer-enhanced materials ordering and inventory control methods. It enhances speed and accuracy in issuing raw materials to factory workstations. It is clear that linking materials with production demand schedules could optimize the flow of the product as it is being constructed in the factory. This could be done in such a manner that material queue times could be minimized (e.g., have the material show up only when needed), and the amount of material needed throughout the factory at any one time could be reduced ultimately. This is an optimization technique that allocates identified sets of materials (sometimes called kits) to specific jobs as they go through the manufacturing process.

Because it is possible for a computer to keep track of large numbers of kits, it is reserves or mortgages materials for specific jobs in time-order sequences. Linking these sequences with a production plan based on customer need dates allows management to release and track orders through the shop accurately. Prior to releasing orders by means of the kitting process based on the production schedule, it was necessary to obtain supplies. The supplies are based on a gross basis depending on the number of orders expected to be shipped to customers over the selected time period and by having the gross amount of inventory on hand at the start of the period to support production. Obviously, the kit will result in fewer extra materials on hand at any point in the production period. This results in large reductions in raw material and work in process and hence in lower operation costs.

Figure 6 gives a flow diagram of an MRPII system (Waldner 1992).

3.2.1.3. Just-in-Time Another method that has received much attention for production planning and control is just-in-time theory. In contrast to MRPII, which is "push" oriented, the JIT philosophy of management is "pull" oriented—that is, it calls for something to be manufactured only when there is a firm order for it. JIT is a productivity enhancer based on the simple proposition that all waste in the manufacturing process must be eliminated.

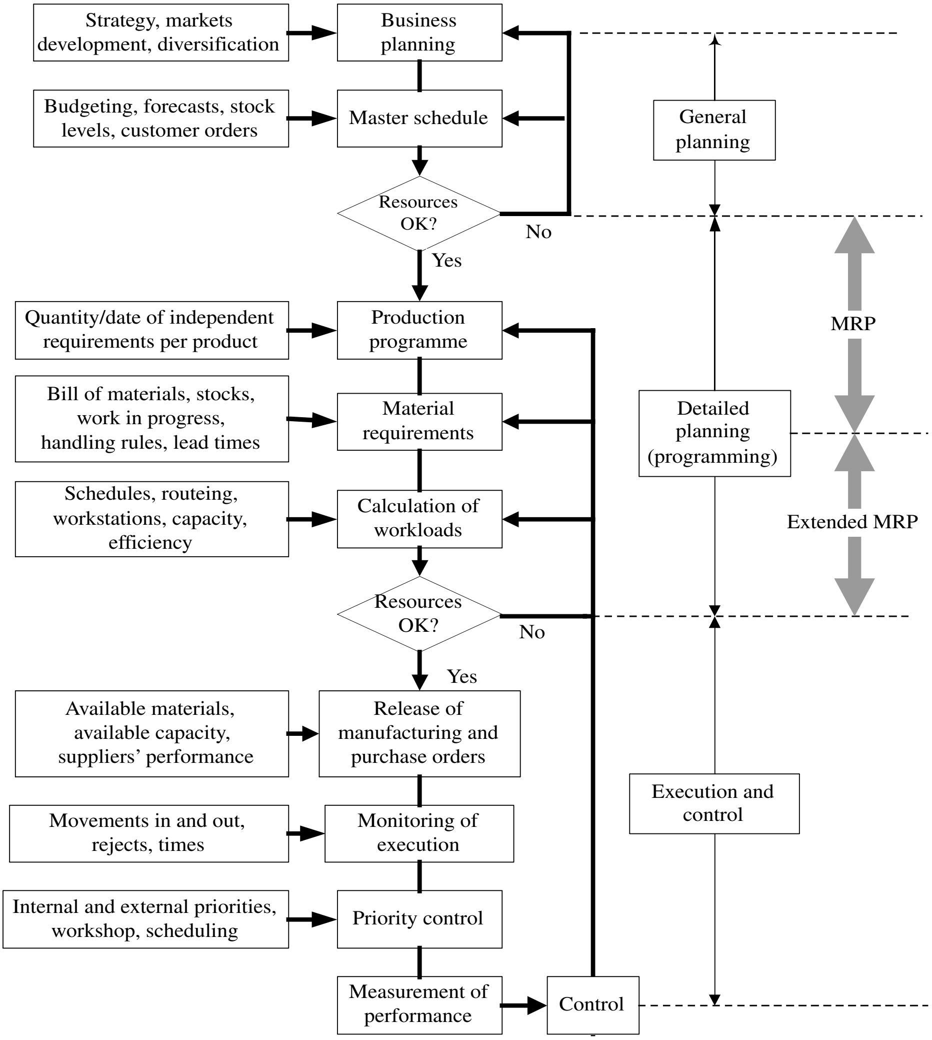

Strategy, markets development, diversification
Business planning
Budgeting, forecasts, stock levels, customer orders
Master schedule
Resources OK?
No
Yes
General planning
Quantity/date of independent requirements per product
Production programme
MRP
Bill of materials, stocks, work in progress, handling rules, lead times
Material requirements
Detailed planning (programming)
Schedules, routeing, workstations, capacity, efficiency
Calculation of workloads
Extended MRP
Resources OK?
No
Yes
Available materials, available capacity, suppliers' performance
Release of manufacturing and purchase orders
Execution and control
Movements in and out, rejects, times
Monitoring of execution
Internal and external priorities, workshop, scheduling
Priority control
Measurement of performance
Control

Figure 6 Flow Diagram of MRPII System. (From J. B. Waldner, *Principles of Computer-Integrated Manufacturing,* copyright 1992 John Wiley & Sons Limited. Reproduced by permission)

JIT theory states that wastes can only begin to be eliminated if the push production control system is replaced with a pull production control system. Inventory levels contain a very large volume of waste. Therefore, a way must be found to minimize inventory levels. If this is done without the analytical capability of the computer, it would be logical to conceive a system that would not let material move or be used until it is necessary. This is what Toyota did. They instituted a backward scheduling technique that started with the desired ship date. They had to know when the product needed to be at final assembly and before that when it needed to be at the subassembly levels and so forth, back through component part manufacturing. Ultimately, this means determining precisely when the raw materials should show up at the receiving dock. This in itself is not unusual or unique.

Although JIT proposes ways to reduce a great deal of waste, it cannot be implemented without the help of CIM and MRPII systems. For example, the means for producing products only at the rate at which the customer wants them can best be realized using the feedback control system production schedule of MRPII. By using the MRPII system, we can monitor the progress of all workstations carrying out the dictates of the strategic plan and thus speed up or slow down the preceding operation to optimize the usage of materials and labor. Koenig (1990) explains in detail the relationship of JIT with the MRPII and CIM systems.

Because JIT and MRPII have their advantages as well as limitations in applications, the combination of JIT and MRPII systems in the common framework of CIM may produce excellent results in production scheduling and control.

3.2.2. CAD/CAPP/CAM System

CAD/CAPP/CAM stands for computer-aided design/computer-aided process planning/computer-aided manufacturing. The system is sometimes called the design automation system, meaning that CAD/CAPP/CAM is used to promote the design automation standard and provide the means to design high-quality products faster.

3.2.2.1. Computer-Aided Design CAD is a process that uses computers to assist in the creation, modification, analysis, or optimization of a product design. It involves the integration of computers into design activities by providing a close coupling between the designer and the computer. Typical design activities involving a CAD system are preliminary design, drafting, modeling, and simulation. Such activities may be viewed as CAD application modules interfaced into a controlled network operation under the supervision of a computer.

A CAD system consists of three basic components: hardware, which includes computer and input–output devices, application software, and the operating system software (Figure 7). The operating system software acts as the interface between the hardware and the application software system.

The CAD system function can be grouped into three categories: geometric modeling, engineering analysis, and automated drafting.

Geometric modeling constructs the graphic images of a part using basic geometric elements such as points, lines, and circles under the support of CAD software. Wire frame is one of the first geometric modeling methods. It uses points, curves, and other basic elements to define objects. Then the surface modeling, solid modeling, and parametric modeling methods are presented in the area of geometric modeling area. Saxena and Irani (1994) present a detailed discussion of the development of geometric modeling methods.

Engineering design completes the analysis and evaluation of product design. A number of computer-based techniques are used to calculate the product's operational, functional, and manufacturing parameters, including finite-element analysis, heat-transfer analysis, static and dynamic analysis, motion analysis, and tolerance analysis. Finite-element analysis is the most important method. It divides an object into a number of small building blocks, called finite elements. Finite-element analysis will fulfill the task of carrying out the functional performance analysis of an object. Various methods and packages have been developed to analyze static and dynamic performance of the product design. The objectives and methods can be found in any comprehensive book discussion of CAD techniques. After the analysis, the product design will be optimized according to the analysis results.

The last function of the CAD system is automated drafting. The automated drafting function includes 2D and 3D product design drafting, converting a 3D entity model into a 2D representation.

3.2.2.2. Computer-Aided Process Planning CAPP is responsible for detailed plans for the production of a part or an assembly. It acts as a bridge between design and manufacturing by translating design specifications into manufacturing process details. This operation includes a sequence of steps to be executed according to the instructions in each step and is consistent with the controls indicated in the instructions. Closely related to the process-planning function are the functions that determine the cutting conditions and set the time standards. The foundation of CAPP is group technology (GT),

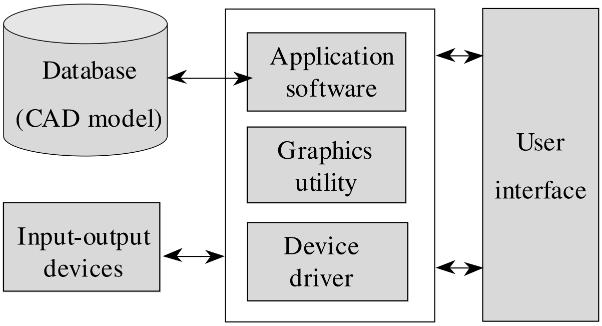


Figure 7 Basic Components of CAD.

which is the means of coding parts on the basis of similarities in their design and manufacturing attributes. A well-developed CAPP system can reduce clerical work in manufacturing engineering and provide assistance in production.

One of the first tasks of the CAPP system is to complete the selection of the raw workpiece. According to the functional requirements of the designed part, it determines the attributes of the raw workpiece, such as shape, size (dimension and weight), and materials. Other jobs for the CAPP system are determining manufacturing operations and their sequences, selecting machine tools, and selecting tools, fixtures, and inspection equipment. Determination of manufacturing conditions and manufacturing times are also part of the work of CAPP. These conditions will be used in optimizing manufacturing cost.

The CAPP system consists of computer programs that allow planning personnel interactively to create, store, edit, and print fabrication and assembly planning instructions. Such a system offers the potential for reducing the routine clerical work of manufacturing engineers. Figure 8 presents the classification of various CAPP systems.

3.2.2.3. Computer-Aided Manufacturing In this section, *computer-aided manufacturing* (CAM) refers to a very restricted area that does not include general production control functions. The production control functions will be introduced in the manufacturing automation subsystem (MAS) section. Here, CAM includes preparing data for MAS, including producing NC code for NC machines, generating tool position, planning tool motion route, and simulating tool movement. Automatic NC code generation is very important for increasing work efficiency. Before the NC code for NC machine centers can be generated, a number of parameters regarding machine tool specification, performance, computer numerical control system behavior, and coding format should be determined first. The manufacturing method and operations will be selected according to these parameters, geometric dimensions, solid forms, and designed part specifications. The CAM system will calculate the tool position data. Then the data regarding the part dimension, the tool motion track, cutting parameters, and numerical control instructions are generated in a program file. This file, called the NC program, is used by the machine tool to process part automatically.

3.2.2.4. CAD/CAPP/CAM Integration Besides the utilization of CAD, CAPP, and CAM technology alone, the integration of CAD, CAPP, and CAM is an important way to enhance the company's product design standards. Three methods can be used in the integration of CAD/CAPP/CAM: exchange product data through specific defined data format; exchange product data through standard data format, such as STEP, IGES, and DXF; and define a unified product data model to exchange product information.

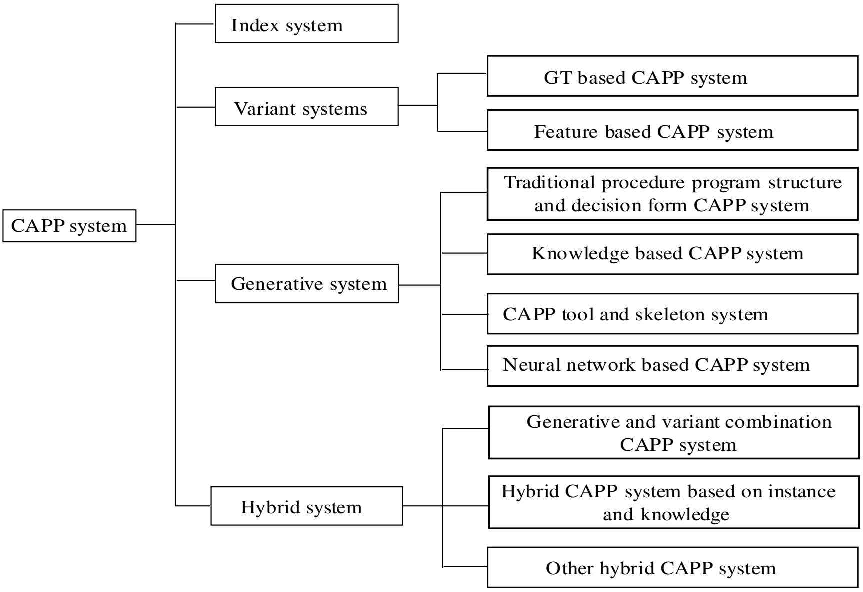

CAPP system
Index system
Variant systems
GT based CAPP system
Feature based CAPP system
Generative system
Traditional procedure program structure and decision form CAPP system
Knowledge based CAPP system
CAPP tool and skeleton system
Neural network based CAPP system
Hybrid system
Generative and variant combination CAPP system
Hybrid CAPP system based on instance and knowledge
Other hybrid CAPP system

Figure 8 Classification of CAPP System.

Figure 9 is a STEP-based CAD/CAPP/CAM integration system developed at the State CIMS Engineering Research Center of China (located at Tsinghua University, Beijing). It was developed as a part of the CIMS application integration platform (Fan and Wu 1997) for manufacturing enterprises. This system focuses on part-level CAD/CAPP/CAM integration. XPRESS language and the STEP development tool ST-developer are used to define and develop the integration interfaces. Different kinds of CAD, CAPP, and CAM systems can be integrated using the interfaces provided.

3.2.3. Manufacturing Automation System

Manufacturing automation system is a value-added system. The material flow and information flow come together in MAS. For a discrete manufacturing company, MAS consists of a number of manufacturing machines, transportation systems, high-bay stores, control devices, and computers, as well as MAS software. The whole system is controlled and monitored by the MAS software system. For the process industry, MAS consists of a number of devices controlled by DCS, the monitor system, and the control software system. The objectives of MAS are to increase productivity, reduce cost, reduce work-in-progress, improve product quality, and reduce production time.

MAS can be described from three different aspects: structural description, function description, and process description. *Structural description* defines the hardware and the software system associated with the production processes. *Function description* defines the MAS using a number of functions that combine to finish the task of transforming raw material into products. The input–output mapping presented by every function is associated with a production activity of the MAS. *Process description* defines the MAS using a series of processes covering every activity in the manufacturing process.

In the research field of MAS, a very important topic is the study of control methods for manufacturing devices, from NC machines to automatic guided vehicles. But the focus of this chapter is on studying MAS from the CIM system point of view. We will describe the shop-floor control and management system functions and components below.

The shop-floor control and management system is a computer software system that is used to manage and control the operations of MAS. It is generally composed of several modules as shown in Figure 10. It receives a production plan from the MRPII (ERP) system weekly. It optimizes the sequence of jobs using production planning and scheduling algorithms, assigns jobs to specific devices and manufacturing groups, controls the operation of the material-handling system, and monitors the operations of the manufacturing process.

Task planning decomposes the order plan from MRPII system into daily tasks. It assigns job to specific work groups and a set of machines according to needed operations. Group technology and optimization technology are used to smooth the production process, better utilize the resources, reduce production setup time, and balance the load for manufacturing devices. Hence, good task planning is the basis for improving productivity and reducing cost of production.

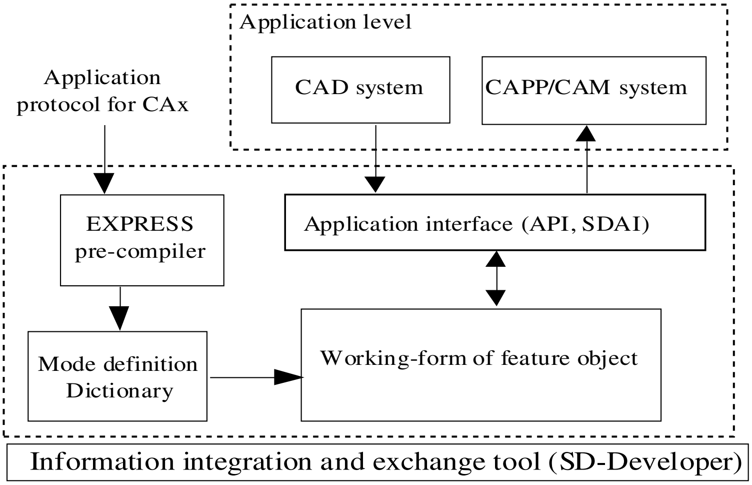

Application level
CAD system
CAPP/CAM system
Application protocol for CAx
EXPRESS pre-compiler
Application interface (API, SDAI)
Mode definition Dictionary
Working-form of feature object
Information integration and exchange tool (SD-Developer)

Figure 9 CAD/CAPP/CAM Integration System Structure.

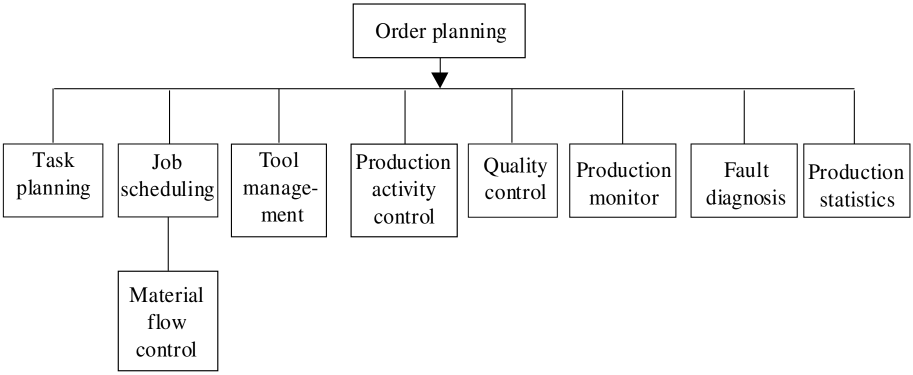


Figure 10 Function Modules of Shop-Floor Control and Management System.

Job scheduling is used to determine the entry time and sequence for different production jobs. It consists of three main functions: static scheduling, dynamic scheduling, and real-time resource scheduling. Material-flow control is one of the tasks for real-time resource scheduling. Static scheduling is a off-line scheduling method. It determines operation sequences before the production starts. The aim of static scheduling is to reduce the makespan (the time duration between when the first task enters the system and when the last task leaves the system). Operations research is an important method for generating static scheduling. Because errors and uncertainties may be caused by machine breakdown, task priorities change and dynamic scheduling is needed for rescheduling the operation sequences and production routes. It is the best method for increasing the flexibility of production system. Heuristic rules are normally used in generating dynamic scheduling. Job scheduling aims to optimize the operation of production system and increase the system flexibility.

Production activity control is used to control the operations of tasks, material flow, and manufacturing resources. Real-time data collecting, processing, and decision making are important tasks of production activity control, which aims to regulate and smooth the production processes even when errors and disturbances occur.

Tool management is also a very important task for the shop-floor control and management system. In a manufacturing system, a large number of tools are needed and the supply of necessary tools on time is vital for improving productivity. Tool quality is important to product quality. The parameters of every tool should be maintained in a correct and real-time fashion because these parameters will be used by machine centers in controlling manufacturing processes.

Quality control, production monitoring, fault diagnosis, and production statistics are important supplementary functions for the shop-floor control and management system to be operated efficiently and effectively.

3.2.4. Computer-Aided Quality-Management System

Since the 1970s, quality has become an extremely important factor for a company to win market competition. Customers always want higher product quality for their investment. The computer-aided quality-management system of CIMS is a system used to guarantee the product quality. It covers a wide range, from product design to material supply to production quality control. The International Standards Organization (ISO) has established a series of quality assurance standards, such as ISOs 9000, 9001, 9002, 9003, and 9004. The computer-aided quality-management system has also been called the integrated quality system.

The computer-aided quality system consists of four components: quality planning, inspection and quality data collection, quality assessment and control, and integrated quality management.

The quality-planning system consists of two kinds of functions: computer-aided product-quality planning and inspection-plan generating. According to the historical quality situation and production-technology status, computer-aided product-quality planning first determines the quality aims and assigns responsibility and resources to every step. Then it determines the associated procedure, method, instruction file, and quality-inspection method and generates a quality handbook. Computer-aided inspection planning determines inspection procedures and standards according to the quality aims, product model, and inspection devices. It also generates automatic inspection programs for automatic inspection devices, such as a 3D measuring machine.

Guided by the quality plan, the computer-aided quality inspection and quality data collection receive quality data during different phases. The phases include purchased-material and part-quality inspection, part-production-quality data collection, and final-assembly quality inspection. The methods and techniques used in quality inspection and data collection are discussed in special books on quality control (Taguchi et al. 1990).

Quality assessment and control fulfills the tasks of manufacturing process quality assessment and control and supply part and supplier quality assessment and control. Integrated quality management includes the functions of quality cost analysis and control, inspection device management, quality index statistics and analysis, quality decision making, tool and fixture management, quality personnel management, and feedback information storage on quality problems, and quality problems backtrack into manufacturing steps.

Quality cost plays an important role in a company's operation. The quality cost analysis needs to determine the cost bearer and the most important cost constituent part to generate a quality cost plan and calculate real cost. It also optimizes the cost in the effort to solve quality problems. Figure 11 presents a quality cost analysis flowchart.

3.2.5. *Computer Network and Database Management Systems*

Computer network and database management systems are supporting systems for CIMS. The computer network consists of a number of computers (called nodes in the network) and network devices, as well as network software. It is used to connect different computers together to enable the communication of data between different computers. The computer network can be classified as a local area network (LAN) or a wide area network (WAN). *LAN* normally means a restricted area network, such as in a building, factory, or campus. *WAN* means a much wider area network, across a city or internationally. Network technology is developing rapidly. The Internet concept has changed manu-

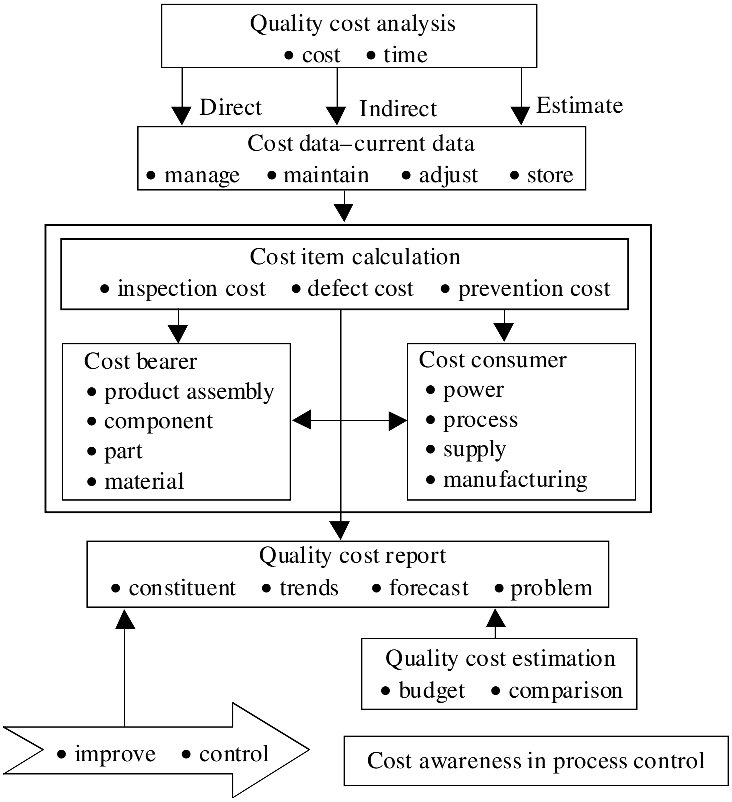

Quality cost analysis
• cost • time
Direct
Indirect
Estimate
Cost data–current data
• manage • maintain • adjust • store
Cost item calculation
• inspection cost • defect cost • prevention cost
Cost bearer
• product assembly
• component
• part
• material
Cost consumer
• power
• process
• supply
• manufacturing
Quality cost report
• constituent • trends • forecast • problem
Quality cost estimation
• budget • comparison
• improve • control
Cost awareness in process control

Figure 11 Quality Cost Analysis Flowchart.

facturing companies' operation method greatly. Global manufacturing, agile manufacturing, and network-based manufacturing paradigms have seen rapid development. A computer network is the infrastructure for these new manufacturing paradigms to be realized in a cost-effective way.

The database management system provides basic support for the data storage and information sharing of manufacturing companies. Currently, relational database management systems are the principal databases used. Information integration of a company is concerned with integration data sources in different locations and with different kinds of database management systems. The heterogeneous properties of computer operating systems and database management systems are the major difficulties in information integration. Advanced software techniques have been developed to cope with the heterogeneity problem. Techniques include CORBA, as well as OLE/DCOM, developed by Microsoft, and the Java language, developed by Sun.

Hundreds of books discussing computer network and database techniques can be found in almost any bookstore.

4. FLEXIBLE MANUFACTURING SYSTEMS

Flexible Manufacturing Systems (FMS) is a manufacturing system with a high degree of flexibility. It was developed due to the need to increase productivity, improve product quality, and reduce cost for product production under the constraints of various uncertainties or disturbances both internal and external to the manufacturing system.

4.1. Flexibility and Components of FMS

4.1.1. Flexibility of Manufacturing System

A number of papers have studied different aspects of FMS. Gupta and Goyal (1989) provide a comprehensive review of the literature on flexibility. Flexibility can be defined as a collection of properties of a manufacturing system that supports changes in production activities or capabilities (Carter 1986).

In a manufacturing system, various types of flexibility are needed to fulfill different requirements. The types most discussed are machine flexibility, routing flexibility, process flexibility, product flexibility, production flexibility, and expansion flexibility. Machine flexibility is the capability of a machine to perform a variety of operations on a variety of part types and sizes. Machine flexibility can reduce the changeover frequency, setup time, and tool-changing time, hence reducing the lead time and making small-lot-size production more economic. Machine flexibility is the basis for routing and process flexibility.

Routing flexibility provides the chance for a part to be manufactured or assembled along alternative routes. Routing flexibility is required to manage shop-floor uncertainties caused by such problems as machine breakdown, tool error, and controller failure. It can also be used to tackle the problems caused by external events such as change of product mix or product due date and emergency product introduction. These changes alter machine workloads and cause bottlenecks. The use of alternative routing helps to solve these problems and finally increase productivity.

Process flexibility, also called mix flexibility, is the ability to absorb changes in the product mix by performing similar operations or producing similar produces or parts on multipurpose, adaptable CNC machining centers.

Product flexibility, also known as mix-change flexibility, is the ability to change over to a new set of products economically and quickly in response to markets or engineering changes or even to operate on a market-to-order basis. In the current global market, high product flexibility is a very important factor for a company to compete.

Expansion flexibility is the ability to change a manufacturing system with a view to accommodating a changed product envelope. It has become more important in the current agile manufacturing era. Improving expansion flexibility can significantly reduce system expansion or change cost, shorten system reconfiguration time, and hence shorten the delivery time for new products.

4.1.2. FMS Definition and Components

An FMS is an automated, mid-volume, mid-variety, central computer-controlled manufacturing system. It can be used to produce a variety of products with virtually no time lost for changeover from one product to the next. Sometimes FMS can be defined as "a set of machines in which parts are automatically transported under computer control from one machine to another for processing" (Jha 1991).

A more formal definition of FMS is that it consists of a group of programmable production machines integrated with automated material-handling equipment and under the direction of a central controller to produce a variety of parts at nonuniform production rates, batch sizes, and quantities (Jha 1991).

From this definition, it can be seen that an FMS is composed of automated machines, material-handling systems, and control systems. In general, the components of an FMS can be classified as follows:

1. *Automated manufacturing devices* include machining centers with automatic tool interchange ability, measuring machines, and machines for washing parts. They can perform multiple functions according to the NC instructions and thus fulfill the parts-fabrication task with great flexibility. In an FMS, the number of automated machining centers is normally greater than or at least equal to two.
2. *Automated material-handling systems* include load/unload stations, high-bay storage, buffers, robots, and material-transfer devices. The material-transfer devices can be automatic guided vehicles, transfer lines, robots, or a combination of these devices. Automated material-handling systems are used to prepare, store, and transfer materials (raw materials, unfinished parts, and finished parts) between different machining centers, load/unload stations, buffers, and high-bay storage.
3. *Automated tool systems* are composed of tool setup devices, central tool storage, tool-management systems, and tool-transfer systems. All are used to prepare tools for the machining centers as well as transfer tools between machining centers and the central tool storage.
4. *Computer control systems* are composed of computers and control software. The control software fulfills the functions of task planning, job scheduling, job monitoring, and machine controlling of the FMS.

Figure 12 shows the FMS layout at the State CIMS Engineering Research Center (CIMS-ERC) of China. (HMC stands for horizontal machining center and VMC stands for vertical machining center.)

Another example of FMS is shown in Figure 13, from Kingdream. This system produces oil well drill bits, mining bits, hammer drills, high-pressure drills, and so on.

4.2. General FMS Considerations

Although FMS was originally developed for metal-cutting applications, its principles are more widely applicable. It now covers a wide spectrum of manufacturing activities, such as machining, sheet metal working, welding, fabricating, and assembly.

The research areas involved in the design, implementation, and operation of an FMS are very broad. Much research has been conducted and extensive results obtained. In this section, we present the research topics, problems to be solved, and methods that can be used in solving the problems.

4.2.1. FMS Design

FMS is a capital investment-intensive and complex system. For the best economic benefits, an FMS should be carefully designed. The design decisions to be made regarding FMS implementation cover

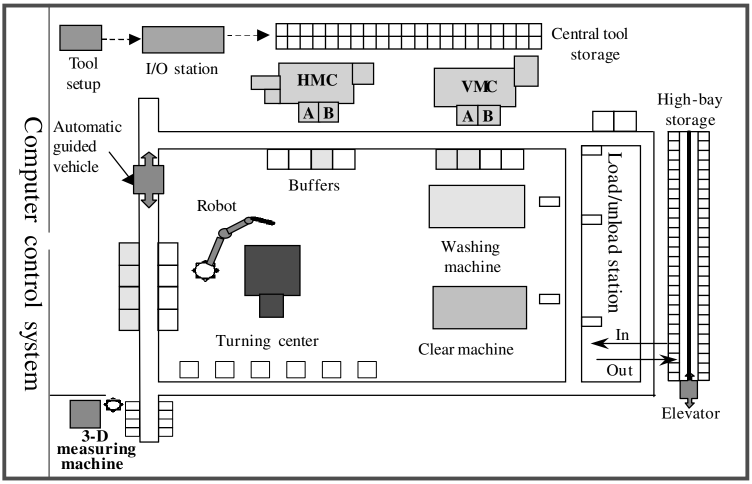


Figure 12 FMS Layout at State CIMS-ERC of China.

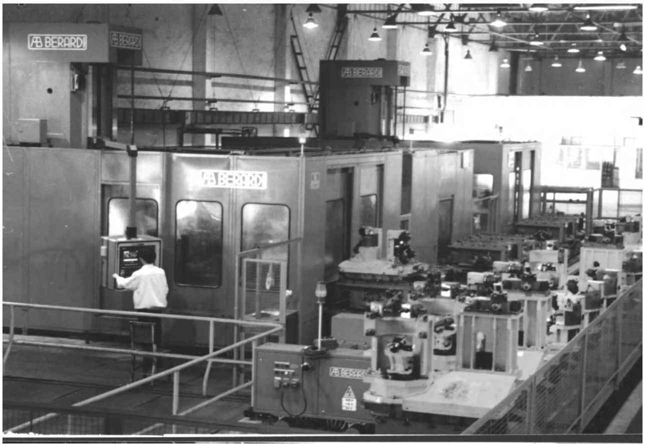


Figure 13 FMS for Oil Well Drill Production.

system configuration and layout, manufacturing devices, material-handling systems, central tool storage, buffers, and high-bay storage.

Before these decisions can be made, the part types to be made, the processes needed to make them, and the possible numbers of processing parts (workload) should first be determined. Based on these basic requirements, the number of machines and their abilities, tools, buffers, and storage system can be roughly determined. A rough system layout and material-handling system can be designed. The designed FMS can be simulated using an FMS simulation tool to test its ability to fulfill the requirements.

The design of an FMS is a system approach. Besides the above-mentioned basic requirements for part manufacturing, many other factors should be considered in designing an FMS. An economic assessment should always be done for every FMS plan obtained. System reliability, productivity, and performance evaluation should also be done for every FMS plan. The design of FMS is an iterative process that requires many experts from different disciplines to work together. Many alternative plans are compared and modified before an optimized plan is decided upon.

Talavage and Hannam (1988) summarize the work of other researchers in FMS design methodology and present a five-step approach to FMS design:

1. Development of goals
2. Establishment of criteria on which goal achievement can be judged
3. Development of alternative candidate solutions
4. Ranking of alternatives by applying the criteria to the alternate solutions
5. Iteration of the above four steps to obtain a deeper analysis of alternate solutions and to converge on an acceptable solution

Other considerations regarding FMS design can be found in Tetzlaff (1990).

4.2.2. FMS Planning, Scheduling, and Control

Planning, scheduling, and control are important and difficult problems in FMS operations. A good planning and scheduling system will improve FMS operation efficiency and yield economic benefits. Extensive research and development of FMS planning and scheduling has been done. The general optimization indexes are:

1. Maximizing the productivity at certain period of time
2. Minimizing the makespan for a group of parts
3. Minimizing the cost for parts manufacturing
4. Maximizing the utility of key manufacturing devices
5. Minimizing the work in progress
6. Minimizing the production time for certain parts
7. Satisfying the due dates of parts

Figure 14 presents a function model for FMS planning, scheduling, and resource management.

The resource management and real-time control functions of FMS are closely related to the dynamic scheduling system. The resource-management system should be activated by a dynamic scheduling system to allocate resources to production process to achieve real-time control for FMS. The resources to be controlled involve tools, automatic guided vehicles, pallets and fixtures, NC files, and human resources.

4.2.2.1. Planning Planning seeks to find the best production plan for the parts entered into the FMS. Its aim is to make an optimized shift production plan according to the shop-order and part-due dates. The FMS planning system receives the shop-order plan in the weekly time scale from the MRPII system. According to the product due dates, it analyzes the shop order and generates a daily or shift production plan. Group technology is used for grouping parts into families of parts. The capacity requirement is calculated for every shift plan generated. Capacity balance and adjustment work should be carried out if the required capacity is higher than that provided by machines.

After feasibility analysis, capacity balancing, and optimization, a shift plan is generated. The shift plan gives detailed information for the following questions:

1. What kind of parts will be machined?
2. In what sequence will the parts enter the FMS?
3. What operations are needed to process the parts? What is the operation sequence?
4. What are the start time and complete time for processed parts?
5. What materials are needed? In what time?
6. What kinds of tool are needed?

4.2.2.2. Static Scheduling Static scheduling is the refinement of the shift production plan. It seeks to optimize machine utility and reduce system setup time. Three functions are performed by a static scheduling system: part grouping, workload allocating and balancing, and part static sequencing. Because all these functions are performed before production starts, static scheduling is also called off-line sequencing.

A number of factors affecting production sequence should be taken into account for static scheduling, such as the part process property, FMS structure, and optimization index. The part process property determines what kind of production method should be used. Flow-shop, flexible-flow-line, and job-shop are three major strategies for producing parts. Different methods can be used to generate static scheduling for the different production strategies.

The second factor affecting static scheduling is FMS structure. The main structural properties are whether a central tool system, a fixture system, or bottleneck devices are present. The third factor is the optimization index chosen. The general optimization index is a combination of several optimization indexes, that is, the FMS static scheduling is a multiobjective optimization process.

The following parameters have an important effect on implementing optimal static scheduling.

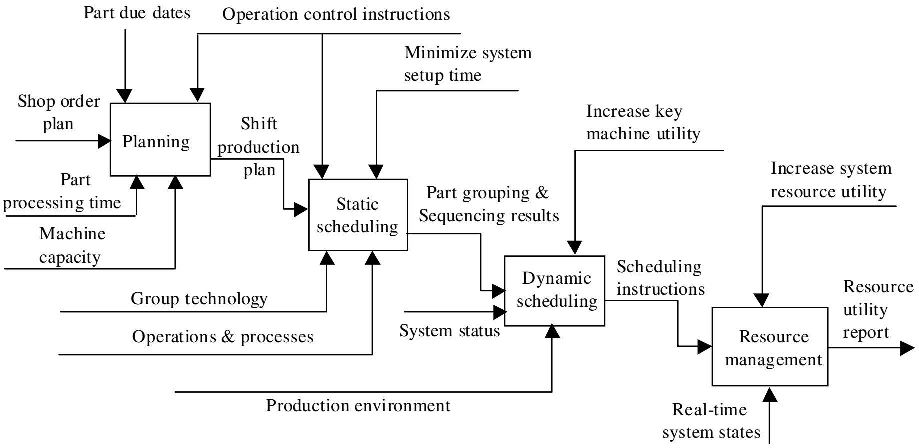

Part due dates
Operation control instructions
Minimize system setup time
Shop order plan
Planning
Shift production plan
Increase key machine utility
Part processing time
Static scheduling
Part grouping & Sequencing results
Increase system resource utility
Machine capacity
Dynamic scheduling
Scheduling instructions
Resource utility report
Group technology
Resource management
Operations & processes
System status
Production environment
Real-time system states

Figure 14 Function Model of FMS Planning and Scheduling.

1. *Time distribution,* such as time distributions for part arrival, tool setup, part fixture, part transfer, machine failure, and delivery time
2. *Shop conditions,* such as device type, transfer system, storage method, shop layout, and device condition
3. *Shop control conventions,* such as priority rule, operation method, hybrid processing route, task decomposition, performance-evaluation method, and workload
4. *Alternate processing route,* such as alternate processing device, alternate processing routing, and alternate processing sequence

A great deal of literature about static scheduling algorithms and systems can be found in the academic journals on FMS, operations research, and manufacturing technology and IEEE magazines on systems, robotics, and automatic control.

4.2.2.3. Dynamic Scheduling Dynamic scheduling is used to control the operations of FMS according to the real-time status of the AFMS. It is a real-time (online) system that focuses on solving uncertainty problems such as device failures, bottlenecks on certain machines, workload unbalance, and resource-allocation conflict. These problems are not anticipated by off-line static scheduling. They can only be solved using real-time dynamic scheduling or rescheduling.

Three strategies can be used to complete the rescheduling functions. The first is periodical scheduling. A certain time interval must be set as a production cycle. A periodical scheduling system calculates a period operation sequence before the next period starts. The sequence is the job list execution instructions followed by the FMS. The second strategy is continuous scheduling, which monitors the FMS and executes scheduling whenever an event (such as the arrival of a new part or a machine completing the production of a part) occurs and the system states has been changed. Since the calculation of work content is effective for rescheduling FMS operations for every event (so as to get optimal scheduling at every point), the third strategy, hybrid scheduling, is frequently used. The hybrid strategy combines periodical and continuous scheduling so that only when an unexpected event occurs is the continuous scheduling algorithm used. Otherwise, periodical scheduling is executed at certain intervals.

For a dynamic manufacturing environment with possible disturbances both internal and external to the FMS, dynamic scheduling seeks to optimize the sequencing for the queue before the device is manufactured. Because the dynamic scheduling of an FMS is an NP-hard problem, it is impossible to find the optimal solution in a short time, especially for continuous scheduling with a very high speed requirement. Normally a suboptimal solution is used in real-time FMS operations. A number of heuristic rules are frequently used for finding the suboptimal solutions in dynamic scheduling. The heuristic rules that are frequently used are:

1. *RANDOM:* assigns a random priority to every part entering the queue and selects a part with smallest priority to be processed
2. *FIFO* (*LIFO*)*:* first-in-first-out (last-in-first-out)
3. *SPT* (*LPT*)*:* selects the part that has the smallest (largest) current operation processing time to be processed
4. *FOPNR* (*MOPNR*)*:* selects the part that has the fewest (most) remaining operations to be processed
5. *LWKR* (*MWKR*)*:* selects the part that has the smallest (largest) remaining processing time to be processed
6. *DDATE:* selects the part that has the earliest due date to be processed
7. *SLACK:* selects the part that has the smallest slack time (due date minus remaining processing time) to be processed

In most cases, several rules will be used in a dynamic scheduling system to reach the satisfied sequencing solution. Besides rule-based scheduling, simulation-based and knowledge-based scheduling systems are also widely used.

4.2.3. FMS Modeling and Simulation

Modeling and simulation are important topics for both design and operation of FMS. FMS modeling is the basis for simulation, analysis, planning, and scheduling. Because FMS is a typical discrete event dynamic system (DEDS), a number of methods for DEDS modeling and analysis can be used to model an FMS, such as Petri nets, network of queues (Agrawal 1985), and activity cycle diagram (Carrie 1988). This section briefly introduces Petri nets, their application in FMS modeling, and the FMS simulation method.

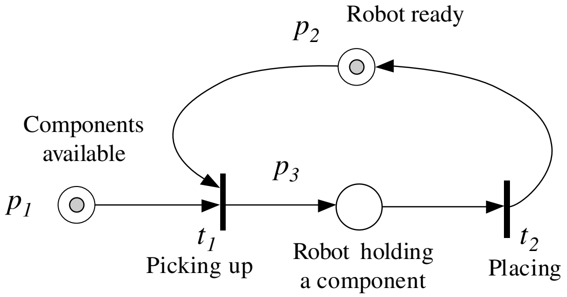


Figure 15 A Simple Petri Net Example. (From M. C. Zhou, Ed., *Petri Nets in Flexible and Agile Automation,* Figure 1, copyright 1995, with kind permission from Kluwer Academic Publishers)

4.2.3.1. Petri Nets and Their Application in FMS Modeling A Petri net (PN) may be identified as a particular kind of bipartite directed graphs populated by three types of objects. These objects are places, transitions, and directed arcs connecting places to transitions and transitions to places. Pictorially, places are depicted by circles, transitions by bars or boxes. A place is an input place to a transition if a directed arc exists connecting this place to the transition. A place is an output place of a transition if a directed arc exists connecting the transition to the place. Figure 15 represents a simple PN. Where places p_1 and p_2 are input places to transition t_1, place p_3 is the output place of t_1.

Formally, a PN can be defined as a five-tuple PN = (P, T, I, O, m_0), where

1. $P = \{p_1, p_2, \ldots, p_n\}$ is a finite set of places
2. $T = \{t_1, t_2, \ldots, t_m\}$ is a finite set of transitions, $P \cup T \neq \phi$, $P \cap T = \phi$.
3. $I:(P \times T) \mapsto N$ is an input function that defines directed arcs from places to transitions, where N is a set of nonnegative integers.
4. $O:(P \times T) \mapsto N$ is an output function that defines directed arcs from transitions to places.
5. $m_0:P \mapsto N$ is the initial marking.

The state of the modeled system is represented by the tokens (small dots within the places) in every place. For example, in Figure 15, a small dot in place p_1 means components available. The change of the states represents the system evolution. State changing is brought by firing a transition. The result of firing a transition is that for every place connected with the transition, after the firing of the transition, a token will be removed from its input place and a token will be added to its output place. In the example of Figure 15, the firing of transition t_1 will cause the tokens in places p_1, p_2 to disappear and a token to be added to place p_3.

Due to the advantages of its formal theory background, natural link with DEDS, and mature simulation tool, PN is well suited to FMS modeling. Figure 16 shows a two-machine production line to demonstrate the modeling of FMS using PN.

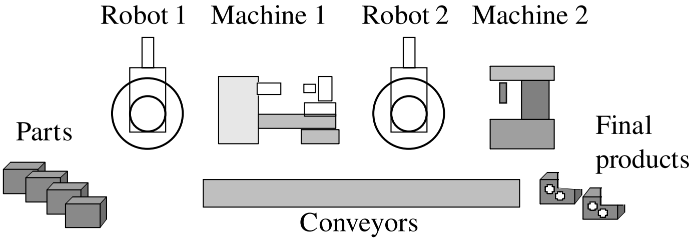


Figure 16 A Two-Machine Production Line.

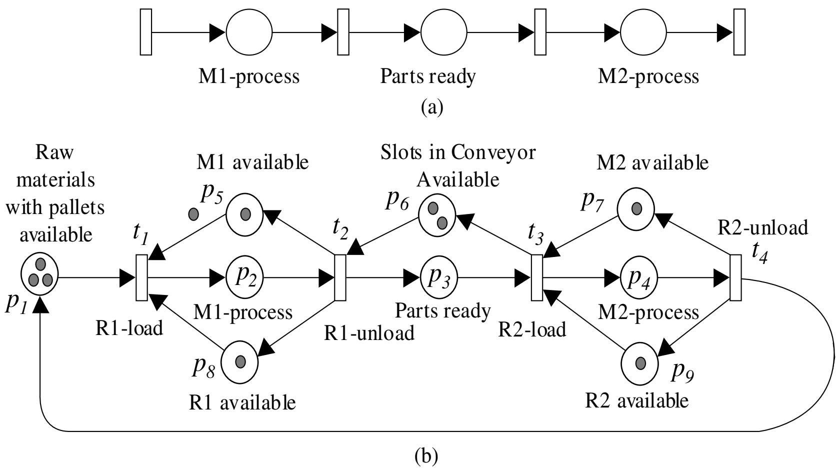
The production line consists of two machines (M1 and M2), two robots (R1 and R2), and two conveyors. Each machine is serviced by a dedicated robot that performs the load/unload task. One conveyor is used to transport workpieces, a maximum two at one time. The other conveyor is used to transport empty pallets. Three pallets are available in the system. Each workpiece is machined on M1 and then on M2. The machining time is 10 time units on M1 and 16 time units on M2. The load and unload tasks takes 1 time unit.

As with modeling general FMS or other systems, the modeling of this system using PN takes several steps:

1. Major activities are identified. In this example, they are R1 loading, M1 processing, R1 unloading, R2 loading, M2 processing, and R2 unloading. The resources are raw materials with pallets, conveyors, M1, M2, R1, R2.
2. The relationships between the four major activities form a sequential order.
3. A partial PN model is defined to describe the four major activities and their relations as shown in Figure 17(*a*), where four transitions are used to represent four short operations, i.e., R1 loading, R1 unloading, R2 loading, and R2 unloading. Two places are used to represent two long operations, i.e., M1 and M2 processing.
4. Through a stepwise process, gradually adding resources, constraints, and links to the partial PN model will finally form the refined model as shown in Figure 17(*b*).
5. The model is checked to see whether it satisfies the specification. The PN simulation tool can also be used in this phase to check the model. If some problems are found, the model will be modified.

4.2.3.2. FMS Simulation Simulation is a useful computer technology in FMS modeling, design, and operation. Simulation modeling allows real-world objects to be described in FMS, such as moving of workpieces from one place to another. There are three approaches to simulation modeling for FMS. The first is network or graphical models, where some objects (such as machines) may be represented by graphical symbols placed in the same physical relationship to each other as the corresponding machines are in the real world. The graphical aspects of this kind of models are relatively easy to specify, and once completed they also provide a communication vehicle for the system design that can be readily understood by a variety of people. SLAM (Pritsker 1984) and SIMAN (Pegden 1982) are two widely used network modeling tools for FMS.

The second approach to FMS modeling is data-driven simulation. The model consists of only (or mainly) numerical data, usually representing, for example, a simple count of machines in a system or a table of operation times for each process on the route of a given part type. The nature of these data is such that, if they were collected in the factory information system, it would only be necessary to access them and place them in proper format in order to run a simulation of the corresponding



(a)

(b)

Figure 17 Petri Net Model for the Two-Machine Production Line. (From M. C. Zhou, Ed., *Petri Nets in Flexible and Agile Automation,* Figure 8, with kind permission of Kluwer Academic Publishers)

real-world system. This concept is quite close to automated simulation. It has the ultimate ease of use. The first such program for FMS was developed at Purdue, the general computerized manufacturing system (GCMS) simulator (Talavage and Lenz 1977).

The third approach for FMS modeling uses a base programming language, such as SIMULA and SIMSCRIPT, which provides more model-specific constructs that can be used to build a simulation model. This approach thus has a much stronger modeling capability. Unfortunately, it is not widely used. One reason may be that few people know it well enough to use it.

Another method for DEDS simulation, called activity cycle diagram (ACD), can also be used in FMS simulation. This is a diagram used in defining the logic of a simulation model. It is equivalent to a flowchart in a general-purpose computer program. The ACD shows the cycle for every entity in the model. Conventions for drawing ACDs are as follows:

1. Each type of entity has an activity cycle.
2. The cycle consists of activities and queues.
3. Activities and queues alternate in the cycle.
4. The cycle is closed.
5. Activities are depicted by rectangles and queues by circles or ellipses

Figure 18 presents an ACD for a machine shop. Jobs are arriving from the outside environment. Jobs are waiting in a queue for the machine. As soon as the machine is available, a job goes to the machine for processing. Once processing is over, the job again joins a queue waiting to be dispatched.

Because ACDs give a better understanding of the FMS to be simulated, they are widely used for FMS simulation.

4.3. Benefits and Limitations of FMS

FMS offers manufacturers more than just a flexible manufacturing system. It offers a concept for improving productivity in mid-variety, mid-volume production situations, an entire strategy for changing company operations ranging from internal purchasing and ordering procedures to distribution and marketing. The benefits of FMS can be summarized as follows:

1. Improved manufacturing system flexibility
2. Improved product quality, increased equipment utility
3. Reduced equipment cost, work-in-progress, labor cost, and floor space
4. Shortened lead times and improved market response speed
5. Financial benefits from the above

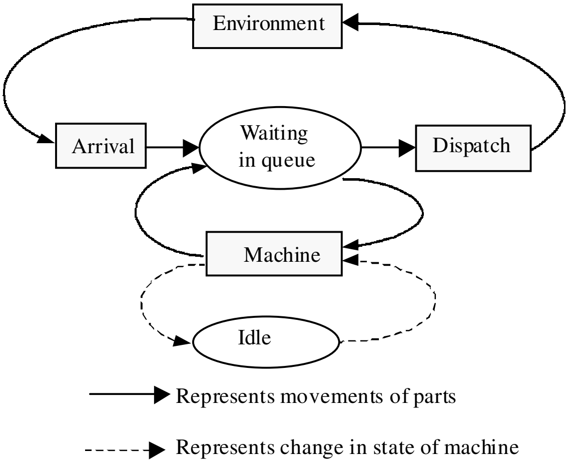

Environment
Arrival
Waiting in queue
Dispatch
Machine
Idle
Represents movements of parts
Represents change in state of machine

Figure 18 Activity Cycle Diagram. (From A. Carrie, *Simulation of Manufacturing Systems,* copyright 1988 John Wiley & Sons Ltd. Reproduced with permission)

Talavage and Hannam (1988) contains a chapter discussing the economic justification for FMS.

The difficulties with FMS should also be given attention. First, FMS is expensive, normally requiring large capital resources. Even if a company is able to afford the investment, FMS may not be financially beneficial if the company does not have much product variety and volume. Second, the design, implementation, and operation of FMS is a quite complex process. Money may be lost if any of the work in the process is not well done. Third, the rapidly changing market may compel the company to change its product. This change may have a negative impact on the production system, causing the large investment in FMS not to be returned before it is abandoned.

5. CIM ARCHITECTURE AND ENTERPRISE MODELING

An enterprise is a complicated social, economic, and physical system. The implementation of CIMS in an enterprise is an even more complicated feat of system engineering. To implement CIMS successfully, engineers with an excellent analytical and technological background and abundant experience are needed, as well as the guidance of advanced CIM system architecture and implementation methodology and powerful support tools. CIM system architecture studies the components and their relationship to CIM to provide the basis and guidelines for the design and implementation of CIMS in a company. A good system architecture can not only act as a basis for describing the system components but also provide a good way to communicate among users, designers, and other parties. A number of CIM system architectures are available, such as the SME wheel structure (Figure 3), CIM open system architecture (CIMOSA), the Purdue enterprise reference architecture (PERA) (Williams 1992), the architecture for information system (ARIS) (Scheer 1992), and GRAI (graphs with results and activities interrelated) integrated methodology (GIM) (Doumeingts et al. 1992).

With the development of CIM reference architecture, a number of enterprise modeling methods have been put forward to describe the enterprise. Because the enterprise is a very complex system, it is difficult to describe it using a simple and unified model. A typical method used by almost all enterprise modeling methods is to describe the enterprise using several view models. Each view defines one aspect from a specific point of view, and then the integration method between the different view models is defined. The general view models now used in describing enterprise are function view, information view, organization view, resource view, and process view. Other views are also presented by researchers are control view, defined in ARIS, decision view, defined in GRAI/GIM, and economic view, proposed by Chen et al. (1994).

5.1. Views of the Enterprise Model

As discussed above, the enterprise model consists of several interrelated view models. Each view describes a specific aspect of the enterprise and has its own modeling method. This section gives a brief description of the aims of each view and the method currently used in building the view model.

5.1.1. Process View

The process view model takes a major role in defining, establishing, analyzing, and extracting the business processes of a company. It fulfills the requirements of transforming the business process, the manufacturing process, and the product-development process into a process view model. The process model is the basis for business process simulation, optimization, and reengineering.

5.1.1.1. Modeling Method for Process View Process view modeling focuses mainly on how to organize internal activities into a proper business process according to the enterprise goals and system restrictions. Traditional function-decomposing-oriented modeling methods such as SADT and IDEF0, which set up the process based on activities (functions), can be used in business process modeling.

The business description languages WFMC (Workflow Management Coalition 1994), IDEF3, and CIMOSA are process-oriented modeling methods. Another modeling method uses object-oriented technology, in which a business process can be comprehended as a set of coordinated request/service operations between a group of objects. Jacobson (1995) presents a method for using object-oriented technology, the use case method, to reengineer the business process. Object-oriented methods offer intrinsic benefits: they can improve systemic extendibility and adaptability greatly, their services based on object-operated mode can assign systemic responsibility easily, existing business processes can be reused easily, and distribution and autonomy properties can be described easily.

The main objective of the process view modeling method is to provide a set of modeling languages that can depict the business process completely and effectively. To depict a business process, it should be able to depict the consequent structure of processes, such as sequence, embranchment, join, condition, and circle, to establish a formal description of the business process. Generally accepted modeling languages today are IDEF3, CIMOSA business process description language, and WFMC workflow description language.

Some business process description methods originating in the concepts and models of traditional project-management tools, such as the PERT chart and other kinds of network chart, are generally adopted in practical application systems because they can be easily extended from existing project management tool software. If the business process is relatively complex, such as existing concurrent or collision activities, some superformal descriptions, such as Petri net, should be used.

Figure 19 is a workflow model of a machine tool-handle manufacturing process. The process model is designed using the CIMFlow tool (Luo and Fan 1999). In Figure 19(*a*), the main sequential process is described and the icon ▦ stands for a subprocess activity. Figure 19(*b*) is the decomposition of the subprocess activity Rough Machining in Figure 19(*a*). After Turning activity is executed, two conditional arcs are defined that split the activity route into two branches. The activity Checking is a decision-making task that is in charge of the product quality or the progress of the whole process.

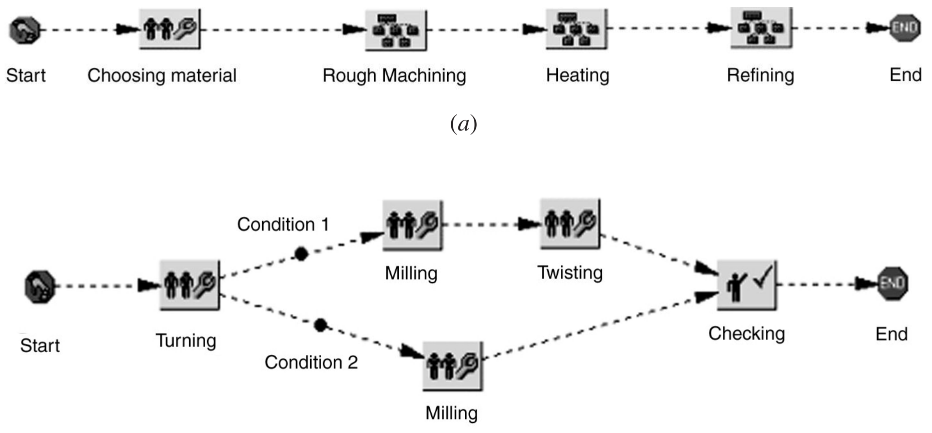
5.1.2. Function View

The function view is used to describe the functions and their relationships in a company. These functions fulfill the objectives of the company, such as sales, order planning, product design, part manufacturing, and human resource management. The efficient and effective operation of these functions contributes to the company's success in competing in the market.

5.1.2.1. Modeling Method for Function View Function view modeling normally uses the top-down structural decomposition method. The function tree is the simplest modeling method, but it lacks the links between different functions, especially the data flow and control flow between different functions, so it is generally used to depict simple function relationships. In order to reflect data and control flow relationships between different functions, SADT and IDEF0 (Colquhoun and Baines 1991) methods are used to model function view. The IDEF0 formalism is based on SADT, developed by Ross (Ross 1985).

The IDEF0 model has two basic elements: activity and arrow. Figure 20 gives the basic graphic symbol used in the IDEF0 method. IDEF0 supports hierarchical modeling so that every activity can be further decomposed into a network of activities. In order to organize the whole model clearly, it is advised that the number of the child activities decomposed from the parent activity be less than 7 and greater than 2. Figure 21 gives a demonstration IDEF0 model (A0 model) of the control shop floor operation function.

In the CIMOSA model, the overall enterprise functions are represented as an event-driven network of domain processes (DPs). An individual domain process is represented as a network of activities. The domain process is composed of a set of business processes (BPs) and enterprise activities (EAs). The BP is composed of a set of EAs or other BPs. An EA is composed of a set of functional operations (FOs). The BP has a behavior property that defines the evolution of the enterprise states over time in reaction to enterprise event generation or conditions external or internal to the enterprise. It is defined by means of a set of rules called procedure rules. The structure property of BP describes the functional decomposition of the enterprise functions of enterprise. This can be achieved by means of a pair of pointers attached to each enterprise function. EA has an activity behavior that defines



(*a*)

(*b*)

Figure 19 Process Model for Tool-Handle Manufacturing.

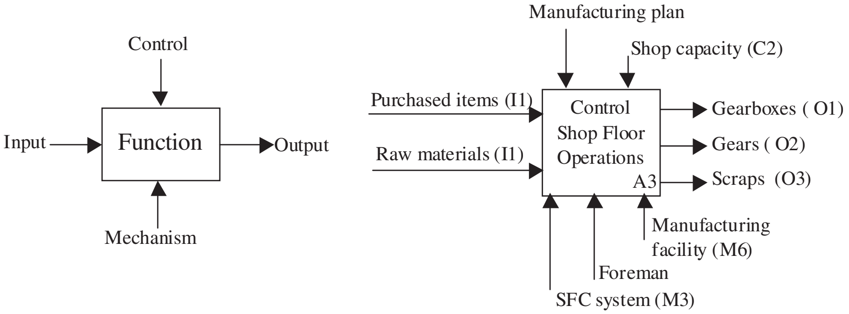

Control
Input
Function
Output
Mechanism
Manufacturing plan
Shop capacity (C2)
Purchased items (I1)
Raw materials (I1)
Control Shop Floor Operations
A3
Gearboxes (O1)
Gears (O2)
Scraps (O3)
Manufacturing facility (M6)
Foreman
SFC system (M3)

Figure 20 General Symbol and a Practical Model of IDEF0. (From Vernadat 1996. Reprinted with kind permission of Kluwer Academic Publishers.)

the internal behavior (or flow of control) of enterprise activities. It specifies how to perform the functionality of an EA in terms of an algorithm making use of FO.

It can be seen that the process view and the function view are closely related to the CIMOSA modeling method. Hence, any tool that supports the CIMOSA modeling methodology should be process oriented and include function decomposition.

5.1.3. Information View

The information view organizes the information necessary to support the enterprise function and process using an information model. Data for a company are an important resource, so it is necessary to provide a method to describe or model the data, such as data structures, repository types, and locations, especially the relationships among different data. It is very important for the company to maintain the data resource consistently, eliminate possible data redundancy, and finally enable data integration.

The information view modeling method provides different models for different phases of a company's information system, from requirement analysis to design specification to implementation. The most commonly used model to express data today is the relational data model, which is the basis

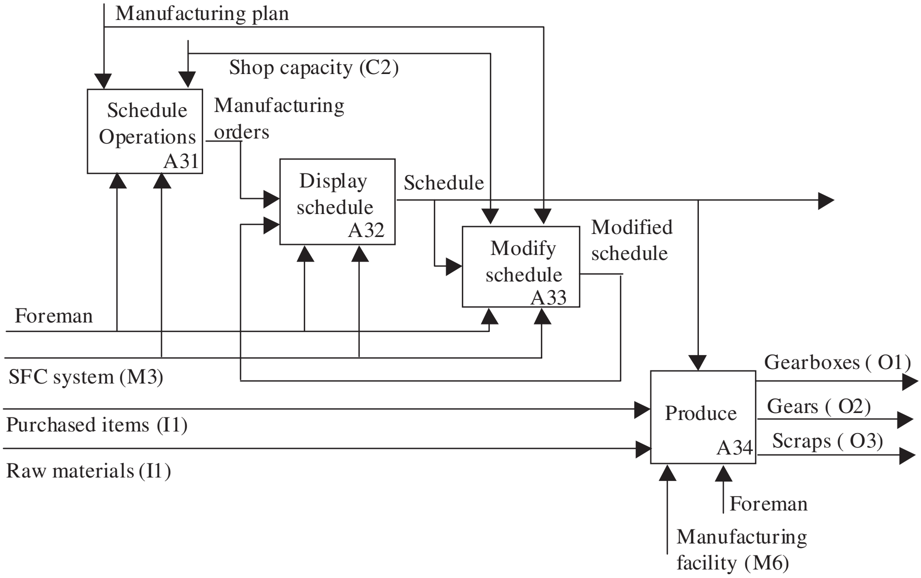

Manufacturing plan
Shop capacity (C2)
Schedule Operations
A31
Manufacturing orders
Display schedule
A32
Schedule
Modify schedule
A33
Modified schedule
Foreman
SFC system (M3)
Purchased items (I1)
Raw materials (I1)
Produce
A34
Gearboxes (O1)
Gears (O2)
Scraps (O3)
Foreman
Manufacturing facility (M6)

Figure 21 IDEF0 Model of the Control Shop-Floor Operation Function. (From Vernadat 1996. Reprinted with kind permission of Kluwer Academic Publishers.)

for the relational database management system. The currently used IDEF1X method is an extension of the entity-relationship model proposed by Chen (1976). Three phases of the modeling process, a conceptual model, a logical model, and a physical model, are used in designing and implementing an information system. Vernadat (1996) gives a good introduction to information modeling in the context of enterprise modeling.

5.1.4. Organization View

The organization view is used to define and represent the organization model of a company. The defined model includes the organization tree, team, faculty, role, and authority. It also creates an organization matrix. In the organization view, the relationships between different organization entities are defined. It provides support for execution of the company's functions and processes.

The hierarchical relationship between the organization units forms the organization tree, which describes the static organization structure. The team describes the dynamic structure of the company. It is formed according to the requirements of business processes. Personnel and organization units are the constituents of the team.

Figure 22 shows the organization view structure. The basic elements of the organization view are organization unit, team, and personnel.

The attributes of the organization unit include organization unit name, position, description, role list, leader, and the organization's associated activities and resources. The leader and subordinate relationships between different units are also defined in the organization unit. In defining a team, the attributes needed are team name, description, project or process ID, associated personnel, and resources.

5.1.5. Resource View

The resource view is similar to the organization view. It describes resources used by the processes to fulfill the company's business functions. Three main objects are defined in the resource view model: resource type object, resource pool object, and resource entity object. Resource type object describes the company's resource according to the resource classification. The resource type object inherits the attributes from its parent object. A resource classification tree is created to describe the company's resource. The resource pool object describes resources in a certain area. All the resources located at this area form a resource pool. Resource entity object defines the atomic resources. An atomic resource is some kind of resource that cannot be decomposed further—that is, the smallest resource entity.

Figure 23 gives the resource classification tree structure. In this tree, the parent node resource consists of all its child node resources. It depicts the static structure of the company's resources.

The resource–activity matrix (Table 1) defines the relationships between resources and process activities. Every cross (×) means that the resource is used by this activity. The resource–activity matrix presents the dynamic structure of the resources.

5.2. Enterprise Modeling Methods

As pointed out in the previous section, there are many enterprise modeling methods. Here we only give a brief introduction to the CIMOSA, ARIS, and GIM methods.

5.2.1. CIMOSA

CIMOSA supports all phases of a CIM system life cycle, from requirements specification, through system design, implementation, operation and maintenance, even to a system migration towards a

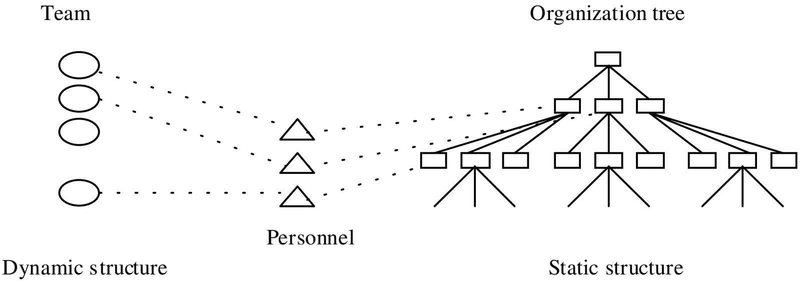


Figure 22 Organization View Structure.

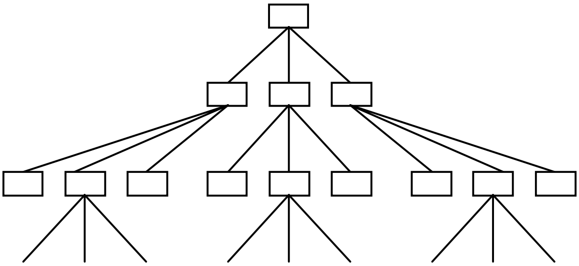


Figure 23 Resource Classification Tree Structure.

CIMOSA solution. CIMOSA provides modeling, analysis, and design concepts in a set of languages and methodologies adapted to enterprise users at different levels and according to different users' viewpoints.

The CIMOSA reference architecture, developed by the AMICE consortium within the European ESPRIT project, is a set of semiformal structures and semantics that is intended to be used as a modeling language environment for any business enterprise. The basic constructs and fundamental control structures of the architecture are collected in volumes called Formal Reference Base II (FRB) and Formal Reference Base III. FRB consists of two parts: the modeling framework and the integrating infrastructure (IIS).

The CIMOSA modeling framework, known as the CIMOSA cube, is shown in Figure 24. The modeling framework provides a reference architecture and a particular architecture. It contains three modeling levels (requirements definition, design specification, implementation description) and four views (function, information, resource, organization). The CIMOSA reference architecture (two left slices of the CIMOSA cube) provides a set of generic building blocks, partial models, and user guidelines for each of the three modeling levels. The particular architecture (right slice of the CIMOSA cube) is the part of the framework that is provided for the modeling of a particular enterprise, that is, it exhibits a given CIM solution.

Generic building blocks or basic constructs are modeling elements with which the requirements and solutions for a particular enterprise can be described. Partial models, which are partially instantiated CIMOSA solutions applicable to one or more industrial sectors, are also provided in the reference architecture. The user can customize partial models to a part of the model of his or her particular enterprise. The CIMOSA modeling framework ensures that partial models from different sources can be used to build a model of one particular enterprise.

The CIMOSA integrating infrastructure (IIS) provides services to integrate all specific application processes of the enterprise into one cooperating system. IIS consists of the following services:

- *Information services:* administering all information required by the various application processes
- *Business process services:* scheduling the provision of resources and dispatching the execution of enterprise activities
- *Presentation services:* representing the various types of manufacturing resources to the business process services in a homogeneous fashion
- *Communication service:* being responsible for system-wide homogeneous and reliable data communication

CIMOSA model creation processes, namely instantiation, derivation and generation, define how generic building blocks and partial models are used to create particular enterprise models.

TABLE 1 Resource-activity Matrix

	Activity 1	Activity 2	–	Activity *n*
Resource 1	×			
Resource 2	×	×		
–			×	
Resource m		×		×

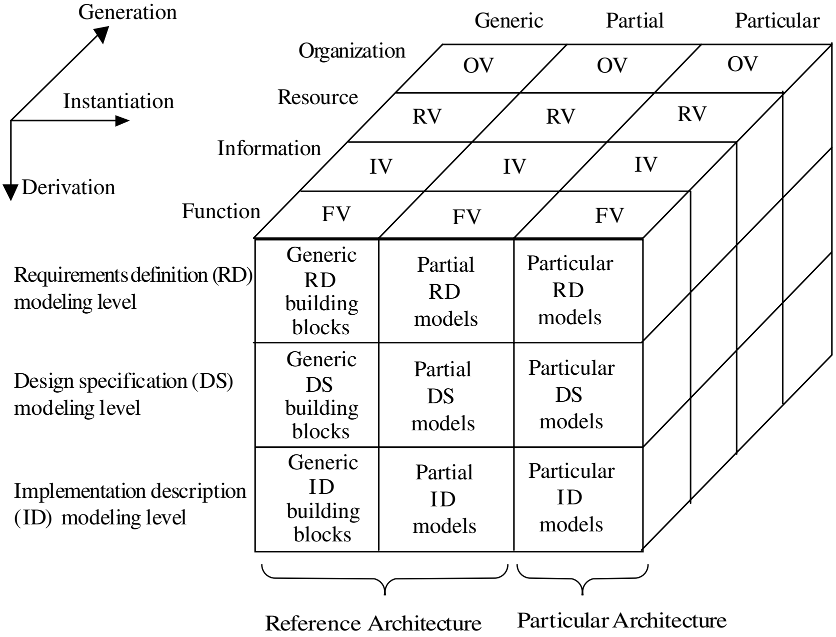

Generation
Instantiation
Derivation
Generic
Partial
Particular
Organization
Resource
Information
Function
OV
RV
IV
FV
Requirements definition (RD) modeling level
Design specification (DS) modeling level
Implementation description (ID) modeling level
Generic RD building blocks
Partial RD models
Particular RD models
Generic DS building blocks
Partial DS models
Particular DS models
Generic ID building blocks
Partial ID models
Particular ID models
Reference Architecture
Particular Architecture

Figure 24 CIMOSA Modeling Framework. (From Esprit Consortium AMICE 1993. Reprinted by permission of Springer-Verlag.)

The instantiation process is a design principle that suggests: (1) going from a generic type to particular type (types are refined into subtypes down to particular instances); and (2) reusing previous solutions (i.e., using particular models or previously defined models) as much as possible. This process applies to all four views. It advocates going from left to right of the CIMOSA cube.

The derivation process is a design principle that forces analysis to adopt a structured approach to system design and implementation, from requirements specification through design specification and finally to full implementation description. This process also applies to all four views. It advocates going from the top to the bottom of the CIMOSA cube.

The generation process is a design principle that encourages users to think about the entire enterprise in terms of function, information, resource, and organization views, in that order. However, the complete definition of the four views at all modeling levels usually requires going back and forth on this axis of the CIMOSA cube.

5.2.2. ARIS

The ARIS (architecture of integrated information systems) approach, proposed by Scheer in 1990, describes an information system for supporting the business process. The ARIS architecture consists of the data view, function view, organization view, and control view. The data view, function view, and organization view are constructed by extracting the information from the process chain model in a relatively independent way. The relationships between the components are recorded in the control view, which is the essential and distinguishable component of ARIS. Information technology components such as computer and database are described in the resource view. But the life-cycle model replaces the resource view as the independent descriptive object. The life-cycle model of ARIS is divided into three levels. The requirement definition level describes the business application using the formalized language. The design specification level transfers the conceptual environment of requirement definition to the data process. Finally, the implement description level establishes the physical link to the information technology. The ARIS architecture is shown in Figure 25.

The ARIS approach is supported by a set of standard software, such as the application systems ARIS Easy Design, ARIS toolset, and ARIS for R/3, which greatly help the implementation of ARIS.

5.2.3. GIM

GIM (GRAI integrated methodology) is rooted in the GRAI conceptual model shown in Figure 26. In this method, an enterprise is modeled by four systems: a physical system, an operating system, an information system, and a decision system.

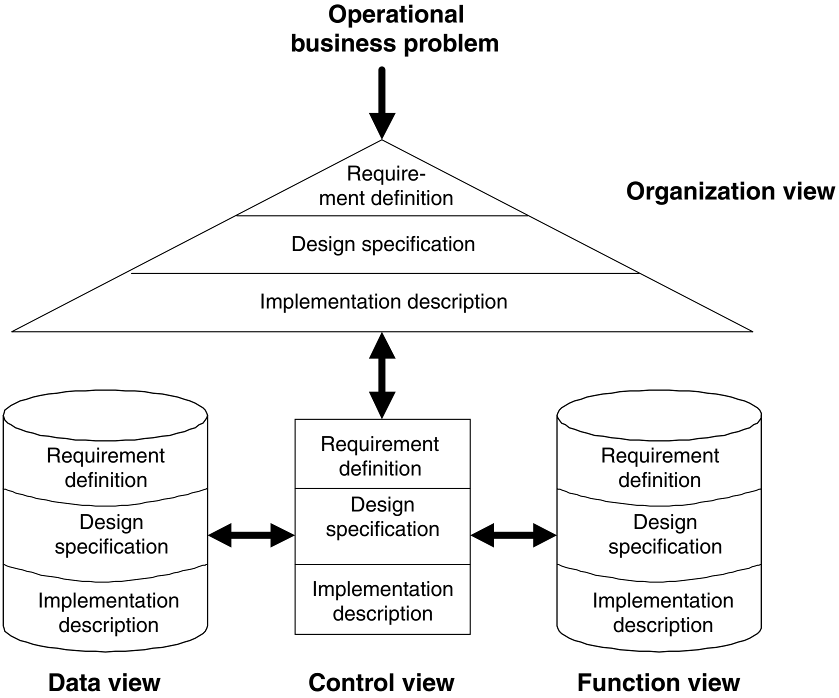

Operational business problem
Require-ment definition
Design specification
Implementation description
Organization view
Requirement definition
Design specification
Implementation description
Requirement definition
Design specification
Implementation description
Requirement definition
Design specification
Implementation description
Data view
Control view
Function view

Figure 25 ARIS Architecture. (From Scheer 1992 by permission of Springer-Verlag New York, Inc.)

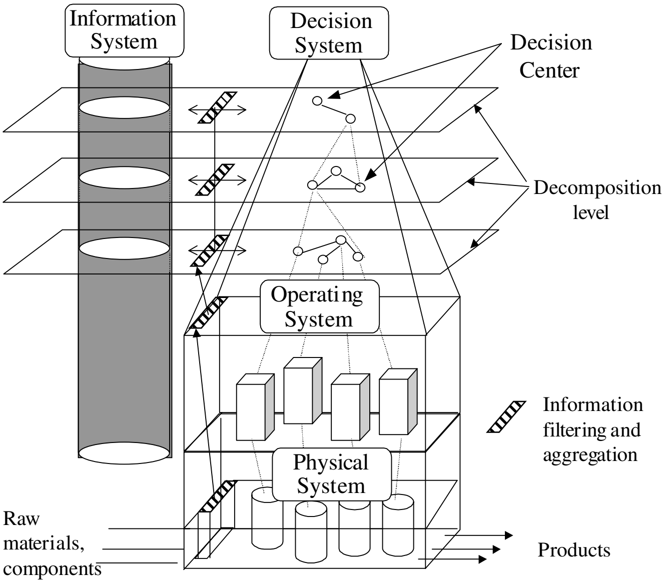

Information System
Decision System
Decision Center
Decomposition level
Operating System
Information filtering and aggregation
Physical System
Raw materials, components
Products

Figure 26 The GRAI Conceptual Model. (From Vernadat 1996)

The GRAI method makes use of two basic modeling tools: the GRAI grid and the GRAI net. The GRAI grid is used to perform a top-down analysis of the domain of the enterprise to be analyzed. It is made of a 2D matrix in which columns represent functions and lines represent decision levels. The decision level is defined by a horizon H and a period P. Long-term planning horizons are at the top and short-term levels are at the bottom of the grid. Each cell in the matrix defines a decision center. The grid is then used to analyze relationships among decision centers in terms of flows of information and flows of decisions.

GRAI nets are used to further analyze decision centers in terms of their activities, resources, and input–output objects. With this method, a bottom-up analysis of the manufacturing systems studied can be made to validate the top-down analysis. In practice, several paths in both ways are necessary to converge to a final model accepted by all concerned business.

GRAI and GIM are supported by a structured methodology. The goal is to provide specifications for building a new manufacturing system in terms of organization, information technology, and manufacturing technology viewpoints. The methodology includes four phases: initialization, analysis, design, and implementation.

6. CIM IMPLEMENTATION

CIM implementation is a very important but also very complex process. It requires the participation of many people with different disciplines. Benefits can be gained from successful implementation, but loss of investment can be caused by inadequate implementation. Therefore, much attention should be paid to CIM implementation.

6.1. General Steps for CIM Implementation

The general life-cycle model discussed in CIM architecture and modeling methodology is the overall theoretical background for CIM implementation. In a practical application, due to the complexity of CIM implementation, several phases are generally followed in order to derive the best effect and economic benefits from CIM implementation. The phases are feasibility study, overall system design, detailed system design, implementation, operation, and maintenance. Each phase has its own goals and can be divided into several steps.

6.1.1. Feasibility Study

The major tasks of the feasibility study are to understand the strategic objectives, figure out the internal and external environment, define the overall goals and major functions of a CIM system, and analyze the feasibility of CIM implementation from technical, economical, and social factors. The aim of this phase is to produce a feasibility study report that will include, besides the above, an investment plan, a development plan, and a cost–benefit analysis. An organization adjustment proposal should also be suggested. A supervisory committee will evaluate the feasibility study report. When it is approved, it will lay the foundation for following up the phases of CIM implementation. Figure 27 presents the working steps for the feasibility study.

6.1.2. Overall System Design

Based on the results of the feasibility study, the overall system design phase further details the objectives and plans regarding proposed CIM system implementation. The tasks of overall system design are to define the CIM system requirements, set up the system function and information model, put forward an overall system design plan, design the system architecture, draft the implementation plan, present the investment plan, carry out the cost–benefit analysis, and finally form the overall system design report. The key technologies and their problem-solving methods should also been given in the overall system design report. Data coding is important work to be done in the overall system design phase.

In order to keep the whole CIM system integrated, in the functional and logical model design, the overall system design follows the top-down decomposition principle. The top level and general functions should be first considered, then decomposed to low-level and detailed operations.

The general procedures and contents of overall system design are as follows:

1. *System requirement analysis:* determines the system requirements of function, performance, information, resource, and organization. This phase's work focuses on the managerial and tactical point of view.
2. *System architecture design:* determines the overall system architecture of the CIM system.
3. *System function and technical performance design:* determines the functions needed to meet the system requirements and system performance.
4. *Information model design:* determines the logical data model of the information system.

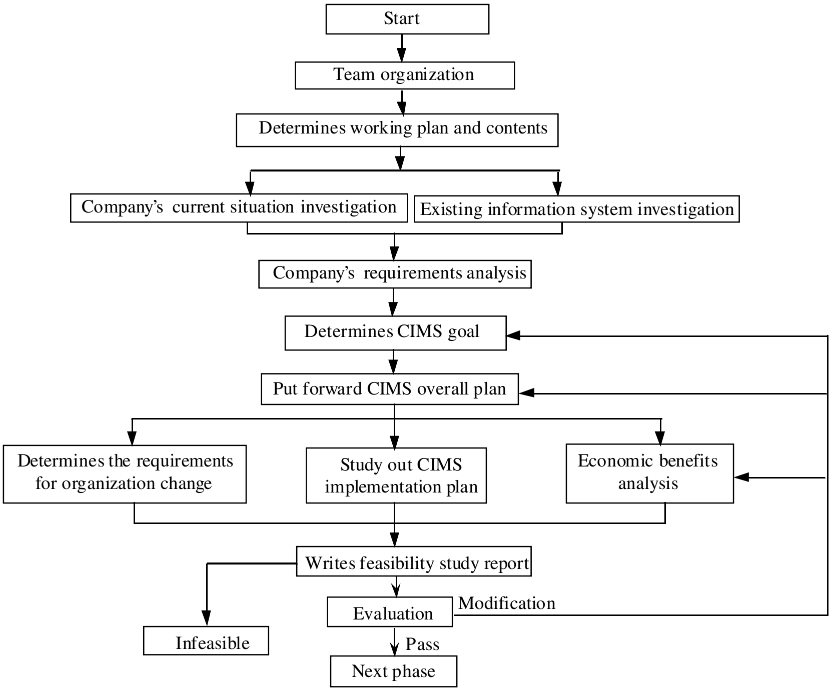

Start
Team organization
Determines working plan and contents
Company's current situation investigation
Existing information system investigation
Company's requirements analysis
Determines CIMS goal
Put forward CIMS overall plan
Determines the requirements for organization change
Study out CIMS implementation plan
Economic benefits analysis
Writes feasibility study report
Evaluation
Modification
Infeasible
Pass
Next phase

Figure 27 Steps of Feasibility Study.

5. *Internal and external interface design:* determines these interfaces for the purpose of system integration, including the functional interfaces between different subsystems and data interfaces between different applications.
6. *Key technology:* lists all key technologies that have important influence on CIM system implementation, gives their solution methods;
7. *System configuration specification:* determines the hardware and software configurations.
8. *Implementation schedule definition:* defines the implementation schedule for the CIM system in the network plan or other forms.
9. *CIM system organization definition:* defines the suitable organization structure for the CIM environment.
10. *Budget making and cost–benefit analysis.*
11. *Overall system design report generation.*

6.1.3. Detailed System Design

The detailed system design phase solves the problem of system specification definition, the associated hardware and software configuration assignment, the functional and data interface definition, the implementation plan and making of steps, the forming of the associated implementation team, and the assignment of responsibility and setting of benchmarks.

In this phase, an important goal is to define the interfaces between different subsystems. The shared data physical model for the CIM system needs to be specified. The number, type, and configuration of hardware systems should be defined. The detailed software products that should meet the requirements defined in overall system design should also be specified. The network scheduling for the implementation plan should be generated and evaluated. A leadership group is formed that will manage the entire CIM implementation. A number of implementation teams with personnel from different disciplines and different business sectors are formed. Each implementation team will be engaged in the implementation of a specific part of the CIM system.

After the detailed system design phase is finished, the CIM system is ready to go into practical implementation.

6.1.4. Implementation and Operation

The implementation phase follows a bottom-up approach. Subsystems are implemented in order according to the implementation scheduling. When the subsystem implementation is finished, integration interfaces between the subsystems are developed and several higher-level subsystems are formed through integration of some low-level subsystems. Finally, the whole CIM system is implemented through an integration.

After the system is built and tested, it becomes an experimental operation, which will last for three to six months. During that period, errors that occur in the operation are recorded and system modifications are carried out. The CIM system is turned to practical use. In the implementation and operation phase, the following steps are generally followed:

1. *Building computer supporting environment:* including computer network, computer room, network and database server, UPS, air conditioner, and fire-proof system
2. *Building manufacturing environment:* including whole system layout setup, installation of new manufacturing devices, and old manufacturing configuration
3. *Application system development:* including new commercial software installment, new application system development, old software system modification
4. *Subsystem integration:* including interface development, subsystem integration, and system operation test
5. *CIM system integration:* including integration and testing of whole CIM system
6. *Software documentation:* including user manual writing, operation rule definition, setting up of system security and data backup strategies
7. *Organization adjustment:* including business process operation mode, organization structure, and operation responsibility adjustment
8. *Training:* including personal training at different levels, from top managers to machine operators
9. *System operations and maintenance:* including daily operations of CIM system, recording of errors occurring in the operation, application system modification, and recording of new requirements for future development

6.2. Integration Platform Technology

6.2.1. Requirements for Integration Platform

The complexity of manufacturing systems and the lack of effective integration mechanisms are the main difficulties for CIMS implementation. Problems include lack of openness and flexibility, inconvenient and inefficient interaction between applications, difficulty in integrating a legacy information system, the long time required for CIMS implementation, and the inconsistency of user interfaces.

To meet the requirements enumerated above, the integration platform (IP) concept has been proposed. IP is a complete set of support tools for rapid application system development and application integration in order to reduce the complexity of CIMS implementation and improve integration efficiency. By providing common services for application interaction and data access, IP fills the gaps between the different kinds of hardware platforms, operating systems, and data storage mechanisms. It also provides a unified integration interface that enables quick and efficient integration of different applications in various computing environments.

6.2.2. The Evolution of Integration Platform Technology

IP has evolved through a number of stages. It was initially considered an application programming support platform that provided a common set of services for application integration through API. A typical structure of the early IPs is the system enabler/application enabler architecture proposed by IBM, shown in Figure 28. Under such a structure, the IP provides a common, low-level set of services for the communication and data transfer (the system enabler) and also provides application domain specific enabling services (the application enabler) for the development of application systems. Thus, the application developer need not start from coding with the operating system primitive services. One disadvantage of the early IP products was that they only provided support for one or a limited number of hardware and operating system and the problem of heterogeneous and distributed computation was not addressed. Also, the released products often covered a specific domain in the enterprises, such as the shop-floor control. These early IPs focused mainly on support for the development of application software, and their support for application integration was rather weak.

Since the 1990s, IP has developed for use in a heterogeneous and distributed environment. An example is shown in Figure 29, where the architecture is divided into several layers, the communication layer, the information management service layer, and the function service layer, providing

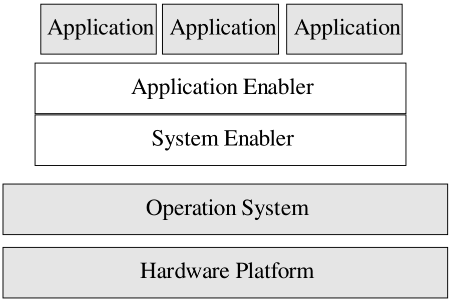

Application
Application
Application
Application Enabler
System Enabler
Operation System
Hardware Platform

Figure 28 IBM System Enabler/Application Enabler.

commonly used system-level services. These services form the middleware layer of IP. The higher layers of IP are classified as general-purpose API, domain-specific API, and application development integration tools. The integration supporting area is extended from a specific domain to the whole enterprise, including management, planning, and manufacturing execution.

6.2.3. MACIP System Architecture

MACIP (CIMS Application Integration Platform for Manufacturing Enterprises) is a Chinese national high-technology R&D key technology research project. The MACIP project is designed to develop a research prototype of an application platform oriented to the new IP technology described above.

The MACIP system architecture is presented in Figure 30. It is a client–server structured, object-oriented platform with a high degree of flexibility. MACIP consists of two layers, the system enabling level and the application enabling level. The system enabling level is composed of two functions, the communication system and the global information system (GIS). The primary function of these components is to allow for the integration of applications in a heterogeneous and distributed computing environment. The communication system provides a set of services that allow transparent communication between applications. The global information system allows applications to have a common means for accessing data sources in a variety of databases and file systems. These functions are implemented in the form of application independent API (AI API). Application independence means that these functions are not designed for specific applications but are general services for communication, data access, and file management. Hence, the system enabling level provides the basic integration mechanisms for information and application integration.

The application enabling level, which utilizes the functions contained within the system enabling level, is composed of three domain sub-integration platforms (SIPs): MIS SIP, CAD/CAM/CAPP SIP, and shop-floor control SIP. Each SIP is designed according to the requirements of a domain application and provides functions for applications in the form of Application Dependent API (AD API). The AD API functions are designed specifically to enable the quick and easy development of

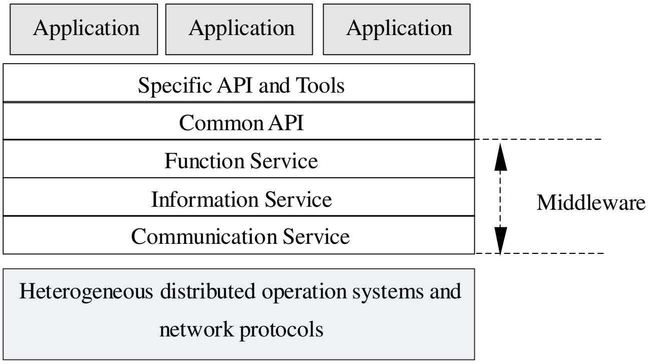

Application
Application
Application
Specific API and Tools
Common API
Function Service
Information Service
Communication Service
Middleware
Heterogeneous distributed operation systems and network protocols

Figure 29 A Multilayer IP Structure.

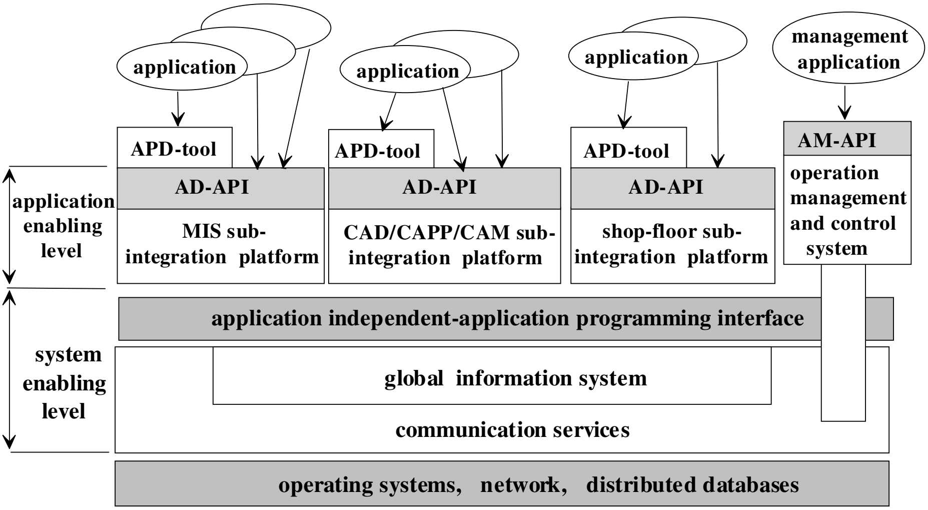

application
application
application
management application
APD-tool
APD-tool
APD-tool
AM-API
AD-API
AD-API
AD-API
application enabling level
MIS sub-integration platform
CAD/CAPP/CAM sub-integration platform
shop-floor sub-integration platform
operation management and control system
application independent-application programming interface
system enabling level
global information system
communication services
operating systems, network, distributed databases

Figure 30 System Architecture of MACIP.

domain specific applications. These functions enable the complete integration of the application. Application development tools (APD tools) are developed using the AD-API. Users can also develop applications using the functions provided by AD-API. Existing applications are integrated by modifying the data exchange interface using AD-API functions. An Internet interface is also included in the application enabling level interfaces and provides access to MACIP through appropriate Internet technologies.

An operation management system was also designed that uses AI API functions to provide an application management API (AM API) for the users. Users use AM API to develop management applications that manage the IP resources and coordinate the operation of different applications.

The development of MACIP was finished in early 1999. It has since been used in several companies to support the rapid implementation of CIMS.

7. CIMS IN PROCESS INDUSTRY

7.1. Introduction

Process industry, by which we refer to continuous or semicontinuous production industry processes, principally includes the petroleum industry, the electric power industry, the metallurgical industry, the chemical industry, the paper industry, the ceramic industry, the glass industry, and the pharmaceutical industry. Process industry is a kind of highly complicated industrial system that not only includes biochemical, physical, and chemical reactions but also transmission or transition of matter and energy. Most process industries are subject to the interlocked relations of enterprise decision making, business marketing, schedule planning, material supplying, repertory transportation, and product R&D, in addition to the characteristics of continuity in wide scope, uncertainty, high nonlinearity, and strong coupling. All these factors are responsible for the unusual difficulty of comprehensive management, scheduling, optimization, and control in process industry enterprises. Therefore, these problems cannot be solved relying on either control and optimization theory, which are based on accurate mathematical models and exact analytical mathematical methods, or automation techniques alone (Ashayberi and Selen 1996). The CIMS technique is one possible solution to complex, comprehensive automation of process industry.

7.1.1. Definitions

- *Process industry:* Those industries in which the values of raw materials are increased by means of mixing and separating, molding, or chemical reaction. Production can be continuous or batch process. The characteristics of process industry must be considered when CIMS is applied to those industries.

- *Architecture structure:* The models that reflect these characteristics of production and business in process industry. The models represent all aspects of CIMS in the multiview and multilayer approach.
- *Models:* The structural representations of object concepts. Models include rules, data, and formal logical methods that are used to depict states, behaviors, and the interactive and inferential relations of objects or events.
- *Reference model:* The model definition for the architecture structure.
- *Modeling method:* According to the architecture descriptions, designers obtain the descriptions of all the states in an enterprise by abstracting the business function, business data, and business period.
- *Information integration:* Information integration activities in the production process or enterprise or even group can be described as a process of obtaining, handling, and processing information so that accurate information can be sent punctually and in the right form to the right people to enable them to make correct decisions.

7.1.2. Key Technologies

Because CIMS in process industry is in the developmental stage, some key technologies still need to be developed further:

1. *Total technology:*
   - Architecture structure and reference model of CIMS in process industry
   - Business reengineering model and managerial modes of enterprise
   - Control modes of Material and cost streams
   - Modeling methods for CIMS in process industry
   - Structural design methods for CIMS in process industry
   - Design specifications for CIMS in process industry
2. *Integration technologies:*
   - Information integration in enterprise and between enterprises
   - Integration of relation database and real-time database systems
   - Core data model, data booting, data compression, and data mining
   - Integration and Utilization of development tools and applications
   - Information integration-based Internet, data navigation, and browser technology
3. *Network technologies:*
   - Architecture structure of computer network system
   - Openess, reliability, safety, expandability, monitoring and management of networks
   - Speed, collisions resolution, concurrency control of networks
4. *Supervisor control technologies:*
   - Distributed intelligent decision-making support system-based intelligent agent
   - Optimization model establishment of large-scale systems
   - Description and analysis of hybrid system
   - Multimode grouping modeling and production operation optimization
   - Advanced process-control strategy and intelligent coordination control
   - Production safety monitoring, fault diagnosis and isolation, failure forecast
   - “Soft” measurement, intelligent data synthesis and coordination

7.2. Reference Architecture of CIMS in Process Industry

CIMS in process industry involves complex systematic engineering. Since its inception, advanced management theories, such as BPR (business process reengineering), CE (concurrent engineering), and TQM (total quality management), have been introduced. Using these theories, managers could reorganize departments that overlapped in function so as to facilitate the development of the enterprise. The realization of CIMS in these enterprises must build a clear reference architecture that can depict all functions in various phases and levels. Under the guidance of the reference architecture, the designers can simulate all potential solutions in an appropriate workbench and determine the total integration solution. The reference architecture of CIMS in process industry can refer to the frame of CIMS-OSA and PERA. The CIMS-OSA frame has many definitions and modeling approaches, in which the concepts are very clear. The PERA frame is very suitable for the definition of every phase in the CIMS life cycle, which considers every human factor that will affect the enterprise integration.

7.2.1. Architecture Structure Model

The architecture structure model of CIMS in process industry has four phases (Aguiar and Weston 1995): the strategic planning, requirement analysis, and definition phase; the conceptual designs phase; the detailed design and implementation phase; and the operation and maintenance phase, as shown in Figure 31. They reflected all the aspects of building process of CIMS. The strategic planning and requirement definition phase relates to senior management. The models in this phase manipulate information with reference to enterprise models and external factors to assess the enterprise's behavior, objective, and strategy in multiview and multidomain so as to support decision making. The conceptual design phase is the domain of system analysis. According to the scope defined in the previous phase, the models in this phase give a detailed description of a system in formalized system modeling technology. A solution will be found that satisfies the demands for performance and includes what and how to integrate. In general, the solution is expressed in the form of functions. Detailed design and implementation will be carried out in the system design phase. In this phase, the physical solutions should be specified, which include all subsystems and components. The models given in this phase are the most detailed. The models in the operation and maintenance phase embody the characteristics of the system in operation. These models, which define all active entities and their interaction, encapsulate many activities in enterprise operation.

The reference models depicted in Figure 31 consist of run-time models, resource models, integration models, system models, and business models, which correspond to the descriptions of the AS-IS system and the TO-BE system in the process of designing CIMS in process industry. Their relationships are abstracted step by step from down to up, opposite to the process of building CIMS.

- *Run-time models* encapsulate the information related to system operation, such as dynamic math models of production, input–output models, and order management models.
- *Resource models* contain the information related to relationships between the resource and the satisfaction of demands. In these models, the resource information that designers will add for some special functions is still included.

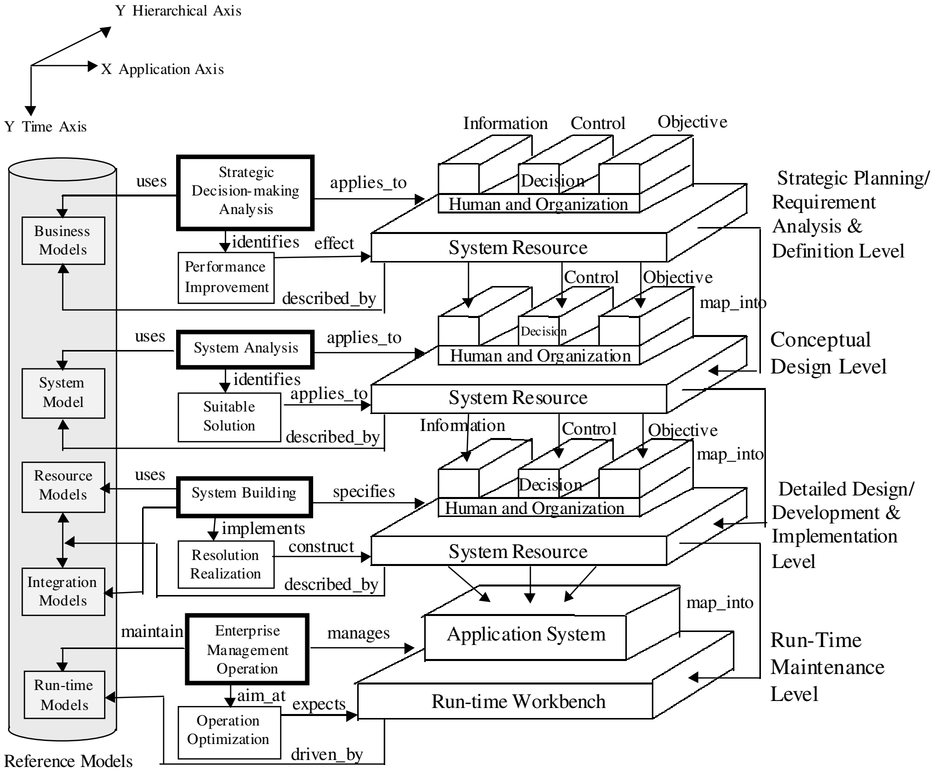

Y Hierarchical Axis
X Application Axis
Y Time Axis
Information
Control
Objective
Business Models
uses
Strategic Decision-making Analysis
applies_to
Decision
Human and Organization
Strategic Planning/ Requirement Analysis & Definition Level
identifies
effect
System Resource
Performance Improvement
described_by
Control
Objective
map_into
uses
System Analysis
applies_to
Decision
Human and Organization
Conceptual Design Level
System Model
identifies
applies_to
Suitable Solution
System Resource
described_by
Information
Control
Objective
map_into
Resource Models
uses
System Building
specifies
Decision
Human and Organization
Detailed Design/ Development & Implementation Level
implements
Resolution Realization
construct
System Resource
Integration Models
described_by
map_into
maintain
Enterprise Management Operation
manages
Application System
Run-Time Maintenance Level
Run-time Models
aim_at
expects
Run-time Workbench
Operation Optimization
driven_by
Reference Models

Figure 31 Architecture Structure Model. (From Aguiar and Weston 1995)

- *Integration models* present the way in which various component elements of the AS-IS system and the TO-BE system could be integrated to complete an integrated system.
- *System models* capture the structure and knowledge of the department in related domains that are currently organized or should be organized to improve the performance of the system. They encapsulate the experiences of system analysts and the descriptions of the prototype.
- *Business models* collectively contain the business knowledge required to accomplish strategic analysis and requirement definition, including business rules and the accumulated experience from analyzing enterprise performance.

With these reference models, CIMS in process industry could be built from up to down. Each phase is dynamic in nature and can be related to each other phase. That is, the implementation in every phase can be modified to adapt to changes in environment and demands.

7.2.2. Hierarchical Structure Model

The hierarchical structure model is a structured description of CIMS engineering. It is an aggregation of models and their relationships in the whole CIMS of an enterprise. It is the foundation of the design and realization of CIMS. A hierarchical structure model used in CIMS in process industry is shown in Figure 32. It has five levels and two supporting systems. The five levels are the direct control system, the supervisory control system, the production scheduling system, the management information system, and the enterprise decision making system. The two supporting systems are the database system and the computer network system.

The main function of the hierarchical structure model is:

- *Direct control system level:* This is the lowest level of automated system in the production process, including the distributed control system used in production devices and the fundamental automated equipment used offsite. The parameters of the production process are measured and controlled by the automated system. They also receive instruction from the supervisory control system level and accomplish the process operation and control.
- *Supervisory control system level:* The system in this level fulfills supervisory control of main production links in the whole production process. According to the instructions from the sched-

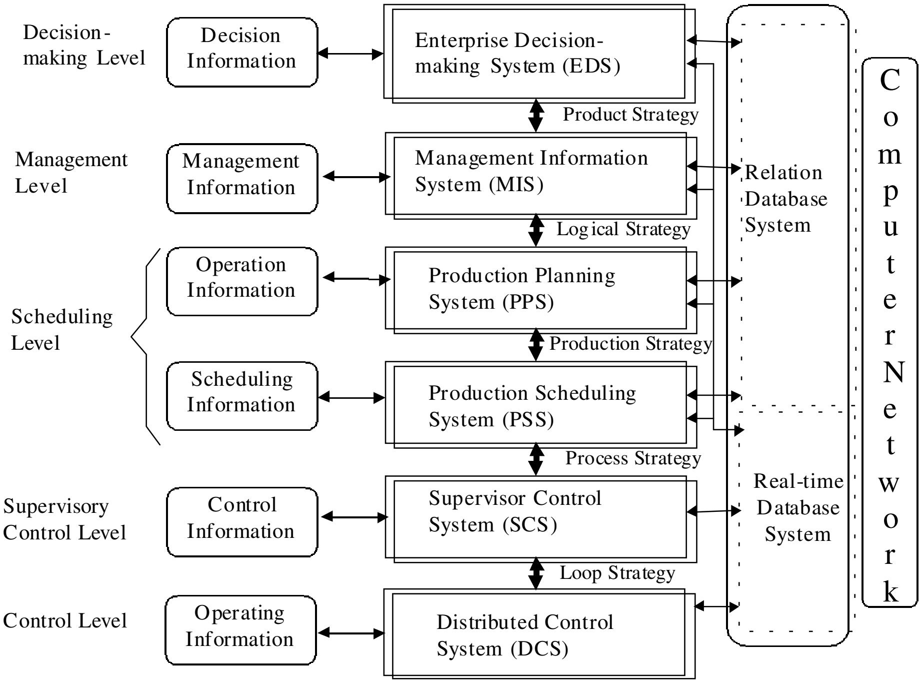

Decision-making Level
Decision Information
Enterprise Decision-making System (EDS)
Product Strategy
Management Level
Management Information
Management Information System (MIS)
Logical Strategy
Scheduling Level
Operation Information
Production Planning System (PPS)
Production Strategy
Scheduling Information
Production Scheduling System (PSS)
Process Strategy
Supervisory Control Level
Control Information
Supervisor Control System (SCS)
Loop Strategy
Control Level
Operating Information
Distributed Control System (DCS)
Relation Database System
Real-time Database System
ComputerNetwork

Figure 32 Hierarchical Structure Model.

uling system level, it formulates process tactics and conducts the actions at the direct control system level, including operation optimization, advanced control, fault diagnosis, and process simulation.

- *Production scheduling system level:* At this level, the production load is determined and the production planning is decomposed into five days rolling work planning for every month, according to the information from the decision-making system and the material-stream and energy-stream data. By optimizing scheduling, allocating energy, and coordinating operations in every workshop, the production becomes balanced, stable and highly efficient.
- *Management information system level:* The system at this level accomplishes the MIS function for the whole enterprise and carries out integrated management of production and business information. According to the instructions from the decision-making system, it makes logical decisions. It is in charge of day-to-day management, including business management and production management.
- *Enterprise decision-making system level:* The system at this level comes up with decisions supporting enterprise business, product strategy, long-term objectives, and developing planning and determines the strategy of production and business. Within the company group, it aims at integration optimization in the whole enterprise so as to yield the maximum benefit.

7.3. Approach to Information Integration for CIMS in Process Industry

The core of CIMS in process industry is the integration and utilization of information. Information integration can be described as follows: The production process of an enterprise is a process of obtaining, processing, and handling information. CIMS should ensure that accurate information is sent punctually on in the right form to the right people to enable them to make correct decisions.

7.3.1. Production Process Information Integration

Production is the main factor to be considered in CIMS design for a process industry. Driven by the hierarchical structure model discussed in Section 7.2.2, these information models of every subsystem at all levels are built. In these models, the modeling of production process information integration is the crux of the matter. This model embodies the design guidance, centering on production in three aspects:

1. Decision → comprehensive planning → planning decomposition → scheduling → process optimization → advanced control
2. Purchase → material management → maintenance
3. Decision → comprehensive planning → planning decomposition → scheduling → product storage and shipment control

The computation of material equilibrium and heat equilibrium and the analysis/evaluation of equipment, material and energy can be done using this model so as to realize the optimized manipulations in the overall technological process.

7.3.2. Model-Driven Approach to Information Integration

Figure 33 depicts the mapping relationship of the models in the building process of CIMS in process industry. It demonstrates that the designed function related to every phase of building CIMS can be depicted from the design view using the structural and model-driven approach (Aguiar and Weston 1995).

The realization of model mapping relies on the building of software tools supporting every phase in a hierarchical way. *Model mapping* refers to the evolutionary relationships of models between the phases of building CIMS. As the enterprise hierarchy is developed downwards, the description in the models becomes more detailed. In contrast, with the increasing widening of modeling scope, the granularity of descriptions in models will be reduced so as to form more abstract models. For example, at the detailed design and implementation level, various dynamic math models should be used, and detailed IDEF0 and static math models should be used in the conceptual design phase. We can conclude that the models of various phases in the building CIMS can be evolved step by step in the model-driven approach from up to down.

In the previous analysis, the realization of the model-driven information integration method requires a workbench. This consists of a series of tools, such as modeling tools from entity to model, simulating tools supporting the simulations in various levels from higher-level strategic planning to lower-level detailed design, and assessing tools appraising the performance of solution simulation at various levels.

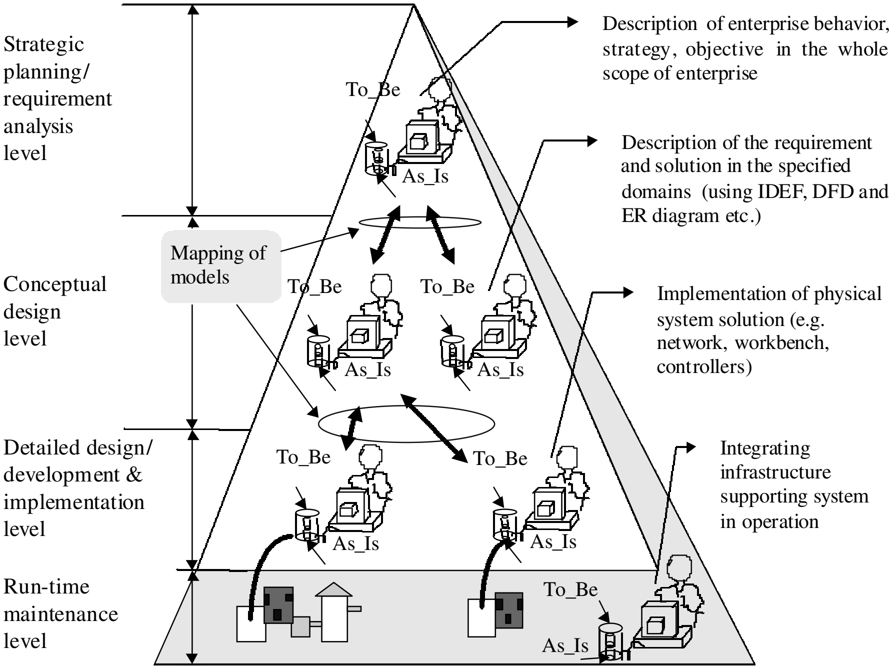

Strategic planning/ requirement analysis level
Conceptual design level
Detailed design/ development & implementation level
Run-time maintenance level
Mapping of models
To_Be
As_Is
Description of enterprise behavior, strategy, objective in the whole scope of enterprise
Description of the requirement and solution in the specified domains (using IDEF, DFD and ER diagram etc.)
Implementation of physical system solution (e.g. network, workbench, controllers)
Integrating infrastructure supporting system in operation

Figure 33 Integrating Map of Models Based on Model-Driven Approach. (From Aguiar and Weston 1995)

7.4. An Application Example

We will discuss an application example of CIMS in a giant refinery enterprise. The technological process of the refinery is continuous, the material stream cannot be interrupted, and strict real-time demands for production manipulation are made. The enterprise aims at the following objectives: material equilibrium, energy equilibrium, safety and high efficiency, low cost and good quality, and optimized operation of the technological process. The realization of CIMS in this type of enterprise requires the consideration not only of problems such as production management, production scheduling, operation optimization, and process control, but also of business, marketing, material supply, oil product transport and storage, development of new products, capital investment, and so on (Fujii et al. 1992). The computer integrated production system of the enterprise is constructed according to changes in crude oil supply, market requirements for products, flexibility of the production process, and different management modes. The integration of business decision making, production scheduling, workshop management, and process optimization is realized in the giant refinery.

7.4.1. Refinery Planning Process

The refinery enterprise consists of many production activities (Kemper 1997). If the blend operation day is called the original day, then the production activities on the day 90 days before that day include crude oil evaluation, making of production strategy, and crude oil purchasing. In the same way, the production activities on the day 10–30 days after the original day include stock transportation and performance adjustment of oil products. Every activity in the production process is relevant to each other activity. For example, in crude oil evaluation, the factors in the activities following the making of production strategy must be analyzed. In another example, people in the activity of crude oil evaluation need to analyze those production activities following the refinery balance in detail. Deep analysis of those activities in the refinery enterprise is the basis of design of CIMS in that enterprise. Figure 34 depicts the refinery planning process.

7.4.2. Integrated Information Architecture

By analyzing of the refinery planning process, we can construct the integration frame depicted in Figure 35.

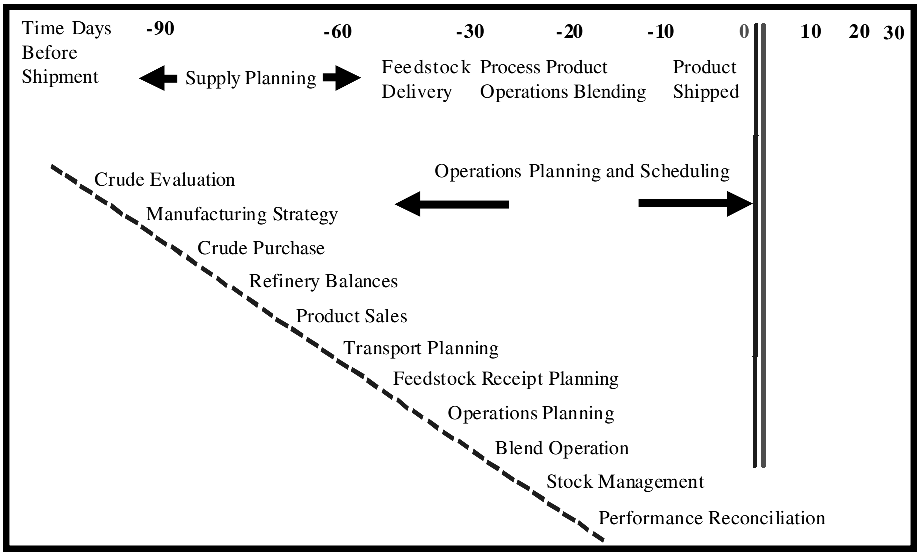

Time Days Before Shipment
-90
-60
-30
-20
-10
0
10
20
30
Supply Planning
Feedstock Delivery
Process Operations
Product Blending
Product Shipped
Operations Planning and Scheduling
Crude Evaluation
Manufacturing Strategy
Crude Purchase
Refinery Balances
Product Sales
Transport Planning
Feedstock Receipt Planning
Operations Planning
Blend Operation
Stock Management
Performance Reconciliation

Figure 34 Refinery Planning Process.

Using the model-driven approach to the modeling of all subsystems, the information integration model in this refinery enterprise could be built as shown in Figure 36 (Mo and Xiao 1999). The model includes the business decision-making level, the planning and scheduling level and the process supervisory control level. Their integration is supported by two database systems.

The relevant information, such as market, costing, financial affairs, and production situation, is synthesized to facilitate business decisions of the enterprise, and crude oil supply and oil product sale planning are both determined at the business decision-making level.

The planning and scheduling level synthesizes management information, decomposes production planning to short-term planning and executes the daily scheduling, and gives instructions directly to process supervisory control level. In the meantime, it accomplishes the management and control of oil product storage and transport, including the management and optimized scheduling control of the harbor area and oil tank area.

The process supervisory control accomplishes process optimization, advanced control, fault diagnosis, and oil product optimized blending.

7.4.3. Advanced Computing Environment

The information integration model of the giant refinery depicted in Figure 36 is built using the model-driven method. The model is the design guidance of the realization of CIMS in the enterprise. Figure

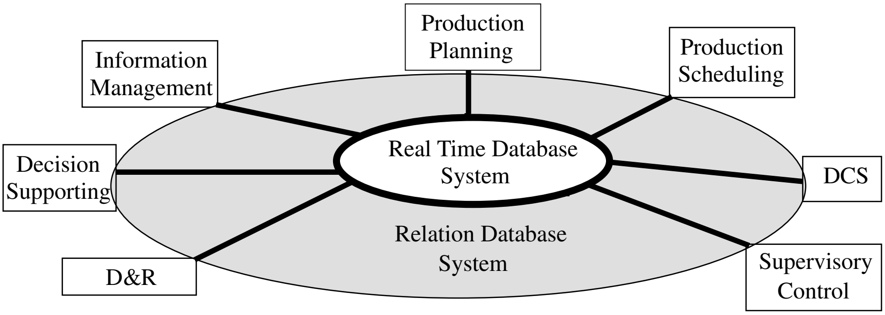

Production Planning
Information Management
Production Scheduling
Real Time Database System
Decision Supporting
DCS
Relation Database System
D&R
Supervisory Control

Figure 35 Integration Frame.

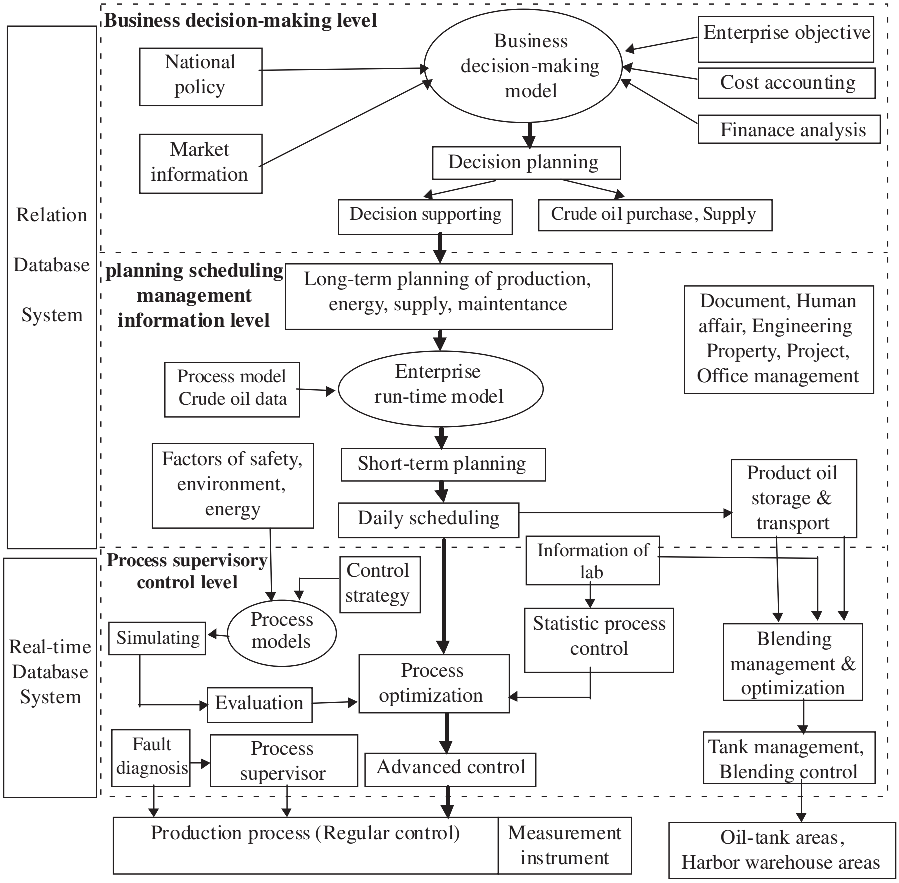

Business decision-making level
National policy
Market information
Business decision-making model
Enterprise objective
Cost accounting
Finanace analysis
Decision planning
Decision supporting
Crude oil purchase, Supply
Relation Database System
planning scheduling management information level
Long-term planning of production, energy, supply, maintentance
Document, Human affair, Engineering Property, Project, Office management
Process model Crude oil data
Enterprise run-time model
Factors of safety, environment, energy
Short-term planning
Daily scheduling
Product oil storage & transport
Process supervisory control level
Control strategy
Information of lab
Real-time Database System
Simulating
Process models
Statistic process control
Blending management & optimization
Evaluation
Process optimization
Fault diagnosis
Process supervisor
Advanced control
Tank management, Blending control
Production process (Regular control)
Measurement instrument
Oil-tank areas, Harbor warehouse areas

Figure 36 Information Integration Model.

37 depicts the computing environment for the realization of the information integration model, using the client–server computing mode (Kemper 1997).

7.5. Conclusions

The reference architecture of CIMS in process industry can instruct designers to optimize solutions by repeatedly optimizing and simulating so as to obtain the final physical system realization in an enterprise. Practical experience indicates that CIMS in process industry is not like selling a car. With the progressive development of technology and the changes in the external environment, CIMS in process industry needs to continue adjusting to yield the maximum economic and social benefit.

8. BENEFITS OF CIMS

Many benefits can be obtained from the successful implementation and operation of a CIM system in a manufacturing company. The benefits can be classified into three kinds: technical, management, and human resources quality.

8.1. Technical Benefits

Technical benefits obtained from implementation CIM system are:

1. *Reducing inventory and work-in-progress:* This can be accomplished through the utilization of an MRPII or ERP system. Careful and reliable material purchasing planning and production planning can to a great extent eliminate high inventory and work-in-progress level, hence reducing capital overstock and even waste through long-term material storage.

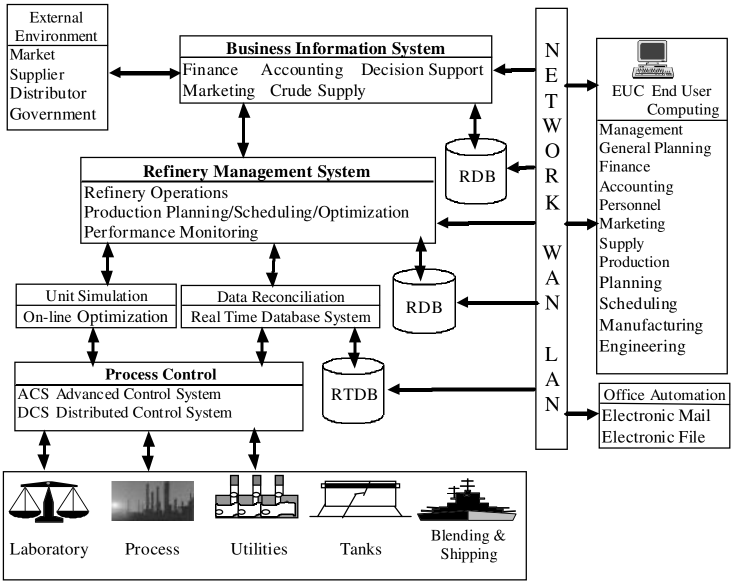

External Environment
Market
Supplier
Distributor
Government
Business Information System
Finance Accounting Decision Support
Marketing Crude Supply
NETWORK WAN LAN
EUC End User Computing
Management
General Planning
Finance
Accounting
Personnel
Marketing
Supply
Production
Planning
Scheduling
Manufacturing
Engineering
Office Automation
Electronic Mail
Electronic File
RDB
Refinery Management System
Refinery Operations
Production Planning/Scheduling/Optimization
Performance Monitoring
Unit Simulation
On-line Optimization
Data Reconciliation
Real Time Database System
RDB
Process Control
ACS Advanced Control System
DCS Distributed Control System
RTDB
Laboratory
Process
Utilities
Tanks
Blending & Shipping

Figure 37 Advanced Computing Environment.

2. *Improving production efficiency:* Through the integration of a production system, planning system, and material supply system, the production processes can be operated in a well-organized way and hence production can be carried out with the shortest possible waiting times and machine utilization greatly increased. Through the integration of CAD, CAPP, and CAM systems, the setup time for NC machines can be reduced significantly. The improvement of production efficiency will bring economic returns from investment in the CIM system.
3. *Improving product quality:* The integration of the company's business processes, design processes, and production processes will help in improving product quality. TQM can be put into effect in the CIM integrated environment.
4. *Reducing cost:* This is the direct effect obtained from the above three benefits.
5. *Improving product design ability:* Through the integration of CAD, CAPP, and CAM systems, by using the current engineering method, the product design ability of the company can be significantly improved. New and improved products can be designed and developed in a shorter time, and the company can win the market competition with these products.

8.2. Management Benefits

The following management benefits can be obtained from the CIM sysem implementation:

1. *Standardizing processes:* The business processes, design processes, and production processes can be standardized. This can help to streamline the company's processes and reduce errors caused by uncontrolled and random operations.
2. *Optimizing processes:* The business processes, design processes, and production processes can be optimized. This can help to locate bottlenecks in the processes and cost-intensive activities and thus to provide methods to reduce the cost.
3. *Improving market response ability:* The traditional pyramid organization structure will be changed to flat structure that can greatly improve the speed of response to market change and user requirements.

8.3. Human Resource Quality

Almost all employees will be involved in the implementation of the CIM system. Through different courses of training, from CIM philosophy to computer operation, the total quality of the employees can be improved at all levels, from management staff to production operators. More importantly, employees will get to know better the company's objectives, situation, technical standards, and manufacturing paradigm, inspiring them to devote their energy to improving the company's operation efficiency.

9. FUTURE TRENDS OF CIM

As a manufacturing paradigm, CIM concepts and practice have developed for more than 20 years. CIM is still in active development and has received much attention from researchers and companies. Some of the development trends for CIM are as follows.

9.1. Agile Manufacturing

In today's continuously, rapidly, and unforeseeably changing market environment, an effective way to keep the company competitive is to use the agile manufacturing strategy. Agile manufacturing has been called the 21st-century manufacturing enterprise strategy (Goldman and Preiss 1991; Goldman et al. 1995). By agile, we mean that the company can quickly respond to market change by quickly reengineering its business processes, reconfiguring its manufacturing systems, and innovating its products.

A number of papers discuss the characteristics of an agile manufacturing company, such as:

- Greater product customization
- Rapid introduction of new or modified products
- Increased emphasis on knowledgeable, highly trained, empowered workers
- Interactive customer relationships
- Dynamic reconfiguration of production processes
- Greater use of flexible production technologies
- Rapid prototyping
- An agile and open system information environment
- Innovative and flexible management structures
- Rapid collaboration with other companies to form a virtual company.

9.2. Green Manufacturing

The increasingly rapid deterioration of environment has caused many problems for society. During the production of products, manufacturing companies also produce pollution to the environment. Pollution produced during the manufacturing processes includes noise, waste gas, wastewater, and waste materials. Another kind of pollution is caused by waste parts at the end of the product's life, such as batteries, printed circuit boards, and plastic covers. Green manufacturing aims at developing a manufacturing paradigm and methods for reducing pollution by a manufacturing company of the environment. The green manufacturing paradigm covers the whole life cycle of a product, from requirements specification, design, manufacturing, and maintenance to final discarding. Research topics in green manufacturing include:

- *Green design* (also called *design for environment*) considers the product's impact on the environment during the design process, designing a product that causes minimal pollution. Multi-life-cycle design, which considers multiple use of most parts and recycling one-time-use parts, has received much attention.
- *Green materials* involves development of materials that can be easily recycled.
- *Green production* involves developing methods to reduce pollution during the production process.
- *Green disposal:* developing new methods to recycle the discarded products.

9.3. Virtual Manufacturing and Other Trends

By using virtual reality and high-performance simulation, virtual manufacturing focuses on building a digital model of the product and studies the dynamic and kinetic performance of the product to reduce product development cost and time.

Many development trends are affecting CIM and its related technologies. Technologies that may have a great influence on CIM include network (Web) technology, distributed object technology, intelligent agent technology, knowledge integration technology, and CSCW technology. CIM systems, using these advanced paradigms and technologies, will have a brilliant future. In the future, a manufacturing company supported by an advanced CIM system may be operated in an Internet environment (Web user interface), running on a virtual dynamic organization structure, using CSCW tools, to design and produce products in a cooperated and integrated way. The company will fully satisfy user requirements and produce products quickly and cheaply. Materials and products will be delivered on time.

REFERENCES

Agrawal, S. C. (1985), *Metamodeling: A Study of Approximations in Queueing Models,* MIT Press, Cambridge, MA.

Aguiar, M. W. C., and Weston, R. H. (1995). "A Model-Driven Approach to Enterprise Integration," *International Journal of Integration Manufacturing,* Vol. 8, No. 3, pp. 210–224.

Ashayberi, J., and Selen, W. (1996), "Computer Integrated Manufacturing in the Chemical Industry," *Production and Inventory Management Journal,* Vol. 37, No. 1, pp. 52–57.

Ayres, R. U. (1991), *Computer Integrated Manufacturing,* Vol. 1, Chapman & Hall, New York.

Bray, O. H. (1988), *Computer Integrated Manufacturing: The Data Management Strategy,* Digital Press, Bedford, MA.

Buffa, E. S. (1984). *Meeting the Competitive Challenge: Manufacturing Strategies for U.S. Companies,* Dow Jones-Irwin, Homewood, IL.

Harrington, J. (1979), *Computer Integrated Manufacturing,* R. E. Krieger, Malabar, FL.

Carrie, A. (1988), *Simulation of Manufacturing,* John Wiley & Sons, New York.

Carter, M. F. (1986), "Designing Flexibility into Automated Manufacturing Systems," in *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems: Operations Research Models and Applications,* K. E. Stecke and R. Suri, Eds. (Ann Arbor, MI, August 12–15), Elsevier, Amsterdam, pp. 107–118.

CCE-CNMA Consortium (1995), *CCE-CNMA: An Integration Platform for Distributed Manufacturing Applications,* Springer, Berlin.

Chen, P. P. S. (1976), "The Entity-Relationship Model: Toward a Unified View of Data," *ACM Transactions on Database Systems,* Vol. 1, No. 1, pp. 9–36.

Chen, Y., Dong, Y., Zhang, W., and Xie, B. (1994), "Proposed CIM Reference Architecture," in *Proceedings of 2nd IFAC/IFIP/IFORS Workshop on Intelligent Manufacturing Systems* (Vienna, June).

Colquhoun, G. J., and Baines, R. W. (1991), "A Generic IDEF0 Model of Process Planning," *International Journal of Production Research,* Vol. 29, No. 11, pp. 239–257.

Doumeingts, G., Vallespir, B., Zanettin, M., and Chen, D. (1992), "GIM, GRAI Integrated Methodology: A Methodology for Designing CIM Systems, Version 1.0," Research Report, LAP/GRAI, University of Bordeaux I.

Esprit Consortium AMICE, Eds. (1993), *CIMOSA: Open System Architecture for CIM,* 2nd Ed. Springer, Berlin.

Fan, Y., and Wu, C. (1997), "MACIP: Solution for CIMS Implementation in Manufacturing Enterprises," in *Proceedings of IEEE International Conference on Factory Automation and Emerging Technology* (Los Angles, September), pp. 1–6.

Fujii, M., Iwasaki, R., and Yoshitake, A. (1992), "The Refinery Management System under the Concept of Computer Integrated Manufacturing," *National Petroleum Refiners Association Computer Conference* (Washington, DC).

Georgakopoulos, D., Hornick, M., and Sheth, A. (1995), "An Overview of Workflow Management: From Process Modeling to Workflow Automation Infrastructure," *Distributed and Parallel Databases,* Vol. 13, No. 2, pp. 119–153.

Goldman, G. L., and Preiss, K., Eds. (1991), *21st Century Manufacturing Enterprise Strategy: An Industry-Led View,* Harold S. Mohler Laboratory, Iacocca Institute, Lehigh University, Bethlehem, PA.

Goldman, G. L., Nagel, R. N., and Preiss, K. (1995), *Agile Competitors and Virtual Organization: Strategies for Enriching the Customer,* Van Nostrand Reinhold, New York.

Greenwood, E. (1989). *Introduction to Computer-Integrated Manufacturing,* Harcourt Brace Jovanovich, New York.

Gupta, Y. P., and Goyal, S. (1989), "Flexibility of Manufacturing Systems: Concepts and Measurements," *European Journal of Operational Research,* Vol. 43, pp. 119–135.

Hall, R. W. (1983), *Zero Inventories,* Dow Jones-Irwin, Homewood, IL.

Hammer, M., and Champy, J. (1993). *Reengineering the Corporation: A Manifesto for Business Revolution,* Nicholas Brealey, London.

Jacobson, I. (1995), *The Object Advantage: Business Process Reengineering with Object Technology,* Addison-Wesley, Reading, MA.

Jha, N. K., Ed. (1991), *Handbook of Flexible Manufacturing Systems,* Academic Press, San Diego.

Keller, G. (1995), "Creation of Business Processes with Event-Driven Process Chains: A Strategic Challenge," in *SAPinfo-Business Reengineering,* SAP AG, Walldorf, pp. 8–13.

Kemper, M. (1997), "IBM Refinery Management System for the Hungarian MOL Refinery," *Erdoel Erdgas Kohle,* Vol. 113, No. 6 (in German).

Kochan, A., and Cowan, D. (1986), *Implementing CIM: Computer Integrated Manufacturing,* IFS, Bedford, UK.

Koenig, D. T. (1990). *Computer Integrated Manufacturing: Theory and Practice,* Hemisphere, New York.

Luo, H., and Fan, Y. (1999), "CIMFlow: A Workflow Management System Based on Integration Platform Environment," in *Proceedings of 7th IEEE International Conference on Emerging Technologies and Factory Automation* (Barcelona, October).

Mayer, R. J., Cullinane, T. P., deWitte, P., Knappenberger, W., Perakath, B., and Wells, S. (1992), *Information Integration for Current Engineering,* IDEF3 Process Description Capture Method Report, AL-TR-1992-0057, Air Force Systems Command, Wright-Patterson Air Force Base, OH.

Mo, Y. W., and Xiao, D. Y. (1999), "A Model-Driven Approach to Information Integration in Continuous Process Industries," in *Proceedings of the 14th World Congress of IFAC* (Beijing), Vol. A, pp. 115–120.

Otte, R., Patrick, P., and Roy, M. (1996), *Understanding CORBA: Common Object Request Broker Architecture,* Prentice Hall, Upper Saddle River, NJ.

Pegden, C. (1982), *Introduction to SIMAN,* System Modeling Corp., Sewickley, PA.

Prasad, B. (1996), *Concurrent Engineering Fundamentals: Integrated Product and Process Organization,* Prentice Hall, Upper Saddle River, NJ.

Pritsker, A. (1984), *Introduction to Simulation and SLAM, II,* John Wiley & Sons, New York.

Rogers, P., Upton, D. M., and Williams, D. J. (1992), "Computer Integrated Manufacturing," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 647–673.

Ross, D. T. (1985), "Applications and Extensions of SADT," *Computer,* Vol. 18, No. 4, pp. 16–24.

Saxena, M., and Irani, R. K. (1994), "A Knowledge-Based Engineering Environment for Automated Analysis of Nozzles," *Concurrent Engineering: Research and Applications,* Vol. 2, No. 1, pp. 45–57.

Scheer, A.-W. (1992), *Architecture for Integrated Information Systems,* Springer, Berlin.

Singh, N. (1996), *Systems Approach to Computer-Integrated Design and Manufacturing,* John Wiley & Sons, New York.

Taguchi, G., Elsayed, A. E., and Hsiang, T. (1990), *Quality Engineering in Production Systems,* McGraw-Hill, New York.

Talavage, J., and Hannam, R. G. (1988), *Flexible Manufacturing Systems in Practice: Application, Design, and Simulation,* Marcel Dekker, New York.

Talavage, J., and Lenz, J. (1977), *General Computerized Manufacturing Systems (GCMS) Simulator,* NSF Report No. 7, August.

Tetzlaff, U. A. W. (1990), *Optimal Design of Flexible Manufacturing Systems,* Springer, New York.

Vernadat, F. B. (1996), *Enterprise Modeling and Integration: Principles and Applications,* Chapman & Hall, London.

Waldner, J. B. (1992), *Principles of Computer-Integrated Manufacturing,* John Wiley & Sons, New York.

Williams, T. J. (1992), *The Purdue Enterprise Reference Architecture,* Instrument Society of America, Research Triangle Park, NC.

Workflow Management Coalition (1994), *The Workflow Reference Model* (WfMC1003), WFMC TC00-1003.

Wright, W. O. (1992), *The Executive's Guide to Successful MRPII,* 2nd Ed., Oliver Wright, Essex Junction, VT.

Zhou, M. C., Ed. (1995), *Petri Nets in Flexible and Agile Automation,* Kluwer Academic Publishers, Boston.

CHAPTER 16

Clean Manufacturing

Julie Ann Stuart
Purdue University

1. INTRODUCTION 530
2. MOTIVATION 531
  2.1. Metrics 531
  2.2. Legal Requirements 531
  2.3. Responsibility Trends 532
    2.3.1. Extended Product Responsibility 532
    2.3.2. Manufacturers as Service Providers 532
    2.3.3. Environmental Information Provided by Manufacturers 532
3. HIERARCHY OF IMPORTANT CONCEPTS 533
4. METHODS 533
  4.1. Waste/Energy Audits and Waste/Energy Minimization 533
    4.1.1. Waste Audits 533
    4.1.2. Energy Audits 534
  4.2. Life-Cycle Design 534
    4.2.1. Product Design 534
    4.2.2. Process Design 536
  4.3. Product Life-Cycle Assessment 536
  4.4. Production Planning with Environmental Considerations 538
    4.4.1. Models for Production Planning over the Product Life Cycle 538
    4.4.2. Production Planning Models for the Manufacturing and Assembly Stage 538
    4.4.3. Disassembly Planning Models 538
    4.4.4. Production Planning Models for Bulk Recycling 538
  4.5. Environmental Management Systems 539
    4.5.1. Corporate Environmental Policies 539
    4.5.2. Environmental Goals and Metrics 539
    4.5.3. ISO 14539 Series Standards 539
5. CONCLUDING REMARKS 539
REFERENCES 540

1. INTRODUCTION

Clean manufacturing focuses on planning, designing, and producing manufactured products to incur minimal environmental impacts during their lifetimes. Implementation of clean manufacturing includes reliable process control, substitution of hazardous materials with nonhazardous materials, reduction of energy consumption, repair for product life extension, and design for materials recycling. Thus, clean manufacturing may apply to materials and process analyses that reach beyond the manufacturing process to include the environmental impacts of the entire product life. Clean manufacturing requires a broad range of multidisciplinary approaches that include local, regional, national and global policies, engineering and technology advances, economics, and management perspectives.

Industrial engineers study industrial metabolism—that is, the linkages between suppliers, manufacturers, consumers, refurbishers, and recyclers. Environmental engineers, on the other hand, study environmental metabolism—that is, the linkages between entities such as biota, land, freshwater, sea water, and atmosphere (Graedel and Allenby 1995). Clean manufacturing requires the study of the interactions between industrial metabolism and environmental metabolism.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

TABLE 1 Types of Environmental Impacts

General Environmental Impacts	Environmental Hazards
Resource depletion	♦ Materials extraction, use, and disposal ♦ Loss of soil productivity ♦ Landfill exhaustion ♦ Loss of species diversity
Pollutants and wastes	♦ Groundwater quality (biological, metals, and toxic contamination, eutrophication,[a] acid deposition,[a] sedimentation) ♦ Atmospheric quality (stratospheric ozone depletion,[a] global warming,[a] toxic air pollution)
Energy consumption	♦ Energy use that results in resource depletion and/or pollutants

Adapted from Norberg-Bohm et al. 1992.

[a] This term is defined in Section 4.3, Table 6.

2. MOTIVATION

The environmental impacts of manufacturing may include resource depletion, energy consumption, air pollutants, water pollutants, and both hazardous and nonhazardous solid waste. Table 1 shows the hazards associated with general environmental impacts.

2.1. Metrics

The simplest metric for environmental impact with respect to material consumption, pollution, waste generation, or energy consumption is the quantity or inventory for a specified period of time. Microlevel inventory may be measured with respect to a geographical area or industrial facility. Sources for macrolevel inventory data for the United States are summarized in Table 2.

Metrics may also be indexed with respect to a manufacturing output, input, throughput, or batch size, as discussed in Allen and Rosselot (1997). Product-based environmental impact analysis requires activity-based inventory assessment (Stuart et al. 1998). Similar to activity-based costing (see Chapter 89), activity-based environmental inventory assessment recognizes the hierarchy of impacts and assigns them proportionately to an activity such as a product or service.

Due to the complexity of environmental impacts, the Swedish Environmental Institute and Volvo recommend consideration of the following characteristics in their environmental priority strategies (EPS) system: scope, extent of distribution, frequency and intensity, duration or permanence, significance of contribution, and remediability (Horkeby 1997; Ryding et al. 1993). Another complexity to consider is the transfer of impacts along the supply chain because materials extraction, assembly, use, reuse, recycling, and disposal may occur in different locations around the world.

2.2. Legal Requirements

Traditional command-and-control requirements that primarily targeted the manufacturing phase of the product life cycle increased significantly in the past 20 years in the United States. For example, the number of environmental laws passed in the United States increased from 7 between 1895 and 1955 to 40 between 1955 and 1995 (Allenby 1999). Similarly, the number of environmental agreements in the European Union has generally escalated from 1982 to 1995, as described in European Environmental Agency (1997). Many of the regulations in the Asia-Pacific region mirror those in the United States and Europe (Bateman 1999a,b).

TABLE 2 Example of Information Sources for Macrolevel Inventory Data for the United States

Medium	Report	Agency
Material	Toxic Release Inventory (TRI)	U.S. EPA
Pollutant	Aerometric Information Retrieval System (AIRS)	U.S. EPA
Waste	Resource Conservation and Recovery Act Biennial Report System (RCRA BRS)	U.S. EPA
Energy	Manufacturing Energy Consumption Survey	U.S. Department of Energy

Manufacturers must also follow local legislation such as mandates for permits to install or operate processes with regulated effluents. In addition to local and federal mandates where manufacturing and sales take place, manufacturers must also keep abreast of global agreements. For example, the manufacture of chlorofluorocarbon (CFC) solvents, which were used for cleaning electronic assemblies, was banned in 1995 (Andersen 1990). Environmental law is discussed in Chapter 19. Additional legal and service trends are discussed in the next section.

2.3. Responsibility Trends

This section outlines three emerging trends that directly affect the manufacturer's responsibility for environmental impacts: extended product responsibility, extended services, and environmental information reporting. The first trend calls for producers to prevent pollution associated with their products over the products' life cycles. For the second trend, rather than solely selling products, some manufacturers are expanding their business to offer service packages that include the use of their products. In the third trend, the availability and mandatory reporting requirements for environmental information for customers are increasing. These trends are discussed in the next three subsections.

2.3.1. Extended Product Responsibility

There is a trend in Europe and East Asia toward product life cycle responsibility legislation that requires manufacturers to minimize environmental impacts from materials extraction to manufacturing to distribution/packaging to repair to recycling to disposal. Essentially, extended product responsibility shifts the pollution prevention focus from production facilities to the entire product life cycle (Davis et al. 1997). For example, proposed legislation may require that manufacturers not only recycle in-plant wastes but also recycle their discarded products (Denmark Ministry of the Environment 1992; Davis 1997). The evaluation of life cycle stages and impacts are discussed further in Section 4.3.

2.3.2. Manufacturers as Service Providers

In recent years, as manufacturers have assumed the additional role of service provider, responsibility for environmental impact has shifted from the user to the manufacturer. For example, a chemical supplier may be reimbursed per total auto bodies cleaned rather than for the procurement of chemicals for auto body cleaning. Under such an arrangement, there is a financial incentive for the supplier to reduce material consumption (Johnson et al. 1997). In another example, an electronic component manufacturer may use a chemical rental program. The supplier provides chemical management from purchasing and inventory management to waste treatment and disposal (Johnson et al. 1997). Thus, chemical suppliers are gaining a broader responsibility for their products throughout their products' life cycles.

Another important service trend is the replacement of products with services. For example, telecommunications providers offer voice mail rather than selling answering machines. Another example is electronic order processing rather than paper processing. These service trends result in dematerialization, the minimization of materials consumed to accomplish goals (Herman et al. 1989).

2.3.3. Environmental Information Provided by Manufacturers

The third trend is the increasing amount of environmental information that manufacturers communicate to customers. Three general approaches for communicating environmental attributes to corporate procurement and consumers have emerged: eco-labels, self-declaration, and life cycle assessment.

Eco-labels are the simplest format for consumers but the most inflexible format for manufacturers in that they require that 100% of their standard criteria be met. Examples of eco-labels include the Energy Star label in the United States and the Blue Angel in Germany. Because over 20 different eco-labels with different criteria are in use around the world, manufacturers may need to consider multiple eco-label criteria sets (Modl 1995).

Another type of label, self-declaration, allows producers to select methods and metrics. However, comparisons among competing products or services are difficult. Self-declaration is the most flexible form for manufacturers, but its use depends on the manufacturer's environmental reputation among customers. The ECMA, a European industry association that proposes standards for information and communication systems, has proposed product-related environmental attribute standards (Granda et al. 1998).

Full life-cycle assessment, a comprehensive method to analyze the environmental attributes of the entire life cycle of a product, requires environmental engineering expertise. Life cycle assessment is described in Section 4.3.

Consumers may learn about environmental impacts from eco-labels, self-declaration, and life cycle assessment studies. Industrial engineers may learn about clean manufacturing as universities integrate industrial ecology concepts into business and engineering programs (Santi 1997; Stuart 2000). Important clean manufacturing concepts are defined in the next section.

3. HIERARCHY OF IMPORTANT CONCEPTS

To provide a framework for clean manufacturing methods, this section will define important basic concepts. Interestingly, many of the concepts, such as recycling, are defined differently by government, societal, industrial, and academic entities (Allen and Rosselot 1997). Other terms are used interchangeably; for example, *pollution prevention* is often defined as source reduction.

In Table 3, sustainable development and industrial ecology top the hierarchy in clean manufacturing. Industrial ecology is an emerging study that attempts to lessen the environmental impacts of manufacturing activities through planning and design. Industrial ecology is a systems approach to optimizing materials and energy cycles of products and processes (Graedel and Allenby 1995). Methods for clean manufacturing and industrial ecology are described in the next section.

4. METHODS

Traditional methods for clean manufacturing focus on waste or energy audits, which are summarized in Section 4.1. New methods focus on life cycle design, life cycle assessment, production planning models with environmental considerations, and environmental management systems, which are described in Sections 4.2, 4.3, 4.4, and 4.5, respectively.

4.1. Waste/Energy Audits for Waste/Energy Minimization

Waste and energy audits require a detailed inventory analysis of waste generation and energy consumption. The point of origin of each waste and the breakdown of the equipment energy consumption patterns must be determined. Audits are used to identify significant sources of waste and energy costs. Because some environmental impacts are interconnected, both individual source and system impacts must be considered. General guidelines are given for waste audits and energy audits in the next two subsections.

4.1.1. Waste Audits

Waste audits may be performed at the waste, product, or facility level. Waste-level audits simply require that each waste stream and its source be identified. Although this approach is the simplest, it ignores the implications and interactions of the waste stream as a whole. Waste audits performed at the product level are product life cycle inventory assessments, which are discussed in Section 4.3. Facility waste audits are the most common type of audit because most environmental laws require discharge reporting by facility. Estimating plant-wide emissions is discussed in Chapter 19.

Facility waste audits require process flow charts, product material information (commonly from the bill of materials), process material information (such as cutting fluids in a machine shop), and

TABLE 3 Hierarchy of Terms in Clean Manufacturing[a]

Term	Definition
Sustainable development	". . . to meet the needs of the present without compromising the ability of future generations to meet their own needs" (President's Council on Sustainable Development 1996)
Industrial ecology	"the self-conscious organization of production and consumption to reduce adverse environmental impacts of human activity" over time (Socolow 1999)
Product life-cycle assessment	assessment of environmental impacts (materials, energy, and waste) from materials extraction to manufacturing to distribution/packaging to repair to recycling to disposal for a specific product
Pollution prevention[a] or source reduction[a]	product, process, or equipment design that emits fewer pollutants to air, water, and/or land
Waste minimization[a]	in-plant activities to reduce gas, liquid or solid waste
In-process recycling[a]	the nonproduct output is treated and fed back into the process
On-site recycling[a]	waste from a process is converted on-site as a raw material for a different product
Off-site recycling[a]	waste from a process is sent off-site, where it is converted to a raw material for a different product
Waste treatment[a]	waste is treated to lessen its toxicity
Secure disposal[a]	waste is sent to a secure landfill
Direct release[a]	waste is released directly into the environment

[a] These terms are adapted/republished with permission of John Wiley & Sons, Inc. from Allen and Rosselot.

environmental information (solid, liquid, and gaseous wastes). Waste auditing guides are available from state-funded programs (e.g., Pacific Northwest Pollution Prevention Resource Center 1999). Allen and Rosselot suggest that waste audits answer the following series of questions: What waste streams are generated by the facility? in what quantity? at what frequency? by which operations? under what legal restrictions or reporting requirements? by which inputs? at what efficiency? in what mixtures? (Allen and Rosselot 1997)

Waste audits require identification of solid wastes, wastewater, direct and secondary emissions. In the United States, solid wastes may be classified as nonhazardous or hazardous according to the Resource Conservation and Recovery Act (RCRA). In general, wastewater is the most significant component of total waste load Allen and Rosselot (1997). Several methods for estimating the rates of direct (fugitive) and secondary emissions are outlined with references for further information in Allen and Rosselot (1997).

Once companies have identified their major wastes and reduced them, they can turn their focus toward prevention. Pollution-prevention checklists and worksheets are provided in U.S. EPA (1992) and Cattanach et al. (1995). Process- and operations-based strategies for identifying and preventing waste are outlined in (Chadha 1994). Case studies sponsored by the U.S. Department of Energy NICE^3 program detail success stories for cleaner manufacturing or increased energy efficiency for several major industries (see Office of Industrial Technologies, NICE^3, www.oit.doe.gov/nice3/).

4.1.2. Energy Audits

Energy audits may be performed at either the facility or equipment level. Plant-wide energy audits are most common because utility bills summarize energy usage for the facility. Facility energy audits focus on characteristics of use such as occupancy profiles, fuel sources, building size and insulation, window and door alignment, ventilation, lighting, and maintenance programs. (Facility audit forms and checklists are available on the Web from the Washington State University Cooperative Extension Energy Program, www.energy.wsu.edu/ten/energyaudit.htm.) Some industries have developed specialized audit manuals. For example, an energy audit manual for the die-casting industry developed with funds from the state of Illinois and the Department of Energy describes how to assess energy use for an entire die casting facility (Griffith 1997). In addition to industry-specific energy consumption information, the U.S. Department of Energy Industrial Assessment Centers provide eligible small- and medium-sized manufacturers with free energy audits to help them identify opportunities to save energy and reduce waste (Office of Industrial Technologies 1999). Energy management is described in Chapter 58.

At the equipment level, energy usage may be determined through engineering calculations or monitors placed on the equipment in question. Identifying equipment with significant energy consumption may lead to actions such as adding insulation or performing maintenance.

Waste and energy audits are performed to identify existing problems. In the next four subsections, new approaches are presented that focus on prevention through life cycle design, life cycle assessment, production planning models with environmental considerations, and environmental management systems.

4.2. Life-Cycle Design*

The design and implementation of manufacturing activities and products have environmental impacts over time. Thus, industrial ecology requires consideration of the materials and energy consumption as well as effluents from resource extraction, manufacturing, use, repair, recycling, and disposal. Environmental considerations for product design and process design are summarized in the next two subsections.

4.2.1. Product Design

Product design guidelines for clean manufacturing are scattered throughout the industrial, mechanical, environmental, and chemical engineering, industrial design, and industrial ecology literature with labels such as “life-cycle design,” “design for environment (DFE),” “environmentally conscious design,” and “green design.” Traditionally, product design and materials selection criteria included geometric, mechanical, physical, economic, service environment, and manufacturing considerations. Industrial ecology introduces criteria such as reducing toxicity, avoiding resource depletion, increasing recyclability, and improving product upgradeability. The product design criteria in Table 4 are categorized by component and assembly level. As design functions for complex products are increasingly distributed, it is important to recognize product level considerations so that local, component design efforts are not cancelled out. For example, if a simple repair module is inaccessible, the design efforts for easy maintenance will be lost.

---

* This section has been adapted and reprinted with permission from Stuart and Sommerville (1997).

TABLE 4 Product Design Guidelines

	Component-Level Guidelines	Product-Level Guidelines
Process Stage	See Table 5	
Distribution Stage	• Minimize component volume and weight to minimize packaging and energy consumption. • Minimize special storage and transport requirements that lead to extra packaging (e.g., reduce fragility, sharp edges, and unusual shapes). • Design to avoid secondary, tertiary, and additional packaging levels. • Design for bulk packaging.	• Minimize product volume and weight to minimize packaging and energy consumption. • Minimize special storage and transport requirements that lead to extra packaging (e.g., reduce fragility, sharp edges, and unusual shapes). • Design to avoid secondary, tertiary, and additional packaging levels. • Design for bulk packaging.
Use Stage	• Design components with multiple functions. • Consider renewable or rechargeable energy sources. • Minimize energy consumption during start-up, use, and standby modes. • Minimize hazardous material content. • Minimize material content of dwindling world supply or requiring damaging extraction.	• Design product with multiple functions. • Consider renewable or rechargeable energy sources. • Minimize energy consumption during start-up, use, and standby modes. • Minimize use of hazardous joining materials. • Minimize toxicity, quantity, and number of different wastes and emissions.
Refurbishment Repair Upgrade	• Use standard components. • Consider replaceable components. • Use repairable components. • Maximize durability/rigidity. • Consider easily replaceable logos for second market. • Maximize reliability of components.	• Maximize ease of disassembly (access and separation). • Minimize orientation of components. • Design upgradeable modules. • Maximize durability/rigidity. • Consider easily replaceable logos for second market. • Maximize reliability of assembly.
Reclamation/materials recycling	• Maximize use of renewable and/or recyclable materials. • Avoid encapsulates, fillers, paint, sprayed metallic, coatings, labels, or adhesives that reduce recyclability. • Avoid hazardous materials.	• Minimize the number of different materials and different colors; minimize incompatible material combinations. • Use easily identifiable, separable materials; make high-value parts and materials easily accessible with standard tools. • Minimize number of different components and fasteners.

Adapted and reprinted with permission from Stuart and Sommerville (1997).

Because design decisions may result in environmental burden transfers from one stage in the life cycle to another or from one medium to another, it is important to recognize life cycle environmental impacts. Therefore, Table 4 is also categorized by five life-cycle stages: process, distribution, use, refurbishment, and recycling. Note that the process design stage for clean manufacturing is detailed separately in Section 4.2.2 and in Table 5.

One of the emerging themes in Table 4 is dematerialization. Dematerialization focuses on using fewer materials to accomplish a particular task (Herman et al. 1989). For example, consumers may subscribe to periodicals and journals on the Web rather than receive printed paper copies. Clearly, miniaturization, information technology, and the World Wide Web are increasing the potential for

TABLE 5 Process Design and Material Selection Guidelines

- Minimize use of materials with extensive environmental impacts.
- Minimize toxic emissions.
- Minimize material and water consumption.
- Minimize energy consumption.
- Consider materials that allow in-process, on-site, and off-site recycling.
- Perform preventive maintenance to reduce environmental impacts over time.
- Minimize secondary processes such as coatings.
- Eliminate redundant processes.
- Minimize cleaning process requirements.
- Capture wastes for recycling, treatment, or proper disposal.

Adapted and reprinted with permission from Stuart and Sommerville (1997).

dematerialization. Evaluating the criteria in Table 4 to avoid resource depletion or to use renewable materials requires assumptions to be made regarding the uncertainties in technological improvement, material substitutability, and rates of extraction and recycling (Keoleian et al. 1997).

The design criterion to increase product life reduces the number of products discarded over time. In the mid-1970s, DEC VT100 terminals could be disassembled quickly without tools to access the processor for maintenance and upgrades (Sze 2000). In 1999, the Macintosh G4 was introduced with a latch on the side cover that provides quick access to upgrade memory and other accessories (Apple 1999). Product life extension is especially important for products with short lives and toxic materials. For example, battery manufacturers extended the life of nickel–cadmium batteries (Davis et al. 1997). An example of product toxicity reduction was the change in material composition of batteries to reduce mercury content while maintaining performance (Tillman 1991). In another example, popular athletic shoes for children were redesigned to eliminate mercury switches when the shoes were banned from landfills (*Consumer Reports* 1994). The criteria related to recyclability may apply to product material content as well as the processing materials described in the next section.

4.2.2. Process Design

The criteria for process design for clean manufacturing focus on minimizing pollution, energy consumption, water consumption, secondary processes, or redundant processes. Table 5 provides a summary of suggested guidelines for materials selection and process design.

Careful process design can reduce environmental impacts and processing costs. For example, many companies eliminated the cleaning step for printed circuit card assembly by changing to low-solids flux and controlled atmospheres. These companies eliminated the labor, equipment, materials, and waste costs as well as the processing time associated with the cleaning step (Gutierrez and Tulkoff 1994; Cala et al. 1996; Linton 1995). Another example of reducing processing material consumption is recycling coolants used in machine shops. Recycling coolant reduces coolant consumption as well as eliminates abrasive metal particles that can shorten tool life or scar product surfaces (Waurzyniak 1999).

4.3. Product Life-Cycle Assessment*

Life-cycle assessment (LCA) is a three-step design evaluation methodology composed of inventory profile, environmental impact assessment, and improvement analysis (Keoleian and Menerey 1994). The purpose of the inventory step is to examine the resources consumed and wastes generated at all stages of the product life cycle, including raw materials acquisition, manufacturing, distribution, use, repair, reclamation, and waste disposal.

Materials and energy balance equations are often used to quantify the inputs and outputs at each stage in the product life cycle. Vigon et al. (1993) defines multiple categories of data for inventory analysis, including individual process, facility-specific, industry-average, and generic data. The most desirable form of data is the first data category, data collected from the process used for a specific product. However, this data category may require extensive personnel, expertise, time, and costs.

A three-step methodology for activity-based environmental inventory allocation is useful in calculating data for the first step of life cycle assessment. First, the process flow and system boundary are determined. Then the activity levels and the activity percentages of the inputs and outputs are identified. Finally, the activity percentages are used to determine the actual quantities of the inputs and outputs and assign them to the product and process design combination responsible for their generation. A detailed example of this method is given in Stuart et al. (1998). As industrial engineers assist companies in calculating the allocation of their wastes to the product responsible, they can help managers make more informed decisions about product and process design costs and environmental impacts.

---

*The following section is adapted and reprinted with permission from Stuart et al. (1998). Copyright MIT Press Journals.

Industry-average and generic data must be used with caution because processes may be run with different energy requirements and efficiencies or may exhibit nonlinear behavior (Barnthouse et al. 1998; Field and Ehrenfeld 1999). For example, different regions have different fuel-producing industries and efficiencies that will have a significant effect on the LCA if energy consumption is one of the largest impacts (Boustead 1995).

Once the inputs and outputs are determined, the second and third steps of LCA, impact analysis and improvement analysis, can be pursued (Fava et al. 1991). For impact analysis, the analyst links the inventory of a substance released to an environmental load factor such as acid deposition, which is defined in Table 6 (Potting et al. 1998). Environmental load factors are a function of characteristics such as location, medium, time, rate of release, route of exposure, natural environmental process mechanisms, persistence, mobility, accumulation, toxicity, and threshold of effect. Owens argues that because inventory factors do not have the spatial, temporal, or threshold characteristics that are inherent to the environmental load, other risk-assessment tools should be used to evaluate a local process (Owens 1997).

LCA software tools and matrices may be used to estimate environmental load factors for impact analysis (Graedel 1998; Graedel et al. 1995). Alting and Legarth (1995) review 18 LCA tools for database availability, impact assessment methodology, and complex product capability.

The results of life-cycle impact assessment (LCIA) provide relative indicators of environmental impact. Eight categories for LCIA are defined in Table 6. Details regarding how to use life cycle impact assessment categories are provided in Barnthouse et al. (1998) and Graedel and Allenby (1995).

Life cycle assessment is a comprehensive, quantitative approach to evaluate a single product. "An extensive survey of the use of mathematical programming to address environmental impacts for air, water, and land is given in [Greenberg (1995)]. A review of applied operations research

TABLE 6 Environmental Categories for Life-Cycle Impact Assessment

Environmental Category	Description
Greenhouse effect Global warming	Lower atmospheric warming from trapped solar radiation due to an abundance of CO_2, CH_4 (methane), N_2O, H_2O, $CFCl_3$, CF_2Cl_2, and O_3 (ozone) (Graedel and Crutzen 1990)
Ozone depletion	Losses in stratospheric ozone due to $CFCl_3$ and CF_2Cl_2 (Graedel and Crutzen 1990)
Photochemical smog	Reactions in the lower troposphere caused by emissions such as NO_x gases and hydrocarbons from automotive exhaust (Theodore and Theodore 1996)
Consumption of abiotic resources	Consumption of nonliving (nonrenewable) substances
Acid deposition	Acidic precipitation that deposits HNO_3 and H_2SO_4 into soil, water, and vegetation (Graedel and Crutzen 1990)
Eutrophication	Large deposits of phosphorous and nitrogen to a body of water that leads to excessive aquatic plant growth, which reduces the water's oxygen levels and capacity to support life (Theodore and Theodore 1996; Riviere 1990)
Ecotoxity	Substances that degrade the ecosystem either directly (acute) or over time (chronic) (Barnthouse et al. 1998; Graedel and Allenby 1995)
Habitat loss	Encroachment by humans into biologically diverse areas, resulting in the displacement and extinction of wildlife
Biodiversity	Living species that constitute the complex food web (Graedel and Allenby 1995)

These categories are reprinted with permission from Barnthouse et al. (1998). Copyright Society of Environmental Toxicology and Chemistry (SETAC), Pensacola, FL.

papers on supply chain analysis and policy analysis with respect to environmental management is in [Bloemhof-Ruwaard et al. (1995)]" (Stuart et al. 1999). Models for production planning with environmental considerations are summarized in the next section.

4.4. Production Planning with Environmental Considerations

"Introduction of product designs and process innovation requires a company to evaluate complex cost and environmental tradeoffs. In the past, these have not included environmental costs" (Stuart et al. 1999). In this section, production planning models are described for the entire product life cycle as well as for different stages of the product life cycle.

4.4.1. Models for Production Planning over the Product Life Cycle

Stuart et al. (1999) developed the first mixed integer linear programming model "to select product and process alternatives while considering tradeoffs of yield, reliability, and business-focused environmental impacts. Explicit constraints for environmental impacts such as material consumption, energy consumption, and process waste generation are modeled for specified assembly and disassembly periods. The constraint sets demonstrate a new way to define the relationship between assembly activities and disassembly configurations through take-back rates. Use of the model as an industry decision tool is demonstrated with an electronics assembly case study in Stuart et al. (1997). Manufacturers may run "what if" scenarios for proposed legislation to test the effects on design selection and the bottom line cost impacts." The effects over time of pollution prevention or product life extension are analyzed from a manufacturer's and potentially lessor's perspective. Several new models explore the relationship between product and component reuse and new procurement. These models include deterministic approaches using mixed integer linear programming (Eskigun and Uzsoy 1998) and stochastic approaches using queueing theory (Heyman 1977) and periodic review inventory models (Inderfurth 1997; van der Laan and Salomon 1997). Location of remanufacturing facilities are analyzed in Jayaraman (1996); Bloemhof-Ruwaard et al. (1994, 1996); Fleischmann et al. (1997). Scheduling policies for remanufacturing are presented in Guide et al. (1997).

4.4.2. Production Planning Models for the Manufacturing and Assembly Stage

Early models focused on reducing the environmental impacts concurrent with process planning for continuous processes in the petroleum and steel industries (Russell 1973; Russell and Vaughan 1974). Recently, models with environmental considerations focus on process planning for discrete product manufacturing (Bennett and Yano 1996, 1998; Sheng and Worhach 1998).

4.4.3. Disassembly Planning Models

Based on graph theory, Meacham et al. (1999) present a fast algorithm, MAXREV, to determine the degree of disassembly for a single product. They are the first to model selection of disassembly strategies for multiple products subject to shared resource constraints. They use their MAXREV algorithm to generate maximum revenue disassembly configurations for their column generation procedure for multiple products. Other disassembly models based on graph theory approaches focus on determining economic manual disassembly sequences for a single product (Ron and Penev 1995; Penev and Ron 1996; Zhang and Kuo 1996; Johnson and Wang 1995; Lambert 1997). A process planning approach to minimize worker exposure hazards during disassembly is given in Turnquist et al. (1996). Disassembly may be economically advantageous for module and component reuse. However, for material recovery, escalating labor costs favor bulk recycling.

4.4.4. Production Planning Models for Bulk Recycling

Production planning for bulk recycling is in the early stages of development. Models include a macro-level transportation model for paper recycling (Glassey and Gupta 1974; Chvatal 1980) and a goal-programming model for recycling a single product (Hoshino et al. 1995). Sodhi and Knight (1998) develop a dynamic programming model for float–sink operations to separate materials by density. Spengler et al. (1997) present a mixed integer linear programming model to determine the manual disassembly level and recycling quantity. Stuart and Lu (2000) develop a multicommodity flow model to select the output purity by evaluating various processing and reprocessing options for bulk recycling of end-of-life products. Isaacs and Gupta (1998) use goal programming to maximize disassembler and shredder profits subject to inventory balance constraints for the automobile recycling problem. Krikke et al. (1998) propose dynamic programming for disassembly planning and an algorithm to maximize revenue from material recycling. Realff et al. (1999) use mixed integer linear programming to select sites and determine the amount of postconsumer material collected, processed, stored, shipped, and sold at various sites.

4.5. Environmental Management Systems

An environmental management system (EMS) is a management structure that addresses the long-term environmental impact of a company's products, services, and processes. An EMS framework should include the following four characteristics:

1. Environmental information collection and storage system
2. Management and employee commitment to environmental performance
3. Accounting and decision processes that recognize environmental costs and impacts
4. Commitment to continuous improvement of environmental performance

Federal U.S. EMS guidelines and information are documented in (Department of Energy 1998). International standards for EMS will be discussed in Section 4.5.3. Environmental management systems (EMS) include environmental policies, goals, and standards, which are discussed in the next three subsections.

4.5.1. Corporate Environmental Policies

Corporate environmental policies require the commitment and resources of senior management. These policies often focus on actions that can prevent, eliminate, reduce, reuse, and recycle, respectively. These policies should be incorporated into all employees' practices and performance evaluations. Communication of environmental policies and information is integral to the success of the policies. Setting viable goals from corporate environmental policies is the subject of the next section.

4.5.2. Environmental Goals and Metrics

Traditional environmental metrics often focus on compliance with legislation. Goals may concentrate on state-dependent benchmark metrics such as reducing emissions, reducing the volume or mass of solid waste, or reducing gallons of waste water to a specified level. On the other hand, goals may focus on non-state-dependent improvement metrics such as reducing the percentage of environmental treatment and disposal costs. It is also important to distinguish between local and aggregate data when developing goals.

Metrics may focus on local product or process goals or system-wide facility or company goals. An example of a local goal might be to lengthen tool life and reduce cutting fluid waste disposal costs. Sometimes local goals may translate to system goals. One machinist's use of a new oil-free, protein-based cutting fluid that eliminates misting and dermatitis problems but provides the necessary lubricity and cooling may be a candidate for a system-wide process and procurement change (Koelsch 1997). With local goals, it is important to investigate their potential positive and negative impacts if implemented throughout the system. An example of a system-wide goal might be to reduce the percentage of polymer sprues and runners discarded at a particular facility by implementing regrinding and remolding or redesigning the mold. For system goals, it is important to identify the most significant contributors through Pareto analysis and target them for improvement.

4.5.3. ISO 14000 Series Standards

ISO 14001, ''an international standard describing the basic elements of an environmental management system, calls for identification of the environmental aspects and impacts of a company's products, processes, and services [Block 1997]. Industrial engineers may develop the information systems to quantify environmental aspects such as input materials, discharges, and energy consumption [Alexander 1996]'' (Stuart et al. 1998).

5. CONCLUDING REMARKS

Clean manufacturing is an important concept to integrate into industrial engineering methodologies. Traditional waste and energy audits help companies identify cost and environmental savings opportunities. New concepts such as life cycle design, product life cycle assessment, production planning with environmental considerations, and environmental management systems help companies to prevent costly negative environmental impacts. In summary, clean manufacturing provides opportunities for increased efficiencies and cost effectiveness as well as movement towards sustainability.

Acknowledgments

This chapter was written while the author was supported by a National Science Foundation CAREER award (NSF DDM-9734310). This chapter benefited considerably from the comments of the reviewers and research assistants Mark Hartman, Qin Lu, Jianhong Qiao, Vivi Christina, and Christiana Kuswanti.

REFERENCES

Alexander, F. (1996), "ISO 14001: What Does It Mean for IEs?" *IIE Solutions,* January, pp. 14–18.

Allen, D. T., and Rosselot, K. S. (1997), *Pollution Prevention for Chemical Processes,* John Wiley & Sons, New York.

Allenby, B. R. (1999), *Industrial Ecology: Policy Framework and Implementation,* Prentice Hall, Upper Saddle River, NJ.

Alting, L., and Legarth, J. B. (1995), "Life Cycle Engineering and Design," *CIRP General Assembly,* Vol. 44, No. 2, pp. 569–580.

Andersen, S. O. (1990), "Progress by the Electronics Industry on Protection of Stratospheric Ozone." in *Proceedings of the 40th IEEE Electronic Components and Technology Conference* (Las Vegas, May 21–23), IEEE, New York, pp. 222–227.

Apple Computer, Inc., "Power Mac G4 Data Sheet," December 1999, www.apple.com/powermac/pdf/PowerMac_G4_DS-e.pdf

Barnthouse, L., Fava, J., Humphreys, K., Hunt, R., Laibson, L., Noesen, S., Norris, G., Owens, J., Todd, J., Vigon, B., Weitz, K., and Young, J. (1998). *Life Cycle Impact Assessment: The State of the Art,* Society of Environmental Toxicology and Chemistry, Pensacola, FL.

Bateman, B. O. (1999a), "Sector-Based Public Policy in the Asia-Pacific Region: Planning, Regulating, and Innovating for Sustainability," United States–Asia Environmental Partnership, Washington, DC.

Bateman, B. O. (1999b), personal communication.

Bennett, D., and Yano, C. (1996), "Process Selection Problems in Environmentally Conscious Manufacturing," *INFORMS Atlanta,* Invited Paper.

Bennett, D., and Yano, C. A. (1998), "A Decomposition Approach for a Multi-Product Process Selection Problem Incorporating Environmental Factors," Working Paper, University of California–Berkeley.

Block, M. R. (1997), *Implementing ISO 14001,* ASQC Quality Press, Milwaukee.

Bloemhof-Ruwaard, J. M., Solomon, M., and Wassenhove, L. N. V. (1994), "On the Coordination of Product and By-Product Flows in Two-Level Distribution Networks: Model Formulations and Solution Procedures," *European Journal of Operational Research,* Vol. 79, pp. 325–339.

Bloemhof-Ruwaard, J. M., Solomon, M., and Wassenhove, L. N. V. (1996), "The Capacitated Distribution and Waste Disposal Problem," *European Journal of Operational Research,* Vol. 88, pp. 490–503.

Boustead, I. (1995), "Life-Cycle Assessment: An Overview," *Energy World,* No. 230, pp. 7–11.

Cala, F., Burress, R., Sellers, R. L., Iman, R. L., Koon, J. F., and Anderson, D. J. (1996), "The No-Clean Issue," *Precision Cleaning,* Vol. 4, No. 2, pp. 15–24.

Cattanach, R. E., Holdreith, J. M., Reinke, D. P., and Sibik, L. K. (1995), *The Handbook of Environmentally Conscious Manufacturing,* Irwin Professional Publishing, Chicago.

Chadha, N. (1994), "Develop Multimedia Pollution Prevention Strategies," *Chemical Engineering Progress,* Vol. 90, No. 11, pp. 32–39.

Chvatal, V. (1980), *Linear Programming,* W.H. Freeman & Co., New York.

Consumer Reports (1994), "Light-up Sneakers: A Dim-bulb Idea Takes a Hike."

Davis, G. A., Wilt, C. A., Dillon, P. S., and Fishbein, B. K. (1997), "Extended Product Responsibility: A New Principle for Product-Oriented Pollution Prevention." *EPA-530-R-97-009,* Knoxville, TN.

Davis, J. B. (1997). "Regulatory and Policy Trends," Cutter Information Corp., Arlington, MA, pp. 10–13.

Denmark Ministry of the Environment (1992), *Cleaner Technology Action Plan.*

Department of Energy (1998), "Environmental Management Systems Primer for Federal Facilities," Office of Environmental Policy & Assistance, es.epa.gov/oeca/fedfac/emsprimer.pdf.

Eskigun, E., and Uzsoy, R. (1998), "Design and Control of Supply Chains with Product Recovery and Remanufacturing." Under revision for *European J. of Operational Res.*

European Environmental Agency (1997), "New Environmental Agreements in EU Member States by Year: Pre-1981 through 1996," in *Environmental Agreements: Environmental Effectiveness,* No. 3, Vol. 1, European Communities, Copenhagen.

Fava, J. A., Denison, R., Jones, B., Curran, M. S., Vigon, B. W., Selke, S., and Barnum, J. (1991). "A Technical Framework for Life-cycle Assessments," Society of Environmental Toxicology and Chemistry Foundation, Pensacola.

Field, F. R., III, and Ehrenfeld, J. R. (1999), "Life-Cycle Analysis: The Role of Evaluation and Strategy," in *Measures of Environmental Performance and Ecosystem Condition,* P. Schulze, Ed., National Academy Press, Washington, DC.

Fleischmann, M., Bloemhof-Ruwaard, J. M., Dekker, R., van der Laan, E., van Nunen, J. A. E. E., and van Wassenhove, L. N. (1997), "Quantitative Models for Reverse Logistics: A Review," *European Journal of Operational Research,* Vol. 103, No. 1, pp. 1–17.

Glassey, C. R., and Gupta, V. K. (1974), "A Linear Programming Analysis of Paper Recycling," *Management Science,* Vol. 21, pp. 392–408.

Graedel, T. E. (1998), *Streamlined Life-Cycle Assessment,* Prentice Hall, Upper Saddle River. NJ.

Graedel, T. E., and Allenby, B. R. (1995), *Industrial Ecology,* Prentice Hall, Englewood Cliffs, NJ.

Graedel, T. E., and Crutzen, P. J. (1990), "The Changing Atmosphere," in *Managing Planet Earth: Readings from Scientific American Magazine,* W.H. Freeman & Co., New York.

Graedel, T. E., Allenby, B. R., and Comrie, P. R. (1995), "Matrix Approaches to Abridged Life Cycle Assessment," *Environmental Science & Technology,* Vol. 29, No. 3, pp. 134A–139A.

Granda, R. E., Hermann, F., Hoehn, R., and Scheidt, L.-G. (1998), "Product-Related Environmental Attributes: ECMA TR/70—An Update," in *Proceedings of the International Symposium on Electronics and the Environment* (Oak Brook, IL), pp. 1–3.

Greenberg, H. J. (1995), "Mathematical Programming Models for Environmental Quality Control," Operations Research, Vol. 43, No. 4, pp. 578–622.

Griffith, L. E. (1997), "Conducting an Energy Audit for a Die Casting Facility," NADCA, Washington, DC.

Guide, V. D. R., Jr., Srivastava, R., and Spencer, M. S. (1997), "An Evaluation of Capacity Planning Techniques in a Remanufacturing Environment," *International Journal of Production Research,* Vol. 35, No. 1, pp. 67–82.

Gutierrez, S., and Tulkoff, C. (1994), "Benchmarking and QFD: Accelerating the Successful Implementation of No Clean Soldering," in *Proceedings of the IEEE/CPMT Int. Electronics Manufacturing Technology Symposium,* pp. 389–392.

Herman, R., Ardekani, S., and Ausubel, J. (1989), "Dematerialization," in *Technology and Environment,* J. H. Ausubel and H. E. Sladovich, Eds., National Academy Press, Washington, DC, pp. 50–69.

Heyman, D. P. (1977), "Optimal Disposal Policies for a Single-item Inventory System with Returns," *Naval Research Logistics Quarterly,* Vol. 24, pp. 385–405.

Horkeby, I. (1997). "Environmental Prioritization," in *The Industrial Green Game: Implications for Environmental Design and Management,* D. J. Richards, Ed., National Academy Press, Washington, DC, pp. 124–131.

Hoshino, T., Yura, K., and Hitomi, K. (1995), "Optimisation Analysis for Recycle-oriented Manufacturing Systems," *International Journal of Production Research,* Vol. 33, No. 8, pp. 2069–2078.

Inderfurth, K. (1997), "Simple Optimal Replenishment and Disposal Policies for a Product Recovery System with Leadtimes," *OR Spektrum,* Vol. 19, pp. 111–122.

Isaacs, J. A., and Gupta, S. (1998), "Economic Consequences of Increasing Polymer Content for the US Automobile Recycling Infrastructure," *Journal of Industrial Ecology,* Vol. 1, No. 4, pp. 19–33.

Jayaraman, V. (1996), "A Reverse Logistics and Supply Chain Management Model within a Remanufacturing Environment," *INFORMS, Atlanta,* Invited Paper.

Johnson, J. K., White, A., and Hearne, S. (1997), "From Solvents to Suppliers: Restructuring Chemical Supplier Relationships to Achieve Environmental Excellence," in *Proceedings of the Int. Symposium on Electronics and the Environment* (San Francisco), pp. 322–325.

Johnson, M., and Wang, M. (1995), "Product Disassembly Analysis: A Cost/Benefit Trade-Off Approach," *International Journal of Environmentally Conscious Design and Manufacturing,* Vol. 4, No. 2, pp. 19–28.

Keoleian, G. A., and Menerey, D. (1994), "Sustainable Development by Design: Review of Life Cycle Design and Related Approaches," *Journal of the Air & Waste Management Association,* Vol. 44, No. 5, pp. 645–668.

Keoleian, G., Kar, K., Manion, M. M., and Bulkley, J. W. (1997), *Industrial Ecology of the Automobile: A Life Cycle Perspective,* Society of Automotive Engineers, Warrendale, PA.

Koelsch, J. R., (1997), "Lubricity vs. the Environment: Cascades of Cleanliness," *Manufacturing Engineering,* May, Vol. 118, pp. 50–57.

Krikke, H. R., Harten, A. V., and Schuur, P. C. (1998), "On a Medium Term Product Recovery and Disposal Strategy for Durable Assembly Products," *International Journal of Production Research,* Vol. 36, No. 1, pp. 111–139.

Lambert, A. J. D. (1997), "Optimal Disassembly of Complex Products," *International Journal of Production Research,* Vol. 35, pp. 2509–2523.

Linton, J. (1995), "Last-Minute Cleaning Decisions," *Circuits Assembly,* November, pp. 30–35.

Meacham, A., Uzsoy, R., and Venkatadri, U. (1999), "Optimal Disassembly Configurations for Single and Multiple Products," *Journal of Manufacturing Systems,* Vol. 18, No. 5, 311–322.

Modl, A. (1995), "Common Structures in International Environmental Labeling Programs," in *Proceedings of the IEEE International Symposium on Electronics and the Environment* (Orlando, FL), pp. 36–40.

Norberg-Bohm, V., Clark, W. C., Bakshi, B., Berkenkamp, J., Bishko, S. A., Koehler, M. D., Marrs, J. A., Nielson, C. P., and Sagar, A. (1992), "International Comparisons of Environmental Hazards: Development and Evaluation of a Method for Linking Environmental Data with the Strategic Debate Management Priorities for Risk Management," CSIA Discussion Paper 92-09, Kennedy School of Government, Harvard University, Cambridge, MA.

Office of Industrial Technologies (1999), "Industrial Assessment Centers," U.S. Department of Energy, www.oit.doe.gov/iac/.

Owens, J. W. (1997), "Life Cycle Assessment: Constraints on Moving from Inventory to Impact Assessment," *Journal of Industrial Ecology,* Vol. 1, No. 1, pp. 37–49.

Pacific Northwest Pollution Prevention Resource Center (1999), "Business Assistance: How to Inventory Your Wastes for Environmental Compliance," Pacific Northwest Pollution Prevention Resource Center, www.pprc.org/pprc/sbap/workbook/toc_all.html.

Penev, K. D., and Ron, A. J. D. (1996), "Determination of a Disassembly Strategy," *International Journal of Production Research,* Vol. 34, No. 2, pp. 495–506.

Potting, J., Schlopp, W., Blok, K., and Hauschild, M. (1998), "Site-Dependent Life-Cycle Impact Assessment of Acidification," *Journal of Industrial Ecology,* Vol. 2, No. 2, pp. 63–87.

President's Council on Sustainable Development (1996), *Sustainable America,* President's Council on Sustainable Development, Washington, DC.

Realff, M. J., Ammons, J. C., and Newton, D. (1999), "Carpet Recycling: Determining the Reverse Production System Design," *Journal of Polymer-Plastics Technology and Engineering,* Vol. 38, No. 3, pp. 547–567.

Riviere, J. W. M. L. (1990), "Threats to the World's Water," in *Managing Planet Earth: Readings from Scientific American Magazine,* W.H. Freeman & Co., New York.

Ron, A. D., and Penev, K. (1995), "Disassembly and Recycling of Electronic Consumer Products: An Overview," *Technovation,* Vol. 15, No. 6, pp. 363–374.

Russell, C. S. (1973), *Residuals Management in Industry: A Case Study in Petroleum Refining,* Johns Hopkins University Press, Baltimore.

Russell, C. S., and Vaughan, W. J. (1974), "A Linear Programming Model of Residuals Management for Integrated Iron and Steel Production," *Journal of Environmental Economics and Management,* Vol. 1, pp. 17–42.

Ryding, S. O., Steen, B., Wenblad, A., and Karlsson, R. (1993), "The EPS System: A Life Cycle Assessment Concept for Cleaner Technology and Product Development Strategies, and Design for the Environment," in *Proceedings of the Design for the Environment: EPA Workshop on Identifying a Framework for Human Health and Environmental Risk Ranking* (Washington, DC), pp. 21–23.

Santi, J. (1997), *Directory of Pollution Prevention in Higher Education: Faculty and Programs,* University of Michigan, Ann Arbor.

Sheng, P., and Worhach, P. (1998), "A Process Chaining Approach toward Product Design for Environment," *Journal of Industrial Ecology,* Vol. 1, No. 4, pp. 35–55.

Socolow, R. (1999), personal communication.

Sodhi, M., and Knight, W. A. (1998), "Product Design for Disassembly and Bulk Recycling," *Annals of the CIRP,* Vol. 47, No. 1, pp. 115–118.

Spengler, T., Puchert, H., Penkuhn, T., and Rentz, O. (1997), "Environmental Integrated Production and Recycling Management," *European Journal of Operational Research,* Vol. 97, pp. 308–326.

Stuart, J. A. (2000), "Integration of Industrial Ecology Concepts into Industrial Engineering Curriculum," *Proceedings of the American Society of Engineering Education Annual Conference,* St. Louis.

Stuart, J. A., and Lu, Q. (2000), "A Model for Discrete Processing Decisions for Bulk Recycling of Electronics Equipment," IEEE Transactions on Electronics Packaging Manufacturing, Vol. 23, No. 4, pp. 314–320.

Stuart, J. A., and Sommerville, R. M. (1997), "A Review of Life Cycle Design Challenges," *International Journal of Environmentally Conscious Design and Manufacturing,* Vol. 7, No. 1, pp. 43–57.

Stuart, J. A., Turbini, L. J., and Ammons, J. C. (1997), "Investigation of Electronic Assembly Design Alternatives through Production Modeling of Life Cycle Impacts, Costs, and Yield," *IEEE Transactions for Components, Packaging, and Manufacturing Technology, Part C: Manufacturing,* Vol. 20, No. 4, pp. 317–326.

Stuart, J. A., Turbini, L. J., and Ammons, J. C. (1998), "Activity-Based Environmental Inventory Allocation," *Journal of Industrial Ecology,* Vol. 2, No. 3, pp. 95–108.

Stuart, J. A., Ammons, J. C., and Turbini, L. J. (1999), "A Product and Process Selection Model with Multidisciplinary Environmental Considerations," *Operations Research,* Vol. 47, No. 2, pp. 221–234.

Sze, Cedric (2000), personal communication.

Theodore, M. K., and Theodore, L. (1996), *Major Environmental Issues Facing the 21st Century,* Prentice Hall PTR, Upper Saddle River, NJ.

Tillman, J. (1991), "Achievements in Source Reduction and Recycling for Ten Industries in the United States," EPA/600/2-91/051, Cincinnati, OH.

Turnquist, M. A., List, G. F., Kjeldgaard, E. A., and Jones, D. (1996), "Planning Tools and Techniques for Product Evaluation and Disassembly," *INFORMS,* Atlanta, Invited Paper.

U.S. Environmental Protection Agency (1992), "Facility Pollution Prevention Guide," EPA/600/R-92/088, Washington, DC.

van der Laan, E., and Salomon, M. (1997), "Production Planning and Inventory Control with Remanufacturing and Disposal," *European Journal of Operational Research,* Vol. 102, pp. 264–278.

Vigon, B. W., Tolle, D. A., Cornaby, B. W., Latham, H. C., Harrison, C. L., Boguski, T. L., Hunt, R. G., Sellers, J. D., and Curran, M. A. (1993), "Life-Cycle Assessment: Inventory Guidelines and Principles," EPA/600/R-92/245, Cincinnati, OH.

Waurzyniak, P. (1999). "Recycle Your Coolant, Chips," *Manufacturing Engineering,* March, pp. 110–116.

Zhang, H., and Kuo, T. (1996), "Graph-Based Approach to Disassembly Model for End-Of-Life Product Recycling," in *Proceedings of the 1996 IEEE/CMPT International Electronic Manufacturing Technology Symposium* (Austin, TX), pp. 247–254.

CHAPTER 17

Just-in-Time, Lean Production, and Complementary Paradigms

Takao Enkawa
Tokyo Institute of Technology

Shane J. Schvaneveldt
Tokyo Institute of Technology/Weber State University

1. INTRODUCTION AND OVERVIEW 544
2. THREE PILLARS OF JUST-IN-TIME AND THE TOYOTA PRODUCTION SYSTEM 545
    2.1. Smoothing of Volume and Variety 545
    2.2. Development of Flexible, Multiskilled Workforce 547
    2.3. Continuous Improvement and *Autonomation* 548
3. KANBAN AS A DECENTRALIZED CONTROL SYSTEM FOR JUST-IN-TIME 549
    3.1. Prerequisites and Role of Kanban System 549
    3.2. Control Parameters of Kanban System 550
    3.3. Kanban's Limitations and Alternatives 550
    3.4. Case Study of JIT/Kanban Implementation 551
4. COMPLEMENTARY PARADIGMS OF JUST-IN-TIME 551
    4.1. 3T's as Mutually Reinforcing Activities for Quality, Cost, Delivery Performance 551
    4.2. Total Quality Management (TQM) 552
    4.3. Total Productive Maintenance (TPM) 553
    4.4. Joint Implementation of the 3T's and Case Study 553
5. LEAN PRODUCTION AND EXTENSIONS OF JUST-IN-TIME 555
    5.1. Lean Production 555
    5.2. Theory of Constraints (TOC) and JIT 557
    5.3. Applications to Service Industries 559
REFERENCES 559

1. INTRODUCTION AND OVERVIEW

The just-in-time (JIT) concept was first developed and implemented over a span of many years by Toyota Motor Corporation under the appellation of the Toyota Production System (TPS). Its overarching goal is to enable production of a variety of end items in a timely and efficient manner, smoothly synchronized with the production and delivery of component materials and without reliance on the conventional stratagem of keeping extra work-in-process and finished goods inventory. By the 1970s, Toyota's just-in-time system had evolved into a remarkable source of competitive advantage, achieving higher quality and productivity and lower cost than traditional mass production systems. Indeed, the Toyota Production System was widely credited with buoying Toyota through the economic turmoil following the 1973 oil crisis. Though originally limited mostly to Toyota and its supplier network, the TPS concept spread to other Japanese manufacturers and by the early 1980s was re-

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

ceiving tremendous attention from manufacturers worldwide. Over the years, a number of labels, such as stockless production and zero inventory, have been applied to TPS-like approaches for managing production and inventory. JIT, however, has become the accepted, albeit imprecise, term.

A broader conceptualization of world-class manufacturing called ''lean production'' was realized through the landmark studies conducted by MIT's International Motor Vehicle Program. This lean production concept was created based on extensive benchmarking of automobile manufacturers' best practices worldwide—arguably, principally those of Japanese manufacturers and the Toyota Production System. From a practical standpoint, lean production may be considered as an expanded view of JIT/TPS that includes additional intraorganizational and interorganizational aspects such as the role of suppliers and information sharing in not only the manufacturing stage but also in product development and distribution.

In the remainder of this chapter, we first provide a more detailed discussion of JIT/TPS in Section 2, including its philosophy and implementation issues. In Section 3, we examine the kanban system, widely used in JIT for control of production and inventory, and present a case study of JIT/kanban implementation. Section 4 follows with an examination of JIT's relation to complementary approaches such as total quality management (TQM) and total productive maintenance (TPM), together with a case study of their joint implementation. In Section 5, we examine lean production as an extension of JIT, explore the relationship of JIT to theory of constraints (TOC), and conclude with a brief consideration of applications to service industries of JIT, TOC, and other manufacturing-based approaches.

2. THREE PILLARS OF JUST-IN-TIME AND THE TOYOTA PRODUCTION SYSTEM

In broad terms, JIT is a management philosophy that seeks manufacturing excellence with an emphasis on eliminating waste from all aspects of the production system. At its most basic level, a JIT system produces only what is used or sold, in the needed quantity and at the needed time. Accordingly, low levels of work-in-process and finished goods inventory are a prominent feature of JIT. Though it can be said that other production management approaches, such as material requirements planning (MRP), share this same overall goal, JIT differs in that it focuses on inventory reduction not only as an *end* but as a purposeful *means* to foster broader improvements in performance. Thus, JIT may be considered a dynamic system, always seeking to achieve still higher levels of performance. In contrast, conventional systems such as MRP or models such as economic order quantity (EOQ) are typically static in that they derive a solution for a given set of conditions (such as lead times or setup costs) and make no consideration for improvement of those conditions.

Another typical element of JIT is the use of a pull method of production coordination, such as the well-known kanban system. In a pull system, production is initiated only to replenish what has been actually used at the next stage of the production system (or sold to the customer). This is a reversal of the push concept, in which production is initiated in anticipation of future demand. Kanban functions as a pull system in that, as materials are used in a downstream stage of the production system, replenishment orders for component materials are relayed to the upstream stages in a progressive cascade upwards. Because actual usage of materials downstream is the only trigger for making more of something upstream, production is initiated only when needed and automatically stopped when the demand ceases. Thus, kanban may be viewed as a decentralized, self-regulating system for controlling production and material flows, even in a complex manufacturing environment with thousands of discrete parts, such as an auto assembly plant. It should be noted, however, that kanban is most appropriate for high-volume, repetitive manufacturing environments and that it is only one means for implementing JIT.

The success of JIT, as well as kanban, is contingent on meeting several prerequisites: (1) smoothing of volume and variety; (2) development of a flexible, multiskilled workforce; and (3) implementation of continuous improvement and *autonomation*. These prerequisites are discussed in the sections below. In addition, achievement of high quality levels is also essential to implement JIT. For this purpose, JIT and TQM efforts should be closely linked with each other, as will be discussed later.

Though some references for further information on JIT are provided in this chapter, no attempt is made to review the thousands of articles and books now available. For more exhaustive reviews of the early JIT literature, the reader is referred to Keller and Kazazi (1993), Golhar and Stamm (1991), and Sohal et al. (1989). Reviews of research focusing on the modeling/design of JIT and kanban systems are provided by Akturk and Erhun (1999) and Groenevelt (1993). General references on Toyota's JIT system include the first publication in 1977 by Sugimori et al., as well as Ohno (1988), Shingo (1989), Monden (1998), and Fujimoto (1999).

2.1. Smoothing of Volume and Variety

Production leveling—in other words, smoothing of volume as well as variety to achieve uniform plant loading—is imperative for JIT implementation. What would happen if the production volume of an item were to fluctuate every day and we blindly pursued a pull policy of withdrawing the needed quantity of parts at the needed time from the preceding process? In that situation, the preceding process must always maintain an inventory and workforce sufficient to meet the highest possible quantity demanded. Consequently, its production volume will exhibit fluctuations larger than those of the downstream process. Since a typical production system involves many sequential processes, there are many stages for these fluctuations to be transmitted from final assembly backward through to the suppliers. At each stage, the fluctuation is transmitted in amplified form to the previous stage, with the result that upstream processes and suppliers may incur vast inventory and waste. This is the so-called bullwhip effect known in the field of supply chain management.

In order to prevent these undesirable effects, an effort must be made to smooth out the production quantities of each item in the final assembly schedule and then keep that schedule fixed for at least some period of time. Smoothing the production schedule minimizes variation in the quantities of parts needed and enables the preceding processes to produce each part efficiently at a constant speed per hour or per day.

Figure 1 shows a simple illustration of production leveling on a daily basis. Based on demand forecasts and firm customer orders, a three-month production schedule is created and shared with relevant suppliers. Each month the production schedule is revised on a rolling basis to reflect the latest information. The first month of the schedule is then frozen and broken down into a daily production schedule as follows.

Suppose that the monthly production schedule for final assembly indicates demand for 2400 units of product A, 1200 units of product B, and 1200 units of product C. The level daily production schedule is determined by dividing these monthly demand quantities by 20 working days per month, that is, 120 A's, 60 B's, and 60 C's per day, along with the corresponding numbers of parts. Minor adjustment of these quantities may occur on, say, a weekly basis, but alterations must be small (within ± 10% empirically) so as to avoid introducing excess fluctuations into the system. Though suppliers use this leveled daily production schedule as a guideline for planning, the actual day-to-day production quantities are determined by actual demand from downstream stages and coordinated through the kanban system or similar approach. By initiating production only as demand occurs, kanban eliminates the possibility of waste due to excess production while absorbing minor, unavoidable fluctuations on the factory floor.

Next, the daily production schedule is smoothed for product variety through the generation of a production sequence. Rather than all A's being made in one batch, then all B's in a following batch, and so on, a mix of items should flow through the system. Because the demand for A's, B's, and C's has a ratio of 2:1:1 in this case, an appropriate mix or sequence for their production would be a cycle of A B A C repeated through the day. For determining sequences in more complex situations, refer to algorithms in Monden (1998) and Vollmann et al. (1997).

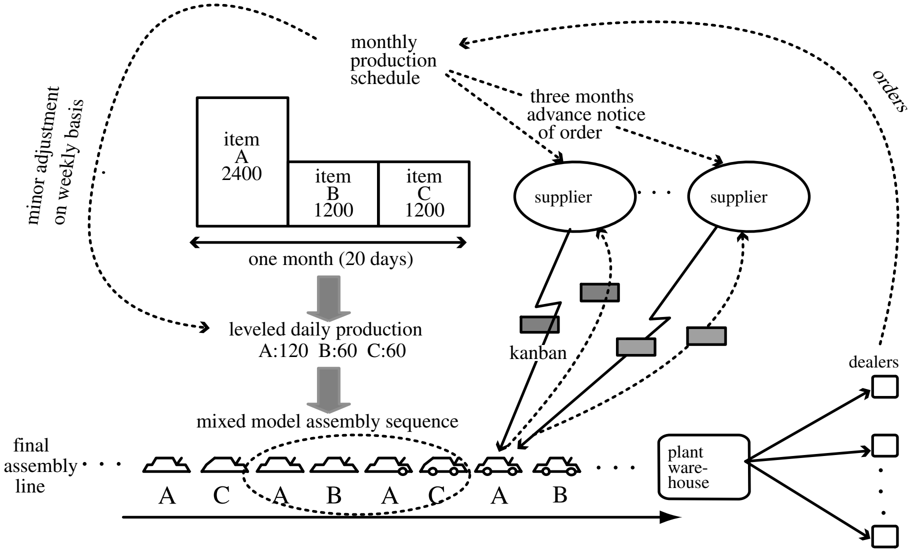

monthly production schedule
minor adjustment on weekly basis
orders
three months advance notice of order
item A 2400
item B 1200
item C 1200
one month (20 days)
supplier
supplier
leveled daily production A:120 B:60 C:60
kanban
dealers
mixed model assembly sequence
final assembly line
A C A B A C A B
plant ware-house

Figure 1 Smoothing Volume and Variety to Achieve One-Piece Flow Production.

Using this kind of mixed model sequencing, the line balance can be kept close to 100 percent even though the standard time or cycle time varies for each product. For example, suppose that the line's *takt* time (cycle time to meet demand) is two minutes and the respective standard times of products A, B, and C are two minutes, three minutes, and one minute for a certain process. Then the cycle time for the A B A C sequence is balanced and equal to eight minutes. The individual time differences between products in the sequence can be accommodated by utilizing a space buffer between adjacent processes.

The ultimate goal is to have a production system so responsive that products can be made in a batch size of one and scheduled according to the market's demand rate. This is also called one-piece flow production. However, to achieve one-piece flow, it is obvious that setup or changeover times must be short for all processes involved. The benchmark of reducing setup times to less than 10 minutes is well known. This is also known as single-minute exchange of dies (SMED) or rapid setup and has a number of associated techniques (e.g., Shingo 1985, 1989). Of these, the most important is to distinguish between internal setup tasks that must be performed while the machine is stopped and external setup tasks that can be done in parallel with continued operation. Once short setup times are realized, production lot size and lead time can be reduced, enabling upstream stages to work in response to the needs of the leveled final assembly schedule.

2.2. Development of Flexible, Multiskilled Workforce

Under one-piece flow production, different operations are potentially required for each succeeding item being produced. For it to be effective, workers must have the skills necessary for handling multiple operations properly. Consequently, a prerequisite for JIT is the development of multiskilled workers who can cope with the frequent changes in products and operations seen in a one-piece flow production environment.

In an automobile engine plant, for example, a multiskilled operator is typically responsible for several semiautomatic machine tools. This system is called multiprocess holding. Combining multiprocess holding together with layout approaches such as U-shaped lines enables production efficiency to be maintained even when production volume varies. To illustrate, consider a fabrication line consisting of 12 machines and 6 operators and suppose that the production quantity is decreased from 100 to 80 per day due to reduced demand. Labor productivity still can be maintained if the operations can be rearranged to be shared by 4 multiskilled operators. This makes sense only with the existence of multiskilled operators capable of handling any of the 12 machines' operations. U-shaped line layouts can facilitate flexible sharing of the operations by minimizing operator movement between the machines.

Figure 2(*a*) illustrates the type of changes seen in automobile final assembly lines. They have evolved from long straight lines into U-shaped lines and more recently have been rearranged into many short lines contrived so that each line consists of similar operations. The major purpose of this arrangement is to enable workers to learn multiple skills quickly and efficiently because they can easily learn the other operations within a line once they master one of its operations. After becoming a multioperation worker, he or she can move to another line to become a multifunction worker.

An extreme case of multiskilled workers is found in self-contained cell lines. Figure 2(*b*) illustrates a self-contained cell line at a Japanese microscope plant, where a single worker moves with his or her workstation from operation to operation. The experienced worker not only assembles and adjusts the entire product him- or herself but also takes responsibility for quality assurance and for production management issues such as scheduling. This facilitates production efficiency for high value-added, low-volume products and also provides workers with intrinsic motivation and satisfaction. At the same time, it is considered an effective means to maintain skills and technologies for a new generation of workers. The concern for preserving worker skills is also illustrated by a common industry practice regarding welding operations: by policy, many companies will automate up to 95% of weld points and reserve the remaining 5% for manual operation. Through the manual welding, skill and knowledge are preserved in the organization and can be applied, for example, to programming the robotic welders.

In today's competitive environment, development of multiskilled operators is universally viewed as a critical issue. Well-organized training programs are seen at virtually every plant in Japan, irrespective of industry. As one means for motivating workers to develop further skills, many plants list workers' names along with their acquired skills on a prominently placed bulletin board. Another important issue is for workers to acquire creative thinking and problem-solving skills and to feel themselves responsible for quality and productivity improvement. For this purpose, small group improvement activities such as QC circles and PM circles are organized. Workers receive training and develop *kaizen* capabilities through these activities. Another effective approach is the implementation of worker suggestion systems, wherein financial rewards are paid based on the value of the suggestion. Not only do these various activities and systems contribute directly to improved performance, they help indirectly to maintain quality and productivity by increasing workforce morale.

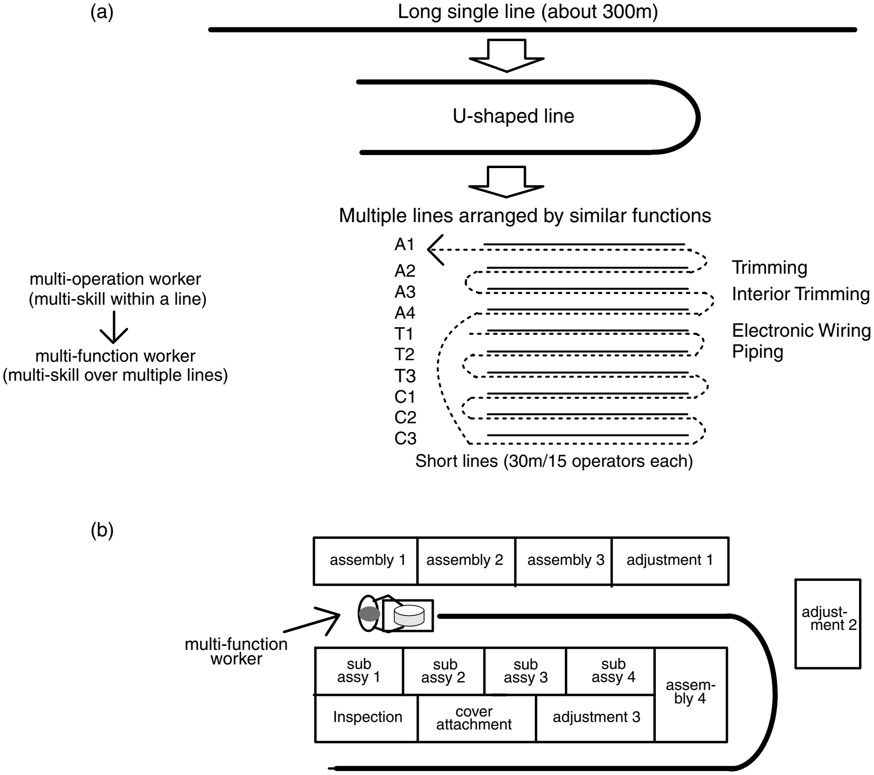

(a)
Long single line (about 300m)
U-shaped line
Multiple lines arranged by similar functions
A1
A2
A3
A4
T1
T2
T3
C1
C2
C3
Trimming
Interior Trimming
Electronic Wiring
Piping
Short lines (30m/15 operators each)
multi-operation worker
(multi-skill within a line)
multi-function worker
(multi-skill over multiple lines)
(b)
assembly 1
assembly 2
assembly 3
adjustment 1
adjust-
ment 2
multi-function
worker
sub
assy 1
sub
assy 2
sub
assy 3
sub
assy 4
assem-
bly 4
Inspection
cover
attachment
adjustment 3

Figure 2 (*a*) Layout Types and Their Evolution in Automobile Final Assembly Lines. (*b*) Layout of Self-Contained Cell Line.

2.3. Continuous Improvement and *Autonomation*

As previously mentioned, JIT is a production system that dynamically seeks ever-higher performance levels. Elimination of waste and defects is an important goal in this regard. Indeed, to achieve true just-in-time, all parts must be defect-free, as there is no excess inventory to draw upon for replacement. To achieve defect-free production, JIT production systems emphasize continuous improvement and utilize autonomation and other related techniques for assuring quality at the source.

The original meaning of autonomation in JIT/TPS is to stop a machine automatically when abnormal conditions are detected so as not to produce defective products. This is made possible by installing automatic detection and stopping devices in machines or equipment, thus making them capable of operating autonomously. The purpose of autonomation is to keep the machine working always for added value only and to prevent it from producing defects or waste. As discussed below, the idea of autonomation is extended in two ways: *poka-yoke* (mistake-proofing) and visual control.

Whereas autonomation is the application of detection and stopping devices to the operation of equipment, poka-yoke can be considered as the application of error detection and stopping devices to the activities of workers and the worker–machine interface. It is assumed that human errors are inevitable, so rather than rely on the vigilance of workers, various poka-yoke devices, such as jigs, sensors, and cognitive aids, are built into the work process (cf. Shingo 1986; Nikkan 1989). A simple example of poka-yoke is the use of color coding. Certain colors can be designated for the position and setting of a machine, so that the presence of any other color alerts the worker to an error and the need to take action. By reducing errors and the resulting defects, poka-yoke assumes an important role in assuring quality. In terms of job safety as well, poka-yoke is an indispensable means for protecting workers and preventing accidents in work operations.

Similarly, visual control systems (cf. Nikkan 1995) may also be viewed as an extension of the autonomation concept because they utilize various devices to share information and make abnormalities evident at a glance. A simple example is the designation of storage positions for tools, with an outline figure drawn for each tool. Any empty spot indicates the absence of a tool as well as its type. Perhaps the best-known example of visual control systems is the *andon*, a light board or stand consisting of red, yellow, and green lamps positioned so as to be visible from anywhere in the plant. A flashing red lamp indicates that a machine or workstation is experiencing some abnormal condition such as a breakdown. A green lamp stands for normal operation, and a flashing yellow lamp may indicate a setup operation or an upcoming tool exchange. The important role of the andon board is that it communicates the working status of the plant to everyone. In conjunction with andon, a common practice at many plants is to stop a production line when the red lamp flashes at any of the workstations so that the cause may be investigated and countermeasures taken.

Inventory level itself may also be considered as an important indicator for visual control. Not only is the inventory considered to be a non-value-adding expense, but high levels of inventory provide visible indication that waste and inefficiency may be hidden within the production system. To reveal the cause of the inefficiency, the inventory level is first decreased by removing some of the kanban containers from the system. The weakest part or bottleneck of the production system is then exposed because it will be the first affected by the removal of the inventory buffer. After reinforcing the weakest point, the inventory level is reduced further to expose the next weak point in an ongoing cycle. In this approach, the number of circulating kanbans is a visible measure of inventory level, and adjusting their number is a means of inventory reduction. More generally, this idea is employed as a continuous improvement cycle for improving production and material flows.

3. KANBAN AS A DECENTRALIZED CONTROL SYSTEM FOR JUST-IN-TIME

3.1. Prerequisites and Role of Kanban System

As discussed above, continuous improvement, reduced lead times, and a multiskilled workforce are crucial in enabling a production system to respond flexibly to the environmental changes confronting manufacturers. Given these prerequisites and a production schedule that has been leveled, a kanban system can function as a self-regulating, decentralized system for controlling the flow of materials from upstream stages through final assembly.

The Japanese term *kanban* simply means card or ticket. They are typically enclosed in a protective vinyl cover and contain the following information: part number and name, process name where the kanban is to be used, number of units in the standard container and type of packing, number of kanban cards issued, preceding process outbound stockpoint number, and subsequent process inbound stockpoint number. There is a one-to-one correspondence between the cards themselves and the standard parts containers that they represent. Furthermore, the cards always circulate together with the actual material flow. Through the kanban system, workers understand their operations' procedures and standards and learn and share the information required for process control. In this way, kanban functions as an information system as well as a means of visual control.

Two kinds of kanbans are used in controlling production and material flows: withdrawal-authorizing kanbans (also called movement kanbans) and production-ordering kanbans. When taken to the preceding process, a withdrawal kanban authorizes the transfer of one standard container of a specific part from the preceding process where it was produced to the subsequent process where it is to be used. A production kanban orders the production of one standard container of a specific part from the preceding process.

Figure 3 illustrates the circulation of kanban cards and containers. Every parts container at an inbound stock point must have a withdrawal kanban attached. When even one of a container's parts is to be used at the subsequent process, the withdrawal kanban is detached from the container and taken to the outbound stock point of the preceding process together with an available empty container. At the outbound stock point of the preceding process, an already full container of the desired parts is located and its production kanban is removed, with the withdrawal kanban now attached in its place. The full container with withdrawal kanban attached is now ready to be transferred to the inbound stockpoint of the subsequent process, while the empty container is left at the preceding process for later use. In addition, the production kanban that was just removed is now placed in a collection box at the preceding process. These production kanbans are frequently collected and become work orders authorizing the production of one more full container of parts. When a new, full container is produced, the production kanban is attached to it and the container is placed in the outbound stockpoint to complete the cycle.

It can be seen that circulation of the kanbans is triggered only by actual usage of parts at the subsequent process. Only what is needed is withdrawn from the preceding process, and then only what is needed for replacement is produced. This chain of usage and production ripples back through upstream stages to the suppliers, with the kanbans functioning as a sort of decentralized, manual

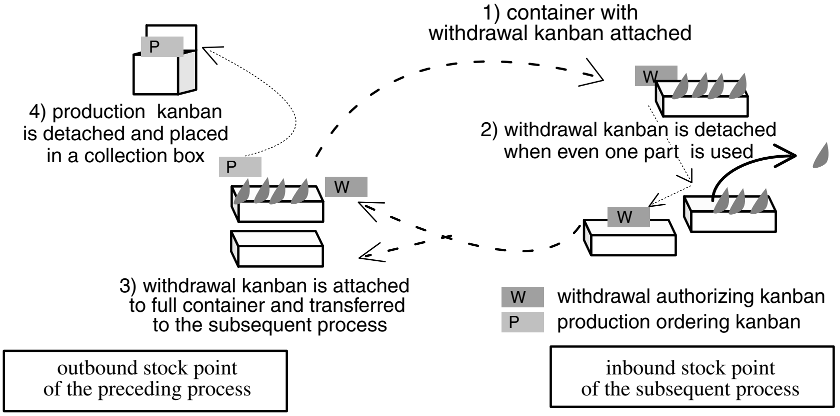

1) container with withdrawal kanban attached
2) withdrawal kanban is detached when even one part is used
3) withdrawal kanban is attached to full container and transferred to the subsequent process
4) production kanban is detached and placed in a collection box
P
W
W withdrawal authorizing kanban
P production ordering kanban
outbound stock point of the preceding process
inbound stock point of the subsequent process

Figure 3 Flow of Kanban Cards and Containers between two Processing Areas.

coordination mechanism. Given that major fluctuations in demand have been smoothed and frozen by the leveled production schedule, the kanbans can self-regulate to maintain the needed production quantities while absorbing the inevitable minor variations in everyday operations.

3.2. Control Parameters of Kanban System

In considering kanban as a decentralized control system, the following control parameters are necessary: number of kanbans in circulation; number of units in the kanban standard container; and kanban delivery cycle *a*-*b*-*c* (where *b* is number of deliveries per *a* days and *c* indicates the delivery delay factor as an indication of replenishment lead time). For example, 1-4-2 means that every 1 day the containers are delivered 4 times and that a new production order would be delivered by the 2nd subsequent delivery (in this case, about a half-day later, given four deliveries per day).

Minimizing work-in-process inventory is a goal of JIT. In keeping with this, the number of units per standard container should be kept as small as possible, with one being the ideal. Furthermore, the number of deliveries per day (b/a) should be set as frequently as possible so as to synchronize with the *takt* time of the subsequent process, and the delivery delay factor c should be kept as short as possible. Ideally, the number of kanbans in circulation between two adjacent workstations also should be minimized. However, in consideration of practical constraints, a tentative number of kanbans may be calculated as follows:

$$\text{Number of kanbans} = \frac{\text{number of units required per day}}{\text{number of units in standard container}} \times \frac{a}{b} \times (1 + c + \text{safety stock factor})$$

This tentative number of kanbans is reconsidered monthly because the daily leveled production requirement may differ under the new monthly production schedule. In addition, the number of kanbans is sometimes further reduced by systematically removing them from circulation in the system. The resultant reduction of work-in-process inventory will stress the production system and reveal the weakest point for further improvement. Thus, the kanban system is part of the approach used in JIT to move toward the goal of stockless production and achieve continuous improvement of the production system.

3.3. Kanban's Limitations and Alternatives

As previously noted, the kanban system is particularly appropriate for high-volume, repetitive manufacturing environments. However, in comparison to the situation when the kanban system was originally created, many industries now face tremendous increases in the variety of products and parts coupled with lower volumes for each individual item. This is seen also in the automobile industry, where more than 50% of parts now have their respective kanbans circulate less than once per day (Kuroiwa 1999). Such a low frequency of circulation leads to undesirably high levels of work-in-process inventory, and the kanban system ends up functioning the same way as the classic double-bin inventory system.

Several alternatives are available when production is not repetitive and/or high volume in nature. For example, when demand is lumpy and cannot be smoothed into a level production schedule, a push method of production coordination may be more appropriate. Karmarkar (1989), for example, discusses this situation and suggests how a hybrid production system utilizing both MRP and kanban may be used. Toyota itself has alternatives to the kanban system (Kuroiwa 1999). For low consumption parts, Toyota uses a push method called *chakko-hiki* (schedule-initiated production). In this approach, production and delivery of the necessary parts is determined in advance, based upon the vehicle final assembly schedule and an explosion of the bills-of-material, rather than through the kanban system. For large components and parts such as engines or seats, yet another system, called *junjo-hiki* (sequence-synchronized production), is used. Here the sequence of the part's production and delivery is synchronized item-by-item with the corresponding vehicle sequence of the final assembly schedule. For example, the engines are produced and delivered from the engine plant to the final assembly line in the exact order and timing needed for final assembly. In determining which of the various production coordination methods is appropriate, the most important factors for consideration are demand volume and delivery lead time. From this standpoint, it should be obvious that kanban is most suitable for parts with relatively high volume and short lead time.

Another important issue in kanban systems is the role of information technology. Kanban is a decentralized, manual system that developed without reliance on computers or other technologies. Given the powerful information technologies presently available, there is some debate over whether to implement electronic kanban systems with the use of bar codes and electronic data interchange (EDI). On the surface, these technologies provide a smart tool in terms of information handling, but it should be noted that the traditional, paper-based kanban has another important role in that workers learn process control and standards through reading the written information on the cards and physically manipulating the system. For this reason, Toyota has limited its use of electronic kanban systems to deliveries involving unavoidably long lead times, such as from distant plants. For example, when a part is consumed at the Kyushu plant, which is located more than 500 kilometers from Toyota's main plants in the Nagoya region, a withdrawal kanban in bar code form is scanned and its information is transferred through EDI to the preceding supplier or distribution center in the Nagoya region. Then a new withdrawal kanban in bar code form is issued in Nagoya for later attachment and delivery with the new parts, while the old card is disposed of in Kyushu. This is physically analogous to a one-way kanban system, though the electronic information is also utilized for many other purposes, such as physical distribution management, financial clearing systems, and so on.

3.4. Case Study of JIT/Kanban Implementation

Aisin Seiki is a Toyota group company that, in addition to manufacturing complex parts for the auto industry, also produces such items as mattresses for sale to retail and institutional customers. As noted in Imai (1997) and Spear and Bowen (1999), Aisin Seiki's Anjo plant achieved remarkable improvements through the introduction of JIT and kanban production systems. In 1986, prior to beginning JIT, the plant produced 200 styles and variations of mattresses with a daily production volume of 160 units and 30 days worth of finished-goods inventory. Production was based on monthly sales projections and scheduled on a weekly cycle. Due to unreliable forecasts and long lead times, the plant maintained high amounts of inventory yet often experienced material shortages, and suppliers had difficulty keeping up with changing orders.

Over the course of several years, the plant introduced various changes to achieve a JIT system with reduced inventories and one-piece production flow. A major change was to produce mattresses only in response to customer demand and based on the sequence in which orders were received. The production schedule cycle was reduced from one week to one day and then eventually to two hours, thereby necessitating many more setups and a reduction in setup times. The quilting machines, for example, now have 60 times as many setup changes. Kanban was introduced for the more popular models so as to produce and hold only the average number of units ordered per day. When a popular model is shipped from the plant to fill an order, its kanban card is returned to the production line as a signal to make a replacement. Lower-volume models are produced only after an order is received. When a large order is received from a hotel, its production is spread out among other orders in order to maintain a level production schedule. As a result of such changes and improvements, the plant is now able to produce a wider variety of end items in higher volumes and with shorter lead times and higher productivity. By 1997, for example, it produced 850 styles of mattresses with a daily production volume of 550 units and only 1.5 days of finished-goods inventory. At the same time, units produced per person had increased from 8 to 26, and overall productivity had increased by a factor of 2.08.

4. COMPLEMENTARY PARADIGMS OF JUST-IN-TIME

4.1. 3T's as Mutually Reinforcing Activities for Quality, Cost, Delivery Performance

Like JIT, total quality management (TQM) and total productive maintenance (TPM) are managerial models that were formulated and systematized in Japan. Each of these managerial paradigms is made up of concepts and tools of great significance for operations management. With TPS used as the name for just-in-time, together with TQM and TPM, a new acronym ‘‘3T’s’’ can be created to denote these three paradigms and their vital contribution to a firm’s competitiveness in terms of quality, cost, and delivery performance (QCD) (Enkawa 1998).

From a historical perspective, it is somewhat difficult to determine when a framework of their joint implementation began to emerge. This certainly took place after the individual advents of JIT/TPS in the 1950s, TQC/TQM in the 1960s, and TPM in the 1970s. If we view these paradigms as a historical sequence, it can be asserted that pressing needs to meet new competitive requirements were the driving impetus behind each of their creations. Being the later-emerging paradigms, TQM and TPM can be considered as systematized activities to support JIT with areas of overlap as well as unique emphasis. This interrelationship is depicted in Figure 4, wherein it may be seen that the three paradigms have roles of mutual complementarity in terms of Q, C, and D.

4.2. Total Quality Management (TQM)

The primary focus of TQM is on achieving customer satisfaction through the design, manufacture, and delivery of high-quality products. This addresses JIT’s requirement for strict quality assurance and elimination of defects. Though having roots in American approaches, TQM broadened and matured as a management paradigm in Japanese industry. By the 1980s, industries worldwide began emulating the Japanese model of quality management (e.g., Ishikawa 1985; Akiba et al. 1992) and have subsequently adapted and reshaped it in new directions.

With the concept of continuous improvement at its core, Japanese-style TQM seeks to boost performance throughout the organization through participation of all employees in all departments and levels. In addition to the tenet of total employee involvement, TQM incorporates the following beliefs and practices: management based on facts and data, policy deployment, use of cross-functional improvement teams, systematic application of the Plan, Do, Check, Act (PDCA) cycle, and the ability of all employees to use fundamental statistical techniques such as the seven basic quality improvement tools. One mainstay of TQM is its involvement of front-line employees in organized QC circle activities wherein the workers themselves identify and correct problems. These small-group activities are instrumental in diffusing continuous improvement capability throughout the company. Besides supporting successful JIT implementation, these activities also improve employee morale and enhance skills.

The essence of TQM can also be understood from the many maxims that have appeared to express new approaches and ways of thinking necessitated by competitive change:

- Quality is built in at the process (instead of by inspection).
- Focus on and correct the process rather than the result (emphasizing prevention).
- Emphasize system design and upstream activities.
- The next process is your customer (to emphasize customer orientation).
- Quality first—the more quality is improved, the lower cost becomes.
- Three-*gen* principle: observe the actual object (*genbutsu*) and actual situation (*genjitsu*) at the actual location (*genba*).

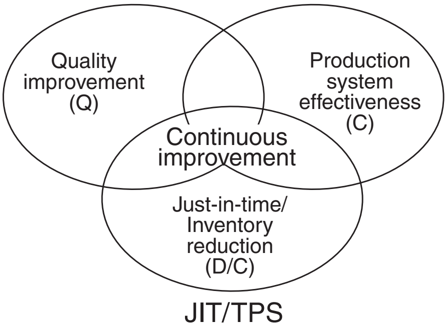

TQM
TPM
Quality improvement (Q)
Production system effectiveness (C)
Continuous improvement
Just-in-time/ Inventory reduction (D/C)
JIT/TPS

Figure 4 Mutual Overlap and Complementarity of the 3T’s (JIT/TPS, TQM, TPM) with respect to quality, cost, and delivery performance (QCD).

4.3. Total Productive Maintenance (TPM)

TPM is a systematic and inclusive approach for achieving maximum utilization of a plant's capability. In particular, TPM focuses on maximization of overall equipment effectiveness, which is defined as the ratio of net value-added time to total available time (where net value-added time is total available time adjusted for time lost due to failures, setup/adjustment, cutting blade changes, start-up, minor stoppage and idling, reduced speed, and defect/rework-related losses). Many companies identify 16 or more major losses, involving not only the equipment-related losses above but also worker and production system factors. These are measured quantitatively as the discrepancy between their current state and ideal state and are often converted into monetary terms as the target of cost reduction goals. To reach these goals, TPM advocates a clearly defined, stepwise approach for eliminating the losses through systematic, plant-wide efforts.

The principal activities of TPM fall into eight areas:

1. Autonomous maintenance
2. Improvement by project team
3. Planned maintenance
4. Quality maintenance
5. Initial-phase management for development and installation of equipment/products
6. Administrative and indirect operations management
7. Education and training
8. Safety, health, and environmental management

Of these, autonomous maintenance (*jishu hozen*) is the original core of TPM and involves the transfer of diverse maintenance-related duties to the machine operators themselves. In fostering operator responsibility for maintenance, autonomous maintenance follows seven formal steps starting from initial cleaning. In addition, small-group activities called PM circles are organized by the company to carry out improvement activities and further develop operators' capabilities. Further discussion of TPM and its implementation is provided by Shirose (1996), Suzuki (1992), and Nakajima et al. (1992).

In instituting the good industrial housekeeping practices required in TPM, many companies use an approach called 5S. This term comes from five Japanese words: *seiri* (sort and clear out), *seiton* (straighten and configure), *seiso* (scrub and cleanup), *seiketsu* (maintain sanitation and cleanliness of self and workplace on an ongoing basis) and *shitsuke* (self-discipline and standardization of these practices) (cf. Imai 1997; Osada 1991). These activities roughly correspond to the initial cleaning step of TPM and involve all employees in an organization.

Like TQM, TPM has an important role in implementing JIT. Specifically, failure-free equipment is a principal factor behind the high quality and short setup times necessary in JIT. At the same time, TPM has benefited from many of JIT's approaches, including its emphasis on visual control and the promotion of education and training for multiskilled operators. As for differences between TPM and the other paradigms, it should be noted that TQM and TPM diverge in their approaches undertaken for problem solving. Whereas TQM stresses the use of empirical, data-oriented methods such as statistical quality control, TPM emphasizes the application of technological theory and principle (*genri-gensoku*). In other words, TQM relies on inductive reasoning while TPM relies on deductive reasoning. These approaches work in complementary fashion with each other because both types of reasoning are indispensable for effective problem solving.

To summarize this discussion of the three paradigms, Table 1 outlines and compares TQM, TPM, and JIT/TPS along the dimensions of focus of attention, methods, and organizational issues.

4.4. Joint Implementation of the 3T's and Case Study

Since the 1980s, the diffusion of the three paradigms from Japanese manufacturers to overseas manufacturers has progressed at an accelerated rate. As evidenced by the coinage of terms such as *3T*'s and *TPQM* (= TPM + TQM), many cases now exist of the simultaneous implementation of part or all of the three paradigms, both inside and outside of Japan. This is underscored by the results of a mail survey of members of American Manufacturing Excellence, an organization whose members are known as leaders in manufacturing practice (White et al. 1999). On average, manufacturers included in the sample had implemented 7 out of 10 listed practices relating to JIT. As a whole, these 10 practices encompass a breadth of activity virtually equivalent to the 3T's. The percentages of manufacturers implementing each practice is as follows, broken down by size of company (small firms sample size $n = 174$; large firms sample size $n = 280$): quality circles (56.3%; 70.4%); total quality control (82.2%; 91.4%); focused factory or reduction of complexities (63.2%; 76.8%); TPM

TABLE 1 **A Comparison of the JIT/TPS, TQM, and TPM Paradigms**

	Criteria	JIT/TPS	TQM	TPM
Focus	Scope	Materials flow	Product life cycle	Equipment life cycle
	Origin	"Supermarket-style" inventory control	Statistical quality control (SQC)	Preventive maintenance
	Direction of expanded focus	Supply chain management (from source through user)	Upstream activities (towards new product development)	Preventive approach (towards maintenance prevention/design issues)
	Control attribute	Inventory (*muri, muda, mura*)	Quality of product and process	Losses related to equipment, workers, materials
Means	Fundamental attitude	Kaizen	Kaizen	Kaizen
	Concept/premise	Load smoothing	Market-in orientation	Utmost equipment effectiveness/zero loss
	Approach to problem solving	Deductive/"simple is best"	Inductive	Deductive
	Means for problem solving	Exposing system weaknesses through inventory reduction	Application of PDCA cycle/data emphasis	Pursuit of ideal/technological theory and principles
	Key methods	Mechanisms that activate improvement (*autonomation*, *andon*, *kanban*, etc.)	Managerial methods with emphasis on statistical thinking	PM analysis, 5S activities, basic technologies, etc.
Organizational Issues	Human resources development	Multifunctional workers	QC education appropriate to level	Basic technologies-related skills
	Vertical integration	Reliance on formal organization/staff	Policy deployment/quality audits by top management	Overlapping small-group activities
	Horizontal integration	–	Cross-functional management	Special committees/project teams
	Bottom-up involvement	–	QC circles	Autonomous maintenance
	Implementation approach	Company-specific	Company-specific	12-step implement-ation program

(51.7%; 65.7%); reduced setup time (82.8%; 90.7%); group technology (66.7%; 69.6%); uniform workload (52.3%; 59.3%); multifunction employees (86.8%; 79.3%); kanban (53.4%; 68.9%); and JIT purchasing (66.1%; 81.8%).

Some caution may be warranted, however, on the question of simultaneously implementing TQM and TPM. One reason is the financial burden arising from their simultaneous introduction. Empirically, there is said to be a delay of up to a few years for TPM, or longer for TQM, before a company can enjoy their financial effects. Another reason relates to organizational issues. Although both TQM and TPM share the same broad objectives of strengthening corporate competitiveness and developing human resources, there are significant differences in their implementation and focus, as shown in Table 1. Thus, the simultaneous introduction of both into a formal organization may be a source of confusion and hamper their success. For this reason, the overwhelming majority of companies in Japan have avoided concurrent implementation. Instead, they have adopted a phased approach wherein one of the paradigms is mastered first before undertaking the second (Miyake and Enkawa 1999; Mathis 1998). As to which of TQM and TPM should be undertaken first, there is no clear consensus in terms of conceptual grounds or actual industry trends, though TPM is seen as having quicker financial benefits.

As the originator of JIT and TPM as well as a leader in TQM, the Toyota group continues to be known for best practice of the 3T's. In Toyota's case, JIT was implemented first, followed by TQM and finally TPM, though this merely reflects the historical order of each paradigm's emergence and the evolving competitive demands that they address. In most Toyota group companies, the management planning department organizes the implementation and education of the 3T's. An exception is Toyota itself, where TQM is administered by the TQM promotion department and JIT and TPM-like activities are coordinated by the operations management consulting department.

Outside of the Toyota group, Sanden Corporation is regarded to be one of the best companies for 3T practices in Japan. As one of the world's three largest manufacturers of car air conditioners and compressors, Sanden is a supplier to most leading auto producers in the world. Beginning in 1993, Sanden introduced TPM to one of its major plants, Sanden Yattajima Manufacturing Site, in response to fierce global competition and pressures to reduce cost. Within three years, it completed the conventional 12-step TPM implementation regimen, enabling Sanden Yattajima to apply successfully in 1996 for the TPM Excellence Award, a prize recognizing outstanding achievements that is granted by the Japan Institute of Plant Maintenance (JIPM). During this same period, Sanden bolstered its quality assurance front through ISO 9001 implementation and certification. Shortly afterwards, JIT was formally introduced in 1997 under the name SPS (Sanden Production System) with the aim of reorganizing process flows and further reducing lead times and inventory. After firmly establishing TPM and JIT, Sanden was poised to complete the triad of the three paradigms through a regeneration of its TQM practices. Its primary motivation was to reinforce the quality of its management systems, as it had already achieved a level of competitiveness in product quality. Sanden's TQM success was evidenced by its receiving of the Deming Prize in 1998 from the Japanese Union of Scientists and Engineers (JUSE) in recognition of outstanding results in the practice of TQM.

5. LEAN PRODUCTION AND EXTENSIONS OF JIT

5.1. Lean Production

The term *lean production* was introduced by Krafcik (1988) and the famous book, *The Machine That Changed the World* (Womack et al. 1990). These publications present the results of a major MIT study to identify systematically best practices of Japanese and other automobile manufacturers worldwide. Lean production is "lean" in that it uses half of the various production resources (labor, manufacturing space, tool investment, engineering hours, inventory, etc.) used in the Ford-style mass production that was prevalent into the 1980s.

The essence of lean production may be summarized into four aspects: (1) lean plant; (2) lean supplier network; (3) lean product development; and (4) relationship with distributors and customers. Of these, the lean plant forms the original core and may be considered as equivalent to the JIT/TPS concept presented earlier. The remaining three aspects delineate how the lean plant interacts with and mutually influences other entities. Thus, lean production may also be considered as an extended paradigm of JIT that includes new intraorganizational and interorganizational aspects. It is interesting to note that Toyota itself was influenced by the systematic, encompassing framework of lean production as presented by the MIT study. Though Toyota's practices indeed formed the basis for the lean production model, many of those practices had evolved gradually in response to local organizational needs or involved historically separate organizational units. The lean production model consequently provided a conceptual thread for understanding its many aspects as a synergistic, integrated whole.

Lean supplier network, referring to various innovative practices in supplier relationships, can be considered as an interorganizational extension of JIT. It is well known that Toyota reorganized its suppliers into functional tiers in the 1950s. Its aim was to provide suppliers with incentive for improvement and promote learning mechanisms among suppliers and workers alike while maintaining long-term relationships. In this system, different responsibilities are assigned to the suppliers in each tier. First-tier suppliers are responsible not only for the manufacture of components, but also for their development while interacting and sharing information with Toyota and other first-tier suppliers. In turn, each first-tier supplier forms its own second tier of suppliers. They are assigned the job of fabricating individual parts and occasionally assist in their design, within the confines of rough specifications provided by the first-tier companies. In many cases, even a third tier of suppliers may exist to manufacture parts according to given specifications.

Roughly only 30% of Toyota's parts production is done in-house, with the remainder allotted to the supplier network. This high degree of outsourcing affords Toyota flexibility in coping with change. At the same time, it underscores the criticality of a competitive supply chain. Toyota's supplier network is neither a rigid, vertical integration nor an arm's-length relationship. Furthermore, it attempts to maintain a delicate balance between cooperative information sharing and the principle of competition. For this purpose, Toyota has employed several organizational strategies.

One approach for promoting information sharing is the formation of supplier associations and the cross-holding of shares with supplier-group firms and between first-tier suppliers. Another is the two-way sharing of personnel with Toyota and between supplier-group firms.

As an approach for encouraging competition, there is ample opportunity for a supplier to move up the ladder if it exhibits very high performance in quality, cost, delivery, and engineering capability. A parts-manufacturing specialist, for example, can expand its role to include the design of parts and even integrated components. Further incentive is provided by a degree of competition between suppliers over their share of business for a given part. Contrary to common assumption, parts are not sole-sourced except for certain complex systems that require considerable investments in tools. Rather, the business is typically divided between a lead supplier and one or more secondary suppliers. Each supplier's relative share of business may grow or shrink depending on performance. By providing substantial incentives to suppliers, this flexible system has spurred a variety of improvements relating to quality, cost, and JIT delivery of parts.

Another aspect of the lean production enterprise concerns the way that new products are designed and developed. Lean product development differs from traditional approaches on four key dimensions: leadership, teamwork, communication, and simultaneous development (Womack et al. 1990). In the case of automobile development, hundreds of engineers and other staff are involved. Consequently, strong leadership is the first condition for doing a better job with less effort. Toyota adopted the *shusa* system, also known as the heavyweight project manager system (Clark and Fujimoto 1991). The *shusa* (literally the chief engineer) is the leader of a project team whose job is to plan, design, and engineer a new product as well as to ramp up production and deliver it to market. Great responsibility and power are delegated to the *shusa*. Most importantly, the *shusa* assembles a flat team with members coming from the various functional departments throughout the company, from marketing to detail engineering. Though retaining ties to their original department, the team members are now responsible on a day-to-day basis to the *shusa*.

The development process used to explore the optimum design solution has been called by names such as set-based engineering (Liker et al. 1995). It can be defined as an approach for engineers to explicitly consider and communicate sets of design alternatives at both conceptual and parametric levels within clearly defined constraints. They gradually narrow these design sets by eliminating inferior alternatives until they eventually freeze the detailed product specifications based on feasibility studies, reviews, and analysis. This broad, sweeping approach helps the design team avoid premature design specifications that might appear attractive based on narrow considerations but that would suboptimize the overall design. Furthermore, set-based engineering facilitates sharing of the product's overall image among all members involved in the project, making it possible to develop different aspects of the product simultaneously.

This simultaneous development approach is now generally known as concurrent or simultaneous engineering. It is defined as the overlapping or parallel development of what otherwise would be sequential jobs. For example, the design of a production die ordinarily is not initiated until the detail engineering is completed for the part to be produced. In concurrent engineering, however, the die design is commenced even before the part design in finalized by utilizing shared knowledge concerning the approximate dimensions of the part, its requirements, and its design process. This overlapping design process is made possible by close communication between the sequential tasks and is somewhat analogous to the exploitation of off-line or external setup in order to reduce production lead time. Concurrent engineering is also employed with outsourced components by sharing personnel and information between Toyota and its suppliers.

Through concurrent engineering and other approaches, lean product development has achieved astonishing results, requiring only one half or less of the time and human resources traditionally needed to develop a new car. Accordingly, it has become a cornerstone of time-based competition. Further means for improvement lie in the full exploitation of information technologies. In particular, CAD, CAE, and electronic interchange of technical data facilitate the sharing of information and have further potential for supporting concurrent engineering, particularly in later development stages.

The final aspect of lean production is the relationship with distributors and customers. As it did with its supplier network, Toyota built a distribution network that incorporated the dealers into the production system. In fact, the dealer is considered the first step in the kanban system. It sends orders for presold cars to the factory for delivery to specific customers about two weeks later, thereby initiating the pull signal for production. For this to be workable, the dealer works closely with the factory to sequence orders in a way that the factory can accommodate, just as the factory works closely with its suppliers. At the same time, dealers have played another role in assessing customer needs and relaying that information to Toyota for consideration in designing new models. Dealers have done this by making house calls directly to customers and increasingly by collecting customer information in databases maintained at the dealers.

Through these innovative practices, Toyota's lean production system formed a complete supply chain from source to user. Its system integrated suppliers and distributors with manufacturing and involved them in product development. In so doing, the lean production model has provided a prototype of the business paradigms of concurrent engineering and supply chain management. For an examination of the implementation of lean production in North American companies, see Liker (1998) and Liker et al. (1999).

5.2. Theory of Constraints (TOC) and JIT

The concept and practice of *kaizen* are features common to TQM, TPM, JIT, and lean production. However, despite great effort expended for continuous improvement throughout their organizations, many Japanese companies in the 1990s faced difficulty in improving their financial bottom line, including even the leaders in JIT practice. This was largely due to Japan's prolonged economic recession, but many companies would have benefited from a more focused approach to their improvement efforts.

One such approach lies in the Theory of Constraints (TOC). Devised by Eli Goldratt (e.g., Goldratt 1990; Goldratt and Cox 1992), TOC can be viewed as an inclusive philosophy and methodology of improvement, which includes bottleneck-focused scheduling methods, performance measurement system, thinking process, and problem-solving tools. Its core idea is that every system, such as a for-profit company, must have at least one constraint. Since a constraint is a factor that limits the system from getting more of whatever it aims to achieve, a company that wants more profits must manage its constraints. See Spencer and Cox (1995) for an analysis of the history and scope of TOC and its relation to optimized production technology (OPT), as well as Umble and Srikanth (1990), Noreen et al. (1995), and Dettmer (1997) for detailed discussions of TOC's sub-areas of scheduling/logistics, performance measurement, and thinking process, respectively. A review of TOC-related journal publications is given by Rahman (1998).

Among TOC's many techniques and tools, the five-step focusing process (e.g., Goldratt and Cox 1992) is the one most closely related to JIT:

Step 1. Identify the system constraint(s).

Step 2. Decide how to exploit the system constraint(s), i.e., better utilize the constraint's existing capacity.

Step 3. Subordinate everything else to the above decision, i.e., align the rest of the system to work in support of the constraint.

Step 4. Elevate the constraint(s), i.e., improve/increase the constraint's capacity.

Step 5. If a constraint has been broken, go back to step 1. Do not allow inertia to cause a system constraint.

Of these, steps 1, 4, and 5 are practically equivalent to the continuous-improvement logic of JIT if *constraint* is considered in the narrow meaning of a production bottleneck. One could even consider these five focusing steps as JIT concepts in combination with steps 3 and 4's unique bottleneck scheduling logic.

Despite this similarity, two important differences should be noted. The first difference relates to the meaning of the term *constraint* and to understanding which goal is being constrained. In JIT, the constraint is tantamount to a production bottleneck, and the primary focus is placed on inventory level. In the case of TOC, the issue is not inventory level per se, but rather achieving the company's goal of making profit by means of increasing throughput, reducing expenses, and/or reducing inventory.

To understand TOC's approach to improving company profitability, consider the often-drawn analogy between a system and a chain, where the strength of the chain corresponds to the company's profit. To improve the strength of the chain, one must identify the weakest link and concentrate effort on strengthening it, instead of applying effort uniformly over all the links. This analogy implies that the focus for continuous improvement must be on the area that will bring about the greatest benefit in relation to the effort extended. Furthermore, this area for improvement may lay outside the manufacturing plant because the constraint may not be a physical one but may involve managerial policies or market factors. Consequently, in such a situation, plant-wide improvement efforts in ‘‘carpet bombing’’ fashion may not lead to an improved bottom line for the company.

The second point of difference concerns the focus of improvement efforts. Whereas JIT, TQM, and TPM concentrate on *changing* the production system as the means to improve it, TOC explicitly considers options for increasing profits while making the constraint work more effectively *as it is*. This concept is embodied in steps 2 and 3 of the five-step focusing process. In particular, step 3 corresponds to the famous scheduling solution called drum-buffer-rope (DBR). As illustrated in Figure 5(*a*), DBR is easily understood using Goldratt’s analogy of a Boy Scout troop. Each Scout stands for one workstation or process, while the distance the troop moves during a specific period is throughput earned and the length of the Scout troop column corresponds to work-in-process. Under the condition that we cannot eliminate statistical fluctuations and disruptions in each Scout’s hiking performance, the objective is to have the Scout troop advance as much distance as possible while keeping the troop’s length short.

The solution to this problem lies in having the slowest Scout (constraint) set the pace of the troop by beating a drum and tying a rope with some slack (work-in-process or time buffer) between him and the leading Scout. The rope prevents the troop length from increasing unnecessarily. At the same time, the slack in the rope allows the slowest Scout to work at his full capability without any interference from disruptions in the preceding Scouts’ performance, thereby enabling the troop as a whole to advance the maximum attainable distance.

For reference, the corresponding analogy for a kanban system is shown in Figure 5(*b*). In a kanban system, all adjacent pairs of Scouts are tied together by a rope whose length represents the number of kanbans in circulation, and the pace of the drum beat must be constant, in keeping with JIT’s prerequisite of leveled production. It should be noted that there is no individual Scout who corresponds to a constraint or bottleneck. Instead, once any Scout experiences difficulty walking with the given length of rope, that Scout is identified as a constraint and countermeasures are taken straightaway to increase his capability. Next, the length of rope between each pair is further shortened and another constraint is discovered to repeat the cycle of continuous improvement. In this sense, kanban should be considered as a means to expose the constraint, rather than a means to schedule it and fully exploit its existing capability.

In sum, TOC adds a beneficial and indispensable viewpoint to the 3T’s in that it helps to clarify the goal of improvement and to identify where improvement should be focused so as to achieve maximum financial benefits from a global optimum standpoint. In addition, TOC complements JIT, TQM, and TPM by emphasizing that improvement should not always be the first option. Rather, quick returns are often possible from first exploiting the constraint as it is, making it work to its own utmost limit.

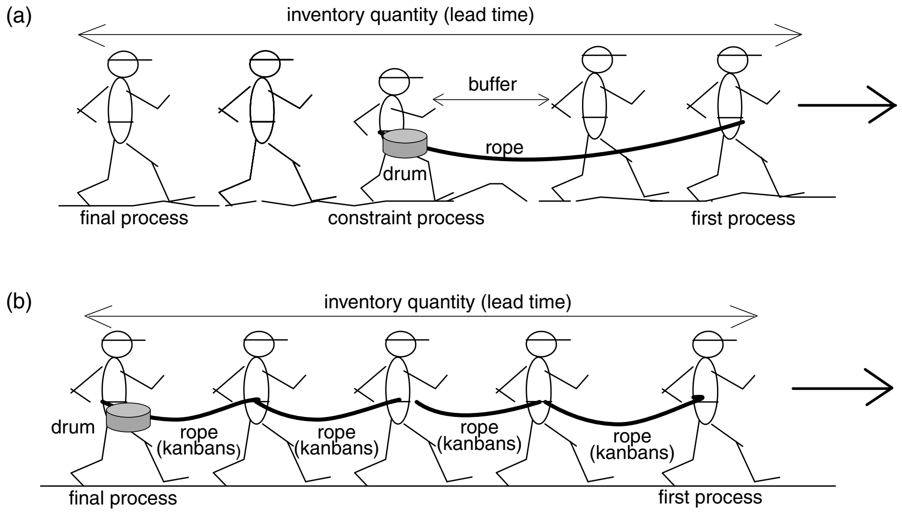


Figure 5 Analogy of Drum-Buffer-Rope in (*a*) Theory of Constraints and (*b*) Kanban System.

5.3. Applications to Service Industries

As exemplified by such terms as the *production-line approach to service* or the *service factory,* manufacturing and service industries have a history of mutually influencing each other's theory and practice. Despite their manufacturing industry origins, JIT, TQM, TPM, and TOC also have applications to service industries. The purpose of this section is briefly to identify some of these application issues and provide references for further information.

JIT and lean production concepts such as rationalization of process flows, batch size reduction, demand leveling, and multifunctional workers are quite applicable to many service environments. Duclos et al. (1995) provide a review of JIT practices in services as well as references for articles describing their implementation. For example, Mehra and Inman (1990) describe an overnight package delivery company's implementation of JIT for its business supplies, and Billesbach and Schneidemans (1989) examine the application of JIT in administrative areas. A broader discussion of the application of lean production principles to service operations is provided by Bowen and Youngdahl (1998), wherein they describe how service businesses such as Taco Bell, Southwest Airlines, and Shouldice Hospital have mastered "lean" service. Southwest Airlines, for example, is known for its rapid turnaround (setup) of planes between flights—about three times faster than the industry average. Service companies should not wrongly assume that production efficiency and customer responsiveness are tradeoffs, as they were in the mass production paradigm. Rather, lean production principles can be used to eliminate non-value-added activities from service processes, integrate the value chain, and increase flexibility and responsiveness.

Of various other JIT and TQM-related concepts, poka-yoke merits special attention. In many services, due to the simultaneity of a service's creation and consumption, mistakes are readily apparent to the customer and it is not possible to carry out inspection or rework. Examples of mistake-proofing in service environments range from color coding of patients' medical files to techniques for differentiating drinks with similar colors. Further discussion and examples of poka-yoke in services are provided by Schvaneveldt (1993) and Chase and Stewart (1994).

While TPM is practiced in the administrative areas of many leading manufacturers, it has received scant attention from service industries. This is understandable given the different nature of technology and equipment in most service environments. However, a precursor to TPM called 5S is very appropriate for services. As briefly described in the TPM section above, 5S is a Japanese acronym for five good housekeeping practices for organizing and cleaning the workplace. These practices involve all employees, not just designated maintenance or custodial staff, and help to establish an environment of self-responsibility and problem awareness among employees.

Finally, several aspects of TOC have application to service industries, particularly the five-step focusing process for system constraint identification and management. In addition to reviewing the emerging literature on TOC in services, Siha (1999) provides a framework for interpreting and applying TOC concepts to different service industries. An interesting case study of TOC in an engineering services firm is given in Motwani and Vogelsang (1996), which describes how the firm identified its constraint in the survey department and took measures to increase throughput. In another case (Olson 1998), a security and alarm company was unable to meet market demand. It determined that installation technicians were the bottleneck resource and consequently redesigned the alarm-installation process to allow technicians to work in teams and perform tasks in parallel. TOC has even been applied in an accounting firm (Green and Larrow 1994). After identifying the constraint to be the tax department, the firm found ways for the rest of the organization to support the work of the tax department better without increasing staff or work hours.

REFERENCES

Akiba, M., Schvaneveldt, S. J., and Enkawa, T. (1992), "Service Quality: Methodologies and Japanese Perspectives," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 2349–2371.

Akturk, M. S., and Erhun, F. (1999), "An Overview of Design and Operational Issues of Kanban Systems," *International Journal of Production Research,* Vol. 37, No. 17, pp. 3859–3881.

Billesbach, T., and Schneidemans, J. (1989), "Application of Just-in-Time Techniques in Administration," *Production and Inventory Management Journal,* Vol. 30, No. 3, pp. 40–45.

Bowen, D. E., and Youngdahl, W. E. (1998), "'Lean Service': In Defence of a Production-Line Approach," *International Journal of Service Industry Management,* Vol. 9, No. 3, pp. 207–225.

Chase, R. B., and Stewart, D. M. (1994), "Make Your Service Fail-Safe," *Sloan Management Review,* Vol. 35, No. 3, pp. 35–44.

Clark, K. B., and Fujimoto, T. (1991), *Product Development Performance: Strategy, Organization, and Management in the World Auto Industry,* Harvard Business School Press, Boston.

Dettmer, H. W. (1997), *Goldratt's Theory of Constraints: A Systems Approach to Continuous Improvement,* ASQC Quality Press, Milwaukee.

Duclos, L. K., Siha, S. M., and Lummus, R. R. (1995), ‘‘JIT in Services: A Review of Current Practices and Future Directions for Research,’’ *International Journal of Service Industry Management,* Vol. 6, No. 5, pp. 36–52.

Enkawa, T. (1998), ‘‘Production Efficiency Paradigms: Interrelationship among 3T—TPM, TQC/TQM, and TPS (JIT),’’ in *1998 World-Class Manufacturing and JIPM-TPM Conference* (Singapore), pp. 1–9.

Fujimoto, T. (1999), *The Evolution of a Manufacturing System at Toyota,* Oxford University Press, New York.

Goldratt, E. M. (1990), *Theory of Constraints,* North River Press, Croton-on-Hudson, NY.

Goldratt, E. M., and Cox, J. (1992), *The Goal* 2nd Rev Ed., North River Press, Croton-on-Hudson, NY.

Golhar, D. Y., and Stamm, C. L. (1991), ‘‘The Just-in-Time Philosophy: A Literature Review,’’ *International Journal of Production Research,* Vol. 29, No. 4, pp. 657–676.

Green, G. C., and Larrow, R. (1994), ‘‘Improving Firm Productivity—Looking in the Wrong Places,’’ *CPA Chronicle,* Summer.

Groenevelt, H. (1993), ‘‘The Just-in-Time System,’’ in *Logistics of Production and Inventory,* S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, Eds., Handbooks in Operations Research and Management Science, Vol. 4, North-Holland, Amsterdam, pp. 629–670.

Imai, M. (1997), *Gemba Kaizen,* McGraw-Hill, New York.

Ishikawa, K. (1985), *What Is Total Quality Control? The Japanese Way,* Prentice Hall, Englewood Cliffs, NJ.

Karmarkar, U. (1989), ‘‘Getting Control of Just-in-Time,’’ *Harvard Business Review,* Vol. 67, No. 5, pp. 122–131.

Keller, A. Z., and Kazazi, A. (1993), ‘‘Just-in-Time Manufacturing Systems: A Literature Review,’’ *Industrial Management and Data Systems,* Vol. 93, No. 7, pp. 1–32.

Krafcik, J. F. (1988), ‘‘Triumph of the Lean Production System,’’ *Sloan Management Review,* Vol. 30, No. 1, pp. 41–52.

Kuroiwa, S. (1999), ‘‘Just-in-Time and Kanban System,’’ in *Seisan Kanri no Jiten,* T. Enkawa, M. Kuroda, and Y. Fukuda, Eds., Asakura Shoten, Tokyo, pp. 636–646 (in Japanese).

Liker, J. K., Ettlie, J. E., and Campbell, J. C., Eds. (1995), *Engineered in Japan: Japanese Technology-Management Practices,* Oxford University Press, New York.

Liker, J. K. (1998), *Becoming Lean: Inside Stories of U.S. Manufacturers,* Productivity Press, Portland, OR.

Liker, J. K., Fruin, W. M., and Adler, P. S., Eds. (1999), *Remade in America: Transplanting and Transforming Japanese Management Systems,* Oxford University Press, New York.

Mathis, R. H. (1998), ‘‘How to Prioritize TQC and TPM in the Quest for Overall Manufacturing Excellence,’’ Master’s Thesis, Darmstadt University of Technology/Tokyo Institute of Technology.

Mehra, S., and Inman, R. A. (1990), ‘‘JIT Implementation in a Service Industry: A Case Study,’’ *International Journal of Service Industry Management,* Vol. 1, No. 3, pp. 53–61.

Miyake, D. I., and Enkawa, T. (1999), ‘‘Matching the Promotion of Total Quality Control and Total Productive Maintenance: An Emerging Pattern for the Nurturing of Well-Balanced Manufacturers,’’ *Total Quality Management,* Vol. 10, No. 2, pp. 243–269.

Monden, Y. (1998), *Toyota Production System: An Integrated Approach to Just-in-Time,* 3rd Ed., Institute of Industrial Engineers, Atlanta.

Motwani, J., and Vogelsang, K. (1996), ‘‘The Theory of Constraints in Practice—At Quality Engineering, Inc.,’’ *Managing Service Quality,* Vol. 6, No. 6, pp. 43–47.

Nakajima, S., Yamashina, H., Kumagai, C., and Toyota, T. (1992), ‘‘Maintenance Management and Control,’’ in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1927–1986.

Nikkan Kogyo Shimbun, Ed. (1989), *Poka–Yoke: Improving Product Quality by Preventing Defects,* Productivity Press, Portland, OR.

Nikkan Kogyo Shimbun, Ed. (1995), *Visual Control Systems,* Productivity Press, Portland, OR.

Noreen, E., Smith, D., and Mackey, J. (1995), *The Theory of Constraints and Its Implications for Management Accounting,* North River Press, Great Barrington, MA.

Ohno, T. (1988), *Toyota Production System: Beyond Large–Scale Production,* Productivity Press, Portland, OR.

Olson, C. (1998), ‘‘The Theory of Constraints: Application to a Service Firm,’’ *Production and Inventory Management Journal,* Vol. 39, No. 2, pp. 55–59.

Osada, T. (1991), *The 5S's: Five Keys to a Total Quality Environment,* Productivity Press, Portland, OR.

Rahman, S. (1998), "Theory of Constraints: A Review of the Philosophy and Its Applications," *International Journal of Operations and Production Management,* Vol. 18, No. 4, pp. 336–355.

Schvaneveldt, S. J. (1993), "Error–Proofing Service Operations: Principles and Practices,"*POM-93 Bulletin: Fourth Annual Meeting of Production and Operations Management Society,* Boston.

Shingo, S. (1985), *A Revolution in Manufacturing: The SMED System,* Productivity Press, Portland, OR.

Shingo, S. (1986), *Zero Quality Control: Source Inspection and the Poka-yoke System,* Productivity Press, Portland, OR.

Shingo, S. (1989), *A Study of the Toyota Production System from an Industrial Engineering Viewpoint,* Rev. Ed., Productivity Press, Portland, OR.

Shirose, K., Ed. (1996), *TPM: New Implementation Program in Fabrication and Assembly Industries,* Japan Institute of Plant Maintenance, Tokyo.

Siha, S. (1999), "A Classified Model for Applying the Theory of Constraints to Service Organizations," *Managing Service Quality,* Vol. 9, No. 4, pp. 255–265.

Sohal, A. S., Keller, A. Z., and Fouad, R. H. (1989), "A Review of Literature Relating to JIT," *International Journal of Operations and Production Management,* Vol. 9, No. 3, pp. 15–25.

Spear, S., and Bowen, H. K. (1999), "Decoding the DNA of the Toyota Production System," *Harvard Business Review,* Vol. 77, No. 5, pp. 97–106.

Spencer, M. S., and Cox, J. F. (1995), "Optimum Production Technology (OPT) and the Theory of Constraints (TOC): Analysis and Genealogy," *International Journal of Production Research,* Vol. 33, No. 6, pp. 1495–1504.

Sugimori, Y., Kusunoki, K., Cho, F., and Uchikawa, S. (1977), "Toyota Production System and Kanban System: Materialization of Just-in-Time and Respect-for-Human System," *International Journal of Production Research,* Vol. 15, No. 6, pp. 553–564.

Suzuki, T. (1992), *New Directions for TPM,* Productivity Press, Portland, OR.

Umble, M., and Srikanth, M. (1990), *Synchronous Manufacturing,* South-Western, Cincinnati.

Vollmann, T. E., Berry, W. L., and Whybark, D. C. (1997), *Manufacturing Planning and Control Systems,* 4th Ed., Irwin, Boston.

White, R. E., Pearson, J. N., and Wilson, J. R. (1999), "JIT Manufacturing: A Survey of Implementations in Small and Large U.S. Manufacturers," *Management Science,* Vol. 45, No. 1, pp. 1–15.

Womack, J. P., Jones, D. T., and Roos, D. (1990), *The Machine that Changed the World: The Story of Lean Production,* Rawson Associates/Macmillan, New York.

CHAPTER 18

Near-Net-Shape Processes

Reimund Neugebauer
Fraunhofer Institute for Machine Tools and Forming Technology

Klaus Herfurth
Technical University of Chemnitz

- **1. PROCESS CHAINS IN PART MANUFACTURING 562**
  - 1.1. Near-Net-Shape Processes: Definition and Limitations 563
  - 1.2. Goals and Benefits 564
  - 1.3. Preconditions for Near-Net-Shape Manufacturing 564
- **2. AN APPROACH TO NEAR-NET-SHAPE PHILOSOPHY: PRINCIPLE AND POSSIBILITIES 565**
  - 2.1. Component Shaping Techniques (Selection) 566
    - 2.1.1. Casting and Powder Metallurgical Techniques (Primary Shaping) 566
    - 2.1.2. Bulk Metal Forming Techniques 567
  - 2.2. Special Applications 568
- **3. NEAR-NET-SHAPE MANUFACTURING EXAMPLES IN ADVANCED PROCESS CHAINS 568**
  - 3.1. Casting: Selected Examples 568
  - 3.2. Powder Metallurgy: Manufacturing Examples 572
    - 3.2.1. Hot Isostatic Pressing (HIP) 572
    - 3.2.2. Powder Forging 574
  - 3.3. Cold-formed Near-Net-Shape Parts: Examples of Applications 575
    - 3.3.1. Extrusion 575
    - 3.3.2. Swaging 577
    - 3.3.3. Orbital Pressing 579
  - 3.4. Semihot-formed Near-Net-Shape Components: Examples of Applications 580
    - 3.4.1. Semihot Extrusion 580
    - 3.4.2. Semihot Forging 581
  - 3.5. Hot-formed Near-Net-Shape Components: Examples of Applications 581
    - 3.5.1. Precision Forging 581
    - 3.5.2. Hot Extrusion 582
    - 3.5.3. Axial Die Rolling 584
  - 3.6. Special Technologies: Manufacturing Examples 584
    - 3.6.1. Thixoforging 584
    - 3.6.2. Near-Net-Shape Processes for Prototyping 586
- **4. NEAR-NET-SHAPE PRODUCTION: DEVELOPMENT TRENDS 586**
- **REFERENCES 587**
- **ADDITIONAL READING 587**

1. PROCESS CHAINS IN PART MANUFACTURING

Part manufacturing is aimed at producing components for various destinations and functionalities. These components can be either autonomous products or parts of a more complex product. Components are characterized by their geometry and material properties. They have to cope with specific quality requirements depending on the corresponding part function and the individual application. The number of required components which determines the type of manufacturing—e.g., in small and

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

medium series or large-lot production—depends on the part's destination. In conclusion, the type of manufacturing mainly results from a wide range of possible preconditions.

The nature of manufacturing itself is to transform in steps the different variants of the initial material that is available for manufacturing technology into the final state predefined by the component's target geometry. For this procedure, very different process stages, depending on the component itself and the manufacturing conditions, are required. These steps can be realized by various mechanical, thermal, and chemical manufacturing techniques. The process chain represents the technological sequence of the individual process steps, which guarantees adequate fabrication of the component (see Figure 1).

As a rule, to manufacture a given component, different process chains are possible, posing the question of optimizing manufacturing processes and subsequently the process chains that should be applied. The first goal is to keep the process costs to a minimum at constant or even enhanced product quality. Another goal is to improve process productivity. Environmental protection, careful use of resources, and use of reproductive materials are increasingly important considerations.

Within this context, *near-net-shape* (nns), which involves optimizing the manufacturing process and particularly shortening the process chains, describes one trend in part manufacturing development.

1.1. Near-Net-Shape Processes: Definition and Limitations

For a better understanding of the near-net-shape principle, all explanations should be based on the different characteristics of the manufacturing techniques employed for shaping within the fabrication of components. These are methods that, on the one hand, create shapes by removing material from an initial state, including cutting methods such as turning, drilling, milling, and grinding as well as some special techniques. In contrast, shaping can be realized by distributing or locating material in an intentional way. The primary shaping methods (e.g., casting, powder metallurgy), metal-forming technology, as well as similar special techniques are based on this principle.

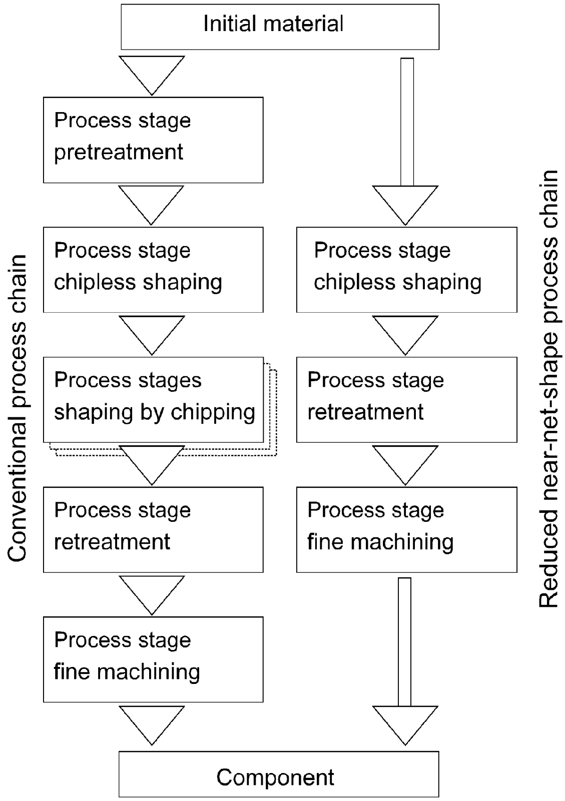

Initial material
Process stage pretreatment
Process stage chipless shaping
Process stages shaping by chipping
Process stage retreatment
Process stage fine machining
Process stage chipless shaping
Process stage retreatment
Process stage fine machining
Component
Conventional process chain
Reduced near-net-shape process chain

Figure 1 Process Chain Variants: Examples.

Starting with this distinction and bearing in mind the heterogeneous terminology in technical literature, we can define near-net-shape production as follows:

> Near-net-shape production is concerned with the manufacture of components, whereby shaping is realized mostly by nonchipping manufacturing techniques and finishing by cutting is reduced to a minimum. Near-net-shape manufacturing involves such objectives as reducing the number of process stages, minimizing costs, and guaranteeing enhanced product quality.

Essentially nonchipping shaping is necessary if materials such as compounds and ceramics are characterized by poor machinability.

The near-net-shape principle carried to its limits finally results in net-shape production that makes finishing by cutting totally unnecessary. In net-shape manufacturing, shaping is performed only with nonchipping techniques. The decision on the extent of residual cutting work—that is, the extent to which finishing by cutting can be reasonably diminished or even completely eliminated—depends on the associated manufacturing costs.

Thus, in some cases it may be more economical to manufacture a shape element by cutting than by metal forming. Maintaining a certain minimum value for finishing by cutting may be necessary for technical reasons as well. For example, a disturbing surface layer (skin after casting or decarburized skin caused by hot working) that results from a nonchipping shaping procedure will have to be removed by cutting.

Near-net-shape manufacturing can cover the whole component, but also only parts of it, such as functional surfaces of outstanding importance.

1.2. Goals and Benefits

To keep the manufacturing industry competitive in today's global markets, innovative products must constantly be developed. But high product quality also has to be maintained at reasonable costs. At the same time, common concerns such as protecting the environment and making careful use of resources must be kept in mind. Near-net-shape manufacturing has to cope with these considerations.

The goals of near-net-shape production are to reduce the manufacturing costs and make the manufacturing procedure related to a given product more productive. These goals can be achieved by employing highly productive nonchipping shaping techniques and minimizing the percentage of cutting manufacturing, thus generating more efficient process chains consisting of fewer process steps. Value adding is thus shifted to the shaping methods as well. For that reason, enhancement of shaping procedures involves great development potential.

An increasing variety of advanced products with new functionality and design are being developed. However, novel components characterized by higher manufacturing demands are emerging. Lightweight manufacturing, reduction of moving mass at energy-consuming assemblies or products, and the application of sophisticated multifunctional parts that save space and weight are typical manifestations. Frequently, when nonchipping shaping methods are used, geometrically complex components can be realized much better than would be possible with conventional cutting techniques perhaps used in conjunction with joining operations. With enhancing near-net-shape manufacturing as a must, the specific features of the different chipless shaping methods, such as low-waste and environmentally compatible production, increase in product quality, and better characteristics for use, can be taken advantage of.

1.3. Preconditions for Near-Net-Shape Manufacturing

As a precondition for developing near-net-shape manufacturing for and applying it to a given manufacturing task, efficient nonchipping shaping techniques that enable working accuracy commensurate with that of cutting techniques are needed. For guidelines for some achievable cutting accuracy values, see Table 1. Table 2 includes the corresponding values for nonchipping shaping techniques.

When the values in Tables 1 and 2 have been compared, a wide variety of feasible nns methods is available. At the same time, the differences that are still apparent pose a challenge to further development and improvement of nonchipping shaping methods. A trend toward still-higher accuracy and productivity of cutting techniques can also be observed. Thus, synergy effects that act mainly on the design of nns processes are emerging. The new quality of part-manufacturing development towards nns technologies is also related to the complex consequences of all those factors influencing quality and efficiency in part manufacturing. Figure 2 represents diagrammatically the principal dependencies and relationships.

From their origin, the influencing factors mentioned above can be assigned to workpiece, process, or equipment. In component design and when choosing the part material, apart from the component's function, the requirements resulting from the foreseen near-net-shape production the workpiece must fulfill must be considered. In cases where shaping has to be realized by a metal forming procedure, the material should possess sufficient formability. Profound knowledge of the theoretical and practical

TABLE 1 Shaping: Cutting Techniques/Finishing

Manufacturing Techniques	Obtainable Accuracy Values/IT Qualities										
	3	4	5	6	7	8	9	10	11	12	13
Turning				░	▓	▓	▓	▓			
Hard metal turning	░	░	░								
Milling				░	▓	▓	▓	▓	▓	▓	
Cylindrical grinding			▓	▓	▓	▓					
Lapping		▓	▓								
Honing	▓	▓	▓								

▓ Usually obtainable
░ Obtainable by special measures

fundamentals of the manufacturing technique to be employed within the process chain is also very important. This expertise is a fundamental precondition for optimizing the entire process chain as well as each technique individually with respect to the manufacturing result and working efficiency. For such optimization tasks, FE simulation of manufacturing operations and production sequences is becoming more and more important. Finally, interaction between the development of manufacturing techniques and expanding the range of application of near-net-shape technologies exists.

With regard to equipment, the chosen parameters and manufacturing constraints must be maintained in order to guarantee the required process reliability.

2. AN APPROACH TO NEAR-NET-SHAPE PHILOSOPHY: PRINCIPLE AND POSSIBILITIES

The basic idea underlying the near-net-shape philosophy is to approximate, as far as possible, an initial material to the target shape of the final product by means of innovative techniques of shaping and structure formation. As a rule, all measures are estimated by their contribution to achieving economic effects such as minimized costs and higher productivity. With respect to the production of components, the trend is to displace chipping operation stages with much more efficient casting, powder metallurgy, and bulk metal forming processes. Inspired by technological progress, the term *near-net-shape* has also been widened to include other ranges such as the fabrication of special and semifinished materials. The criteria for deciding upon the techniques to be applied for creating a special shape and structure do not have to be those known from part manufacturing. Thus, for instance, in the process chain of near-net-shape casting of semifinished materials, the percentage of casting vs. bulk metal forming is greater than in conventional fabrication of semifinished material.

TABLE 2 Shaping: Nonchipping Techniques

Manufacturing Techniques	Obtainable Accuracy Values/IT Qualities										
	6	7	8	9	10	11	12	13	14	15	16
Metal forming											
Die forging		░	░	░	░	░	▓	▓	▓	▓	▓
Hot extrusion					░	░	▓	▓	▓	▓	▓
Semi-hot extrusion				▓	▓	▓					
Cold extrusion	░	▓	▓	▓	▓	▓					
Swaging		▓	▓	▓	▓	▓					
Rolling		░	▓	▓	▓	▓					
Primary shaping											
Sintering	░	░	░	▓	▓	▓	▓	▓			
PM injection molding			▓								
Powder forging			▓	▓	▓						
Investment casting				▓	▓	▓	▓	▓			
Die casting				▓	▓	▓	▓	▓			

▓ Usually obtainable
░ Obtainable by special measures

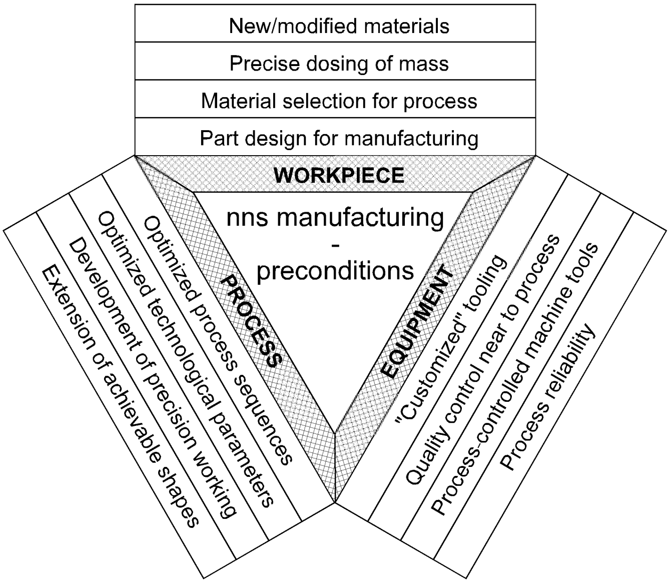

New/modified materials
Precise dosing of mass
Material selection for process
Part design for manufacturing
WORKPIECE
nns manufacturing
-
preconditions
PROCESS
Optimized process sequences
Optimized technological parameters
Development of precision working
Extension of achievable shapes
EQUIPMENT
"Customized" tooling
Quality control near to process
Process-controlled machine tools
Process reliability

Figure 2 Preconditions for Developing Near-Net-Shape Processes.

2.1. Component Shaping Techniques (Selection)

2.1.1. Casting and Powder Metallurgical Techniques (Primary Shaping)

Primary shaping is the manufacturing of a solid body from an amorphous material by creating cohesion. Thus, primary shaping serves originally to give a component an initial form starting out from a material in an amorphous condition. Amorphous materials include gases, liquids, powder, fibers, chips, granulates, solutions, and melts. With respect to a product shape and follow-up processing, primary shaping can be divided into three groups:

1. Products made by primary shaping that will be further processed by forming, cutting, and joining. With respect to shape and dimensions, the final product is no longer similar to the product originally made by primary shaping. In other words, shape and dimensions are also essentially altered by means of techniques from other main groups of manufacturing processes. The manufacturing of flat products from steel, which is widely realized by casting, is one near-net-shape application. Thereby, later shaping by rolling can be kept to a minimum.
2. Products made by primary shaping whose forms and dimensions are essentially similar to those of the finished components (e.g., machine parts) or final products. The shape of these products corresponds to the product's purpose to the maximum extent. Obtaining the desired final form as well as final dimensions mostly requires only a few, as a rule chipping, operations to be carried out at functional surfaces.
3. Metal powders produced by primary shaping, whereby the powders are atomized out of the melt. From powder, sintering parts are produced as a result of powder metallurgical manufacturing.

The fabrication of moldings from metals in foundry technology (castings) and powder metallurgy (sintered parts) as well as of high-polymer materials in the plastics processing industry yields significant economic advantages, such as:

- The production of moldings is the shortest way from raw material to finished product. Forming, including all related activities, is thereby bypassed. A form almost equal to the component's final shape is achieved in only one direct operation, and masses ranging from less than 1 g to hundreds of tons can thus be handled.

- Maximum freedom for realizing shapes that cannot be achieved by any other manufacturing technique is enabled by manufacturing of moldings that are primarily shaped from the liquid state.
- Primary shaping can also be applied to materials that cannot be processed with other manufacturing techniques. An advantageous material and energy balance is ensured by the direct route from raw material to the molding or the final product.
- Components and final products of better functionality, such as moldings of reduced wall thickness, diminished allowances, fewer geometric deviations, and enhanced surface quality (near-net-shape manufacturing), can be produced to an increasing extent due to the constantly advancing primary shaping methods.
- Great savings in materials and energy can be realized with castings and sintered parts. Positive consequences for environmental protection and careful use of natural resources can be traced back to these effects.

2.1.1.1. Primary Shaping: Manufacturing Principle In general, the technological process of primary shaping methods can be divided into the following steps:

- Supply or production of the raw material as an amorphous substance
- Preparation of a material state ready for primary shaping
- Filling of a primary shaping tool with the material in a state ready for primary shaping
- Solidification of the material in the primary shaping tool, e.g., usually the casting mold
- Removal of the product of primary shaping from the primary shaping tool

For a survey of primary shaping methods, see Figure 3.

2.1.2. Bulk Metal Forming Techniques

Aiming at the production of parts described by defined geometry and dimensions, all bulk metal forming techniques are based on the plastic alteration of the form of a solid body (whose raw material is massive in contrast to sheet metal forming) whereby material cohesion is maintained. The form is altered (forming procedure) due to forces and moments applied from the outside and generating a stress condition able to deform the material—that is, to transform it into a plastic state, permanently deformable inside the region to be formed (forming zone). The form to be produced (target form) is

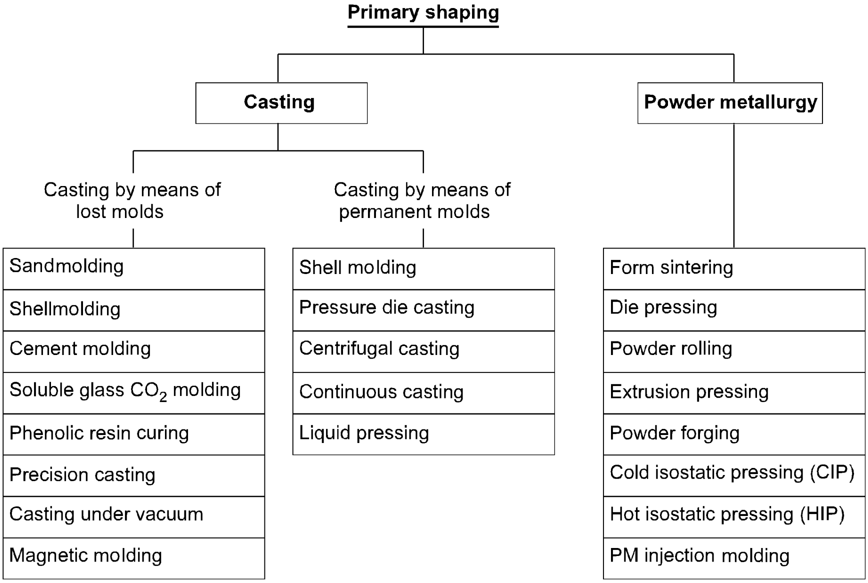

Primary shaping
Casting
Powder metallurgy
Casting by means of lost molds
Casting by means of permanent molds
Sandmolding
Shellmolding
Cement molding
Soluble glass CO2 molding
Phenolic resin curing
Precision casting
Casting under vacuum
Magnetic molding
Shell molding
Pressure die casting
Centrifugal casting
Continuous casting
Liquid pressing
Form sintering
Die pressing
Powder rolling
Extrusion pressing
Powder forging
Cold isostatic pressing (CIP)
Hot isostatic pressing (HIP)
PM injection molding

Figure 3 Survey of Primary Shaping Techniques.

mapped by the form of the tool's active surfaces and the kinematic behavior of each forming method. The requirements to be fulfilled by the material to be formed—summarized as the material formability property—result from the precondition that internal material cohesion not be destroyed during manufacturing. Formability of a material is influenced by the conditions under which the metal forming procedure is performed. Forming temperature, strain rate, and the stress condition within the forming zone are essential factors influencing formability.

A selection of essential bulk metal forming techniques to be used for realizing near-net-shape technologies in part manufacturing is given in Table 3.

All bulk metal forming techniques can be run at different temperature ranges dictated by the characteristics of each material (which in turn acts on the forming conditions), design of the process chain, and the forming result. According to their temperature ranges, cold, semihot, and hot forming processes can be distinguished.

2.1.2.1. Cold Forming In cold forming, the forming temperature is below the recrystallization temperatures of the materials to be formed. The workpiece is not preheated. In this temperature range, workpieces with close dimensional tolerances and high surface qualities can be manufactured. The cold forming process results in some work-hardening of the metal and thereby in enhanced mechanical components' properties. As a minus, the forces and energy required for cold forming as well as the tool stresses are much higher than for hot forming, and the formability of the material in cold forming is less.

2.1.2.2. Semihot Forming The warm forming temperature range starts above room temperature and ends below the recrystallization temperature from which the grain of metal materials such as steels begin to be restructured and the material solidification due to deformation degraded. For most steel materials, the semihot forming temperature ranges from 600–900°C. In comparison to hot forming, higher surface qualities and closer dimensional tolerances are obtained. The forces and energy required for semihot forming as well as the tool stresses are less than for cold forming. However, the additional thermal stress acting on the tool is a disadvantage. Consequently, in manufacturing, the requirements for exact temperature control are high.

2.1.2.3. Hot Forming In hot forming processes, the workpiece is formed at temperatures above the temperature of recrystallization of the corresponding metal. For the majority of materials, formability is higher due to higher forming temperatures. The forces and energy required for hot forming, as well as the tool stresses, are essentially lower than those for cold and semihot forming. Surfaces of poor quality due to scaling and decarburized skin (for steel materials: reduced carbon content in marginal layers near the surface) and wider tolerances at the component's functional surfaces, which in turn result in increased allowances for the necessary follow-up chipping operation, are disadvantageous. Further limitations are caused by the forming tools' grown thermal stress and the high expenditure involved in heating and reworking the part.

2.2. Special Applications

The near-net-shape technologies of part manufacturing also involve applications whose type of shaping can be assigned to none of the conventional primary shaping or forming technologies. This concerns, for instance, the thixoforming method, which involves processing metals at temperatures between the molten and partially solidified state so that the working conditions are between hot forming and casting. Depending on whether the component is shaped on a die-casting or forging press, the procedure is termed *thixocasting* or *thixoforging*. A specific globular structure of the initial material is required to transform metals into a thixotropic state. During heating to produce the thixotropic condition, the matrix phase, which becomes liquid at a lower temperature, is molten first and the metal becomes pulpy and easily formed.

Thixoforming is able to produce complex components of sophisticated form with minimum use of material and energy in only one operation.

Generic part manufacturing, represented by the rapid prototyping techniques, occupies a special position. The rapid prototyping principle is based on successive sedimentation of material (shaping layer by layer). These techniques are used not only for manufacturing of patterns but also for producing tools in a near-net-shape manner, such as for casting.

3. NEAR-NET-SHAPE MANUFACTURING EXAMPLES IN ADVANCED PROCESS CHAINS

3.1. Casting: Selected Examples

The main advantage of shaping by casting or powder metallurgy is the realization of near-net-shape production of castings and sintered parts, thereby minimizing cutting processing and drastically shortening the process chains due to fewer process stages. The process chain is dominated up to the finished part by chipless shaping.

TABLE 3 Bulk Metal Forming Techniques (Selection)

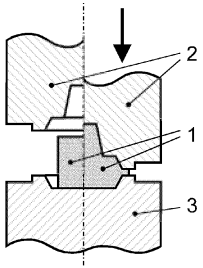
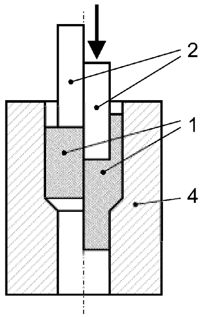
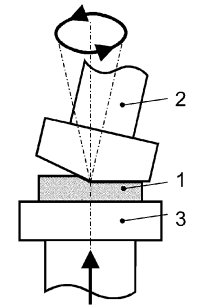
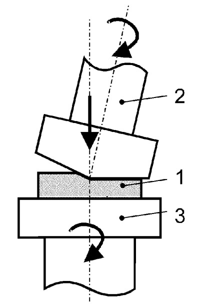
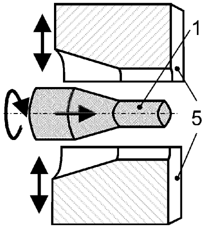
Manufacturing Technique	Process Principle	Technological Background	Obtainable Manufacturing Results
Die forging		• Forming under compressive conditions, translational tool motion • Shaping by means of tools replicating the part shape(dies) • with or even without loss of material (with flash, low flash, flashless) • Primarily hot forming, but also warm forming • Different design types depending on the forming machine (hammer, press, horizontal forging machine)	• Manifold of shapes (flat, cramped long parts with/without secondary shape elements) • Components of higher accuracy (close tolerance forging) • Components with functional surfaces ready to be installed • Surface roughness $R_a = 1.6 \ldots 12.5$ μm and $R_z = 16 \ldots 400$ μm
Extrusion		• Forming under compressive conditions, translational tool motion • Shaping by squeezing the material from almost closed tool installations • Varying kinematics of metal flow • Cold, semihot, hot forming	• Wide variety of forms, also due to the combination with other techniques (upsetting, pressing in) • Very close manufacturing tolerances can be kept • Components with functional surfaces ready to be installed (e.g., gearings) • Surface roughness $R_a = 0.8 \ldots 12.5$ μm and $R_z = 6.3 \ldots 100$ μm

TABLE 3 (*Continued*)

Manufacturing Technique	Process Principle	Technological Background	Obtainable Manufacturing Results
Orbital pressing		• Incremental forming under compressive conditions • One die moves on rolling contact in a tumbling manner, the other is fixed • Varying tumbling motion such as circular, helical, linear orientation • Shaping: one workpiece face is moving on rolling contact + translationally, another face is moving translationally • Semihot and cold forming	• Flat parts, wide variety of forms, high accuracy and most complicated contours on the face with only translational shaping • Manufacturing tolerances in the range from 0.05–0.2 mm can be maintained • Components with functional surfaces ready to be installed (e.g., gearings)
Axial die rolling		• Incremental forming under compressive conditions • Two rotating dies (both driven in the same direction or one driven, the other following) with crossing rotary axes • Shaping: one workpiece face moves on rolling contact + translationally, another workpiece face translationally • Primarily hot forming	• Flat rotationally symmetrical parts (ring- and disk-shaped) • Obtainable accuracies and surface qualities commensurate with die forging
Swaging		• Incremental forming under compressive conditions, tool moved in translational direction • Shaping by two or more tools bound to path in conjunction with the workpiece motion (rotation + translation or rotation, only) • Cold and hot forming	• Bar- and rod-shaped parts with straight or shouldered outer and inner contours, different profiles • Precision parts ready to be installed as a result of cold forming

1: workpiece; 2: upper tool/die; 3: lower tool; 4: bottom die; 5: plasticizing jaw.

Development in shaping by casting is focused on two directions. First, the components become increasingly closer to the finished parts. Second, many single parts are aggregated to one casting (integral casting). Both directions of development are realized in all variants of casting technology. However, in precision casting, casting by low pressure, and pressure die casting, the material and energy savings to be achieved are especially pronounced.

For evaluation, the manufacturing examples in Figures 4 and 5 were considered starting from melting up to a commensurable part state.

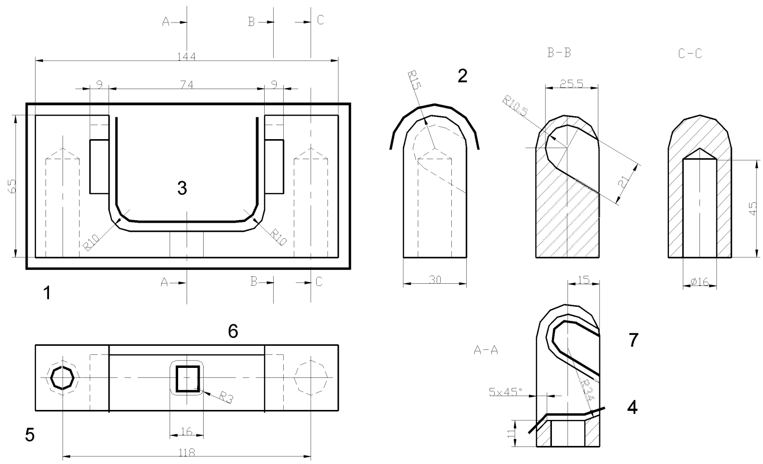
Figure 4 shows a technical drawing of a flat part that had previously been produced by cutting starting from a bar, that is now made as a casting (malleable cast iron) using the sand-molding principle. In cutting from the semifinished material, material is utilized at only 25.5%. As a result of shaping by casting, utilization of material was increased to 40%. The effects of shaping by casting become evident in the energy balance (primary energy). For cutting the flat part from the solid, 49.362 GJ/t parts are required. For shaping by casting, 17.462 GJ/t parts are required. Consequently, 64.6% of the energy can be saved. Compared to cutting of semifinished steel material, for part manufacturing about a third as much primary energy is required.

The doorway structure of the Airbus passenger door (PAX door: height about 2100 mm; width about 1200 mm) is illustrated in Figure 5.

In conventional manufacturing of the doorway structure as practiced until now, apart from the standard parts such as rivets, rings, and pegs, 64 milling parts were cut from semifinished aluminum materials with very low utilization of material. Afterwards, those parts were joined by about 500 rivets.

As an alternative technological variant, it is proposed that the doorway structure be made of three cast segments (die casting—low pressure). Assuming almost the same mass, in production from semifinished materials, the ratio of chips amounted to about 63 kg, whereas in casting, it can be reduced to about 0.7 kg. Thus, in casting, the chip ratio amounts to only 1% in comparison to the present manufacturing strategy. In the method starting from the semifinished material, about 175 kg of materials have to be molten, however, in shaping by casting, this value is about 78 kg—that is, 44.6%.

As a result of the energy balance (primary energy), about 34,483 MJ are required for manufacturing the doorway structure from the semifinished material. However, in shaping by casting, 15,002



Designation of shape elements:

1 Cuboid (sawing, milling: roughing and finishing)
2 External radius (milling: roughing and finishing)
3 Pocket (rough milling: roughing and finishing)
4 Pocket (milling, finishing)
5 Hole (drilling into the solid, boring)
6 Square profile (drilling into the solid, broaching)
7 Pocket element (milling)

Figure 4 Example: Flat Part.

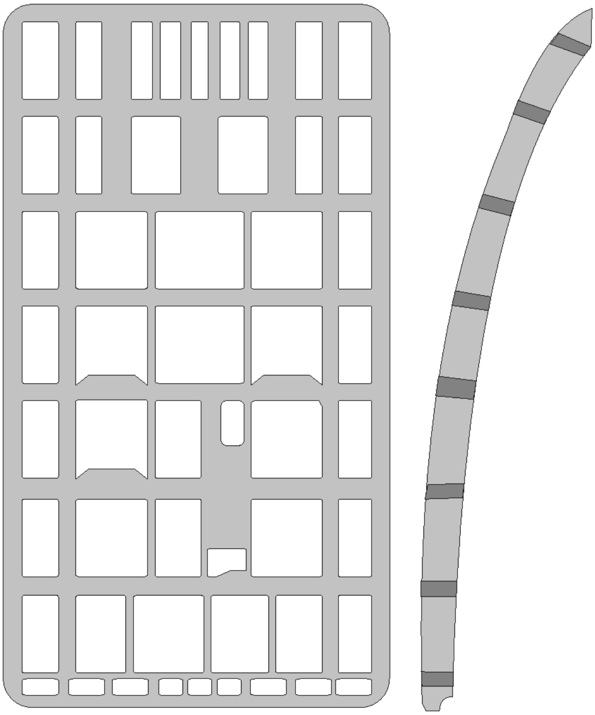


Figure 5 Example: Airbus Doorway.

MJ are needed—that is about 46%. The result of having drastically diminished the cutting volume due to nns casting can be clearly proven in the energy balance: in the variant starting from the semifinished material, 173 MJ were consumed for cutting; in casting, less than 2 MJ.

Today, the Airbus door constructions that have, in contrast to the studies mentioned above, cast as one part only are tried out. In this variant, 64 parts are aggregated to one casting (integral casting).

Figure 6 illustrates a welding-constructed tool holder consisting of seven steel parts (on the left), for which one consisting of two precision castings has been substituted (steel casting, GS-42CrMo4, on the right). Comprehensive cutting and joining operations can be saved through substitution by investment casting. In the variant where the nns design is based on precision casting, only a little cutting rework is required.

3.2. Powder Metallurgy: Manufacturing Examples

3.2.1. Hot Isostatic Pressing (HIP)

Since the 1950s, hot isostatic pressing technology has been developing rapidly in the field of powder metallurgy. Assuming appropriate powders are used, precision parts of minimum allowances can be produced. Applications provide solutions especially in cases of maraging steels, wear steels with high carbon content, titanium and superalloys, and ceramic materials.

As usual in this process, after the production of the appropriate powder, powder is enclosed in an envelope under an almost complete vacuum or inert gas atmosphere. Within this envelope, the powder is compacted under high pressure and temperature (hot isostatic pressing or compaction, see Figure 7). Almost isotropic structural properties are achieved by identical pressure acting from all sides. Thereby, the pressing temperature is partially far below the usual sintering temperature, which results in a fine-grained structure apart from the almost 100% density.

HIP technology can be applied to

- Postcompaction of (tungsten) carbide
- Postcompaction of ceramics

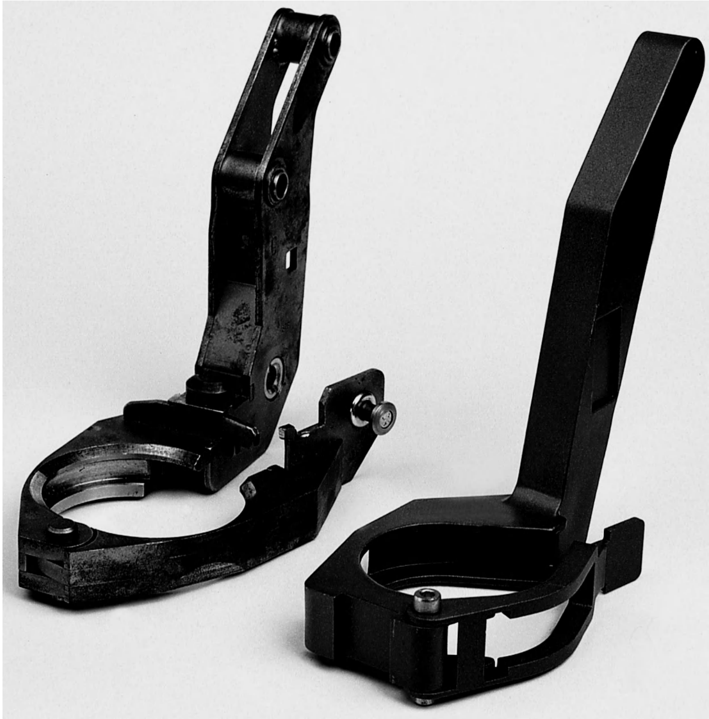


Figure 6 Tool Holder for a CNC Machine Tool (with Hinge).

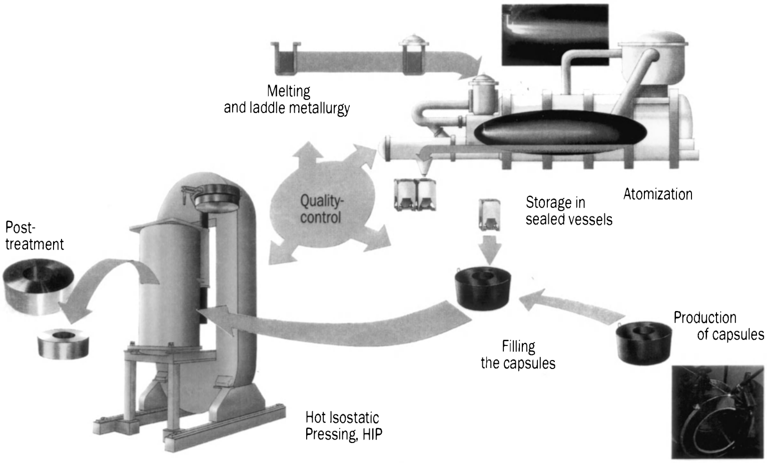
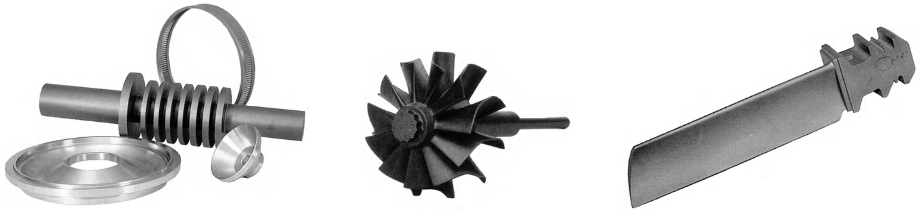

Melting
and laddle metallurgy
Atomization
Storage in
sealed vessels
Quality-
control
Post-
treatment
Production
of capsules
Filling
the capsules
Hot Isostatic
Pressing, HIP

Figure 7 Hot Isostatic Pressing (HIP): Process Flow. (From firm prospectus, courtesy ABB, Västerås, Sweden, 1989)



Metal powder parts Ceramic part Casting

Figure 8 Workpieces Made Using HIP Technology. (From firm prospectus, courtesy ABB, Västerås, Sweden, 1989)

- Powder compaction
- Postcuring of casting structures
- Reuse and heat treatment of gas turbine blades
- Diffusion welding
- Manufacturing of billets from powders of difficult-to-form metals for further processing

Workpieces made using the HIP process are shown in Figure 8. These parts are characterized by high form complexity and dimensional accuracy. With regard to the mechanical properties, parts made using HIP are often much better than conventionally produced components due to their isotropic structure.

3.2.2. Powder Forging

The properties of powder metallurgical workpieces are greatly influenced by their percentage of voids and entrapments (see Figure 9). For enhanced material characteristics, a density of 100% density must be striven for.

The application of powder or sintering forging techniques can provide one solution to achieving this objective. The corresponding process flow variants are given in Figure 10. Regarding the process sequence and the equipment applied, the powder metallurgical background of the process basically conforms to shaping by sintering.

According to their manufacturing principle, the powder-forging techniques have to be assigned to the group of precision-forging methods. Depending on part geometry, either forging starting from the sintering heat or forging with inductive reheating (rapid heating up) is applied. When forging the final shape, an almost 100% density is simultaneously generated in the highly stressed part sections. The part properties resulting from manufacturing by powder forging are equivalent to or frequently even more advantageous than those of usual castings. This becomes evident especially in the dynamic characteristics.

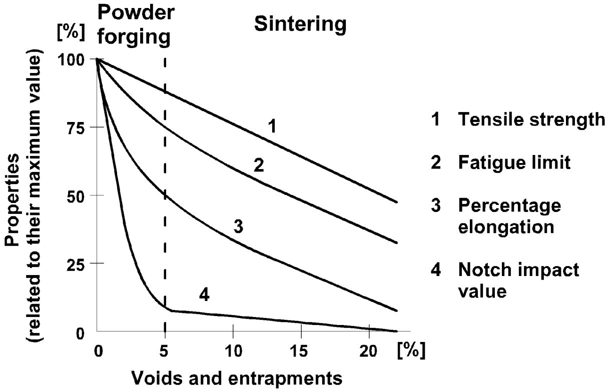


Figure 9 Workpiece Properties vs. the Percentage of Voids and Entrapments. (From Lorenz 1996)

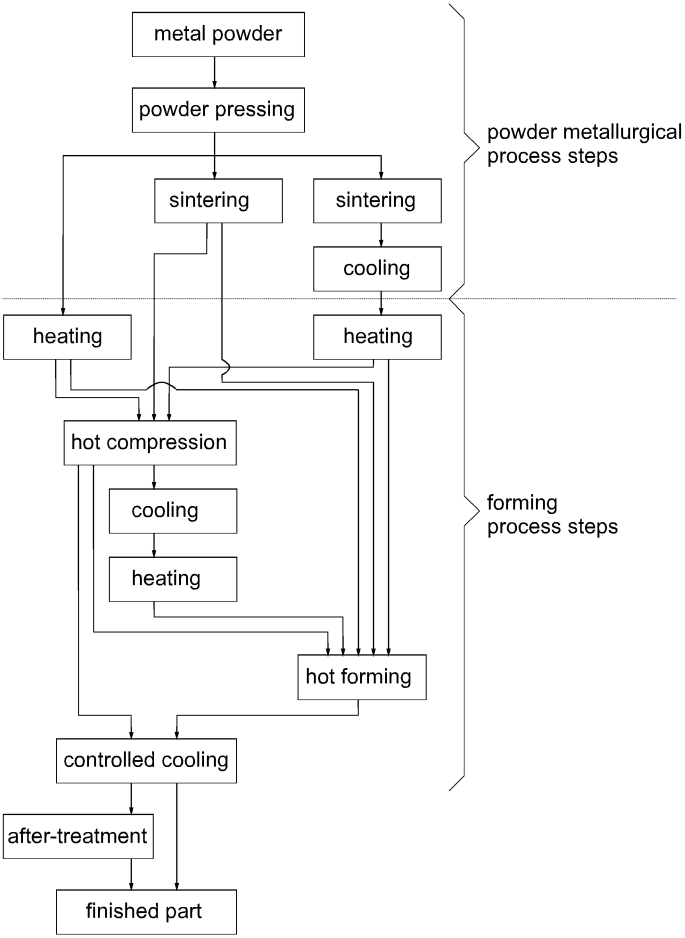

metal powder
powder pressing
sintering
sintering
cooling
heating
heating
hot compression
cooling
heating
hot forming
controlled cooling
after-treatment
finished part
powder metallurgical process steps
forming process steps

Figure 10 Powder and Sinter Forging. (From Lorenz 1996)

Figure 11 shows the stages of powder-forging processes for forming rotationally symmetrical parts.

The advantages of powder-forged parts can be summarised as follows:

- Material savings (flashless process)
- Low mass fluctuations between the individual parts
- High accuracy (IT8 to IT11)
- High surface quality
- Low distortion during heat treatment due to the isotropic structure
- High static and dynamic characteristic values

Some typical powder forging parts are shown in Figure 12.

3.3. Cold-formed Near-Net-Shape Parts: Examples of Applications

3.3.1. Extrusion

Cylindrical solid and hollow pieces including an axial direction slot of defined width and depth at the face are examples of typical mechanical engineering and car components. Parts like these can be found in diesel engines and pumps (valve tappets), in braking devices (as pushing elements), and for magnetic cores.

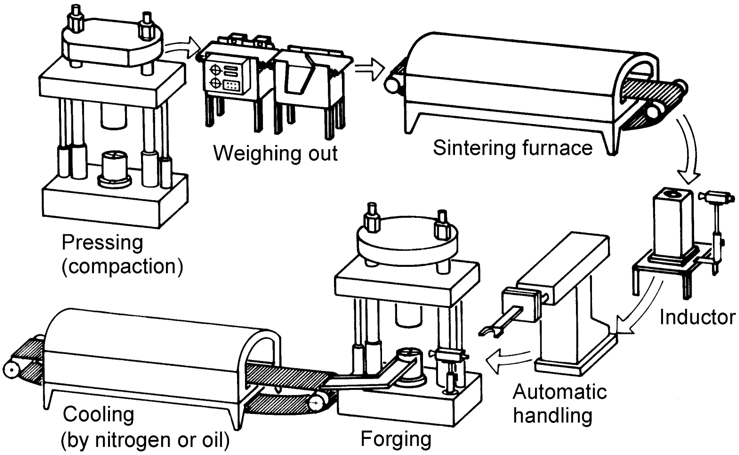


Figure 11 Powder Forging: Process Sequence. (From firm prospectus, courtesy KREBSÖGE, Radevormwald, Germany, 1994)

The valve tappet for large diesel motors made of case-hardening steel (16MnCr5), illustrated in Figure 13, is usually produced by cutting from the solid material, whereby the material is utilized at a rate of ≤40%. A near-net-shape cold-extrusion process was developed to overcome the high part-manufacturing times and low utilization of material and cope with large part quantities.

Starting from soft annealed, polished round steel, the valve tappet was cold extruded following the process chain given in Figure 14. The initial shapes were produced by cold shearing and setting, but also by sawing in the case of greater diameters. Surface treatment (phosphatizing and lubricating) of the initial forms is required for coping with the tribological conditions during forming. Near-net-shape extrusion can be carried out both in two forming steps (forward and backward extrusion) and in a one-step procedure by combined forward/backward extrusion. The number of forming steps depends on the workpiece geometry, particularly the dimensional ratios b/d and D/d.

Modified knee presses up to a nominal force of 6300 kN are employed to extrude the valve tappets. The output capacity is 25–40 parts per minute. As a result of the near-net-shape process, the material is utilized to a higher extent—about 80%. The sizes of slot width and hole diameter are delivered at an accuracy ready for installation, whereas the outer diameter is pressed at IT quality 9–10 ready for grinding. The number of process steps and part manufacturing time were reduced.

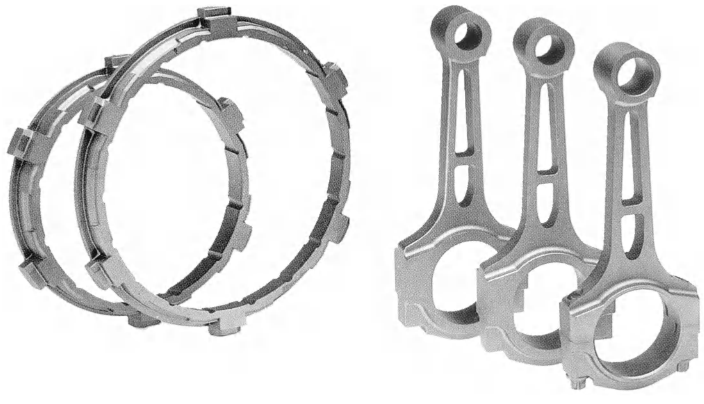


Figure 12 Workpieces Made Using Powder Forging. (From prospectus, courtesy KREBSÖGE, Radevormwald, Germany, 1994)

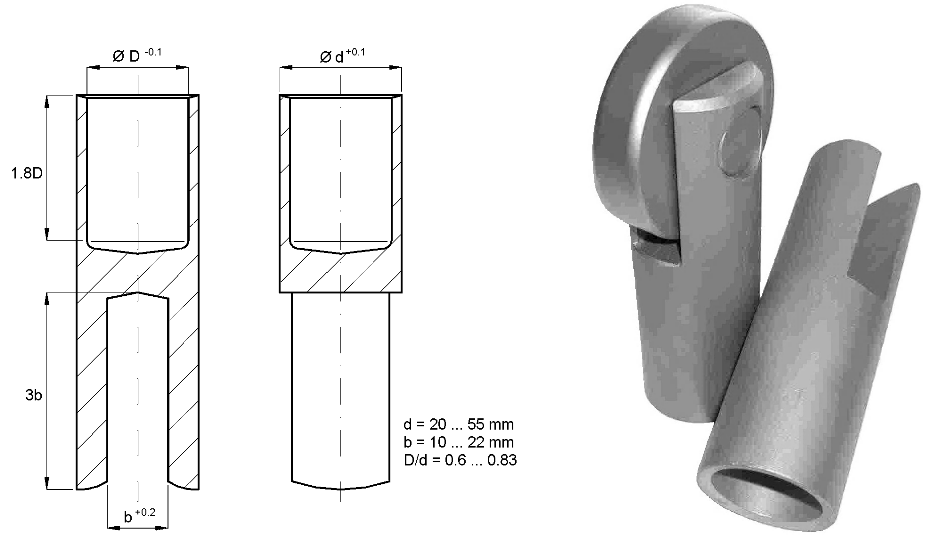

Ø D -0.1
Ø d +0.1
1.8D
3b
b +0.2
d = 20 ... 55 mm
b = 10 ... 22 mm
D/d = 0.6 ... 0.83

Figure 13 Valve Tappet for Large Diesel Motors: Ranges of Dimensions, Stamping Die, and Finished Part. (Courtesy ZI-Kaltumformung Oberlungwitz, Germany)

Apart from these effects, material cold solidification—caused by the nature of this technology—and continuous fiber orientation resulted in an increase in internal part strength.

3.3.2. Swaging

Swaging is a chipless incremental forming technique characterized by tool segments radially quickly pressing in an opposite direction on the closed-in workpiece. A distinction is made between the feeding method for producing long reduced profiles at relatively flat transition angles and the plunge method for locally reducing the cross-section at steep transition angles (see Figure 15).

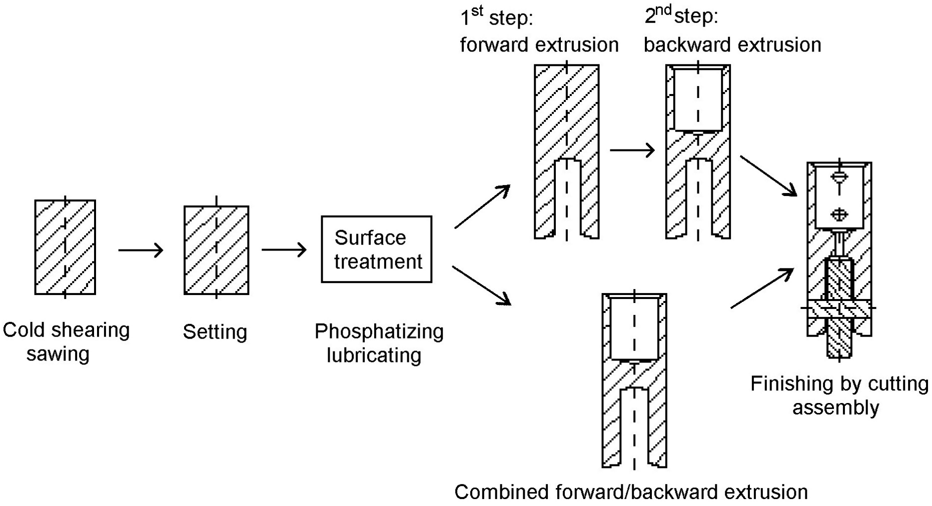

1st step: forward extrusion
2nd step: backward extrusion
Surface treatment
Cold shearing sawing
Setting
Phosphatizing lubricating
Finishing by cutting assembly
Combined forward/backward extrusion

Figure 14 Process Chain to Manufacture Valve Tappets for Large Diesel Motors by Cold Forming. (Courtesy ZI Kaltumformung GmbH, Oberlungwitz, Germany)

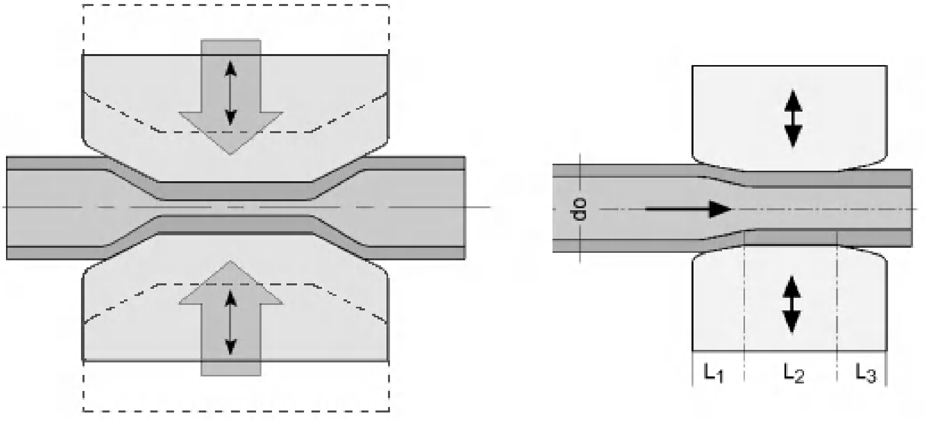


Figure 15 Schematic Diagram of Plunge and Feed Swaging.

A sequence as used in practice, from blank to nns finished part, consisting of six operations is described with the aid of the process chain shown in Figure 16. In the first process step, feed swaging is performed upon a mandrel to realize the required inner contour. The blank is moved in the working direction by oscillating tools. The forming work is executed in the intake taper. The produced cross-

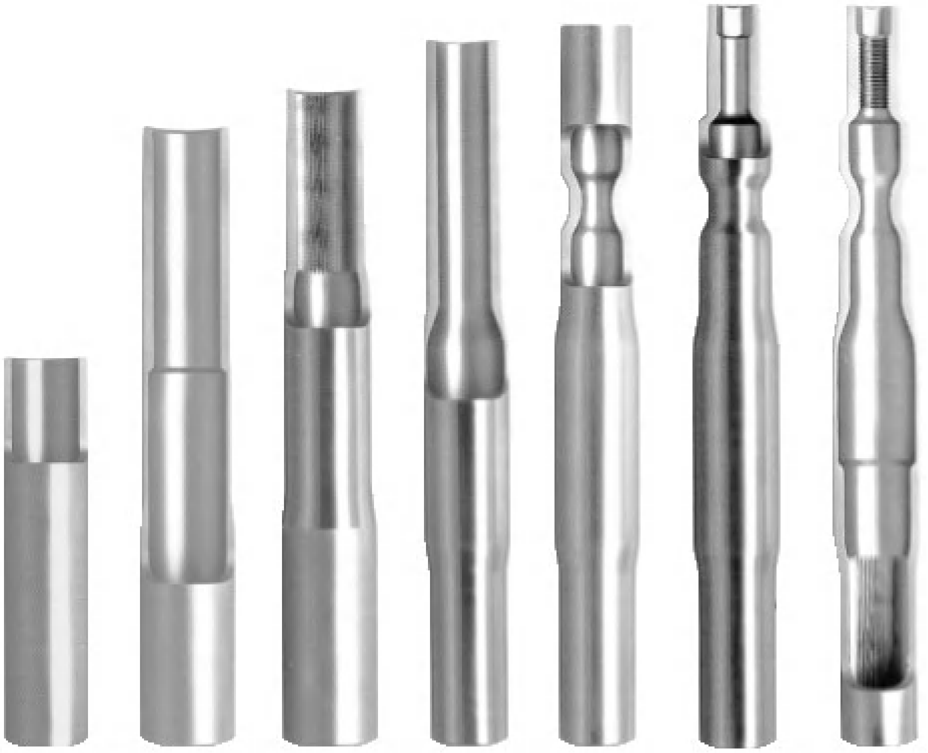


Figure 16 Process Chain Illustrated by Part Stages in Car Transmission Shaft Production. (Courtesy FELSS, Königsbach-Stein, Germany)

section is calibrated in the following cylinder section. Feeding is carried out when the tools are out of contact. In the second stage of operation, the required internal serration is formed by means of a profiled mandrel. The semifinished material having been turned 180° in the third process step, the opposite workpiece flank is formed upon the mandrel by feed swaging. To further diminish the cross-section in the marked section, plunge swaging is following as the fourth process step. This is done by opening and closing the dies in a controlled way superimposed upon the actual tool stroke. The workpiece taper angles are steeper than in feed swaging. The required forming work is performed in both the taper and the cylinder. In the plunge technique, the size of the forming zone is defined by the tools' length. A mandrel shaped as a hexagon is used to realize the target geometry at one workpiece face by plunge swaging (fifth step). In the sixth and final process step, an internal thread is shaped.

The obtainable tolerances to guarantee the nns quality at the outer diameter are basically commensurate with the results achieved on precise machine tools by cutting. Depending on material, workpiece dimensions, and deformation ratio, the tolerances range from ±0.01 mm to ±0.1 mm. When the inner diameter is formed upon the mandrel, tolerances of <0.03 mm are obtainable.

The surfaces of swaged workpieces are characterized by very low roughness and a high bearing percentage. As a rule, for plunging, the roughness is usually $R_a < 0.1 \mu$m, whereas in feed swaging, R_a is less than 1.0 μm. During plastic deformation, fiber orientation is maintained rather than interrupted as in cutting. Thus, enhanced functionality of the final product is guaranteed.

The material that can be saved and the optimization of weight are additional key issues in this forming procedure. The initial mass necessary for the production of bars and tubes diminishes because the usual cutting procedure is replaced by a procedure that redistributes the material. For an abundance of workpieces, it is even possible to substitute tubes for the bar material, thus reducing the component mass.

In swaging, the material has an almost constant cold work hardening over the entire cross-section. Nevertheless, even at the highest deformation ratios, a residual strain remains that is sufficient for a follow-up forming process.

Swaging rotationally symmetric initial shapes offers the following advantages:

- Short processing time
- High deformation ratios
- Manifold producible shapes
- Material savings
- Favourable fiber orientation
- Smooth surfaces
- Close tolerances
- Environmental friendliness because lubricant film is unnecessary
- Easy automation

Swaging can be applied to almost all metals, even sintered materials, if the corresponding material is sufficiently strained.

3.3.3. Orbital Pressing

Orbital pressing is an incremental forming technique for producing demanding parts that may also have asymmetric geometries. The method is particularly suited to nns processing due to its high achievable dimensional accuracy. The principle design of orbital pressing process chains is illustrated in Figure 17.

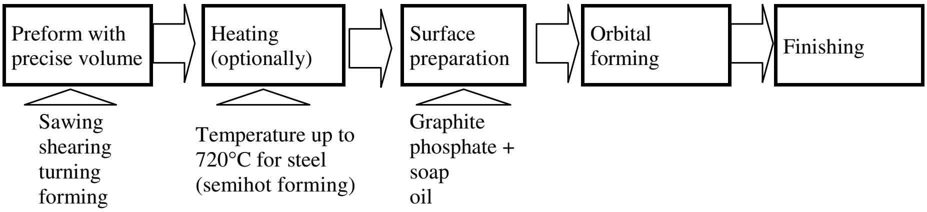


Figure 17 Process Chain for Orbital Pressing.

An accuracy of about IT 12 is typical for the upper die region. In the lower die, an accuracy up to IT 8 is achievable. Parts are reproducible at an accuracy of ±0.05 mm over their total length. Roughness values of $R_a = 0.3$ μm can be realized at functional surfaces. The accuracies as to size and shape are mainly determined by the preform's volumetric accuracy, whereas the obtainable roughness values depend on die surface and friction.

The advantages of orbital pressing compared to nonincremental forming techniques such as extrusion are a reduced forming force (roughly by factor 10) and higher achievable deformation ratios. As a result of reducing forming forces, the requirements for die dimensioning are lowered at diminished tool costs. Orbital pressing can be economically employed for the production of small to medium part quantities. For lower dies, tool life quantities of 6,000 to 40,000 parts are obtainable. The lifetime is about four times higher for upper dies. Concerning the usual machine forces, orbital presses are usually available at 2,000, 4,000, and 6,300 kN. For special applications, manufacturing facilities up to a pressing force of 1,600,000 kN are also available.

Generally, workpieces of up to 5 kg maximum weight, 250 mm diameter, and 220 mm maximum blank height are formed of steel types with low carbon content and nonferrous metals, such as copper and aluminum alloys.

The example in Figure 18 indicates an extremely high-stressed component of a large diesel motor (part weight 1.1 kg; flange diameter 115 mm; flange height 16 mm; total height 55 mm). With the use of the orbital pressing technique, for this workpiece, the process chain—originally consisting of turning, preforging, sandblasting, and cold coining of the profile—could be reduced to only two steps. Furthermore, the cavity could be formed at a greater depth and at higher accuracy, thus yielding significantly higher load capacity, reliability, and component life. In this example, the part costs were reduced by about 60% due to the application of orbital forging.

3.4. Semihot-formed Near-Net-Shape Components: Examples of Applications

3.4.1. Semihot Extrusion

In the automotive industry, the functional parts in particular are subjected to high-quality requirements. The successful use of semihot extrusion is demonstrated, for example, by the manufacture of triple-recess hollow shafts used in homokinematic joints in the drive system. The complex inner contours of the shaft have to be produced to high accuracy and surface quality. Producing these parts by cutting is not economically viable due to the high losses of material and the high investment volume required, so the inner contour is shaped completely (net shape) by forming. In forming technology, the required inner contour tolerance of ±0.03 mm can only by achieved by calibrating a preform of already high accuracy at room temperature. However, the material Cf 53, employed due to the necessary induction hardening of the inner contour, is characterized by poor cold formability. For that reason, shaping has to be performed at increased forming temperature but still with a high demand for accuracy. Shaping can be realized by means of a multistep semihot extrusion procedure carried out within a temperature interval between 820°C and 760°C in only one forming step. Semihot extrusion is performed on an automatic transfer press at the required accuracy, at optimum cost and also avoiding scaling. The component is shaped by semihot extrusion starting out from a cylindrical rod section without thermal and chemical pretreatment. The corresponding steps of operation are reducing the peg, upsetting the head, centering the head, and pressing and compacting the cup (see Figures 19 and 20). The advantages most relevant for calculation of profitability in comparing semihot

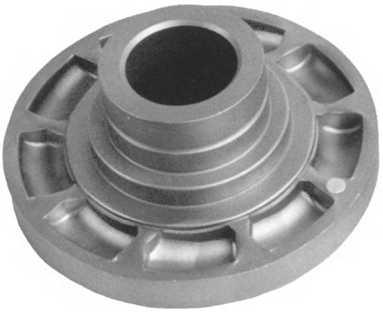


Figure 18 Basic Body of a Rotocup for Large Diesel Motors. (Courtesy ZI-Kaltumformung Oberlungwitz, Germany)

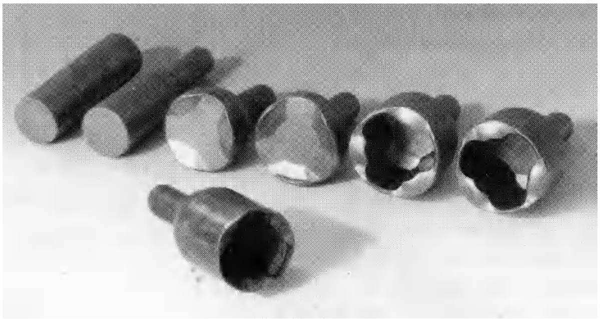


Figure 19 Original Production Stages of a Triple-Recess Hollow Shaft. (Courtesy Schuler Pressen, Göppingen, Germany)

extrusion of shafts to hot extrusion are caused by the low material and energy costs as well as postprocessing and investment expenditures.

3.4.2. Semihot Forging

Forging within the semihot forming temperature interval is usually performed as precision forging in a closed die. With closed upper and lower dies, a punch penetrates into the closed die and pushes the material into the still-free cavity. Bevel gears completely formed, including the gear being ready for installation (net shape process), are a typical example of application. Such straight bevel gears (outer diameter 60 mm or 45 mm and module about 3.5 mm) are employed in the differential gears of car driving systems. For tooth thickness, manufacturing tolerances of ±0.05 mm have to be maintained. In forming technology, the required tolerance can only be obtained by calibrating the teeth preformed at high accuracy at room temperature. The preform's tooth flank thickness is not to exceed about 0.1 mm compared to the final dimension. The shaping requirements resulting from accuracy, shape complexity, and the solid, alloyed case-hardening steel can only be met by semihot forming. Processing starts with a rod section that is preshaped without the teeth in the first stage (see Figure 21). The upset part is of a truncated cone shape. The actual closed die forging of the teeth is carried out as a semihot operation in the second stage. The bottom (thickness about 8 mm) is pierced in the last semihot stage. The cost efficiency effects of the semihot forming process compared to precision forging carried out at hot temperatures are summarized in Table 4.

Today, differential bevel gears of accuracy classes 8 to 10 are still produced by cutting, despite their large quantity. Low-cost forming technologies perspectively represent the better alternative, even though they require relatively high development costs.

3.5. Hot-formed Near-Net-Shape Components: Examples of Applications

3.5.1. Precision Forging

Which hot-forming technique to select for a given workpiece from a low-cost manufacturing viewpoint depends mainly on the part quantity and the extent to which the forming part is similar to the finished shape. In precision forging, where the forming part can be better approximated to the finished

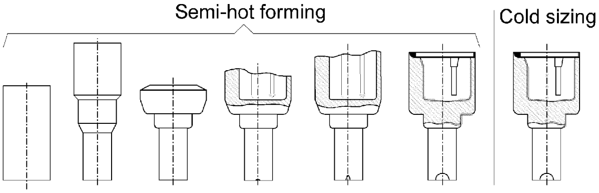


Figure 20 Semihot Forming Process Stages of a Triple-Recess Hollow Shaft. (From Körner and Knödler 1992)

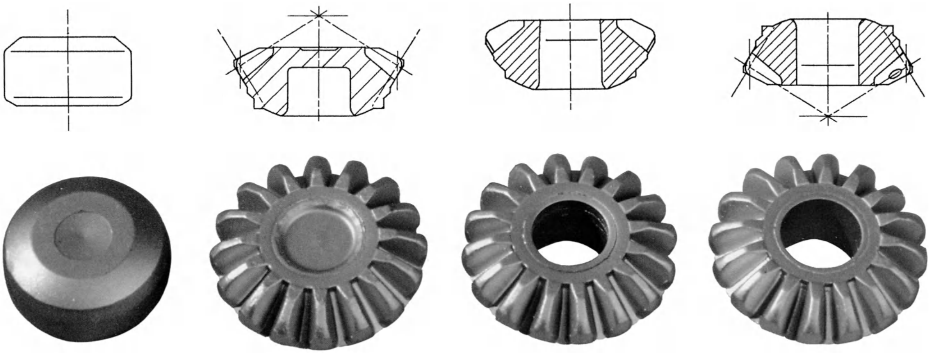


Figure 21 Production Stages of Axle-Drive Bevel Pinions. (From Körner and Knödler 1992)

part, the total cost, including tool and material costs as well as costs for the forming and cutting processes, can be decreased despite increasing blank costs.

The accuracy values obtained are above the die forge standards (e.g. in Germany, DIN 7526, forging quality E). As a goal, one process step in finishing by cutting is to be saved (this means either grinding or finishing is necessary). In forging, some dimensions of importance for the near-net-shape have closer tolerances.

The steering swivel shown in Figure 22 is an example of nns manufacturing using precision forging. The fixing points determining the spatial location are fitted to the dimensions of the finished part by additional hot calibrating.

Forged parts (hot forming) that make the subsequent cutting operation unnecessary or whose accuracy need only be enhanced at certain shape elements of the part (IT9 to IT6) by follow-up cold forming can also be produced by precision forging. Thus, formed components of even cold-pressed parts' accuracy can be made this way. This is especially important if cold forming cannot be realized due to too-high forming forces and tool stresses or too-low formability of the part material. From an economic viewpoint, precision forging is also more advantageous if the surfaces to be generated could otherwise only be produced by relatively expensive cutting techniques.

3.5.2. Hot Extrusion

Field spiders are usually made by die forging with burr and subsequent cold calibrating. High pressing forces are required due to formation of burr. The forged shape is approximated to the final shape only to a minor extent. For that reason, high forces also have to act in cold calibration. A nns manufacturing variant for field spiders consists of combining hot extrusion and cold forming.

TABLE 4 Cost-efficiency: Hot and Semihot-Forged Axle Drive Bevel Pinions in Comparison

	Hot-Forged Bevel Gear	Cost Share	Semihot-Forged Bevel Gear	Cost Share
Material	0.6 kg	16%	0.5 kg	13%
Manufacturing steps	1 Sawing, chamfering	8%	1 Sawing	6%
	2 Heating (with inert gas) Forming	33%	2 Heating (without inert gas) Forming	27%
	3 B-type graphite annealing (with inert gas)	7%	3 Controlled cooling	1%
	4 Cold calibrating	10%	4 Cold calibrating	10%
	5 Turning	17%	5 Turning	11%
Tool wear		9%		9%
Production costs		100%		77%

From Hirschvogel 1997.

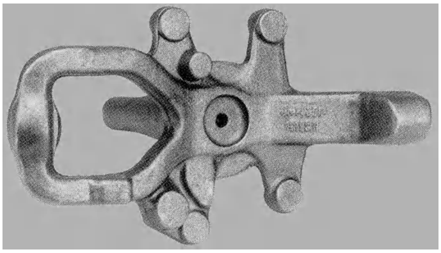


Figure 22 Precision Forging of a Steering Swivel (hot forged, simultaneously deburred, pierced and hot calibrated, weight 8.8 kg, material 41Cr4). (From Adlof 1995)

The process chain is characterized by the following procedural stages:

1. Shearing a rod section of hot-rolled round steel (unalloyed steel with low carbon content) with a mass tolerance of ± 1.5%
2. Heating in an induction heating device (780°C)
3. Hot extrusion on a 6,300-kN knee press performing the following forming stages (see also Figure 23):
    - Upsetting
    - Cross-extrusion
    - Setting of the pressed arms
    - Clipping
4. Cold forming (see Figure 24):
    - Backward extrusion of the hub
    - Forming of erected arms
    - Piercing of the hub
    - Calibrating.

An automated part-handling and special tool-cooling and lubricating system are installed to keep the forming temperatures and the tribologic conditions exactly according to the requirements for nns processes.

The advantages of hot extrusion/cold calibrating vs. die forging with burr/cold calibrating of field spiders can be summarized as follows:

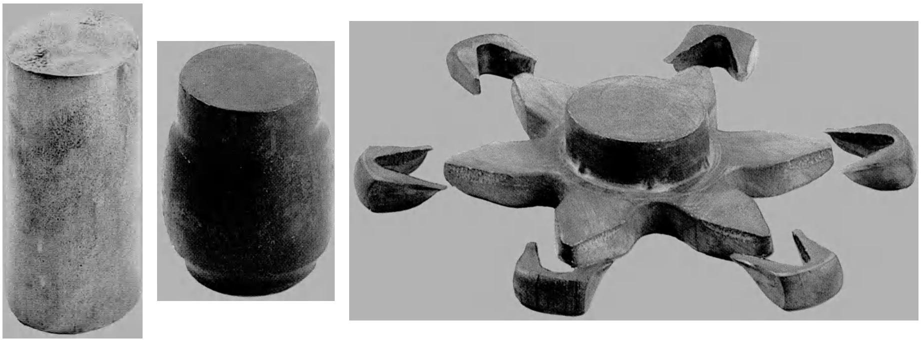


Figure 23 Hot Extrusion of Field Spiders. (From Körner and Knödler 1992)

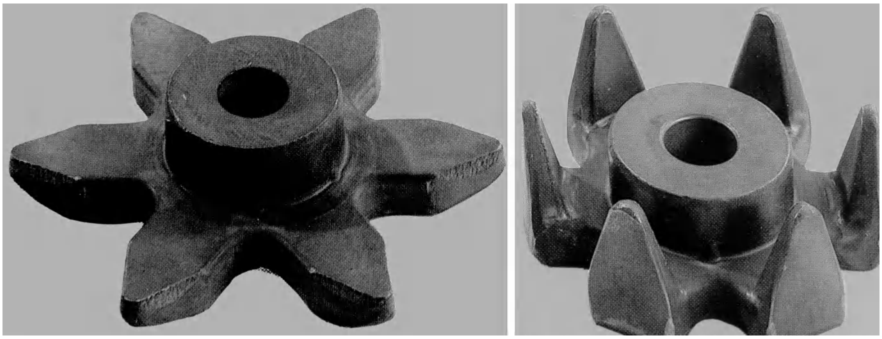


Figure 24 Cold Forming to Enhance Accuracy by Calibrating (applied to field spiders). (From Körner and Knödler 1992)

Hot extrusion:	*Die forging with burr:*
Pressing forces demand a 6,300-kN press	Pressing forces demand a 20,000-kN press
Low heating temperature (780°C)	High heating temperature (1100°C)
Descaling unnecessary	Descaling required
Cold calibrating: at 3,300 kN	Cold calibrating: at 10,000 kN

In calibrating the hot extruded part, forming is required to a lower extent due to better approximation to the final shape. This excludes the danger of cracks in the field spider.

3.5.3. Axial Die Rolling

In axial forging–die rolling, the advantages of die forging and hot rolling are linked. Disk-like parts with or without an internal hole can be manufactured. On the component, surfaces (e.g., clamping surfaces) can be shaped in nns quality, thus enabling later part finishing by cutting in only one clamping. Accuracy values of IT9 to IT11 are obtained because the axial die rolling machines are very stiff and the forming forces relatively low. Furthermore, the desired component shape can be approximated much better because no draft angles are necessary and small radii or sharp-edged contours can be formed. Within the workpiece, a favorable fiber orientation, resulting in reduced distortion after heat treatment, is achieved by burrless forming. The advantages of the nns hot-forming technology using axial die rolling vs. conventional die forging can be shown by a comparison of manufacturing costs of bevel gears for a car gear unit (see Figure 25). Cost reductions up to 22% as well as mass savings up to 35% are achieved by the nns technology.

3.6. Special Technologies: Manufacturing Examples

3.6.1 Thixoforging

Thixoforming processes make use of the favorable thixotropic material behavior in a semiliquid/ semi solid state between the solid and liquid temperatures. Within this range, material cohesion, which makes the metal still able to be handled as a solid body and inserted into a die mold, is kept at a liquid rate of 40–60%. This suspension is liquefied under shear stresses during pressing into a die mold. Due to the suspension's high flowability, complex geometries can be filled at very low forces, which represents the main benefit of this technique because part shapes of intricate geometry (thin-walled components) can be produced at low forming forces.

The process stages between thixocasting and thixoforging are different only in their last stage of operation (see Figure 26). With an appropriately produced feedstock material, the strands are cut into billets, heated up quickly, and formed inside a forging or casting tool.

A component made of an aluminum wrought alloy by thixoforging is illustrated in Figure 27. The forging has a very complex geometry but it was reproducible very accurately. The thick-walled component regions as well as parts with essential differences in wall thickness are characterized by a homogeneous structure free of voids.

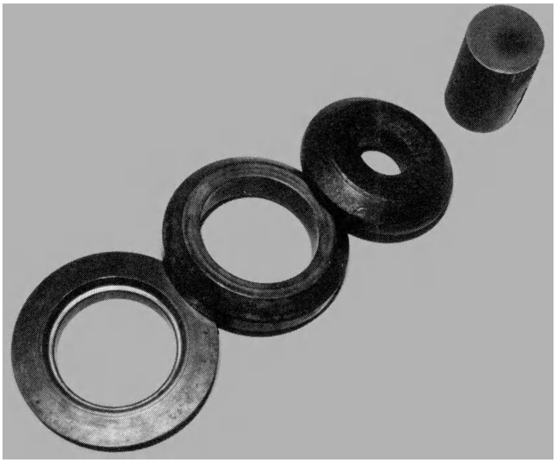


Figure 25 Axial Die Rolling of Car Bevel Gears. (From firm prospectus, courtesy SMS Eumuco, Leverkusen, Germany)

Industrial experience shows that components with filigree and complex part geometries, unfavorable mass distribution, undercuts, and cross-holes can be produced in one stage of operation in a near-net-shape manner by means of thixoprocesses.

Comparison of thixoforming to die casting shows the following results:

- 20% improvement in cycle times
- About 20% increase in tool life
- Raw material costs still up to 20% higher

The thixoprocesses are regarded as an interesting option for both the forging industries and foundries. With them near-net-shape technology is added to conventional procedures.

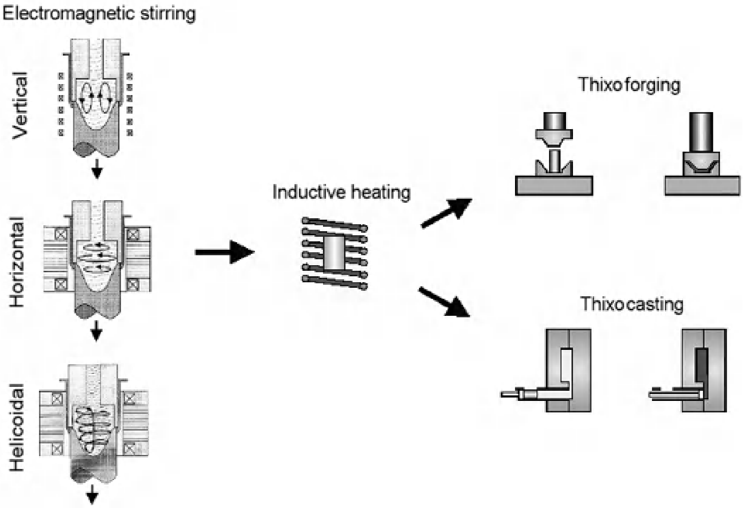

Electromagnetic stirring
Vertical
Horizontal
Helicoidal
Inductive heating
Thixoforging
Thixocasting

Figure 26 Thixoforging and Thixocasting: Process Stages. (Courtesy EFU, Simmerath, Germany)

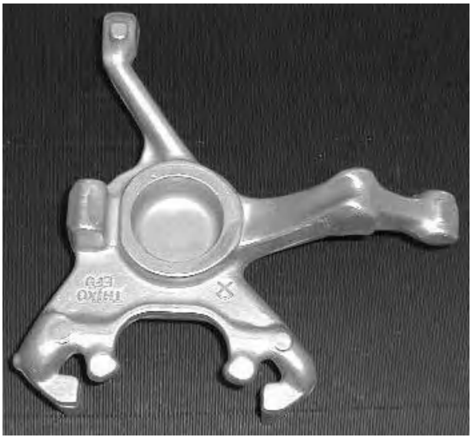


Figure 27 Steering Knuckle Made by Thixoforging. (Courtesy EFU, Simmerath, Germany)

3.6.2 Near-Net-Shape Processes for Prototyping

The ability to create prototypes of properties conforming totally with the series properties quickly at low cost has become increasingly important because the engineering lead time dictated by the market has drastically diminished. Apart from the classical methods of generating prototypes, such as cutting and casting, generic manufacturing techniques are increasingly being used. In cutting or casting, tools, dies, or fixtures are required for fabricating the prototype, whereas prototyping by means of generic techniques is based on joining incremental solid elements. Thus, complex components can immediately be made from computerized data, such as from a CAD file, without comprehensive rework being necessary.

In contrast to cutting technologies, where the material is removed, in generic manufacturing techniques (rapid prototyping), the material is supplied layer by layer or is transformed from a liquid or powder into a solid state (state transition) by means of a laser beam. The most essential rapid prototyping techniques (also called solid freeform manufacturing or desktop manufacturing) are:

- Stereolithography (STL)
- Fused deposition modeling (FDM)
- Laminated object manufacturing (LOM)
- Selective laser sintering (SLS)

These techniques are able to process both metals and non metals. As a result of intensive research, disadvantages of these techniques, such as high fabrication time, insufficient accuracy, and high cost, have been widely eliminated. Today, maximum part dimension is limited to less than 1000 mm.

Near-net-shape processes for prototyping and small-batch production by means of rapid prototyping techniques can be characterised by the process chains listed in Figure 28.

In principle, prototypes or patterns that in turn enable the manufacture of prototypes and small batches in combination with casting processes can be produced directly by means of rapid prototyping techniques. Rapid prototyping can also be applied to manufacturing primary shaping and forming tools to be used for the fabrication of prototypes.

4. NEAR-NET-SHAPE PRODUCTION: DEVELOPMENT TRENDS

Near-net-shape production will continue to expand in the directions of application described in this chapter. An important question for decision making is how far the objectives of industrial production and common global interests in reduction of costs, shorter process chains and production cycles, and environment- and resource-protecting manufacturing can be made compatible one with each other and whether these objectives are feasible in an optimal manner. Besides chipless shaping, for near-net-shape process planning, the state of the art of both chipless shaping and cutting techniques has to be paid attention to because a reasonable interface between chipless and cutting manufacturing is highly significant. Thus, for instance, in the case of a nns strategy consisting of chipless shaping, heat treatment, and grinding, in future, grinding could be substituted for to a greater extent by

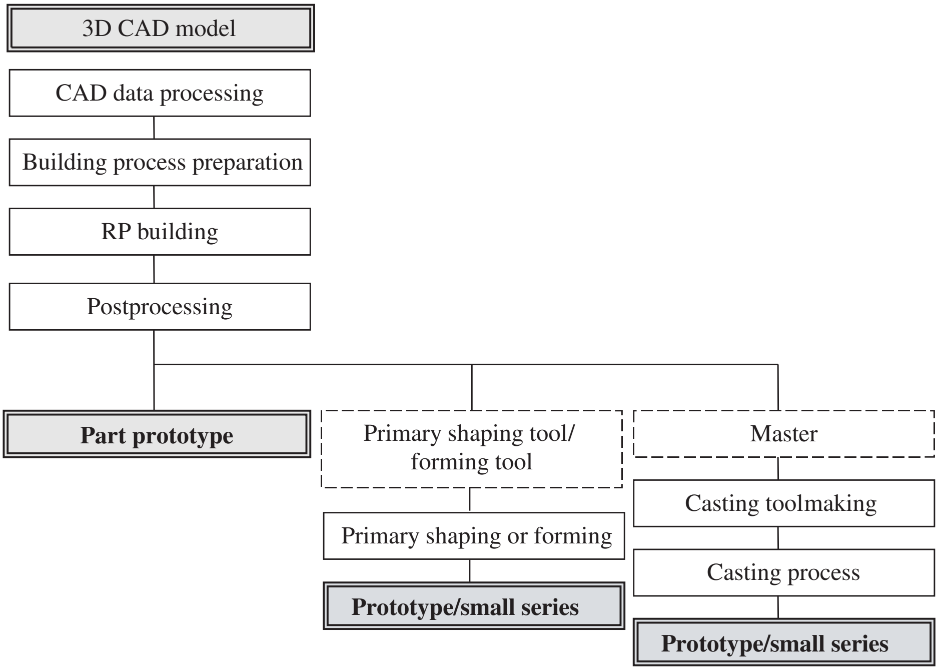

3D CAD model
CAD data processing
Building process preparation
RP building
Postprocessing
Part prototype
Primary shaping tool/ forming tool
Primary shaping or forming
Prototype/small series
Master
Casting toolmaking
Casting process
Prototype/small series

Figure 28 Near-Net-Shape Process Chain for Producing Prototypes and Small Series by Means of Generic Manufacturing Techniques (RP techniques).

machining of hard materials (turning, milling). As a result, reduced manufacturing times and lower energy consumption could be achievable and the requirements for chipless shaping could possibly be diminished. In cases where the final contours can be economically produced to the required accuracy using chipless shaping, the change to net-shape manufacturing is a useful development goal.

REFERENCES

Adlof, W. W. (1995), *Schmiedeteile: Gestaltung, Anwendung, Beispiele,* Informationsstelle Schmiedestück-Verwendung im Industrieverband Deutscher Schmieden e.V., Hagen.

Hirschvogel, M. (1997), "Potentiale der Halbwarmumformung: ein Verfahrensvergleich," in *Tagungsband des 6. Umformtechnischen Kolloquiums* (Darmstadt), pp. 195–205.

Körner, E., and Knödler, R. (1992), "Möglichkeiten des Halbwarmfliesspressens in Kombination mit dem Kaltfliesspressen," *Umformtechnik,* Vol. 26, No. 6, pp. 403–408.

Lorenz, B. (1996), "Ein Beitrag zur Theorie der Umformung pulvermetallurgischer Anfangsformen," *Freiberger Forschungshefte,* Vol. 281, pp. 1–127.

ADDITIONAL READING

Altan, T., and Knörr, M., "Prozeß- und Werkzeugauslegung für die endkonturnahe Fertigung von Verzahnungen durch Kaltfließpressen," in *Konferenzbericht, 5. Umformtechnisches Kolloquium* (Darmstadt, 1994), pp. 20.1–20.11.

Beitz, W., and Küttner, K.-H., *Dubbel: Taschenbuch für den Maschinenbau,* 17th Ed. Springer, Berlin, 1990.

Doege, E., Thalemann, J., and Brüggemann, K., "Massnahmen zur Qualitätssicherung beim Präzisionsschmieden," *Umformtechnik,* Vol. 29, No. 4, 1995, pp. 243–249.

Doege, E., Thalemann, J., and Westercamp, C., "Präzisionsschmieden von Zahnrädern," *wt-Produktion und Management,* Vol. 85, No. 3, 1995, pp. 85–89.

Dueser, R., "Gesenkwalzen," Technische Information 1/91, Thyssen Maschinenbau GmbH Wagner, Dortmund, 1991.

Föllinger, H., and Meier, R., "Near-Net-Shape-Umformung: Chance und Herausforderung," *Umformtechnik,* Vol. 26, No. 6, 1992, pp. 422–424.

Gabathuler, J. P., "Thixoforming: Ein neues Verfahren für die Produktion von Near-Net-Shape-Formteilen," in *Konferenzbericht, Praxis-Forum, Automobil-Arbeitskreise,* Vol. 2, 1998, pp. 155–168.

Gärtner, R., Hemyari, D., Müller, F., and Rupp, M., "Näher zur Endkontur," *Schweizer Maschinenmarkt,* No. 17, 1998, pp. 24–26.

Gebhardt, A. (1996), *Rapid Prototyping,* Carl Hanser, Munich.

Geiger, M. (1999), *Advanced Technology of Plasticity,* University of Nuremberg-Erlangen.

Geiger, R., and Hänsel, M., "Von Near-Net-Shape zu Net-Shape beim Kaltfließpressen: Stand der Technik," in *Konferenzbericht, Neue Entwicklungen in der Massivumformung* (Fellbach, 1995), pp. 435–456.

Hänsel, M., and Geiger, R., "Net-Shape-Umformung," *Umformtechnik,* Vol. 29, No. 4, 1995, pp. 218–224.

Heinze, R. (1996), "Taumelpressen geradverzahnter Zylinderräder," Dissertation, Rhine-Westphalian Technical University, Aachen, 1996.

Ketscher, N., and Herfurth, K., "Vorteile der Formgebung durch Giessen," *Konstruktion,* No. 9, 1999, pp. 41–44.

Kirch, W., "Radnaben wirtschaftlich Taumelfließpressen," *Werkstatt und Betrieb,* No. 11, 1994, pp. 892–894.

Kobayashy, M., Nakane, T., Kamada, A., and Nakamura, K. (1979), "Deformation Behaviour in Simultaneous Extrusion-Upsetting by Rotary Forming," in *Proceedings of 1st International Conference on Rotary Metal-Working Processes* (London), pp. 291–294.

König, W., and Lennartz, J., "Fertigungsfolge Fließpressen—Zerspanen," *VDI-Zeitschrift,* Vol. 135, No. 3, 1993, pp. 73–78.

Koenig, W., Leube, H., and Heinze, R., "Taumelpressen von geradverzahnten Zylinderrädern," Industrie-Anzeiger, No. 91, 1986, pp. 25–35.

Kunze, H.-D., *Competitive Advantages by Near-Net-Shape Manufacturing,* DGM-Informationsgesellschaft mbH, Frankfurt am Main, 1997.

Lange, K., "Moderne Umformtechnik: ein Grundpfeiler der industriellen Produktion," in *Konferenzbericht,1. Sächsische Fachtagung Umformtechnik* (Chemnitz, 1994), pp. 2.1–2.18.

Lange, K., and Meyer-Nolkemper, H., *Gesenkschmieden,* Springer, Berlin, 1977.

Müller, F., and Gärtner, R. (1997), "Hohle Getriebewellen: eine fertigungstechnische Herausforderung," in *Konferenzbericht, 6. Umformtechnisches Kolloquium* (Darmstadt, 1997), pp. 239–248.

Müller, F., and Heislitz, F., "Verkürzte Prozeßketten in der Massivumformung," *Werkstatt und Betrieb,* Vol. 130, No. 10, 1997, pp. 911–915.

Rebholz, M., "Prinzip und Einsatzmöglichkeit des Taumelpressens," in *Konferenzbericht,1. Sächsische Fachtagung Umformtechnik* (Chemnitz, 1994), pp. 19./1–8.

Scheipers, P., *Handbuch der Metallbearbeitung,* Europa-Lehrmittel, Haan-Gruiten, 2000.

Schmoeckel, D., Rupp, M., and Müller, F., "Auf dem Weg zur Net-Shape-Umformung," *wt-Werkstattstechnik,* Vol. 87, 1997, pp. 310–314.

Seilstorfer, H., and Moser, G., "Die heissisostatische Presstechnik (HIP)," *Metallwissenschaft und Technik,* Vol. 34, No. 10, 1980, pp. 925–929.

Spur, G., "Wettbewerbsdruck setzt Kreativität frei: Entwicklungstendenzen beim Zerspanen," *Fertigung,* No. 11, 1997, pp. 68– 70.

Standring, P., and Arthur, A., "A Structured Approach to the Design Selection and Development of Components for Rotary Forging," in *Proceedings of the 27th General Meeting of the International Cold Forging Group* (Padua, 1994).

Steffens, K. (1997), "Umformtechnik 2000: Erwartungen voll erfüllt?," *Konferenzbericht, 6. Umformtechnisches Kolloquium* (Darmstadt, 1997), pp. 11–21.

Werning, H. (1993), "Optimal spanend Bearbeiten: Giessen schafft die notwendigen Voraussetzungen," *konstruieren + giessen,* Vol. 18, No. 2, 1993, pp. 33–38.

Westerlund, J., "Four Decades of HIP Progress," *Metal Powder Report,* Vol. 55, No. 2, 2000, pp. 14–21.

CHAPTER 19

Environmental Engineering: Regulation and Compliance

Robert B. Jacko
Timothy M. C. LaBreche
Purdue University

1. OVERVIEW **589**

2. ENVIRONMENTAL LAW **589**
- 2.1. Overview 589
- 2.2. Environmental Protection Act 590
- 2.3. Clean Air Acts 590
  - 2.3.1. Air Pollution Control Concept in the United States 590
  - 2.3.2. The 1970 Clean Air Act 592
  - 2.3.3. The 1977 Clean Air Act Amendments 592
  - 2.3.4. The 1990 Clean Air Act Amendments 592
- 2.4. Worker Right to Know 593
- 2.5. Resource Conservation and Recovery Act 593
- 2.6. Hazardous Materials Transportation Act 594
- 2.7. Comprehensive Environmental Response, Compensation and Liability Act (CERCLA), Superfund Amendments and Reauthorization Act (SARA) 594
- 2.8. Clean Water Act (CWA) 595

3. COMPLYING WITH ENVIRONMENTAL LAWS **595**
- 3.1. Overview 595
- 3.2. Permits 595
  - 3.2.1. Air Permits 595
  - 3.2.2. Water Permits 596

4. ESTIMATING PLANT-WIDE EMISSIONS **596**
- 4.1. Overview 596
- 4.2. Estimating Methods 596
  - 4.2.1. Mass Balance 596
  - 4.2.2. Emission Factors 597

5. TOTAL-ENCLOSURE CONCEPT FOR FUGITIVE AIR EMISSIONS **598**
- 5.1. Criteria for 100% Permanent Total Enclosure of Air Pollutants 598

6. GREEN ENGINEERING **598**
- 6.1. Product Design 598
- 6.2. Process Design 599
- 6.3. Total Cost Analysis 599

REFERENCES **599**

1. OVERVIEW

This chapter provides industrial, plant, and facilities engineers with a brief overview of some environmental engineering considerations in process and manufacturing operations. This information will be helpful to engineers who are designing or operating new industrial processes or involved in the modification of existing facilities.

2. ENVIRONMENTAL LAW

2.1. Overview

Regulations designed to protect and improve the environment play a substantial role in the design and management of industrial processes. These regulations should be incorporated into the initial

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

process design and not as an afterthought. This section provides a brief overview of some major environmental laws and organizations.

2.2. Environmental Protection Act

Prior to the Environmental Protection Act, environmental regulations were divided along media lines (air, water, earth). In 1970, President Nixon submitted to Congress a proposal to consolidate many of the environmental duties previously administered by agencies including the Federal Water Quality Administration, the National Air Pollution Control Administration, the Bureau of Solid Waste Management, the Bureau of Water Hygiene, the Bureau of Radiological Health; certain functions with respect to pesticides carried out by the Food and Drug Administration; certain functions of the Council on Environmental Quality; certain functions of the Atomic Energy Commission and the Federal Radiation Council; and certain functions of the Agricultural Research Service (Nixon 1970a).

President Nixon recognized that some pollutants exist in all forms of media and that successful administration of pollution-control measures required the cooperation of many of the federal bureaus and councils. A more effective management method would be to recognize pollutants, observe their transport and transformation through each medium, observe how they interact with other pollutants, note the total presence of the pollutant and its effect on living and nonliving entities, and determine the most efficient mitigation process. This multimedia approach required the creation of a new agency to assume the duties of many existing agencies, thus eliminating potential miscommunication and interdepartmental biases that could hinder environmental protection. Thus, the President recommended the establishment of an integrated federal agency that ultimately came to be called the Environmental Protection Agency (EPA).

The roles and functions of the EPA were to develop and enforce national standards for the protection of the environment; research the effects of pollutants, their concentrations in the environment, and ways of controlling them; provide funding for research and technical assistance to institutions for pollutant research, and propose new legislation to the President for protection of the environment (Nixon 1970b).

In the words of William D. Ruckelshaus, EPA's first administrator,

> EPA is an independent agency. It has no obligation to promote agriculture or commerce; only the critical obligation to protect and enhance the environment. It does not have a narrow charter to deal with only one aspect of a deteriorating environment; rather it has a broad responsibility for research, standard-setting, monitoring and enforcement with regard to five environmental hazards; air and water pollution, solid waste disposal, radiation, and pesticides. EPA represents a coordinated approach to each of these problems, guaranteeing that as we deal with one difficulty we do not aggravate others. (Ruckelshaus 1970)

The EPA has instituted numerous programs and made significant changes in the way businesses operate in the United States. A brief summary of the EPA's milestones (Table 1) shows the many ways the course of business and the environment have been altered since the agency's inception in 1970.

2.3. Clean Air Acts

2.3.1. Air Pollution Control Concept in the United States

Air pollution control in the United States is said to be a "command and control" regulatory approach to achieving clean air. That is, regulations are promulgated at the federal level and via state implementation plans (SIPs) at the state level and air pollution sources are required ("commanded") to comply. These regulations have been set up to operate on the air we breathe as well as the sources that produce the air pollution. The regulations established to control the air we breathe are called National Ambient Air Quality Standards (NAAQSs). Their engineering units are concentration based, that is, micrograms-pollutant per cubic meter of air or parts per million by volume (ppm_v). These NAAQSs also have a time-averaging period associated with them such as a 24-hour or an annual averaging period. Additionally, the NAAQSs have primary standards and secondary standards associated with them. The primary standards are for the protection of human health and the secondary standards are for the protection of things. For example, the primary 24-hour average standard for sulfur dioxide is 365 $\mu g/m^3$ and the secondary standard is a 3-hour average standard of 1300 $\mu g/m^3$. The 365 $\mu g/m^3$ standard protects humans from a high-level short-term dose of sulfur dioxide, while the 1300 $\mu g/m^3$ standard could protect a melon crop from a high-level, short-term dose of sulfur dioxide. Table 2 contains the NAAQSs for the six EPA criteria pollutants.

The other thrust of the air pollution regulations applies to the sources of the pollutants. These regulations are emission standards called the New Source Performance Standards (NSPSs). These sources include stationary sources such as power plants and industrial manufacturing operations as well as mobile sources such as automobiles, trucks, and aircraft. These regulations are mass flow rate-based, that is, grams-pollutant/hr or lb-pollutant/10^6 Btu. Pollutant-specific emission limits were

TABLE 1 Timeline of Environmental Regulation

Year	Event
1970	Agency established December 2, 1970. Clean Air Act Amendments set national health-based standards.
1971	Bans use of lead-containing interior paints in residences built or renovated by the federal government.
1972	Bans the use of DDT. Commits to a national network of sewage-treatment plants to provide fishable and swimmable waterways. United States and Canada sign the International Great Lakes Water Quality Agreement.
1973	Phase-out of lead-containing gasoline begins. First industrial wastewater discharge permit issued.
1974	New Safe Water Drinking Act establishes health-based standards for water treatment. Standards limiting industrial water pollution established.
1975	Fuel economy standards allow consumers to include fuel efficiency in their purchasing considerations. EPA emission standards require catalytic convertors to be installed on automobiles.
1976	Resource Conservation and Recovery Act established to track hazardous waste from “cradle to grave.” Toxic Substance Control Act developed to reduce public exposure to pollutants that pose an unreasonable threat of injury. Included in this act is a ban on polychlorinated biphenyls (PCBs).
1977	Air quality in pristine natural areas is addressed by Clean Air Act Amendments.
1978	Chlorofluorocarbons (CFCs) banned for use as a propellant in most aerosol cans. CFCs are implicated as ozone-depleting chemicals.
1979	Dioxin is indicated as a carcinogen. Two herbicides containing dioxins banned.
1980	Under the new Superfund law a nationwide program for the remediation of hazardous waste sites is established.
1984	Amendments to the RCRA Resource Conservation and Recovery Act enacted to prevent contamination from leaking underground storage tanks (USTs) and landfills and require treatment of hazardous wastes prior to land disposal.
1985	EPA joins international movement calling for worldwide ban on the use of Ozone-Depleting Chemicals (ODCs).
1986	A major toxic chemical release in Bhopal, India, stimulates the passage of the Community Right to Know Law, requiring that individuals in possession of chemicals maintain records of location, quantity, and use and any releases of chemicals. The EPA is charged with maintaining a public database of these records. The EPA begins assisting communities in the development of Emergency Response Plans.
1987	The United States and 23 other countries sign the Montreal Protocol to phase out the production of chloroflourocarbons (CFCs).
1989	The new Toxic Releases Inventory provides the public with the location and character of toxic releases in the United States.
1990	EPA assesses $15 billion fine to a single company for PCB contamination at 89 sites. New Clean Air Act Amendments require states to demonstrate continuing progress toward meeting national air quaility standards.
1992	EPA bans dumping of sewage sludge into oceans and coastal waters.
1993	EPA’s Common Sense Initiative announces a shift from pollutant-by-pollutant regulation to industry-by-industry regulation.
1994	Grants initiated to revitalize inner-city brownfields. Clinton administration nearly doubles the list of toxic chemicals that must be reported under the Community Right to Know laws.
1995	Two thirds of metropolitan areas that did not meet air quality standards in 1990 now do meet them, including San Francisco and Detroit. EPA requires municipal incinerators to reduce toxic emissions by 90%.

From EPA OCEPA 1995.

assigned to both existing sources and proposed new sources under the 1970 Clean Air Act. The existing-source emission limits were less stringent than the new source limits because new sources had the opportunity to incorporate the latest technology into the process or manufacturing design.

To determine what emission standards apply to your new proposed operation, refer to the Code of Federal Regulations (CFR) part 60, which is also Title I under the 1990 Clean Air Act Amend-

TABLE 2 National Ambient Air Quality Standards

Criteria Pollutant	Averaging Period	Primary NAAQS $\mu g/m^3$	Secondary NAAQS $\mu g/m^3$
PM-10	Annual	50	50
	24 hour	150	150
Sulfur dioxide (SO_2)	Annual	80	
	24 hour	365	
	3 hour		1,300
Nitrogen dioxide (NO_2)	Annual	100	100
Ozone	1 hour	235	235
Carbon monoxide (CO)	8 hour	10,000	
Lead	Quarterly	1.5	1.5

ments. The NSPSs are shown for a large number of industrial classifications. Look for your type of industry and you will find the applicable new source emission standards.

2.3.2. The 1970 Clean Air Act

Control of outdoor air pollution in the United States certainly grew its regulatory roots prior to the 1970 Clean Air Act. However, for the practicing industrial engineer, the passage of the 1970 Clean Air Act is the real beginning of contemporary outdoor or ambient air pollution control. Therefore, this section will focus on those aspects of the Clean Air Act of 1970 and the subsequent amendments of 1977 and 1990 that will be most meaningful to the industrial engineer.

The 1970 Clean Air Act was the first piece of air pollution legislation in this country that had any real teeth. The establishment of an autonomous federal agency that ultimately came to be known as the EPA was a result of the 1970 Act. Prior to this time, matters relating to outdoor air pollution had their home in the United States Public Health Service. Furthermore, it is historically interesting to note that the EPA went through a number of name changes in the months following promulgation of the 1970 Clean Air Act before the current name was chosen.

2.3.3. The 1977 Clean Air Act Amendments

The 1970 Clean Air Act had a deadline of July 1, 1975, by which all counties in the United States were to comply with concentration-based standards (NAAQSs) for specified air pollutants. However, when July 1, 1975, arrived, very few counties in industrialized sections of the country met the standards. As a consequence, much debate took place in Congress and the result was the promulgation of the 1977 Clean Air Act Amendments. In some quarters, this amendment was viewed as the strongest piece of land use legislation the United States had ever seen. This legislation required each state to evaluate the air sheds in each of its counties and designate the counties as either attainment or non-attainment counties for specified air pollutants, according to whether the counties attained the National Ambient Air Quality Standards (NAAQS) for those pollutants. If the state had insufficient data, the county was designated unclassified. A company in a non-attainment county had to take certain steps to obtain an air permit. This ultimately added a significant increased cost to a manufactured product or the end result of a process.

Under this same amendment were provisions for ‘‘protecting and enhancing’’ the nation’s air resources. These provisions, called the Prevention of Significant Deterioration (PSD) standards, prevented an industrial plant from locating in a pristine air shed and polluting that air shed up to the stated ambient air standard. They classified each state’s counties as to their economic status and air polluting potential and established ceilings and increments more stringent than the ambient air standards promulgated under the 1970 Clean Air Act. No longer could a company relocate to a clean air region for the purpose of polluting up to the ambient air standard. In essence, the playing field was now level.

2.3.4. The 1990 Clean Air Act Amendments

Due to a number of shortcomings in the existing air pollution regulations at the end of the 1980s, the Clean Air Act was again amended in 1990. The 1990 Clean Air Amendments contain a number of provisions, or titles, as they are referred to. These titles, some of which carry over existing regulations from the 1970 and 1977 amendments, are listed below, along with a general statement of what they cover:

- Title I:
  - National Ambient Air Quality Standards (NAAQS), Clean Air Act sections 105–110 and 160–193; 40 CFR Parts 50–53, 55, 58, 65, 81, and 93. Establishes concentration-based ambient air standards.
  - New Source Performance Standards (NSPSs), Clean Air Act section 111; 40 CFR Part 60. Establishes emission limitations for specific categories of new or modified sources.
- Title II:
  Mobile Sources Program, Clean Air Act sections 202–250; 40 CFR Parts 80 and 85–88. Covers tailpipe emission standards for aircraft, autos, and trucks, including fuel and fuel additives, clean fuel vehicles, and Hazardous Air Pollutants (HAPS) research for mobile sources.
- Title III:
  National Emission Standards for Hazardous Air Pollutants (NESHAPS), Clean Air Act section 112; 40 CFR Parts 61, 63, and 68. Includes an accidental release program, list of HAPS and sources, residual risk standards, and maximum achievable control technology (MACT) standards.
- Title IV:
  Acid Rain Program, Clean Air Act sections 401–416; 40 CFR Parts 72–78. Acid deposition control via sulfur and nitrogen oxide controls on coal- and oil-burning electric utility boilers.
- Title V:
  Operating permit program, Clean Air Act sections 501–507; 40 CFR Parts 70–71. Requires operating permits on all sources covered under the Clean Air Act.
- Title VI:
  Stratospheric Ozone Protection Program, Clean Air Act sections 601–618; 40 CFR Part 82. Contains a list of ozone-depleting substances. Bans certain freons, requires freon recycling.
- Title VII:
  Enforcement Provisions, Clean Air Act sections 113, 114, 303, 304, 306, and 307. Compliance certification, enhanced monitoring, record keeping and reporting, $25,000/day fines, civil and criminal penalties, entry/inspector provisions, citizen lawsuits and awards up to $10,000, and public access to records.

2.4. Worker Right to Know

The Hazard Communication Standard of the Occupational Safety and Health Act requires that all employees be informed of the hazards associated with the chemicals they are exposed to or could be accidentally exposed to. In addition to chemical hazards, OSHA requires workers be trained to recognize many other types of hazards. Hazards may include chemical, explosion and fire, oxygen deficiency (confined spaces, for example), ionizing radiation, biological hazards, safety hazards, electrical hazards, heat stress, cold exposure, and noise. Compliance with Worker Right to Know laws general requires a written plan that explains how the Hazard Communication Standard will be executed. The only operations where a written Hazard Communication Plan is not required are handling facilities where workers contact only sealed chemical containers and laboratories. These facilities must still provide hazard training to employees, retain the original labeling on the shipping containers, and make material safety data sheets (MSDSs) available to all employees. In general, no employees should be working with any chemical or equipment that they are not familiar with. OSHA statistics indicate that failure to comply with the Hazard Communication Standard is its most cited area (OSHA 1999b).

The Hazard Communication Standard requires each facility to conduct a hazard assessment for each chemical in the workplace, maintain an inventory of chemicals in the workplace, retain MSDSs for each chemical in the workplace, properly label each chemical according to a uniform labeling policy, train each employee to understand the MSDSs, product labels, and Hazard Communication Standard, and develop a written program that explains how the Hazard Communication Standard is to be implemented at the facility.

2.5. Resource Conservation and Recovery Act

The Resource Conservation and Recovery Act (RCRA) of 1976, an amendment to the Solid Waste Disposal Act of 1965, and the Hazardous and Solid Waste Amendments of 1984, which expanded RCRA, were enacted to protect human health and environment, reduce waste and energy, and reduce or eliminate hazardous waste as rapidly as possible. Three programs addressing hazardous waste, solid waste, and underground storage tanks are included in RCRA.

Subtitle C requires the tracking of hazardous waste from “cradle to grave,” which refers to the requirement that hazardous waste be documented by a paper trail from the generator to final disposal, whether it be incineration, treatment, landfill, or some combination of processes. Subtitle D establishes criteria for state solid waste management plans. Funding and technical assistance are also provided for adding recycling and source reduction implementation and research. Subtitle I presents rules for the control and reduction of pollution from underground storage tanks (USTs).

A RCRA hazardous waste is any substance that meets physical characteristics such as ignitibility, corrosivity, and reactivity or may be one of 500 specific hazardous wastes. They may be in any physical form: liquid, semisolid, or solid. Generators and transporters of hazardous waste must have federally assigned identification numbers and abide by the regulations pertinent to their wastes. Individuals treating and disposing of hazardous wastes must meet stringent operating guidelines and be permitted for treatment and disposal technologies employed. RCRA hazardous waste regulations apply to any commercial, federal, state, or local entity that creates, handles, or transports hazardous waste.

A RCRA solid waste is any sludge, garbage, or waste product from a water treatment plant, wastewater treatment plant, or air pollution control facility. It also includes any discarded material from industrial, mining, commercial, agricultural, and community activities in contained gaseous, liquid, sludge, or solid form. RCRA solid waste regulations pertain to owners and operators of municipal solid waste landfills (EPA OSW 1999a,b).

2.6. Hazardous Materials Transportation Act

The Hazardous Materials Transportation Act of 1975 (HMTA) and the 1990 Hazardous Materials Uniform Safety Act were promulgated to protect the public from risks associated with the movement of potentially dangerous materials on roads, in the air, and on waterways. They do not pertain to the movement of materials within a facility. Anyone who transports or causes to be transported a hazardous material is subject to these regulations, as is anyone associated with the production or modification of containers for hazardous materials. Enforcement of the HMTA is delegated to the Federal Highway Administration, Federal Railway Administration, Federal Aviation Administration, and Research and Special Programs Administration (for enforcement of packaging rules).

The regulations of the HMTA are divided into four general areas: procedures and policies, labeling and hazard communication, packaging requirements, and operational rules. Proper labeling and hazard communication requires the use of the standard hazard codes, labeling, shipping papers, and placarding. Hazardous materials must be packaged in containers compatible with the material being shipped and be of sufficient strength to prevent leaks and spills during normal transport (DOE OEPA 1996).

2.7. Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) and Superfund Amendments and Reauthorization Act (SARA)

The Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) was enacted in 1980 to provided a federal Superfund for the cleanup of abandoned hazardous waste sites. Funding for the Superfund was provided through fines levied or lawsuits won against potentially responsible parties (PRPs). PRPs are those individuals having operated at or been affiliated with the hazardous waste site. Affiliation is not limited to having physically been on the site operating a process intrinsic to the hazardous waste creation. Affiliation can include those parties that have provided transportation to the site or customers of the operation at the hazardous waste site that transferred raw materials to the site for processing (EPA OPA 1999a).

The Superfund Amendment Reauthorization Act (SARA) of 1986 continued cleanup authority under CERCLA and added enforcement authority to CERCLA (EPA OPA 1999b). Within SARA was the Emergency Planning and Community Right-to-Know Act (EPCRA) also know as SARA Title III. EPCRA established the framework for communities to be prepared for chemical emergencies that could occur at industrial sites in their neighborhoods. EPCRA required states to form State Emergency Response Commissions (SERCs). SERCs divided the states into Emergency Planning Districts and formed Local Emergency Planning Committees. These committees consist of a broad range of community leaders, emergency officials, and health professionals (EPA OPA 1999c). SERCs use information acquired through Section 313 of SARA, the Toxic Releases Inventory (TRI), to make emergency planning decisions. Chemical producers and consumers must annually report releases of chemicals during the year to the EPA. These releases may occur continuously throughout the year or in a single large burst and include releases that a company is legally permitted to make, such as through air permits. Included in these reports are the type and volume of chemical released, the media the chemical was released to, how much of the chemical was transported from the site for recycling or disposal, how the chemical was treated for disposal, the efficiency of treatment, and pollution prevention and recycling activities at the reporting company. Reporting is required if a facility employs 10 or more full-time employees and (EPA OPPT 1999):

Manufactures or processes over 25,000 lb (total) of approximately 600 designated chemicals or 28 chemical categories
OR
Manfactures or processes over 10,000 lb of an individual designated chemical or chemical category
OR
Is among facilities grouped into Standard Industrial Classification Codes 20–39
OR
Is a federal facility ordered to report by August 1995 President Clinton executive decree.

2.8. Clean Water Act (CWA)

The 1972 Water Pollution Control Act, the 1977 Clean Water Act amendments, and the 1987 reauthorization of the Clean Water Act provide the basis for regulation of discharges to receiving water bodies in the United States. Both point sources, discharges such as a pipe or channel, and nonpoint sources, such as runoff from fields or parking lots, are regulated by the Clean Water Act. Both direct dischargers and dischargers to a municipal treatment works must obtain National Pollution Discharge Elimination System (NPDES) permits that govern the character of pollutants in waste streams and mandatory control technologies (EPA OW 1998).

The objectives of the CWA were to achieve fishable and swimmable waters and to eliminate the discharge of pollutants into navigable waterways. To achieve these goals, industries were required to meet performance standards for pollutant emissions, states were charged with developing criteria for their waterways as well as programs to protect them, funding was provided for the construction of public wastewater treatment plants and other technologies to mitigate discharges, and development was regulated via a permit process to protect wetlands and minimize impacts on water resources.

In addition to the permit requirements of the 1977 amendments, the 1987 amendments permitted citizen suits that allowed any individual to bring a legal suit against any party believed to be in conflict with the provisions of the CWA. Citizen suits were also permitted against any party responsible for the enforcement and administration of the CWA that was believed derelict in its responsibilities. This expanded the responsibility of clean water enforcement to include the public as well as government bodies.

3. COMPLYING WITH ENVIRONMENTAL LAWS

3.1. Overview

In the early 1970s, compliance with environmental regulations by most industrial entities was not a top priority. Indeed, some major corporations brought lawsuits and attempted to lessen the impact of the regulations. However, in the intervening years, continued research on the human health effects of water and air pollution has resulted in a heightened awareness of their deleterious impact on the human system. As a result, compliance with environmental regulations is now the norm rather than the exception for the industrial sector. An example of this is the tightening of the ambient air quality standards for particulates over the years. In the early years, dustfall buckets (measured particulates >40 μm diameter) and total suspended particulates (measured particulates <40 μm diameter) were targeted by the NAAQS. Human health research indicated that finer particulates (10 μm and less), which penetrate deeper into the human respiratory system, are more harmful than the larger particulate matter. Fine particulate matter has a much greater surface area-to-volume ratio relative to larger particulate matter, which allows the fine particulate matter to adsorb and absorb hazardous volatile substances and carry them deep into the human lung. This “piggyback” effect of fine particulate matter with hazardous air pollutants has resulted in the diameter of the particulate standard being reduced over the years. The current NAAQS standard for fine particulate matter is PM-10, which means particles less than or equal to 10 μm. But a PM-2.5 NAAQS standard has been proposed by the EPA and awaits promulgation at the time of writing.

3.2. Permits

3.2.1. Air Permits

One of the provisions of the 1970 Clean Air Act initiated the requirement for air pollution sources to obtain a permit for construction of the source and a permit to operate it. The construction permit application must be completed prior to the initiation of construction of any air pollution source. Failure to do so could result in a $25,000 per day fine. In some states, initiation of construction was interpreted as issuance of a purchase order for a piece of equipment; in others, groundbreaking for the new construction. Therefore, to ensure compliance with the air permit requirement, this author suggests that completion of the air permit be given first priority in any project involving air emissions into the atmosphere. The best practice is to have the state-approved permit in hand before beginning construction. Most states have their permit forms on the Internet, and a hard copy can be downloaded. Alternatively, the permit forms can be filled out electronically and submitted.

Recently, many states have offered the option of allowing an air pollution source to begin construction prior to obtaining the approved construction permit. However, the required paperwork is not trivial, and it may still be easier to fill out the permit form unless extenuating circumstances demand that construction begin before the approved permit is in hand. The caveat is the state could disapprove the construction (which the company has already begun) if, after reviewing the permit application, the state disagrees with the engineering approach taken to control the air pollution.

3.2.2. Water Permits

Any person discharging a pollutant from a point source must have a National Pollution Discharge Elimination System (NPDES) permit. This applies to persons discharging both to a public treatment works or directly to a receiving water body. These permits will limit the type of pollutants that can be discharged and state the type of monitoring and reporting requirements and other provisions to prevent damage to the receiving body or treatment facility. In most states, the state department of environmental quality is responsible for issuing permits. In states that have not received approval to issue NPDES permits, you should contact the regional EPA office.

Wastewater permits that regulate the discharge of many pollutants can be broken down into the general categories of conventional, toxic, and nonconventional. Conventional pollutants are those contained in typical sanitary waste such as human waste, sink disposal waste, and bathwater. Conventional wastes include fecal coliform and oil and grease. Fecal coliform is present in the digestive tracts of mammals, and its presence is commonly used as a surrogate to detect the presence of pathogenic organisms. Oils and greases such as waxes and hydrocarbons can produce sludges that are difficult and thus costly to treat. Toxic pollutants are typically subdivided into organics and metals. Organic toxins include herbicides, pesticides, polychlorinated biphenyls, and dioxins. Nonconventional pollutants include nutrients such as phosphorus and nitrogen, both of which can contribute to algal blooms in receiving waters.

The EPA maintains many databases of environmental regulatory data. Among these is the Permit Compliance System (PCS) database http://www.epa.gov/enviro/html/ pcs/pcs_query_java.html. This database provides many of the details governing a facility's wastewater discharges. Specific limitations typically depend on the classification and flow of the stream to which a facility is discharging. Surface waters are classified as to their intended use (recreation, water supply, fishing, etc.), and then the relevant conditions to support those uses must be maintained. Discharge limitations are then set so that pollution will not exceed these criteria.

Total suspended solids, pH, temperature, flow, and oils and grease are typical measures that must be reported. In the case of a secondary sewage treatment, the minimum standards for BOD_5, suspended solids, and pH over a 30-day averaging period are 30 mg/L, 30 mg/L, and 6–9 pH, respectively. In practice, these limits will be more stringent, and in many cases there will be separate conditions for summer and winter conditions as well as case-by-case limitations.

4. ESTIMATING PLANT-WIDE EMISSIONS

4.1. Overview

A part of the overall air pollution requirements is an annual emissions inventory to be submitted to the state in which the source resides. This inventory identifies and quantifies all significant atmospheric discharges from the respective industrial plant. Most states now have this inventory in electronic form and require submission electronically. The Indiana Department of Environmental Management uses a commercial package called i-STEPS, which is an automated tool for storing, reporting, and managing air emissions data. i-STEPS facilitates data compilation for pollution management and reporting emissions data to government agencies.*

4.2. Estimating Methods

4.2.1. Mass Balance

From an engineering viewpoint, the most direct way to estimate pollution emissions from an industrial plant is by mass balance. The concept is "mass in = mass out"—that is, everything the purchasing department buys and is delivered to the plant must somehow leave the plant, whether within the manufactured product; as solid waste to a landfill; as air emissions either through a stack or vent; or as liquid waste either to an on-site treatment plant or to the sewer and the municipal wastewater treatment plant.

---

* i-STEPs Environmental Software is available from Pacific Environmental Services, Inc., 5001 South Miami Boulevard; Suite 300, P.O. Box 12077, Research Triangle Park, NC 27709-2077, www.i-steps.com.

The mass balance approach, however, does not necessarily yield exceptional accuracy. Accuracy is a function of understanding the way in which a particular feed stock or raw material is used and how much of it is released to the atmosphere or perhaps transformed in the process. The mass balance approach would probably be used if measured parameters were not available, such as stack emission data and wastewater effluent data.

An example where a mass balance approach would yield inaccurate emissions to the atmosphere would be an industrial resin coating line. This process uses an organic carrier solvent, such as ethanol, which is volatized from the resin solids in a drying oven. The vapors (volatile organic compounds [VOCs]) are then incinerated. The ethanol can be transformed into other organic compounds in the incinerator. The total mass of applied carrier solvent (ethanol) would not be accounted for in the incinerator exhaust gas stream due to its transformation into other organic compounds.

4.2.2. Emission Factors

Emission factors are unique to the air pollution field. They are usually based on stack emission test data for a specific process and are presented as a ratio of two flow rates. The numerator is the mass flow rate of the air pollutant parameter and the denominator is the flow rate of the process or manufactured product. In the spraypainting of automobiles, for example, the carrier solvent in the paint (VOC) is released to the atmosphere as the paint is dried in an oven. A stack test quantifies the amount of paint VOC released during the painting of some quantity of autos. The resulting VOC emission factor would be lb-VOC/hr divided by # of autos/hr painted. Note that the hour (hr) unit cancels and the emission factor is expressed as lb-VOC/# autos.

The U.S. Environmental Protection Agency publishes AP-42, a well-known compilation of air pollutant emission factors that contains emission factors for many industrial process and manufacturing operations (EPA AP-42 1995).

To illustrate the use of an emission factor, let us turn to the fish-canning industry. From AP-42, section 9.13.1-7 and Table 3, entitled "Uncontrolled Emission Factors for Fish Canning and By-product Manufacture," emission factors are shown for particulate emissions, trimethylamine (fish odor), and hydrogen sulfide (rotten egg odor). Emissions from the fish scrap cookers for both fresh fish and stale fish are shown. Notice that the particulate emissions are negligible for the cookers. Trimethylamine (fish odor) has an emission factor of 0.3 lb-trimethylamine/ton of fresh fish cooked. If we are cooking 5 tons of fresh fish/hr, then the uncontrolled fish odor emission to the atmosphere is 0.3 lb-trimethylamine/ton of fresh fish cooked × 5 tons of fresh fish/hr = 1.5 lb-trimethylamine/hr. Since the chemical responsible for typical fish odor is trimethylamine, 1.5 lb/hr will certainly be noticed by residents' noses downwind of the plant.

For hydrogen sulfide (H_2S, rotten egg odor), the uncontrolled atmospheric emission factor is 0.01 lb-H_2S/ton of fresh fish cooked. If we are cooking 5 tons of fresh fish/hr, then the uncontrolled H_2S odor to the atmosphere is 0.01 lb-H_2S/ton cooked fresh fish × 5 tons cooked fresh fish/hr = 0.05 lb-H_2S/hr. It is interesting to note that if the cooking fish is stale, not fresh, the trimethylamine emission is over 10 times higher. Similarly, for the hydrogen sulfide, if the fish is stale, the hydrogen sulfide emission is 20 times greater. With the concept of the emission factor, uncontrolled emissions

TABLE 3 Uncontrolled Emission Factors for Fish Canning and Byproduct Manufacture[a]

EMISSION FACTOR RATING: C

	Particulate		Trimethylamine $[(CH_3)_3N]$		Hydrogen Sulfide (H_2S)	
Process	kg/Mg	lb/ton	kg/Mg	lb/ton	kg/Mg	lb/ton
Cookers, canning (SCC 3-02-012-04)	Neg	Neg	–[c]	–[c]	–[c]	–[c]
Cookers, scrap						
Fresh fish (SCC 3-02-012-01)	Neg	Neg	0.15[c]	0.3[c]	0.005[c]	0.01[c]
Stale fish (SCC 3-02-012-02)	Neg	Neg	1.75[c]	3.5[c]	0.10[c]	0.2[c]
Steam tube dryer (SCC 3-02-012-05)	2.5	5	–[b]	–[b]	–[b]	–[b]
Direct-fired dryer (SCC 3-02-012-06)	4	8	–[b]	–[b]	–[b]	–[b]

[a] From Prokop (1992). Factors are in terms of raw fish processed. SCC = Source Classification Code. Neg = negligible.

[b] Emissions suspected, but data are not available for quantification.

[c] Summer (1963).

from fish cookers can be estimated without costly measurements. The expected accuracy of the estimate is alluded to at the top of the table as having an emission factor rating of C. This means that on a scale of A to E, with A being the relative best estimate, the C rating is mediocre. It is important to refer to the references associated with each emission factor so a value judgment can be made when using the factor.

5. TOTAL-ENCLOSURE CONCEPT FOR FUGITIVE AIR EMISSIONS

Most if not all industrial operations result in the generation of air pollution emissions to the atmosphere. The emissions may be partially captured by a ventilation hood and be taken to an oven, incinerator, baghouse, or other air pollution control device. The air pollutants not captured by the ventilation hood escape into the workplace and are referred to as fugitive emissions. These fugitive emissions eventually reach the outdoor atmosphere either through roof ventilation fans, open windows, doors, or exfiltration through the walls of the building structure itself. Ultimately, all the air pollutants generated by the industrial process reach the outdoor atmosphere. For this reason, the EPA and state regulatory agencies require an industrial process that generates air pollutants to be mass-balance tested prior to approving an air pollution operating permit. In other words, either the fugitive emissions must be quantified and figured in the overall destruction and removal efficiency (DRE) of the air pollution controls or it must be demonstrated that the ventilation hood(s) form a 100% permanent total enclosure of the pollutants. With 100% permanent total enclosure demonstrated, it is known that all air pollutants generated by the process are captured by the hood(s) or enclosure and taken to the "as-designed" air pollution control device, whether it be an integral oven, external incinerator, baghouse filter, electrostatic precipitator, or some other tail gas cleanup device.

5.1. Criteria for 100% Permanent Total Enclosure of Air Pollutants

The United States Environmental Protection Agency has developed a criterion for determining 100% total enclosure for an industrial operation, EPA method 204 (EPA M204 2000). It requires that a series of criteria be met in order for a ventilation hooded industrial process to qualify as 100% permanent total enclosure. These criteria are:

1. The total area of all natural draft openings (NDOs) must be less than 5% of the total surface area of the enclosure.
2. The pollutant source must be at least four equivalent diameters away from any natural draft opening (NDO). The equivalent diameter is defined as $2(LW)/(L + W)$, where L is the length of the slot and W is the width.
3. The static pressure just inside each of the NDOs must be no less than a negative pressure of 0.007 in. of water.
4. The direction of flow must be into the enclosure as demonstrated by a device such as a smoke tube.

6. GREEN ENGINEERING

While many of the earlier legislative efforts were oriented toward "end-of-pipe" technologies, where pollution was captured and disposed of after being generated, the focus is now shifting to legislation that favors recycling, reuse, and reduction of waste generation. The problem with end-of-pipe technologies is that frequently the waste product, while no longer in the air, water, or earth, must still be disposed of in some manner. If a product is designed from its inception with the goal of reducing or eliminating waste not only during the production phase but also for its total life cycle, from product delivery to salvage, resources can be greatly conserved. This philosophy results in the design concept of "cradle to reincarnation" rather than "cradle to grave" because no well-designed product is ever disposed of—it is recycled (Graedel et al. 1995). Many phrases have been adopted into the expanding lexicon of green engineering. Total cost assessment (TCA), design for the environment (DFE), design for recycling (DFR), and life-cycle assessment are some of the many phrases that summarize efforts to look beyond the traditional design scope when designing and manufacturing a product.

6.1. Product Design

There is a threshold for the economic feasibility of recycling. If only a small portion of a product is recyclable or if great effort is required to extract the recyclable materials, then it may not be economically feasible to recycle a product. If, however, the product is designed to be largely recyclable and is designed in a manner that facilitates easy removal, then reuse and recycling will be economically reasonable. Consider the personal computer. The rapid rate at which computing power has expanded has resulted in a proportionate turnover in computers in the workplace. However, no such improvement in computer externals has accompanied that in computer internals, so last year's computer case could in theory hold this year's computer processor, motherboard, memory, and so on.

TABLE 4 Elements of Designing for Recycling

Element	Rationale
Minimize variety of materials	Multiple material types require more separation processes, thus increasing the cost of recycling.
Minimize use of toxics	Toxins must be removed prior to product reuse or recycling.
Do not plate plastics with metal.	Reduces the value of scrap plastic, may prevent recycling.
Place parts that are frequently improved and parts that frequently fail in accessible areas and make them easy to replace.	Designing a product that can be upgraded and easily repaired reduces the need for totally new products.
Use molded labels and details rather than adhesive labels or printed inks.	Adhesive labels and printed text must be removed prior to recycling.
Avoid polymeric materials with similar specific gravity to plastics used in product.	Gravimetric separation methods may be complicated or negated if incompatible materials with similar specific gravities are used.

The computer is an excellent example where design for recycling could be applied. While the elements in Table 4 are based on computer and electronic production, similar principles can be adopted for other processes.

6.2. Process Design

In many cases, recycling of process material streams is both environmentally and economically sound. While efforts to minimize the use of hazardous air pollutants (HAPs) such as toluene and xylene continue, some applications, such as cleanup solvents for spraypainting, still require the properties of HAP-containing solvents. However, HAP emissions can be minimized through the use of solvent-recovery systems. These systems reduce emissions to the atmosphere, thus reducing permitting costs, and they also reduce purchasing costs by reducing the total volume of virgin cleanup solvent necessary.

In other cases, waste minimization technologies can result in enhanced product quality. A continuous process data analysis system is one example. Rather than operator experience and intuition guiding when to cycle a process bath, computer analysis can be instituted to monitor key indicators of the process solution and automatically adjust the bath conditions. The result can be longer bath life, a more consistent product, and greater product engineer confidence (Dickinson 1995).

6.3. Total Cost Analysis

In some cases, an immediate benefit from a process change or recycling effort will not be evident. Any number of organizational, legal, or economic obstacles could distort the analysis of process options. Alternative budgeting and accounting methods are necessary to provide a more holistic perception of the process alternatives. Total cost analysis assesses the potential profitability of a green engineering proposal with several key differences from traditional methods. These differences are (White and Savage 1999):

1. Included in the inventory of costs, savings, and revenues are compliance training, testing, liability, and product and corporate image.
2. Rather than being lumped into overhead accounts, specific costs and savings are placed into process and product lines.
3. Longer time horizons are used to capture longer-term benefits.
4. Profitability indicators capable of incorporating the time value of money, long-term costs, and savings are used.

REFERENCES

D'Anjou, L. O., Choi, J. H., Glantschnig, W. J., and Stefanacci, E. F. (1995), "Designing with Plastics: Considering Part Recyclability and the Use of Recycled Materials," *AT&T Technical Journal,* November–December, pp. 54–59.

DOE OEPA (1996), "OEPA Environmental Law Summary: Hazardous Materials Transportation Act," Department of Energy, Office of Environmental Policy Analysis, http://tis-nt.eh.doe.gov/oepa/law_sum/HMTA.HTM

Dickinson, D. A., Draper, C. W., Saminathan, M., Sohn, J. E., and Williams, G. (1995), "Green Product Manufacturing," *AT&T Technical Journal,* November–December pp. 26–34.

EPA AP-42 (1995), *AP-42 Compilation of Air Pollutant Emission Factors,* 5th Ed., U.S. Environmental Protection Agency, Research Triangle Park, NC. (U.S. GPO (202) 512-1800 (stock no. 055-000-00500-1, $56.00), individual sections available through CHIEF AP-42 website).

EPA M204 (2000), "EPA Method 204," 40 Code of Federal Regulations, Part 51, Appendix M.

EPA OCEPA (1995), "What Has the Agency Accomplished?," EPA Office of Communication, Education, and Public Affairs, November, http://www.epa.gov/history/faqs/milestones.htm.

EPA OPA (1999a), "Comprehensive Environmental Response, Compensation, and Liability Act 42 U.S.C. 9601 et seq. (1980)," United States Environmental Protection Agency, Office of Public Affairs, http://www.epa.gov/reg5oopa/defs/html/cercla.htm.

EPA OPA (1999b), "Superfund Amendments and Reauthorization Act 42 U.S.C. 9601 et seq. (1986)," United States Environmental Protection Agency, Office of Public Affairs, http://www.epa.gov/reg5oopa/defs/html/sara.htm.

EPA OPA (1999c), "Emergency Planning and Community Right-to-Know Act, 42 U.S.C. 11001 et seq. (1986)," United States Environmental Protection Agency, Office of Public Affairs, http://www.epa.gov/reg5oopa/defs/html/epcra.htm.

EPA OPPT (1999), "What Is the Toxics Release Inventory," United States Environmental Protection Agency, Office of Pollution Prevention and Toxics, http://www.epa.gov/opptintr/tri/whatis.htm.

EPA OSW (1999), "Frequently Asked Questions about Waste," United States Office of Solid Waste, October 12, 1999, http://www.epa.gov/epaoswer/osw/basifact.htm#RCRA.

EPA OSW (1998) "RCRA Orientation Manual," United States Office of Solid Waste, http://www.epa.gov/epaoswer/general/orientat/.

EPA OW (1998), "Clean Water Act: A Brief History," United States Environmental Protection Agency, Office of Water, http://www.epa.gov/owow/cwa/history.htm.

Graedel, T. E., Comrie, P. R., and Sekutowski, J. C. (1995), "Green Product Design," *AT&T Technical Journal,* November–December, pp. 17–24.

Nixon, R. M. (1970a), "Reorganization Plan No. 3 of 1970," U.S. Code, Congressional and Administrative News, 91st Congress—2nd Session., Vol. 3.

Nixon, R. M. (1970b), *Public Papers of the Presidents of the United States: Richard Nixon,* 1970, GPO, Washington, DC, pp. 578–586.

OSHA (1999a), "Occupational Safety and Health Standards," 29 CFR 1910, *U.S. Code of Federal Regulations.*

OSHA (1999b), "Frequently Cited OSHA Standards," Occupational Safety and Health Administration, U.S. Department of Labor, http://www.osha.gov/oshstats/std1.html.

Ruckelshaus, W. D. (1970), EPI press release, December 16.

White, A., and Savage, D. (1999), *Total Cost Assessment: Evaluating the True Profitability of Pollution Prevention and Other Environmental Investments,* Tellus Institute, http://www.tellus.org.

CHAPTER 20

Collaborative Manufacturing

José A. Ceroni
Catholic University of Valparaíso—Chile

Shimon Y. Nof
Purdue University

1. INTRODUCTION 601
2. MANUFACTURING IN THE CONTEXT OF THE GLOBAL ECONOMY: WHY COLLABORATE? 601
3. COORDINATION AND CONTROL REQUIREMENTS IN COLLABORATIVE MANUFACTURING 603
4. FRAMEWORK FOR COLLABORATIVE MANUFACTURING 604
5. FACILITATING AND IMPLEMENTING COLLABORATION IN MANUFACTURING 604
6. MOVING FROM FACILITATING TO ENABLING COLLABORATION: E-WORK IN THE MANUFACTURING ENVIRONMENT 606
7. CASE EXAMPLES 606
    7.1. Coordination Cost in Collaboration 607
        7.1.1. Job-Shop Model 608
        7.1.2. Coordination Cost 608
        7.1.3. Results 608
    7.2. Collaboration in Distributed Manufacturing 609
        7.2.1. Integrated Optimization 611
        7.2.2. Simulation Results 612
        7.2.3. Local Optimization 612
        7.2.4. Case Remarks 613
    7.3. Variable Production Networks 616
8. EMERGING TRENDS AND CONCLUSIONS 617
REFERENCES 617

1. INTRODUCTION

Manufacturing is a constantly evolving field that is strongly driven by optimization of the resources it employs. Manufacturing optimization is showing a shift from processes to businesses in a more systemic analysis of its nature and role in the whole picture. Current economic conditions throughout the world, characterized by the steady growth of local economies, have contributed a good deal to this trend.

This chapter briefly presents the economic principles that drive collaborative manufacturing and the conditions supporting what manufacturing has become nowadays. Later sections of this chapter examine the coordination feature that enables collaborative manufacturing within and between enterprises and discuss several cases that demonstrate the ideas of collaborative manufacturing.

2. MANUFACTURING IN THE CONTEXT OF THE GLOBAL ECONOMY: WHY COLLABORATE?

To understand collaborative manufacturing in its actual form, we must refer to the current world economic conditions that motivate collaborative manufacturing. Many scholars recognize that we are

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

living in the knowledge revolution. As Sahlman (1999) notes, the new economy markedly drives out inefficiency, forces intelligent business process reengineering, and gives knowledgeable customers more than they want. This new economy, based primarily on knowledge and strong entrepreneurship, is focused on productivity and is profoundly changing the role of distribution. Distribution and logistics must be more efficient, cheaper, and more responsive to the consumer. This trend of new, competitive, and open channels between businesses is geographically dispersed, involving highly technical and rational parties in allocating effort and resources to the most qualified suppliers (even if they are part of another company).

One of the key aspects leading the way in this new knowledge-based world economy is science. There is a well-documented link between science and economic growth (Adams 1990), with a very important intermediate step, technology. Science enables a country to grow stronger economically and become the ideal base for entrepreneurs to start new ventures that will ultimately raise productivity to unprecedented levels. This science–growth relationship could lead to the erroneous conclusion that this new economy model will prevail only in these geographic areas favoring high-level academic and applied R&D and, even more, only to those institutions performing leading research. However, as Stephan (1996) points out, there is a spillover effect that transfers the knowledge generated by the research, and this knowledge eventually reaches underdeveloped areas (in the form of new plants, shops, etc.). This observation is confirmed by Rodriguez-Clare (1996), who examined an early study of collaborative manufacturing, multinational companies and their link to economic development. According to the Rodriguez-Clare model, the multinational company can create a positive or negative linkage effect upon any local economy. A positive linkage effect, for example, is created by forcing local companies to attain higher standards in productivity and quality. An example is the electronics industry in Singapore (Lim and Fong 1982). A negative linkage effect is created by forcing local companies to lower their operational standards. An example is the Lockheed Aircraft plant in Marietta, Georgia (Jacobs 1985).

We are therefore facing a new world of business, business of increasing returns for knowledge-based industries (Arthur 1996). The behavior of increasing-returns products is contrary to the classical economic equilibrium, in which the larger the return of a product or service, the more companies will be encouraged to enter the business or start producing the product or service, diminishing the return. Increasing-returns products or services, on the other hand, present positive feedback behavior, creating instability in the market, business, or industry. Increasing returns put companies on the leading edge further ahead of the companies trailing behind in R&D of new products and technologies. A classical example of this new type of business is the DOS operating system developed by Microsoft, which had a lock-in with the distribution of the IBM PC as the most popular computer platform. This lock-in made it possible for Microsoft to spread its costs over a large number of users to obtain unforeseen margins. The world of new business is one of pure adaptation and limits the use of traditional optimization methods, for which the rules are not even defined.

Reality presents us with a highly complex scenario: manufacturing companies unable to perform R&D seem doomed to disappear. One of the few alternatives left to manufacturing companies is to go downstream (Wise and Baumgartner 1999). This forces companies to rethink their strategy on downstream services (customer support) and view them as a profitable activity instead of a trick to generate sales. Under this new strategy, companies must look at the value chain through the customer's eyes to detect opportunities downstream. This affects how performance is measured in the business. Product margin is becoming more restricted to the manufacturing operation, disregarding services related to the functioning and maintenance of the product throughout its life. A feature that is increasing over time in actual markets is for businesses to give products at a very low price or even for free and wait for compensation in service to the customer or during the maintenance stage of the product's life cycle (e.g., cellphones, cable television markets in the United States). According to Wise and Baumgartner (1999), manufacturing companies follow one of four downstream business models (Table 1).

The ability to respond quickly and effectively to satisfy customers is what is making the difference among manufacturing companies nowadays. Technological advances such as Internet are facilitating ways for companies to meet their customer needs. As DeVor et al. (1997) point out, agile manufacturing focuses on enhancing competitiveness through cooperation and use of information technology to form virtual enterprises. Virtual enterprises are constructed by partners from different companies collaborating with each other to design and manufacture high-quality, customized products (Chen et al. 1999). Agile manufacturing practices are based on five principles (Yusuf and Sarhadi 1999):

- Identifying and creating value
- Enabling the flow-of-value stream
- Allowing customers to pull value uninterrupted
- Responding to unpredictable change
- Forming tactical and virtual partnerships

TABLE 1 Downstream Business Models

Downstream Model	Characteristics	Example
Embedded services	Embedding of downstream services into the product, freeing the customer of the need to perform them.	Honeywell and its airplane information management system (AIMS)
Comprehensive services	Coverage of downstream services not possible to embed into the product, for example financial services.	General Electric in the locomotive market
Integrated solutions	Combination of products and services for addressing customer needs	Nokia's, array of products for mobile telephony
Distribution control	Moving forward over the value chain to gain control over distribution activities	Coca-Cola and its worldwide distribution network

Reprinted by permission of *Harvard Business Review.* The Spectrum of Downstream Models, from "Go Downstream: The New Profit Imperative in Manufacturing," by Richard Wise and Peter Baumgartner, Sept.–Oct. 1999, p. 136. Copyright © 1999 by the Harvard Business School Publishing Corporation.

Agile manufacturing can be considered as the integration of technologies, people, and business processes.

3. COORDINATION AND CONTROL REQUIREMENTS IN COLLABORATIVE MANUFACTURING

An increasing number of companies are basing their future on global markets. The globalization of resources and customers has shifted the focus of industrial companies from resource control to customer-focused control over time (Hirsch et al. 1995). Competitiveness among companies nowadays relies on an increasingly important aspect: time-to-market. Pursuing shorter time-to-market requires faster development cycles for the products and close attention to geographically distributed markets. To cope effectively in this demanding environment, companies frequently engage in collaborative partner relationships, which allow them to focus and coordinate their efforts and improve their position in the market. The collaboration results in an integrated, aligned enterprise composed of several independent companies. Partner companies combine their capabilities in generating new business opportunities to which they could not have access otherwise.

Manufacturing covers a wide range of activities, from early design stages to product recycling. Companies often need to use collaboration between designers, technicians, departments, and divisions, or with other companies, to attain the desired results in an efficient way. As the complexity of the problems in manufacturing increases, concurrent engineering teams have resulted the most effective manner in which to tackle them. Concurrent engineering teams are composed of individuals with a wide range of expertise in different areas. This diversity of knowledge and viewpoints provides the team with the view of the manufacturing process necessary for addressing the complexity of the problems. However, to make from concurrent engineering something more than never-ending meetings, support to coordinate and control the collaboration must be provided. Coordination allows the cooperative operation of two or more systems in the pursuit of complementary objectives, as well as the efficient utilization of resources and allocation of efforts in the organization(s). Much of the support for the collaboration effort required by concurrent engineering comes from the information technologies and the great advances they have experienced in the last decade. Computer-supported collaborative work (CSCW) has been implemented largely for engineering collaboration (Phillips 1998), along with more sophisticated techniques, such as conflict resolution in distributed design (Nof and Huang 1998). The advantages from using information technologies in collaborative manufacturing arise from two sources. First, more information can be acquired from teams having computer support. Second, information availability makes possible a more objective analysis of the problem in a system view. However, some drawbacks should be kept in mind: information overload, lack of knowledge integration, cooperation, and coordination among team members may render the utilization of CSCW tools completely counterproductive.

A key aspect that any CSCW tool must consider is reconfiguration. Adaptation to constantly changing conditions in the manufacturing industry must be attained through tools providing the enterprise with reconfiguration capabilities. Numerous methods for quick reconfiguration of collaborative engineering initiatives have been developed so far. Methods range from those based on integration requirements of the activities being performed (Khanna and Nof 1994; Witzerman and Nof

1995; and Kim and Nof 1997, among others) to those based on concepts taken from disciplines other than manufacturing (Khanna et al. 1998; Ceroni 1999; and Ceroni and Nof 1999, who extend the parallel computing problem to manufacturing modeling).

4. FRAMEWORK FOR COLLABORATIVE MANUFACTURING

The distributed environment presents new challenges for the design, management, and operational functions in organizations. Integrated approaches for designing and managing modern companies have become mandatory in the modern enterprise. Historically, management has relied on a well-established hierarchy, but, the need for collaboration in modern organizations overshadows the hierarchy and imposes networks of interaction among tasks, departments, companies, and so on. As a result of this interaction, three issues arise that make the integration problem critical: variability, culture, and conflicts. Variability represents all possible results and procedures for performing the tasks in the distributed organizations. Variability is inherently present in the processes, but distribution enhances its effects. Cultural aspects such as language, traditions, and working habits impose additional requirements for the integration process of distributed organizations. Lastly, conflicts may represent an important obstacle to the integration process. Conflicts here can be considered as the tendency to organize based on local optimizations in a dual local/global environment. Collaborative relationships, such as user–supplier, are likely to present conflicts when considered within a distributed environment. Communication of essential data and decisions plays a crucial role in allowing organizations to operate cooperatively. Communication must take place in a timely basis in order to be an effective integration facilitator and allow organizations to minimize their coordination efforts and costs.

The organizational distributed environment has the following characteristics (Hirsch et al. 1995):

- Cooperation of different (independent) enterprises
- Shifting of project responsibilities during the product life cycle
- Different conditions, heterogeneity, autonomy, and independence of the participants' hardware and software environments

With these characteristics, the following series of requirements for the integration of distributed organizations can be established as the guidelines for the integration process:

- Support of geographically distributed systems and applications in a multisite production environment and, in special cases, the support of site-oriented temporal manufacturing
- Consideration of heterogeneity of systems ontology, software, and hardware platforms and networks
- Integration of autonomous systems within different enterprises (or enterprise domains) with unique responsibilities at different sites
- Provision of mechanisms for business process management to coordinate the information flow within the entire integrated environment

Among further efforts to construct a framework for collaborative manufacturing is Nofs' taxonomy of integration (Figure 1 and Table 2), which classifies collaboration in four types: mandatory, optional, concurrent, and resource sharing. Each of these collaboration types is found along a human–machine integration level and an interaction level (interface, group decision support system, or computer-supported collaborative work).

5. FACILITATING AND IMPLEMENTING COLLABORATION IN MANUFACTURING

During the design stages of the product, *codesign* (Eberts and Nof 1995) refers to integrated systems implemented using both hardware and software components. Computer-supported collaborative work (CSCW) allows the integration and collaboration of specialists in an environment where work and codesigns in manufacturing are essential. The collaboration is accomplished by integrating CAD and database applications, providing alphanumeric and graphical representations for the system's users. Codesign protocols were established for concurrency control, error recovery, transaction management, and information exchange (Figure 2). The CSCW tool supports the following design steps:

- Conceptual discussion of the design project
- High-level conceptual design
- Testing and evaluation of models
- Documentation

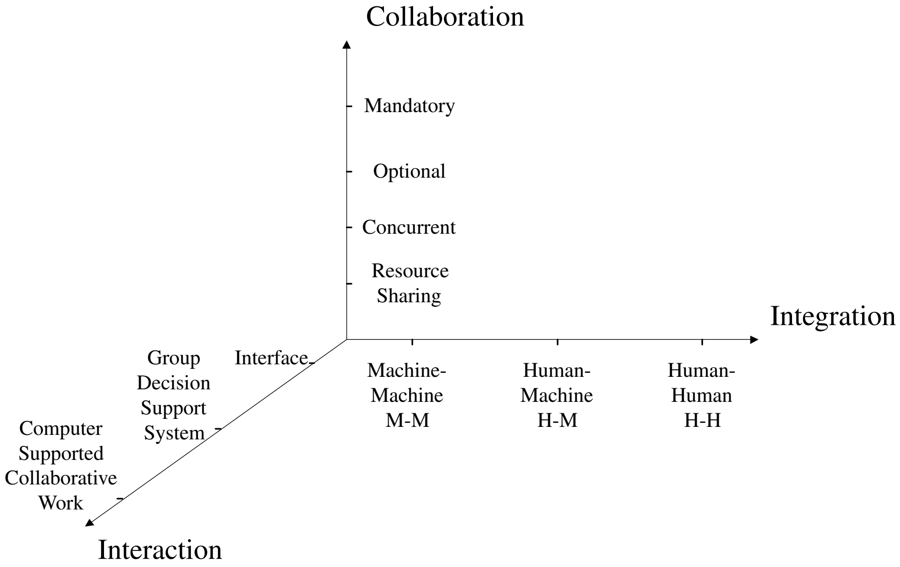

Collaboration
Mandatory
Optional
Concurrent
Resource Sharing
Integration
Machine-Machine M-M
Human-Machine H-M
Human-Human H-H
Interface
Group Decision Support System
Computer Supported Collaborative Work
Interaction

Figure 1 Integration Taxonomy. (From Nof 1994)

When deciding on its operation, every manufacturing enterprise accounts for the following transaction costs (Busalacchi 1999):

1. Searching for a supplier or consumer
2. Finding out about the nature of the product
3. Negotiating the terms for the product
4. Making a decision on suppliers and vendors
5. Policing the product to ensure quality, quantity, etc.
6. Enforcing compliance with the agreement

Traditionally, enterprises grew until they could afford to internalize the transaction costs of those products they were interested in. However, the transaction costs now have been greatly affected by

TABLE 2 Example of Collaboration Types for Integration Problems

	Integration Problem	Example		Collaboration Type
1	Processing of task without splitting and with sequential processing	Concept design followed by physical design	H-H	Mandatory, sequential
2	Processing of task with splitting and parallel processing	Multi-robot assembly	M-M	Mandatory, parallel
3	Processing of task without splitting. Very specific operation	Single human planner (particular)	H-M	Optional, similar processors
4	Processing of task without splitting. General task	Single human planner (out of a team)	H-M	Optional, different processors types
5	Processing of task can have splitting	Engineering team design	H-H	Concurrent
6	Resource allocation	Job-machine assignment	M-M	Competitive
7	Machine substitution	Database backups	M-M	Cooperative

From Nof 1994.

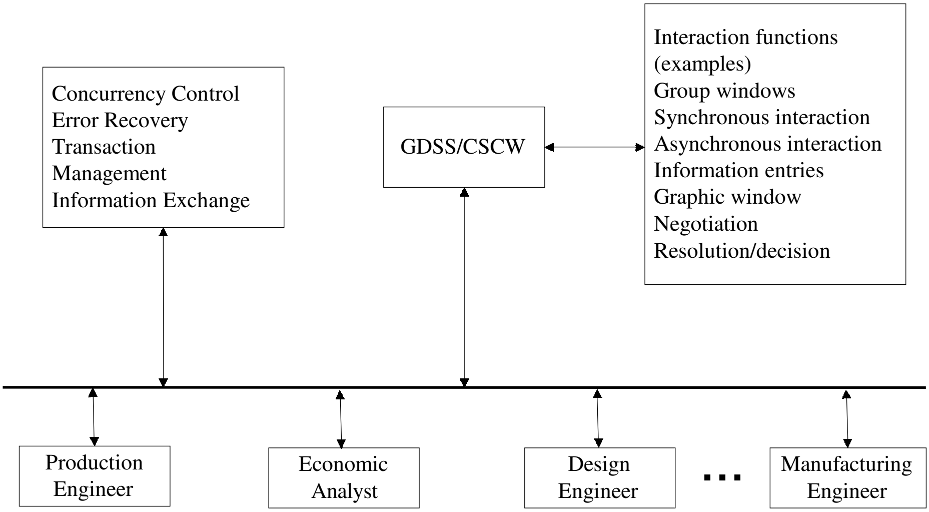

Concurrency Control
Error Recovery
Transaction
Management
Information Exchange
GDSS/CSCW
Interaction functions
(examples)
Group windows
Synchronous interaction
Asynchronous interaction
Information entries
Graphic window
Negotiation
Resolution/decision
Production Engineer
Economic Analyst
Design Engineer
...
Manufacturing Engineer

Figure 2 Codesign Computer-Supported Collaborative Work (CSCW) System.

technologies such electronic data interchange (EDI) and the Internet, which are shifting the way of doing business, moving the transaction costs to:

1. Coordination between potential suppliers or consumers
2. Rapid access to information about products
3. Means for rapid negotiation of terms between suppliers and consumers
4. Access to evaluative criteria for suppliers and consumers
5. Mechanisms for ensuring the quality and quantity of products
6. Mechanisms for enforcing compliance with the agreement

6. MOVING FROM FACILITATING TO ENABLING COLLABORATION: E-WORK IN THE MANUFACTURING ENVIRONMENT

We define e-work as collaborative, computer-supported activities and communication-supported operations in highly distributed organizations of humans and/or robots or autonomous systems, and we investigate fundamental design principles for their effectiveness (Nof 2000a,b). The premise is that without effective e-work, the potential of emerging and promising electronic work activities, such as virtual manufacturing and e-commerce, cannot be fully realized. Two major ingredients for future effectiveness are autonomous agents and active protocols. Their role is to enable efficient information exchanges at the application level and administer tasks to ensure smooth, efficient interaction, collaboration, and communication to augment the natural human abilities.

In an analogy to massively parallel computing and network computing, the teamwork integration evaluator (TIE) has been developed (Nof and Huang 1998). TIE is a parallel simulator of distributed, networked teams of operators (human, robots, agents). Three versions of TIE have been implemented with the message-passing interface (on Intel's Paragon, on a network of workstations, and currently on Silicon Graphics' Origin 2000):

1. TIE/design (Figure 3) to model integration of distributed designers or engineering systems (Khanna et al. 1998)
2. TIE/agent (Figure 4) to analyze the viability of distributed, agent-based manufacturing enterprises (Huang and Nof, 1999)
3. TIE/protocol (Figure 5) to model and evaluate the performance of different task administration active protocols, such as in integrated assembly-and-test networks (Williams and Nof 1995)

7. CASE EXAMPLES

Global markets are increasingly demanding that organizations collaborate and coordinate efforts for coping with distributed customers, operations, and suppliers. An important aspect of the collaboration

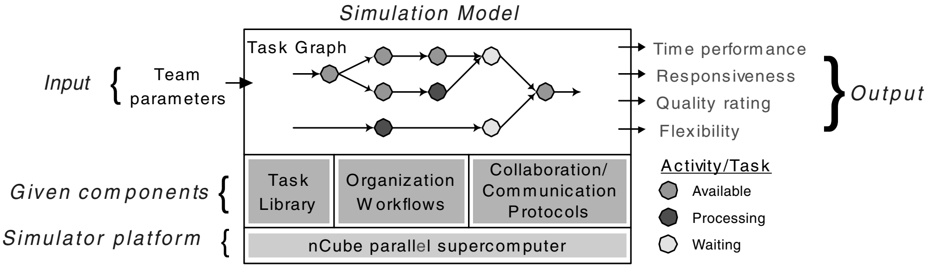
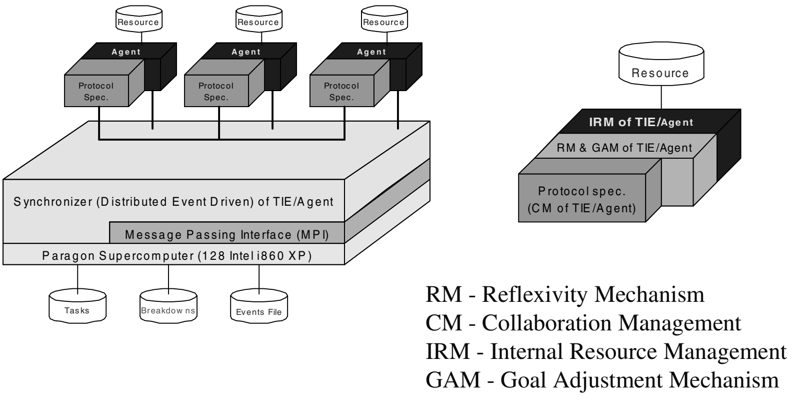

Simulation Model
Input
Team parameters
Task Graph
Time performance
Responsiveness
Quality rating
Flexibility
Output
Given components
Task Library
Organization Workflows
Collaboration/ Communication Protocols
Simulator platform
nCube parallel supercomputer
Activity/Task
Available
Processing
Waiting

Figure 3 Integration of Distributed Designers with TIE/Design.

process of distributed, often remote organizations is the coordination cost. The coordination equipment and operating costs limit the benefit attainable from collaboration. In certain cases, this cost can render the interaction among distributed organizations nonprofitable. Previous research investigated a distributed manufacturing case, operating under a job-shop model with two distributed collaborating centers, one for sales and one for production. A new model incorporating the communication cost of coordination has been developed (Ceroni et al. 1999) yields the net reward of the total system, determining the profitability of the coordination. Two alternative coordination modes are examined: (1) distributed coordination by the two centers and (2) centralized coordination by a third party. The results indicate that distributed and centralized coordination modes are comparable up to a certain limit; over this limit, distributed coordination is always preferred.

7.1. Coordination Cost in Collaboration

In a modern CIM environment, collaboration among distributed organizations has gained importance as companies try to cope with distributed customers, operations, and suppliers (Papastavrou and Nof 1992; Wei and Zhongjun 1992). The distributed environment constrains companies from attaining operational efficiency (Nof 1994). Furthermore, coordination becomes critical as operations face real-time requirements (Kelling et al. 1995). The role of coordination is demonstrated by analyzing the coordination problem of sales and production centers under a job-shop operation (Matsui 1982, 1988). Optimality of the centers' cooperative operation and suboptimality of their noncooperative operation have been demonstrated for known demand, neglecting the coordination cost (Matsui et al. 1996). We introduce the coordination cost when the demand rate is unknown and present one model of the coordination with communication cost. The communication cost is modeled by a message-passing protocol with fixed data exchange, with cost depending on the number of negotiation iterations for reaching the system's optimal operating conditions. The model developed here is based on the research in Matsui et al. (1996) and the research developed on the integration of parallel distributed production systems by the Production Robotics and Integration Software for Manufacturing Group (PRISM) at Purdue University (Ceroni 1996; Ceroni and Nof 1999).


Resource
Resource
Resource
Agent
Agent
Agent
Protocol Spec.
Protocol Spec.
Protocol Spec.
Synchronizer (Distributed Event Driven) of TIE/Agent
Message Passing Interface (MPI)
Paragon Supercomputer (128 Intel i860 XP)
Tasks
Breakdowns
Events File
Resource
IRM of TIE/Agent
RM & GAM of TIE/Agent
Protocol spec. (CM of TIE/Agent)

RM - Reflexivity Mechanism
CM - Collaboration Management
IRM - Internal Resource Management
GAM - Goal Adjustment Mechanism

Figure 4 Viability Analysis of Distributed Agent-Based Enterprises with TIE/Agent.

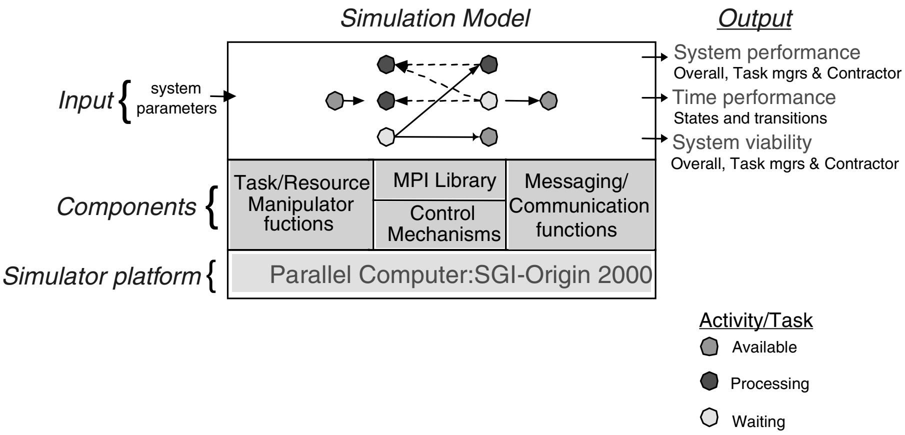

Simulation Model
Output
Input
system parameters
System performance
Overall, Task mgrs & Contractor
Time performance
States and transitions
System viability
Overall, Task mgrs & Contractor
Components
Task/Resource Manipulator fuctions
MPI Library
Control Mechanisms
Messaging/ Communication functions
Simulator platform
Parallel Computer:SGI-Origin 2000
Activity/Task
Available
Processing
Waiting

Figure 5 Modeling and Evaluation of Task Administration Protocols with TIE/Protocol.

7.1.1. *Job-Shop Model*

The job-shop model consists of two distributed centers (Figure 6). Job orders arrive at the sales center and are selected by their marginal profit (Matsui 1985). The production center processes the job orders, minimizing its operating cost (Tijms 1977).

7.1.2. *Coordination Cost*

Two basic coordination configurations are analyzed: (1) a distributed coordination model in which an optimization module at either of the two centers coordinates the optimization process (Figure 7) and (2) a centralized coordination model where a module apart from both centers optimizes all operational parameters (Figure 8).

The distributed model requires the centers to exchange data in parallel with the optimization module. The centralized model provides an independent optimization module.

Coordination cost is determined by evaluating (1) the communication overhead per data transmission and (2) the transmission frequency over the optimization period. This method follows the concepts for integration of parallel servers developed in Ceroni and Nof (1999). Communication overhead is evaluated based on the message-passing protocol for transmitting data from a sender to one or more receptors (Lin and Prassana 1995). The parameters of this model are exchange rate of messages from/to the communication channel (t_d), transmission startup time (t_s), data packing/unpacking time from/to the channel (t_l), and number of senders/receptors (p).

7.1.3. *Results*

The coordination of distributed sales and production centers is modeled for investigating the benefits of different coordination modes. Results show that the coordination cost and the number of negotiation iterations should be considered in the decision on how to operate the system. The numerical results indicate:

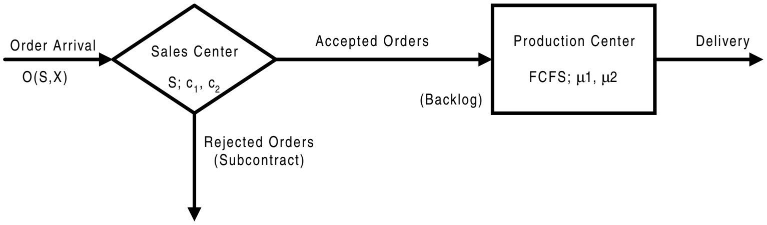

Order Arrival
O(S,X)
Sales Center
S; c1, c2
Accepted Orders
(Backlog)
Production Center
FCFS; μ1, μ2
Delivery
Rejected Orders
(Subcontract)

Figure 6 Distributed Job-shop Production System.

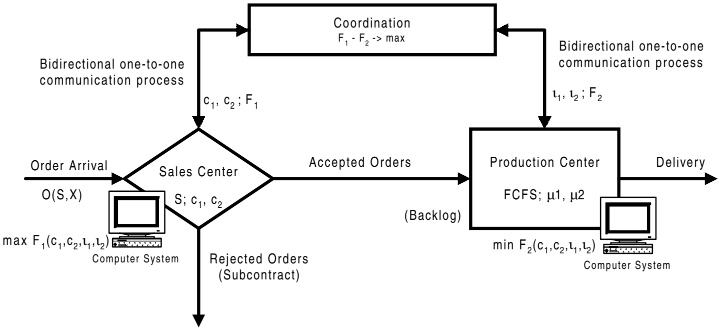

Coordination
$F_1 - F_2$ -> max
Bidirectional one-to-one communication process
$c_1, c_2 ; F_1$
Bidirectional one-to-one communication process
$\iota_1, \iota_2 ; F_2$
Order Arrival
O(S,X)
Sales Center
S; c_1, c_2
max $F_1(c_1,c_2,\iota_1,\iota_2)$
Computer System
Accepted Orders
(Backlog)
Production Center
FCFS; μ1, μ2
Delivery
min $F_2(c_1,c_2,\iota_1,\iota_2)$
Computer System
Rejected Orders
(Subcontract)

Figure 7 Distributed Configuration for the Coordination Model.

1. Same break-even point for the distributed and centralized modes at 400 iterations and $\lambda = 2$ jobs per time unit.
2. The lowest break-even point for the centralized mode at 90 iterations and $\lambda = 5$ jobs per time unit.
3. Consistently better profitability for the centralized mode. This effect is explained by lower communication requirements and competitive hardware investment in the centralized mode.
4. The distributed mode with consistently better profitability than the centralized mode at a higher hardware cost. This shows that distributed coordination should be preferred at a hardware cost less than half of that required by the centralized coordination mode.

From this analysis, the limiting factor in selecting the coordination mode is given by the hardware cost, with the distributed and centralized modes becoming comparable for a lower hardware investment in the centralized case. Coordination of distributed parties interacting for attaining a common goal is also demonstrated to be significant by Ceroni and Nof (1999) with the inclusion of parallel servers in the system. This model of collaborative manufacturing is discussed next.

7.2. Collaboration in Distributed Manufacturing

Manaco S.A. is a Bolivian subsidiary of Bata International, an Italian-based shoemaker company with subsidiaries in most South American countries as well as Canada and Spain. The company has several plants in Bolivia, and for this illustration the plants located in the cities of La Paz and Cochabamba (about 250 miles apart) are considered. The design process at Manaco is performed by developing prototypes of products for testing in the local market. The prototypes are developed at the La Paz plant and then the production is released to the Cochabamba plant. This case study analyzes the integration of the prototype production and the production-planning operations being performed at distributed locations by applying the distributed parallel integration evaluation model (Ceroni 1999).

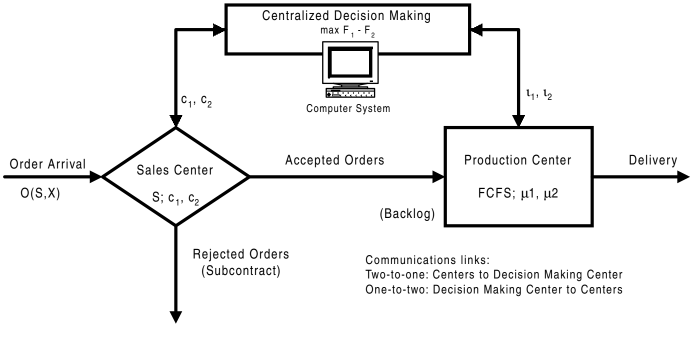

Centralized Decision Making
max $F_1 - F_2$
Computer System
c_1, c_2
ι_1, ι_2
Order Arrival
O(S,X)
Sales Center
S; c_1, c_2
Accepted Orders
(Backlog)
Production Center
FCFS; μ1, μ2
Delivery
Rejected Orders
(Subcontract)
Communications links:
Two-to-one: Centers to Decision Making Center
One-to-two: Decision Making Center to Centers

Figure 8 Centralized Configuration for the Coordination Model.

TABLE 3 Description of Tasks in the Distributed Manufacturing Case

Operation: Production Planning (La Paz plant)		Operation: Prototype Production (Cochabamba plant)	
Task	Description	Task	Description
1A	Market research	1B	CAD drawing
2A	Cutting planning	2B	Cutting
3A	Purchasing planning	3B	Sewing
4A	Capacity planning	4B	Assembly
5A	Assembly planning		
6A	Generation of production plan		

The production-planning operation (operation A) consists of six tasks, with some of them being executed in parallel. The prototype development operation (operation B) consists of four tasks, all of them sequential. Table 3 and Figure 9 show the description and organization of the tasks in each operation.

Means and standard deviations for the task duration are assumed. The time units of these values are work days (8 hours).

To contrast the situations with and without integration, two alternative solutions were developed. The first solution considers the sequential processing of the operations: the prototype was developed at La Paz and then the results were sent to Cochabamba for performance of the production planning. Parallelism is included at each operation for reducing the individual execution cycles. The second solution considers the integration and inclusion of parallelism in both operations simultaneously.

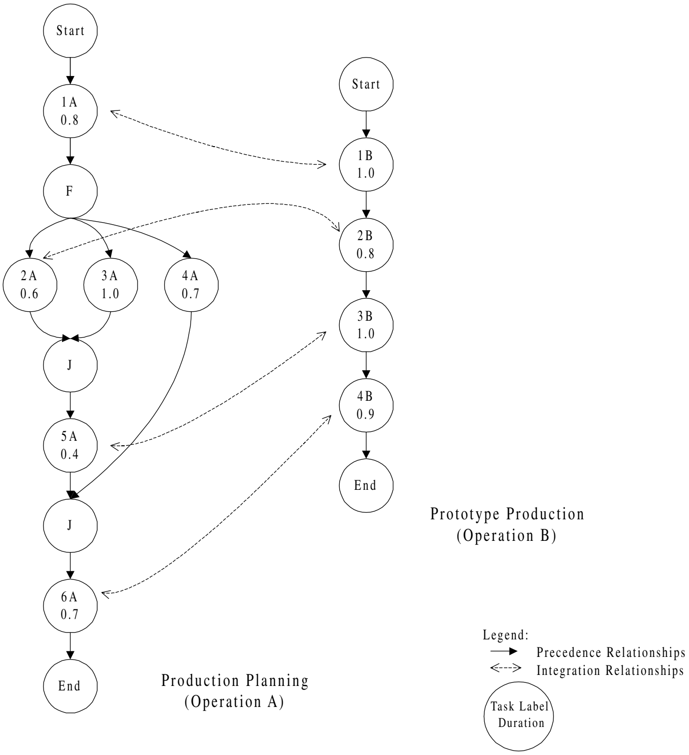

Start
1A 0.8
F
2A 0.6
3A 1.0
4A 0.7
J
5A 0.4
J
6A 0.7
End
Production Planning (Operation A)
Start
1B 1.0
2B 0.8
3B 1.0
4B 0.9
End
Prototype Production (Operation B)
Legend:
Precedence Relationships
Integration Relationships
Task Label Duration

Figure 9 Tasks and Their Precedence Relationship in the Distributed Manufacturing Case.

Assumptions were made for generating an integration model for applying the parallelism optimization.

7.2.1. Integrated Optimization

The integration process in both operations is simplified by assuming relationships between tasks pertaining to different operations. A relationship denotes an association of the tasks based on the similarities observed in the utilization of information, resources, personnel, or the pursuing of similar objectives. Integration then is performed by considering the following relationships:

1A–1B
2A–2B
5A–3B
6A–4B

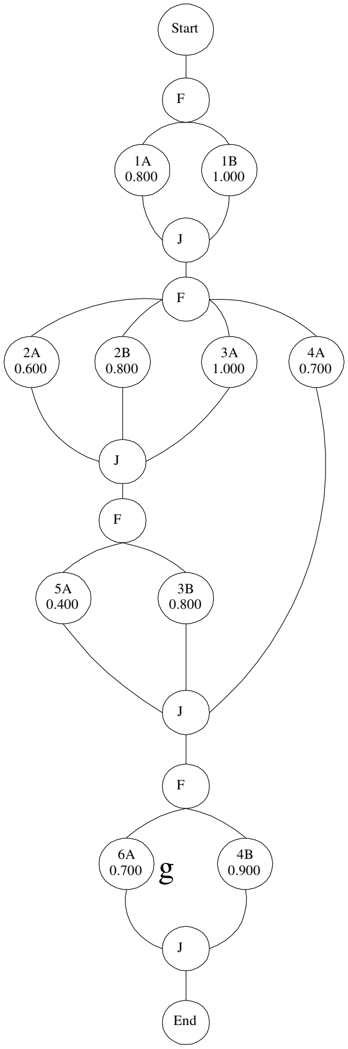

Start
F
1A 0.800
1B 1.000
J
F
2A 0.600
2B 0.800
3A 1.000
4A 0.700
J
F
5A 0.400
3B 0.800
J
F
6A 0.700
g
4B 0.900
J
End

Figure 10 Integrated Model of the Distributed Manufacturing Tasks.

TABLE 4 Simulation Results for Operation A

Task	λ_I Average Communication Delay	$\sigma_{\lambda I}$ Standard Deviation	π Average Duration
1A	0.0008104	0.000129777	0.801865
2A	0.0004519	0.000154528	0.595181
4A	0.0003821	2.87767E-05	0.977848
3A	0.0003552	1.83533E-05	0.734888
5A	0.0277421	0.013688588	0.399918
6A	0.0561208	0.022934333	0.698058

The task relationships allow the construction of an integrated model of the operations. This integrated model preserves the execution order of the tasks as per their local model (Figure 9). Figure 10 shows the integrated model for operations A and B.

Once the integrated schema was generated, the parallelism analysis was performed. In order to evaluate the parallelism in the system, the time of communication and congestion delays needed to be estimated. The estimation of these delays was performed using the software TIE 1.4 (Khanna and Nof 1994; Huang and Nof 1998). TIE 1.4 allows the simulation of a network of distributed tasks with an Intel Paragon Supercomputer, up to a maximum of 132 parallel processors. TIE 1.4 uses a message-passing mechanism for communicating data among the computer nodes simulating the tasks. The data transmission can take place either synchronously or asynchronously. In synchronic data transmission the activity of the sending processor is stalled while waiting for confirmation from the receiver processor. In asynchronic data transmission, the sending processor does not wait for confirmation from receiving nodes and continue with their activity.

The simulation with TIE 1.4 is modeled based on two types of programs: a controller node and a task node. The controller assigns each of the programs to the available computer nodes and starts the execution of the first task in the execution sequence. The implementation in TIE 1.4 will require as many computer nodes as there are tasks in the operation plus the controller node. For example, operation A has 6 tasks, requiring a partition of 7 nodes for its execution on the parallel computer.

7.2.2. Simulation Results

Three models were simulated: operation A, operation B, and integrated operation. A total of 10 simulation runs were performed for each model, registering in each case the production (Π), interaction (T), and total (Φ) times, and the degree of parallelism (Ψ), which is a concurrency measurement for the system. The results obtained for each case are presented in Tables 4 to 6. The simulation of the individual operations allows us to generate an estimate of the delay times due to communication and congestion, both required to optimize the operations locally.

The results obtained from the simulation of the integrated operation were utilized for determining the parallelism of the tasks. The parallelism optimization assumes the tasks' duration and communication times as per those generated by the TIE 1.4 simulation (Table 5) and congestion time as $0.02e^{0.05*\Psi}$. The results obtained are presented in Table 7 and Figures 11 and 12.

The solution generated includes the number of parallel servers for the tasks shown in Figure 12.

7.2.3. Local Optimization

For generating the local optimal configurations of the tasks, the PIEM model was applied to both cases with the congestion delays computed according to the expression $0.02e^{0.05*\Psi}$. The results obtained for each operation are presented in Figures 13 and 14.

TABLE 5 Simulation Results for Operation B

Task	λ_I Average Communication Delay	$\sigma_{\lambda I}$ Standard Deviation	π Average Duration
1B	0.000779	0.000073	0.971070
2B	0.000442	0.000013	0.793036
3B	0.000450	0.000013	0.790688
4B	0.000433	0.000017	0.877372

TABLE 6 Simulation Results for the Integrated Operation

Task	λ_I Average Communication Delay	$\sigma_{\lambda I}$ Standard Deviation	π Average Duration
1A	0.000523	0.033373	0.777950
1B	0.000428	0.030206	0.726785
2A	0.000423	0.040112	0.725435
2B	0.000409	0.023629	0.726027
3A	0.000382	0.053346	0.737365
3B	0.000411	0.036727	0.720308
4A	0.000348	0.079189	0.689378
4B	0.000422	0.024282	0.725532
5A	0.040830	0.037069	0.714532
6A	0.102065	0.015836	0.748419

7.2.4. *Case Remarks*

The numerical results obtained from the local and integrated scenarios are presented in Table 8.

The results in Table 8 show a slight difference in the total production time, which seems to contradict the hypothesis that the integrated scenario will benefit from a reduction of the cycle time. However, it must be noted that the number of subtasks required by the local scenario for achieving a comparable total production time is double that required by the integrated scenario. This situation is the result of no constraint being imposed on the maximum number of subtasks for each task in each scenario (infinite division of tasks). In particular for operation B, the number of subtasks in which each task is divided is nine, which can be considered excessive given the total number of four tasks in the operation.

For evaluating the comparative performance of the integrated and local scenarios, the local scenario for each operation was chosen according to the final number of subtasks in the integrated

TABLE 7 PIEM Model Result for the Integrated Operation

Iteration #	Task Modified	Ψ	Π	T	Φ
0	–	4	3.7000	0.2006	3.9006
1	1B	4	3.5000	0.2010	3.7010
2	3A	5	3.3000	0.2169	3.5169
3	4B	5	3.1000	0.2173	3.3173
4	1A	5	2.8000	0.2178	3.0178
5	2B	6	2.2000	0.2380	2.4380
6	6A	6	1.9500	0.3401	2.2901
7	2A	7	1.8500	0.3660	2.2160
8	1B	7	1.7500	0.3664	2.1164
9	3A	8	1.6500	0.3995	2.0495
10	4B	8	1.5500	0.3999	1.9499
11	1A	8	1.4833	0.4004	1.8838
12	2B	9	1.4167	0.4428	1.8595
13	5A	9	1.38333	0.4837	1.8670
14	4A	10	1.28333	0.5379	1.8212
15	6A	10	1.23333	0.6400	1.8733
16	1B	10	1.16667	0.6404	1.8071
17	3A	11	1.13333	0.7100	1.8433
18	2A	12	1.10000	0.7993	1.8993
19	4B	12	1.03333	0.7997	1.8330
20	1A	12	1.01667	0.8002	1.8169
21	2B	13	0.93333	0.9147	1.8480
22	6A	13	0.92500	1.0168	1.9418
23	1B	13	0.87500	1.0172	1.8922
24	3A	14	0.82500	1.1641	1.9891
25	4B	14	0.78000	1.1645	1.9445

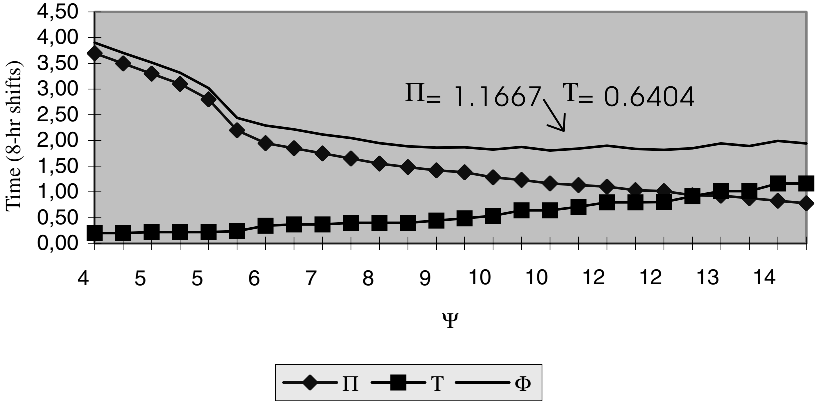


Figure 11 Changes in the Total Production Time per Unit (Π) and Congestion Time (T) for Different Values of the Degree of Parallelism (Ψ) at the Integrated Operation.

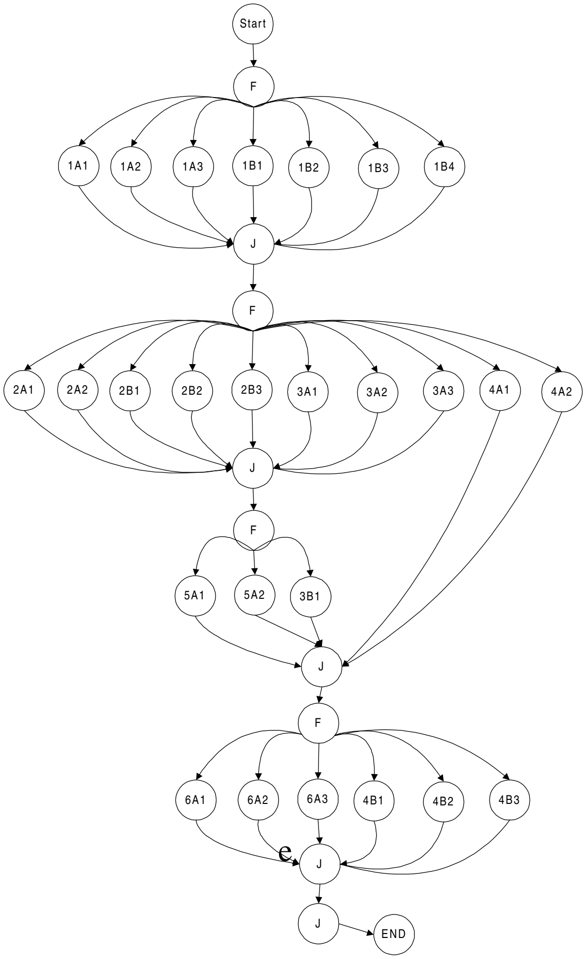

Start
F
1A1
1A2
1A3
1B1
1B2
1B3
1B4
J
F
2A1
2A2
2B1
2B2
2B3
3A1
3A2
3A3
4A1
4A2
J
F
5A1
5A2
3B1
J
F
6A1
6A2
6A3
4B1
4B2
4B3
e
J
J
END

Figure 12 Configuration of Parallel Tasks for the Integrated Operation.

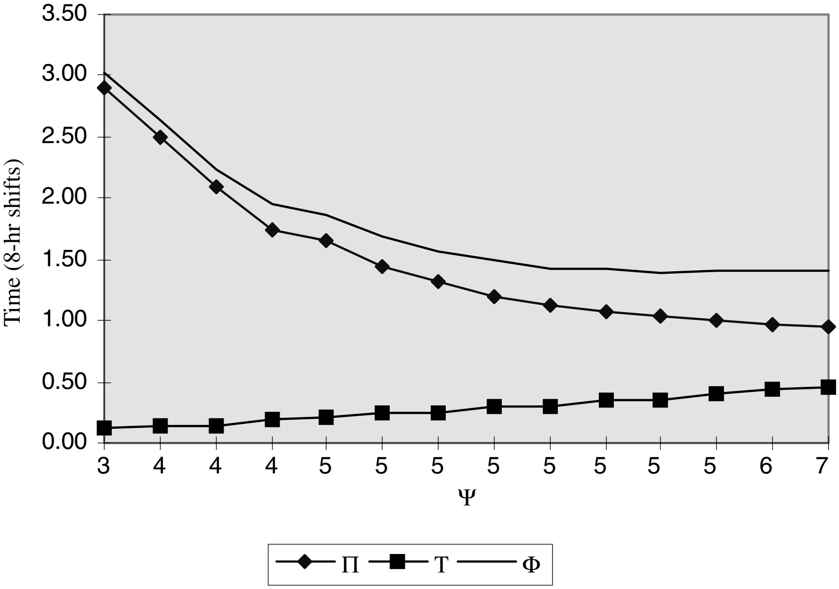


Figure 13 Total Production Time (Π) and Congestion Time (T) for Different Values of the Degree of Parallelism (Ψ) at Operation A.

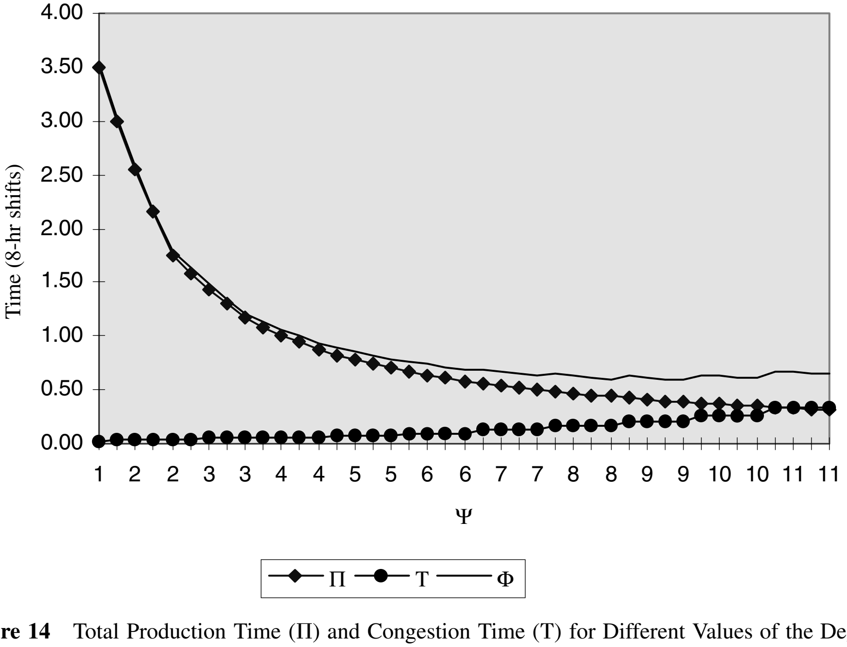


Figure 14 Total Production Time (Π) and Congestion Time (T) for Different Values of the Degree of Parallelism (Ψ) at Operation B.

TABLE 8 Summary of the Results for the Local and Integrated Scenarios

	Φ	Π	T	Number of Subtasks
Local Scenario	1.9846	1.4239	0.5607	52
Operation A	1.3908	1.0350	0.3558	16
Operation B	0.5938	0.3889	0.2049	36
Integrated Scenario	1.8071	1.1666	0.6404	26

scenario. Therefore, the number of subtasks was set at 15 for operation A and 11 for operation B. This makes a total of 26 subtasks in both local operations, which equals the number of subtasks for the integrated scenario. The values for the performance parameters show a total production time (Π + Φ) of 2.7769 shifts, direct-production time (Π) of 2.3750, and interaction time (Φ) of 0.4019 shifts. These values represent an increment of 54% for the total production time, an increment of 104% for the production time, and a decrement of 37% for the interaction time with respect to the integrated solution.

The results obtained from this case study reinforce the hypothesis that exploiting potential benefits is feasible when optimizing the parallelism of integrated distributed operations. Key issues in PIEM are the communication and congestion modeling. The modeling of time required by tasks for data transmission relates to the problem of coordination of cooperating servers. On the other hand, the congestion modeling relates to the delays resulting from the task granularity (number of activities being executed concurrently).

7.3. Variable Production Networks

The trend for companies to focus on core competencies has forced enterprises to collaborate closely with their suppliers as well as with their customers to improve business performance (Lutz et al. 1999). The next step in the supply chain concept is the production or supply networks (Figure 15), which are characterized by intensive communication between the partners. The aim of the system is to allocate among the collaborating partners the excess in production demand that could not be faced by one of them alone. This capability provides the entire network with the necessary flexibility to respond quickly to peaks in demand for the products. A tool developed at the Institute of Production Systems at Hanover University, the FAS/net, employs basic methods of production logistics to provide procedures for the efficient use of capacity redundancies in a production network. The tool satisfies the following requirements derived from the capacity subcontracting process:

- Monitoring of resource availability and order status throughout the network
- Monitoring should be internal and between partners

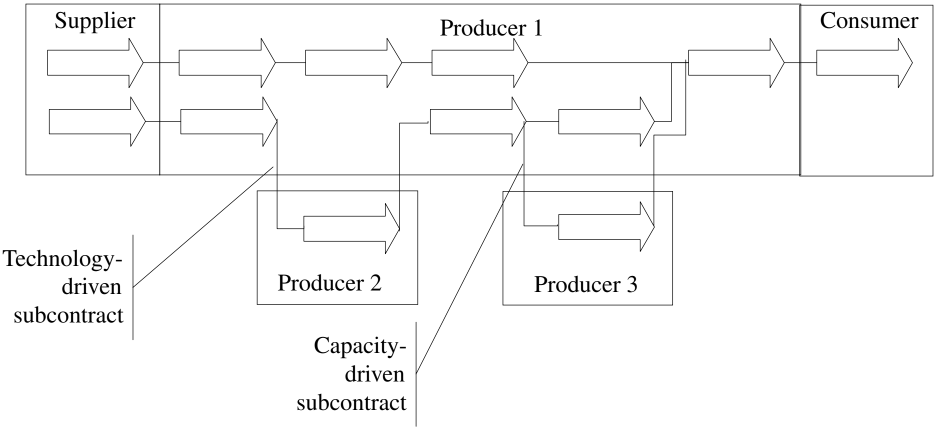


Figure 15 Capacity- and Technology-Driven Subcontracting in Production Networks. (Adapted from Lutz et al. 1999)

- Support of different network partner perspectives (supplier and producer) for data encapsulation
- Detection of the logistics bottlenecks and capacity problems

A key aspect of the system is the identification of orders the partner will not be able to produce. This is accomplished by detecting the bottlenecks through the concept of degree of demand (a comparison of the capacity needed and available in the future, expressed as the ratio between the planned input and the capacity). All the systems with potential to generate bottlenecks are identified by the FAS/net system and ranked by their degree of demand. The subcontracting of the orders can be performed by alternative criteria such as history, production costs, and throughput time.

The system relies on the confidence between partners and the availability of communication channels among them. Carefully planned agreements among the partners concerning the legal aspects, duties, responsibilities, and liability of the exchanged information are the main obstacles to implementing production networks.

8. EMERGING TRENDS AND CONCLUSIONS

The strongest emerging trend that we can observe in the collaborative manufacturing arena is partnership. In the past, the concept of the giant, self-sufficient corporation with presence in several continents prevailed. Emerging now and expected to increase in the years to come are agile enterprises willing to take advantage of and participate in partnerships. The times of winning everything or losing everything are behind us. What matters now is staying in business as competitively as possible. Collaborative efforts move beyond the manufacturing function downstream where significant opportunities exist. Manufacturing is integrating and aligning itself with the rest of the value chain, leaving behind concepts such as product marginal profit that controlled operations for so long. The future will be driven by adaptive channels (Narus and Anderson 1996) for acquiring supplies, producing, delivering, and providing after-sale service. Companies will continue to drive inefficiencies out and become agile systems, forcing companies around them to move in the same direction.

REFERENCES

Adams, J. (1990), ‘‘Fundamental Stocks of Knowledge and Productivity Growth,’’ *Journal of Political Economy,* Vol. 98, No. 4, pp. 673–702.

Arthur, W. B. (1996), ‘‘Increasing Returns and the New World of Business,’’ *Harvard Business Review,* Vol. 74, July–August, pp. 100–109.

Busalacchi, F. A. (1999), ‘‘The Collaborative, High Speed, Adaptive, Supply-Chain Model for Lightweight Procurement,’’ in *Proceedings of the 15th International Conference on Production Research* (Limerick, Ireland), pp. 585–588.

Ceroni, J. A. (1996), ‘‘A Framework for Trade-Off Analysis of Parallelism,’’ MSIE Thesis, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Ceroni, J. A. (1999), ‘‘Models for Integration with Parallelism of Distributed Organizations,’’ Ph.D. Dissertation, Purdue University, West Lafayette, IN.

Ceroni, J. A., and Nof, S. Y. (1997), ‘‘Planning Effective Parallelism in Production Operations,’’ Research Memo 97-10, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Ceroni, J. A., and Nof, S. Y. (1999), ‘‘Planning Integrated Enterprise Production Operations with Parallelism,’’ in *Proceedings of the 15th International Conference on Production Research* (Limerick, Ireland), pp. 457–460.

Ceroni, J. A., Matsui, M., and Nof, S. Y. (1999), ‘‘Communication Based Coordination Modeling in Distributed Manufacturing Systems,’’ *International Journal of Production Economics,* Vols. 60–61, pp. 29–34

Chen, X., Chen, Q. X., and Zhang, P. (1999), ‘‘A Strategy of Internet-Based Agile Manufacturing Chain,’’ in *Proceedings of the 15th International Conference on Production Research* (Limerick, Ireland), pp. 1539–1542

DeVor, R., Graves, R., and Mills, J. (1997), ‘‘Agile Manufacturing Research: Accomplishments and Opportunities,’’ *IIE Transactions,* Vol. 29, pp. 813–823.

Eberts, R. E., and Nof, S. Y. (1995), ‘‘Tools for Collaborative Work,’’ in *Proceedings of IERC 4* (Nashville), pp. 438–441.

Hirsch, B., Kuhlmann, T., Marciniak, Z. K., and Massow, C. (1995), ‘‘Information System Concept for the Management of Distributed Production,’’ *Computers in Industry,* Vol. 26, pp. 229–241.

Huang, C. Y., and Nof, S. Y. (1999), ‘‘Viability—A Measure for Agent-Based Manufacturing Systems,’’ in *Proceedings of the 15th International Conference on Production Research* (Limerick, Ireland), pp. 1869–1872.

Jacobs, J. (1985), *Cities and the Wealth of Nations,* Vintage Books, New York.

Kelling, C., Henz, J., and Hommel, G. (1995), “Design of a Communication Scheme for a Distributed Controller Architecture Using Stochastic Petri Nets,” in *Proceedings of the 3rd Workship on Parallel and Distributed Real-Time Systems* (Santa Barbara, CA, April 25), pp. 147–154.

Khanna, N., and Nof, S. Y. (1994), “TIE: Teamwork Integration Evaluation Simulator—A Preliminary User Manual for TIE 1.1,” Research Memo 94-21, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Khanna, N., Fortes, J. A. B., and Nof, S. Y. (1998), “A Formalism to Structure and Parallelize the Integration of Cooperative Engineering Design Tasks,” *IIE Transactions,* Vol. 30, pp. 1–15.

Kim, C. O., and Nof, S. Y. (1997), “Coordination and Integration Models for CIM Information,” in *Knowledge Based Systems,* S. G. Tzafestas, Ed., World Scientific, Singapore, pp. 587–602.

Lim, L., and Fong, P. (1982), “Vertical Linkages and Multinational Enterprises in Developing Countries,” *World Development,* Vol. 10, No. 7, pp. 585–595

Lin, C., and Prassana, V. (1995), “Analysis of Cost of Performing Communications Using Various Communication Mechanisms,” in *Proceedings of 5th Symposium Frontiers of Massively Parallel Computation* (McLean, VA), pp. 290–297.

Lutz, S., Helms, K., and Wiendahl, H. P. (1999), “Subcontracting in Variable Production Networks,” in *Proceedings of the 15th International Conference on Production Research* (Limerick, Ireland), pp. 597–600.

Matsui, M. (1982), “Job-Shop Model: A M/(G,G)/1(N) Production System with Order Selection,” *International Journal of Production Research,* Vol. 20, No. 2, pp. 201–210.

Matsui, M. (1985), “Optimal Order-Selection Policies for a Job-Shop Production System,” *International Journal of Production Research,* Vol. 23, No. 1, pp. 21–31.

Matsui, M. (1988), “On a Joint Policy of Order-Selection and Switch-Over,” *Journal of Japan Industrial Management Association,* Vol. 39, No. 2, pp. 87–88 (in Japanese).

Matsui, M., Yang, G., Miya, T., and Kihara, N. (1996), “Optimal Control of a Job-Shop Production System with Order-Selection and Switch-Over” (in preparation).

Narus, J. A., and Anderson, J. C. (1996), “Rethinking Distribution: Adaptive Channels,” *Harvard Business Review,* Vol. 74, July–August, pp. 112–120.

Nof, S. Y. (1994), “Integration and Collaboration Models,” in *Information and Collaboration Models of Integration,* S. Y. Nof, Ed., Kluwer Academic Publishers, Dordrecht.

Nof, S. Y., and Huang, C. Y. (1998), “The Production Robotics and Integration Software for Manufacturing (PRISM): An Overview,” Research Memorandum No. 98-3, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Nof, S. Y. (2000a), “Modeling e-Work and Conflict Resolution in Facility Design,” in *Proceedings of the 5th International Conference on Computer Simulation and AI* (Mexico City, February).

Nof, S. Y. (2000b), “Models of e-Work,” in *Proceedings of the IFAC/MIM-2000 Symposium* (Patras, Greece, July).

Papastavrou, J., and Nof, S. Y. (1992), “Decision Integration Fundamentals in Distributed Manufacturing Topologies,” *IIE Transactions,* Vol. 24, No. 3, pp. 27–42.

Phillips, C. L. (1998), “Intelligent Support for Engineering Collaboration,” Ph.D. Dissertation, Purdue University, West Lafayette, IN.

Rodriguez-Clare, A. (1996), “Multinationals, Linkages, and Economic Development,” *American Economic Review,* Vol. 86, No. 4, pp. 852–873.

Sahlman, W. A. (1999), “The New Economy Is Stronger Than You Think,” *Harvard Business Review,* Vol. 77, November–December, pp. 99–106.

Stephan, P. E. (1996), “The Economics of Science,” *Journal of Economic Literature,* Vol. 34, pp. 1199–1235.

Tijms, H. C. (1977), “On a Switch-Over Policy for Controlling the Workload in a System with Two Constant Service Rates and Fixed Switch-Over Costs,” *Zeitschrift für Operations Research,* Vol. 21, pp. 19–32.

Wei, L., Xiaoming, X., and Zhongjun, Z. (1992), “Distributed Cooperative Scheduling for a Job-Shop,” in *Proceedings of 1992 American Control Conference* (Green Valley, AZ), Vol. 1, pp. 830–831.

Williams, N. P., and Nof, S. Y. (1995), “TestLAN: An Approach to Integrated Testing Systems Design,” Research Memorandum No. 95-7, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Wise, R., and Baumgartner, P. (1999), “Go Downstream: The New Profit Imperative in Manufacturing,” *Harvard Business Review,* Vol. 77, September–October, pp. 133–141.

Witzerman, J. P., and Nof, S. Y. (1995), ‘‘Integration of Cellular Control Modeling with a Graphic Simulator/Emulator Workstation,’’ *International Journal of Production Research,* Vol. 33, pp. 3193–3206.

Yusuf, Y. Y., and Sarhadi, M. (1999), ‘‘A Framework for Evaluating Manufacturing Enterprise Agility’’, *Proceedings of the 15th International Conference on Production Research* (Limerick, Ireland), pp. 1555–1558.

II.C
Service Systems

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 21

Service Industry Systems and Service Quality

Martin Wetzels
Ko de Ruyter
Maastricht University

1. INTRODUCTION 623
2. THE NATURE OF SERVICES 624
3. THE SERVICE ENCOUNTER 624
4. DEFINING SERVICE QUALITY 625
5. THE CONCEPTUAL MODEL OF SERVICE QUALITY 626
6. THE SERVQUAL INSTRUMENT 627
7. THE SERVQUAL INSTRUMENT: A CRITICAL REVIEW 628
REFERENCES 631

1. INTRODUCTION

Since the beginning of the 20th century, most economically advanced western societies have evolved from predominantly manufacturing-based to predominantly service-based economies (Bell 1973; Fitzsimmons and Fitzsimmons 1998; Heskett 1987; Mills 1986; Rust et al. 1996; Zeithaml and Bitner 2000). This transition has been especially dramatic in the United States (Ginsberg and Vojta 1981), where from 1948 to 1978 the service sector increased from 54% to 66% of GNP. A similar trend can be discerned in employment. From 1948 to 1977, employment in the service sector has risen from 27.2 million to 54.4 million, more than the total number of people employed in 1948. Currently, the services sector employs approximately 80% of the workforce and accounts for about 75% of GNP (Fitzsimmons and Fitzsimmons 1998; Zeithaml and Bitner 2000). Apart from national indicators, trade in services is growing globally. For the United States, the positive trade balance for services has helped to offset the negative trade balance for goods (Henkoff 1994; Zeithaml and Bitner 2000).

The increased importance of the service sector regarding both national economies and international trade has led to increased attention by marketers on the marketing of services during the last three decades (Swartz and Iacobucci 2000). The marketing field has moved quite rapidly beyond mere definitional issues to the development of models of service management and organization (Swartz et al. 1992). A major research topic in the marketing field currently is service quality (Parasuraman et al. 1985, 1988, 1991; Rust and Oliver 1994). This emphasis on service quality in marketing can be explained to a large extent by the fact that several authors have demonstrated a positive relationship between (service) quality and economic performance (Anderson and Fornell 1994; Buzzell and Gale 1987; Reichheld and Sasser 1990; Rust et al. 1995).

In Section 2, we will explore the nature of services. In Section 3, we will discuss the service encounter, which is at the heart of the majority of service organizations. During the service encounter, service quality is rendered to the customer in the interplay among customer, customer-contact service employee, and service organization. In Section 4, we will focus on defining service quality and discuss the conceptual model of service quality, a framework for the management of service quality. In Section 6, we will describe a measurement instrument, SERVQUAL, that has been derived from this model. In Section 7, we will present a critical review of the conceptual model of service quality and the SERVQUAL instrument.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2. THE NATURE OF SERVICES

Initially, in the marketing field it was assumed that the marketing of goods and services were essentially identical. However, marketers increasingly realized that the marketing of services is separated from the marketing of goods by a number of attributes (Grönroos 1978; Shostack 1977a, b; Zeithaml 1981). Generally, the following four attributes are used to distinguish goods from services (Zeithaml et al. 1985):

1. *Intangibility:* Services are much less tangible than physical goods. Services are experiences rather than objects that can be possessed.
2. *Inseparability of production and consumption:* Goods are first produced and then consumed. Services, on the other, hand, are characterized by simultaneous production and consumption.
3. *Heterogeneity:* The quality of a service may vary from service provider to service provider, from consumer to consumer, and from day to day.
4. *Perishability:* Because services are experiences rather than objects, they cannot be stored. As a result, service providers may find it difficult to synchronize supply and demand.

Intangibility is generally recognized as critical to the dichotomy between goods and services (Zeithaml et al. 1985). The other three attributes can be viewed as consequences of intangibility. Each attribute leads to specific problems for service marketers, which in turn necessitate special marketing strategies to solve them. For instance, intangibility may affect the marketing communications of an organization because services cannot be easily communicated to consumers.

Quality management in service organizations is especially strongly affected by these attributes of services vis-à-vis goods. First, because services are performances rather than objects, service organizations might find it difficult to understand how consumers perceive and evaluate service quality. Consequently, uniform and consistent quality standards can rarely be set (Berry 1980; Zeithaml 1981). Secondly, services are characterized by simultaneous production and consumption. Thus, services are not manufactured at a plant but are generally the result of the interaction between customer and service provider. Consequently, quality control will be rather difficult to ensure (Grönroos 1978). Thirdly, services, especially those with high labor content, are heterogeneous. As a result, consistent and uniform quality will be a serious problem because it is contingent on the interaction between customer and customer-contact service employee (Bitner 1990; Czepiel et al. 1985). Finally, perishability means that services cannot be stored and hence quality cannot be verified in advance of the sale (Shostack 1977a). In the next section, we will discuss the service encounter, which is at the heart of the majority of service organizations.

3. THE SERVICE ENCOUNTER

Findings from the American Customer Satisfaction Index and other national indexes reveal that services are consistently the lowest-scoring sector on customer satisfaction, with public services scoring lowest (Anderson and Fornell 1994). In service organizations, customer satisfaction is often determined by the quality of individual encounters—the service encounter (Bitner 1990; Solomon et al. 1985). The service encounter has been defined as "a period of time during which a consumer directly interacts with a service" (Shostack 1985, p. 243). This definition emphasizes that the service encounter encompasses all elements of the interaction between consumer and service organization: the intangible as well as the tangible elements. Others, however, indicate that the service encounter is mainly conceived as interpersonal interaction between service provider and customer (Chase 1978; Solomon et al. 1985). Solomon et al. (1985, p. 100) define the service encounter as "the face-to-face encounter between a buyer and a seller in a service setting." Although we acknowledge the importance of the personal elements in the service encounter, we feel that tangible elements need to be included in the service encounter (cf. Bitner 1990). For instance, the use of advanced technology and equipment may bestow a feeling of trust and a connotation of high quality on the customer.

The nature of the service encounter is succinctly depicted in Figure 1. Essentially, the service organization consists of two parts: a visible and an invisible part (Chase 1978; Langeard et al. 1981; Shostack 1985). The invisible part is concerned with all organizational processes in the service organization that support the visible part in delivering the service to the customer. The visible part consists of the tangible elements (Bitner 1990) and intangible elements—the customer-contact service employee.

The evaluation of the service encounter can be approached from several perspectives (Bateson 1985; Czepiel et al. 1985): (1) an organizational perspective, (2) a customer perspective, and (3) a customer-contact service employee perspective. The service organization is mainly interested in the performance of the customer-contact service employee because this perspective allows the service organization to attain its objectives. It is therefore essential for the service organization to identify organizational factors that affect the performance of service employees. The customer is mainly

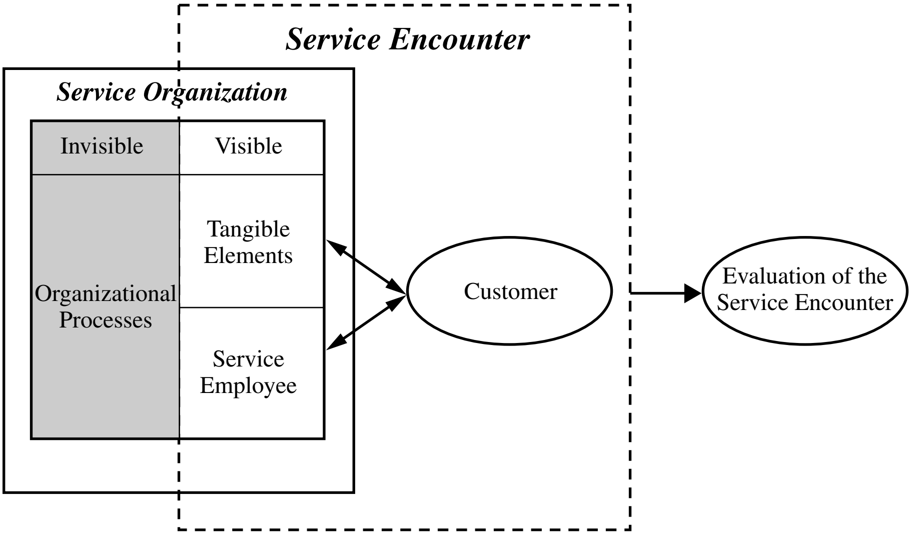


Figure 1 The Service Encounter. (Partly based on the SERVUCTION system model developed by Langeard et al. 1981)

interested in the service quality and customer satisfaction derived from the service encounter. If the evaluation of service quality and customer satisfaction is positive, the customer may decide to remain loyal to the service organization (Bateson 1985; Czepiel et al. 1985). The customer-contact service employee perspective is mainly concerned with the primary rewards of the service encounter, such as pay, promotion, job satisfaction, and recognition from the employee's colleagues and supervisor. These primary rewards are mainly contingent on the employee's performance in the service encounter. However, it should be noted that customer-contact personnel generally really care about the customer and are willing to exert the greatest possible effort to satisfy the customer's needs (Schneider 1980; Schneider and Bowen 1985).

The marketing field has concentrated mainly on the service customer. The customer-contact service employee has been relatively neglected in marketing academia (cf. Hartline and Ferrell 1996). One notable exception is the conceptual model of service quality (Parasuraman et al. 1985), where perceived service quality is determined by four organizational gaps. This model was later extended by adding organizational control and communication processes (Zeithaml et al. 1988). In the next section, we will explore the definition of service quality, which is used as the basis for conceptual model of service quality.

4. DEFINING SERVICE QUALITY

Quality is an elusive and fuzzy concept and as a result is extremely hard to define (Garvin 1984a, b; Parasuraman et al. 1985; Reeves and Bednar 1994; Steenkamp 1989). This may partly be caused by the different perspectives taken by scholars from different disciplines in defining quality. Multiple approaches to defining quality have been identified by various authors (Garvin 1984a, b; Reeves and Bednar 1994). Garvin (1984b) distinguishes five major approaches to define quality.

The *transcendent approach* defines quality as "innate excellence" (Garvin 1984b, p. 25): Proponents of this approach contend that quality cannot be precisely defined but rather is absolute and universally recognizable. This approach finds its origins in philosophy, particularly in metaphysics (Reeves and Bednar 1994; Steenkamp 1989). However, its practical applicability is rather limited because quality cannot be defined precisely using this perspective (Garvin 1984a, b; Steenkamp 1989).

The *product-based approach* posits that quality differences amount to differences in the quantity of a particular desired attribute of the product (Steenkamp 1989). Garvin (1984b, p. 26) provides the following illustration of this approach: "[H]igh-quality ice cream has a high butterfat content, just as fine rugs have a large number of knots per square inch." The assumption underlying this approach suggests two corollaries (Garvin 1984b). First, higher quality products can only be obtained at higher cost because quality is reflected in the quantity of a particular desired attribute. Secondly, quality can be evaluated against an objective standard, namely quantity.

The *user-based approach* to defining quality is based on the notion that "quality lies in the eye of the beholder" (Garvin 1984b, p. 27). In essence, this approach contends that different consumers have different needs. High quality is attained by designing and manufacturing products that meet the specific needs of consumers. As a result, this approach reflects a highly idiosyncratic and subjective view of quality. Juran (1974, p. 2-2), a proponent of this approach, defines quality as "fitness for use." This approach is rooted in the demand side of the market (cf. Dorfman and Steiner 1954). Two issues should be addressed with regard to this approach (Garvin 1984b). The first issue concerns the aggregation of individual preferences at the higher level of the market. The second issue deals with the fact that this approach essentially equates quality with (a maximization) of satisfaction. In other words, as Garvin (1984b, p. 27) puts it: "A consumer may enjoy a particular brand because of its unusual taste or features, yet may still regard some other brand as being of higher quality." As opposed to the user-based approach, the *manufacturing-based approach* to quality originates from the supply side of the market, the manufacturer. This approach is based on the premise that meeting specifications connotes high quality (Crosby 1979). The essence of this approach boils down to this: quality is "conformance to requirements." (Crosby 1979, p. 15). This approach to defining quality is quite elementary, being based on an objective standard or specification (Reeves and Bednar 1994). A critical comment on this approach is articulated by Garvin (1984b), who finds that although a product may conform to certain specifications or standards, the content and validity of those specifications and standards are not questioned. The perspective taken in this approach is predominantly inward. As a result, firms may be unaware of shifts in customer preferences and competitors' (re)actions (Reeves and Bednar 1994).

The ultimate consequence of this approach is that quality improvement will lead to cost reduction (Crosby 1979; Garvin 1984b), which is achieved by lowering internal failure costs (e.g., scrap, rework, and spoilage) and external failure costs (e.g., warranty costs, complaint adjustments, service calls, and loss of goodwill and future sales) through prevention and inspection.

The *value-based approach* presumes that quality can be defined in terms of costs and prices. Garvin (1984b, p. 28) uses the following example to clarify this perspective: "[A] $500 running shoe, no matter how well constructed, could not be a quality product, for it would find few buyers." Reeves and Bednar (1994) emphasize that this definition of quality forces firms to concentrate on internal efficiency ("internal conformance to specifications") and external effectiveness ("the extent to which external customer expectations are met"). However, this approach mixes two distinct, though related, concepts: quality and value (Reeves and Bednar 1994). Because of its hybrid nature, this concept lacks definitional clarity and might result in incompatible designs when implemented in practice.

Reeves and Bednar (1994) propose one additional approach to quality: quality is meeting and/or exceeding customers' expectations. This approach is based on the definition of perceived service quality by Parasuraman et al. (1988, p. 17): "Perceived service quality is therefore viewed as the degree and direction of discrepancy between consumers' perceptions and expectations." This definition was initially conceived in the services marketing literature and as such takes an extremely user-based perspective (Grönroos 1990; Parasuraman et al. 1985). Grönroos (1990, p. 37) in this respect emphasizes: "It should always be remembered that *what counts is quality as it is perceived by the customers*" (emphasis in original).

Relying on only a single approach to defining quality might seriously impede the successful introduction of high-quality products; a synthesis of the above approaches is clearly needed. Garvin (1984b) proposes a temporal synthesis, in which emphasis shifts from the user-based approach to the product-based approach and finally to the manufacturing-based approach as products move from design to manufacturing and to the market. The user-based approach is the starting point because market information must be obtained using marketing research to determine the features that connote high quality. Next, these features must be translated into product attributes (the product-based approach). Finally, the manufacturing approach will need to ensure that products are manufactured according to specifications laid down in the design of the product. This notion is readily recognizable in the conceptual model of service quality conceived by Parasuraman et al. (1985).

5. THE CONCEPTUAL MODEL OF SERVICE QUALITY

Parasuraman et al. (1985) distinguish three premises concerning service quality:

1. Service quality is more difficult for the consumer to evaluate than goods quality.
2. Service quality perceptions result from a comparison of consumer expectations with the actual service performance.
3. Quality evaluations are not made solely on the outcome of a service; they also involve evaluations of the process of service delivery.

From these three, Parasuraman et al. (1985) develop the conceptual model of service quality based on executive interviews and focus group interviews. In this model, GAP5 (perceived service quality) is defined as perception (P) minus expectations (E) and is determined by the magnitude and direction of internal gaps, GAP1–GAP4. The four internal gaps can be described as follows:

1. *Marketing information gap* (*GAP1*)*:* the difference between actual customer expectations and management perception of customer expectations
2. *Standards gap* (*GAP2*)*:* the difference between management perception of customer expectation and service quality specifications
3. *Service performance gap* (*GAP3*)*:* the difference between service quality specifications and the service actually delivered
4. *Communication gap* (*GAP4*)*:* the difference between the service delivered and what is communicated about the service to customers

GAP5 (perceived service quality) is multidimensional in nature. Parasuraman et al. (1985) distinguish 10 underlying dimensions of perceived service quality. These dimensions are summarized in Table 1. Using these original 10 dimensions, Parasuraman et al. (1988) developed a measurement instrument for perceived service quality: SERVQUAL.

6. THE SERVQUAL INSTRUMENT

The SERVQUAL instrument for measuring service quality has evolved into a kind of gospel for academics and practitioners in the field of service quality. With the 10 dimensions in Table 1 as a starting point, 97 items were generated (Parasuraman et al. 1988). Each item consisted of two components: one component reflected perceived service or perceptions and the other component reflected expected service or expectations. Both components were measured on seven-point Likert scale with only the ends of scale anchored by "Strongly disagree" (1) and "Strongly agree" (7). The items were presented in a two consecutive parts. The first part contained the expectation components for the items, while the second part contained the perception components for the items. In order to prevent distortion of the responses by acquiescence bias or "yea-saying or nay-saying" tendencies, about half of the items were negatively worded and the other half positively worded—reverse statement polarization.

Two stages of data collection and scale purification were subsequently carried out. The first stage of data collection and scale purification, using coefficient α, item-to-total correlations and principal components analysis, resulted in a reduction of the number of factors to seven. Five of the original factors were retained in this configuration (see Table 1): (1) tangibles, (2) reliability, (3) responsiveness, (4) understanding/knowing the customer, and (5) access. The remaining five dimensions (communication, credibility, security, competence and courtesy), were collapsed into two dimensions. The number of factors was further reduced in the second stage of data collection and scale purification. The results of principal components analysis suggested an overlap between the dimensions understanding/knowing the customer and access and the dimensions communication, credibility, security, competence and courtesy. Consequently, the overlapping dimensions were combined to form two separate dimensions: (1) assurance and (2) empathy.

Parasuraman et al. (1991) present a replication and extension of their 1988 study. In particular, they propose a number of modifications to the original SERVQUAL instrument (Parasuraman et al. 1988). The first modification is concerned with the expectations section of SERVQUAL. Confronted with extremely high scores on the expectations components of the individual statements, Parasuraman et al. (1991) decided to revise the expectation part of the instrument. Whereas the original scale

TABLE 1 Dimensions of Service Quality

1. Reliability involves consistency of performance and dependability.
2. Responsiveness concerns the willingness or readiness of employees to provide service.
3. Competence means possession of the required skills and knowledge to perform the service.
4. Access involves approachability and ease of contact.
5. Courtesy involves politeness, respect, consideration, and friendliness of contact personnel.
6. Communication means keeping customers informed in language they understand and listening to them.
7. Credibility involves trustiness, believability and honesty.
8. Security is freedom from danger, risk, or doubt.
9. Understanding/knowing the customer involves making the effort to understand the customer's needs.
10. Tangibles include the physical evidence of the service.

Adapted from Parasuraman et al. 1985, p. 47.

reflected normative or ideal expectations, the revised instrument reflected predictive expectations relative to an excellent firm in the industry. For example, with regard to statement no. 5, the expectation item (E5) of the original instrument is formulated as follows: "When these firms promise to do something by a certain time, they *should* do so" (Parasuraman et al. 1988, p. 38). In the revised SERVQUAL instrument, the wording of expectation item no. 5 (E5) has been changed to "When *excellent telephone companies* promise to do something by a certain time, they *will* do so" (Parasuraman et al. 1991, p. 446).

A second modification related to the use of negatively worded items for the responsiveness and empathy dimensions in the original instrument. For the modified instrument, all negatively worded items were replaced by positively worded items. Moreover, in their 1991 study, Parasuraman and his colleagues suggest adding an importance measure to instrument in order to be able to calculate "a composite, weighted estimate of overall service quality" (Parasuraman et al. 1991, p. 424). Parasuraman et al. (1991) propose that importance should be measured by allocating 100 points to the individual dimensions of service quality in accordance with their perceived importance.

7. THE SERVQUAL INSTRUMENT: A CRITICAL REVIEW

Several conceptual and measurement concerns have been raised with regard to the SERVQUAL instrument. The single most important strength of the conceptual model underlying the SERVQUAL instrument is its inherent parsimony. However, Iacobucci et al. (1994) argue that this strength is simultaneously its major shortcoming. The conceptual model of service quality is based on relative evaluations. The absolute level of perceptions (P) and expectations (E) does not enter the equation. The ultimate consequence of this definition is that service quality will be evaluated favorably as long as expectations are met or exceeded. The absolute level of either perceptions or expectations is not explicitly taken into consideration. Iacobucci et al. (1994, p. 16) use the example of Joe's Diner, a truck-stop restaurant with very low expectations, to illustrate the importance of absolute levels of expectations: "[T]he customer who enters Joe's Diner with appropriately low expectations and indeed experiences poor food and rude service. It is unlikely that predicting a favorable evaluation is valid, even though the customer's prior expectations had been met."

Another consequence of this conceptualization of service quality is that services exceeding expectations in the same magnitude (and direction) are predicted to lead to similar levels of perceived service quality. For example, assume that a seven-point Likert-type rating scale is used to capture both perceptions and expectations. Further, assume service A scores 2 on expectations and service B scores a 6 on expectations. The corresponding perception scores are a 3 for service A and a 7 for service B. These results in the same perceived service quality score for both service A ($7 - 6 = 1$) and service B ($3 - 2 = 1$). However, it would be rather unrealistic to assume that both services result in the same level of perceived service quality, since service B exhibited both a higher level of perceptions and expectations.

The multidimensional nature of service quality has been acknowledged in the European as well as the North American research traditions in services marketing (Grönroos 1990; Parasuraman et al. 1988, 1991). Although the exact number of dimensions remains open to discussion, the general notion of multidimensionality seems to be generally accepted. The generalizability of the five SERVQUAL dimensions in other than the original industries (Parasuraman et al. 1988, 1991) is still rather problematic (e.g., Babakus and Boller 1992; Carman 1990; Cronin and Taylor 1992; Paulin and Perrien 1996).

Furthermore, it could be argued that a service encounter consists of two major components: (1) the service process and (2) the service outcome (Grönroos 1990; De Ruyter and Wetzels 1998). The five dimensions identified in the conceptual model of service quality are directed towards the interaction between customer and service provider and therefore focus on the service process (Lapierre 1996). De Ruyter and Wetzels (1998) find in an experimental study that service process and service outcome interact. Although service process is an important determinant of evaluative judgments (e.g., customer satisfaction), it may not wholly compensate for an unfavorable service outcome.

An additional problem with the conceptual model of service is its implicit assumption that each service encounter consists of only a single stage (Lemmink et al. 1998; De Ruyter et al. 1997). In particular, if a service encounter consisted of multiple stages, the dimensions of the conceptual model of service quality would require an extension at the level of the individual stage level. Rust et al. (1995) take this train of thought to an even more extreme position. They are structuring service quality as structures around the business process. Apart from introducing stages into the model, such an approach also ensures managerial relevance.

An issue that has received ample attention in academia in this respect is the differentiation between service quality and customer satisfaction (Cronin and Taylor 1992, 1994; Iacobucci et al. 1994, 1996; Oliver 1993). The confusion surrounding these two constructs in services research can be accounted for by the fact that both are based on the same canonical model (Iacobucci et al. 1996). Service quality and customer satisfaction models share the following characteristics (Iacobucci et al. 1996):

1. Customers are thought to hold the expectations prior to their purchases.
2. Customers make perceptions regarding their purchases.
3. Customers compare their perceptions to their expectations.
4. This comparative process results in evaluations of quality and/or satisfaction (and subsequent effects, e.g., future purchase intentions).

Oliver (1993) proposes that service quality and customer satisfaction differ in four fundamental characteristics. To begin with, the dimensions underlying quality are quite specific, while customer satisfaction can result from any dimension related to the service encounter. Secondly, service quality is based on ideals or "excellence" (Parasuraman et al. 1988, 1991), whereas customer satisfaction can be based on a host of alternative standards, such as predictive norms and experience-based norms (Iacobucci et al. 1994, 1996). Iacobucci et al. (1994, p. 15), following similar reasoning, propose a similar approach: "Perhaps satisfaction is indeed judged by consumers against their own internal standards, whereas quality is would be better defined as judgment relative to managerial or competitive standards." In fact, they propose that service quality is based on external standards and customer satisfaction on internal standards.

This classification of standards is closely related to the following third characteristic suggested by Oliver (1993), who contends that service quality does not require experience with service or service provider. Customer satisfaction, on the other hand, is an experiential construct. Customer satisfaction can only be evaluated by actually experiencing the service encounter (Anderson and Fornell 1994; Iacobucci et al. 1994). Anderson and Fornell (1994) suggest that in general, customer satisfaction is influenced by price, whereas service quality is viewed as independent from price. Price or costs incurred are often modeled using value; value is thus operationalized as the ratio of perceived quality relative to price (cf. Zeithaml 1988).

Finally, service quality and customer satisfaction are based on different sets of antecedents. The antecedents of service quality are mainly limited to communication, both personal and impersonal, and situational characteristics (Zeithaml et al. 1993). Customer satisfaction has been hypothesized to be influenced by a number of cognitive and affective processes, such as disconfirmation, equity, attribution, mood states, and emotions (Oliver 1993).

Another major weakness of the conceptual model of service quality is its omission of financial factors (Anderson and Fornell 1994; Iacobucci et al. 1994; Lemmink et al. 1998; De Ruyter et al. 1997). During the past decade, various competing models have been advanced to explain consumer evaluations of services (Iacobucci et al. 1996). Many of these models include service quality and satisfaction as their basic focal constructs, departing from a comparison between customer expectations and service provider performance (Iacobucci et al. 1994).

Value has frequently been conceptualized as the outcome of a price/quality ratio or the comparison of what one receives with the cost of acquisition (Anderson et al. 1994; Zeithaml 1988). According to this point of view, service customers will attempt to maximize the level of quality in relation to the disutility of price. Price in this case may be interpreted as a psychological price in terms of time and distance. It has been argued that customers will favor service providers that maximize quality minus the disutility from prices. Therefore, while the quality of a service may be conceived of as good, its net or marginal value may still be rated poor if the price of that service is perceived to be too high (Rust and Oliver 1994). This conceptualization of value as a proxy for the quality price ratio may be labeled the value-for-money approach. This approach closely focuses on value as a cognitive construct because an explicit comparison between price and quality is made by consumers. However, it has been emphasized recently that affect should also be considered in determining postpurchase responses (Oliver 1993). If value is perceived as a summary cognitive and affective response then an affective component should also be incorporated in a conceptualization of value. De Ruyter et al. develop a conceptualization of value, in which they introduce three dimensions: (1) emotional, (2) practical, and (3) logical (e.g., Lemmink et al. 1998; De Ruyter et al. 1997).

Moreover, several empirical concerns have also been raised with regard to the conceptual model of service quality, particularly with regard to the measurement instrument SERVQUAL instrument. The dimensionality of the SERVQUAL instrument is a well-researched issue (Asubonteng 1996; Buttle 1996; Paulin and Perrien 1996).

In their initial study, Parasuraman et al. (1988) report relatively high values of coefficient α for the individual dimensions of SERVQUAL. Moreover, using exploratory factor analysis to assess the convergent and discriminant validity, they find that each item loads high only on the hypothesized factor for the four participating companies.

These favorable results, however, seem not to have been replicated in their 1991 study (Parasuraman et al. 1991). Although the reliabilities in terms of coefficient α were still relatively high, their factor-analytic results seemed somewhat problematic. In particular, the tangibles dimension loaded on two dimensions (one representing equipment and facilities and one representing personnel and communication materials), thus casting considerable doubt on the unidimensionality of this dimension. Furthermore, responsiveness and assurance (and to some degree reliability) loaded on the same factor, while in general interfactor correlations were somewhat higher than in their 1988 study. These divergent results might be caused by an artifact: restraining the factor solution to five factors. Therefore, Parasuraman et al. (1991) proposed that a six-factor solution might lead to a more plausible result. Although responsiveness and assurance seemed to be slightly more distinct, tangibles still loaded on two factors, whereas the interfactor correlations remained high.

Replication studies by other authors have fared even less well than the studies by the original authors. Carman (1990) carried out a study using an adapted version of SERVQUAL in four settings: (1) a dental school patient clinic, (2) a business school placement center, (3) a tire store, and (4) an acute care hospital, and found similar factors (although not an equal number) as compared to Parasuraman et al. (1988, 1991). However the item-to-factor stability appeared to be less than in the original studies. Carman (1990) also notes that the applicability in some of the settings requires quite substantial adaptations in terms of dimensions and items. Babakus and Boller (1992) report on a study in which they applied the original SERVQUAL to a gas utility company. Using both exploratory and confirmatory (first-order and second-order) factor analysis, Babakus and Boller (1992) were unable to replicate the hypothesized five-dimensional structure of the original SERVQUAL instrument. Finally, Cronin and Taylor (1992, 1994) used confirmatory factor analysis and found that a five-factor structure did not provide an adequate fit to the data. Subsequently, they carried out exploratory factor analysis and a unidimensional factor structure was confirmed.

Authors using a more metaanalytic approach have reported similar results (Asubonteng et al. 1996; Buttle 1996; Parasuraman et al. 1991; Paulin and Perrien 1996). In general, they report relatively high reliabilities in terms of coefficient α for the individual dimensions. However, results differ considerably when looking at different service-quality dimensions. Paulin and Perrien (1996) find that for the studies included in their overview, coefficient α varies from 0.50 to 0.87 for the empathy dimension and from 0.52 to 0.82 for the tangibles dimension. However, the number of factors extracted and factor loading patterns are inconsistent across studies. Furthermore, interfactor correlations among the responsiveness, assurance and reliability dimensions are quite high (Buttle 1996; Parasuraman et al. 1991). Finally, Paulin and Perrien (1996) suggest that the limited replicability of the SERVQUAL instrument may be caused by contextuality (cf. Cronbach 1986). They find that studies applying SERVQUAL differ in terms of units of study, study observations, and type of study.

Empirical research, however, has found that the inclusion of an importance weight as suggested by Parasuraman et al. (1991) may only introduce redundancy (Cronin and Taylor 1992, 1994). Therefore, the use of importance weights is not recommended; it increases questionnaire length and does not add explanatory power.

The SERVQUAL instrument employs a difference score (perception minus expectation) to operationalize perceived service quality. However, the use of difference scores is subject to serious psychometric problems (Peter et al. 1993). To begin with, difference scores per se are less reliable than their component parts (perceptions and expectations in the case of SERVQUAL). Because reliability places an upper limit on validity, this will undoubtedly lead to validity problems. Peter et al. (1993) indicate that low reliability might lead to attenuation of correlation between measures. Consequently, the lack of correlations between measures might be mistaken as evidence of discriminant validity. Furthermore, difference scores are closely related to their component scores. Finally, the variance of the difference score might potentially be restricted. Peter et al. (1993) point out that this violates the assumption of homogeneity of variances in ordinary least-squares regression and related statistical techniques. Finally, an alternative might be the use of a nondifference score, which allows for the direct comparison of perceptions to expectations (Brown et al. 1993; Peter et al. 1993).

Empirical evidence indicates that the majority of the respondents locate their service quality score at the right-hand side of the scale (Brown et al. 1993; Parasuraman et al. 1988; 1991; Peterson and Wilson 1992). This distribution is referred to as negatively skewed. A skewed distribution contains several serious implications for statistical analysis. To begin with, the mean might not be a suitable measure of central tendency. In a negatively skewed distribution, the mean is typically to the left of the median and the mode and thus excludes considerable information about the variable under study (Peterson and Wilson 1992). Skewness also attenuates the correlation between variables. Consequently, the true relationship between variables in terms of a correlation coefficient may be understated (Peterson and Wilson 1992). Finally, parametric tests (e.g., t-test, F-test) assume that the population is normally or at least symmetrically distributed.

A less skewed alternative for measuring service quality is the nondifference score for service quality. Brown et al. (1993) report that the nondifference score for service quality is approximately normally distributed. Moreover, several authors have suggested that the number of scale points might have considerably contributed to the skewness of satisfaction measures (Peterson and Wilson 1992). Increasing the number of scale points may increase the sensitivity of the scale and consequently reduce skewness.

REFERENCES

Anderson, E. W., and Fornell, C. (1994), "A Customer Satisfaction Research Prospectus," in *Service Quality: New Directions in Theory and Practice,* R. T. Rust and R. L. Oliver, Eds., Sage, Thousand Oaks, CA, pp. 1–19.

Asubonteng, Mcleary, K. J., and Swan, J. E. (1996), "SERVQUAL Revisited: A Critical Review of Service Quality," *Journal of Services Marketing,* Vol. 10, No. 6, pp. 62–81.

Babakus, E., and Boller, G. W. (1992), "An Empirical Assessment of the SERVQUAL Scale," *Journal of Business Research,* Vol. 24, pp. 253–268.

Bateson, J. E. G. (1985), "Perceived Control and the Service Encounter," in *The Service Encounter: Managing Employee/Customer Interaction in the Service Businesses,* J. A. Czepiel, M. R. Solomon and C. F. Surprenant, Eds., Lexington Books, Lexington, MA, pp. 67–82.

Bell, D. (1973), *The Coming of Post-Industrial Society: A Venture in Social Forecasting,* Basic Books, New York.

Berry, L. L. (1980), "Services Marketing Is Different," *Business,* Vol. 30, May–June, pp. 24–28.

Bitner, M. J. (1990), "Evaluating Service Encounters: The Effects of Physical Surroundings and Employee Responses," *Journal of Marketing,* Vol. 54, January, pp. 71–84.

Brown, T. J., Churchill, G. A., and Peter, J. P. (1993), "Improving the Measurement of Service Quality," *Journal of Retailing,* Vol. 69, Spring, pp. 127–139.

Buttle, F. (1996), "SERVQUAL: Review, Critique, Research Agenda," *European Journal of Marketing,* Vol. 30, No. 1, pp. 8–32.

Buzzell, R. D., and Gale, B. T. (1987), *The PIMS Principles: Linking Strategy to Performance,* Free Press, New York.

Carman, J. M. (1990), "Consumer Perceptions of Service Quality: An Assessment of the SERVQUAL Dimensions," *Journal of Retailing,* Vol. 66, No. 1, pp. 33–55.

Chase, R. B. (1978), "Where Does the Customer Fit in a Service Operation," *Harvard Business Review,* Vol. 56, November–December, pp. 137–142.

Cronbach, L. J. (1986), "Social Inquiry by and for Earthlings," in *Metatheory in Social Science: Pluralisms and Subjectivities,* D. W. Fiske and R. A. Schweder, Eds., University of Chicago Press, Chicago.

Cronin, J. J., Jr. and Taylor, S. A. (1992), "Measuring Service Quality: A Reexamination and Extension," *Journal of Marketing,* Vol. 56, July, pp. 55–68.

Cronin, J. J., Jr., and Taylor, S. A. (1994), "SERVPERF versus SERVQUAL: Reconciling Performance-Based and Perceptions-Minus-Expectations Measurement of Service Quality," *Journal of Marketing,* Vol. 58, January, pp. 132–139.

Crosby, P. B. (1979), *Quality Is Free: The Art of Making Quality Certain,* New American Library, New York.

Czepiel, J. A., Solomon M. R., Surprenant, C. F., and Gutman, E. G. (1985), "Service Encounters: An Overview," in *The Service Encounter: Managing Employee/Customer Interaction in the Service Businesses,* J. A. Czepiel, M. R. Solomon, and C.. Surprenant, Eds., Lexington Books, Lexington, MA, pp. 3–15.

De Ruyter, K., and Wetzels, M. G. M. (1998), "On the Complex Nature of Patient Evaluations of General Practice Service," *Journal of Economic Psychology,* Vol. 19, pp. 565–590.

De Ruyter, J. C., Wetzels, M. G. M., Lemmink J., and Mattsson, J. (1997), "The Dynamics of the Service Delivery Process: A Value-Based Approach," *International Journal of Research in Marketing,* Vol. 14, No. 3, pp. 231–243.

Dorfman, R., and Steiner P. O. (1954), "Optimal Advertising and Optimal Quality," *American Economic Review,* Vol. 44, December, pp. 822–836.

Fitzsimmons, J. A., and Fitzsimmons, M. J. (1998), *Service Management: Operations, Strategy, and Information Technology,* Irwin/McGraw-Hill, Boston.

Garvin, D. A. (1984a), "Product Quality: An Important Strategic Weapon," *Business Horizons,* Vol. 27, March–April, pp. 40–43.

Garvin, D. A. (1984b), "What Does 'Product Quality' Really Mean?" *Sloan Management Review,* Vol. 26, Fall, pp. 25–43.

Ginsberg, E., and Vojta, G. (1981), "The Service Sector of the U.S. Economy," *Scientific American,* March, pp. 31–39.

Grönroos, C. (1978), "A Service-Oriented Approach to Marketing of Services," *European Journal of Marketing,* Vol. 12, No. 8, pp. 588–601.

Grönroos, C. (1990), *Service Management and Marketing: Managing the Moments of Truth in Service Competition,* Lexington Books, Lexington, MA.

Hartline, M. D., and Ferrell, O. C. (1996), ''The Management of Customer-Contact Service Employees,'' *Journal of Marketing,* Vol. 60, October, pp. 52–70.

Henkoff, R. (1994), ''Service Is Everybody's Business,'' *Fortune,* June 24, pp. 48–60.

Heskett, J. L. (1987), ''Lessons in the Service Sector,'' *Harvard Business Review,* Vol. 65 March–April, pp. 118–126.

Iacobucci, D., Grayson, K. A., and Ostrom, A. L. (1994), ''The Calculus of Service Quality and Customer Satisfaction: Theoretical and Empirical Differentiation,'' in *Advances in Services Marketing and Management: Research and Practice,* T. A. Swartz, D. E. Bowen, and S. W. Brown, Eds., Vol. 3, JAI Press, Greenwich, pp. 1–67.

Iacobucci, D., Ostrom, A. L., Braig, B. M., and Bezjian-Avery, A. (1996), ''A Canonical Model of Consumer Evaluations and Theoretical Bases of Expectations,'' in *Advances in Services Marketing and Management: Research and Practice,* T. A. Swartz, D. E. Bowen, and S. W. Brown, Eds., Vol. 5, JAI Press, Greenwich, pp. 1–44.

Juran, J. M. (1974), ''Basic Concepts,'' in *Quality Control Handbook,* J. M. Juran, F. M. Gryna, Jr., and R. S. Bingham, Jr., Eds., McGraw-Hill, New York, pp. 2-1–2-24.

Langeard, E., Bateson, J., Lovelock C., and Eiglier, P. (1981), ''Marketing of Services: New Insights from Customers and Managers,'' Report No. 81-104, Marketing Sciences Institute, Cambridge.

Lapierre, J. (1996), ''Service Quality: The Construct, Its Dimensionality and Its Measurement,'' in *Advances in Services Marketing and Management: Research and Practice,* T. A. Swartz, D. E. Bowen, and S. W. Brown, Eds., Vol. 5, JAI Press, Greenwich, 45–70.

Lemmink, J., de Ruyter, J. C., and Wetzels, M. G. M. (1998), ''The Role of Value in the Service Delivery Process of Hospitality Services,'' *Journal of Economic Psychology,* Vol. 19, pp. 159–179.

Mills, P. K. (1986), *Managing Service Industries: Organizational Practices in a Post-industrial Economy,* Ballinger, Cambridge.

Oliver, R. L. (1993), ''A Conceptual Model of Service Quality and Service Satisfaction: Compatible Goals, Different Concepts,'' in *Advances in Services Marketing and Management: Research and Practice,* T. A. Swartz, D. E. Bowen, and S. W. Brown, Eds., Vol. 3, JAI Press, Greenwich, pp. 65–85.

Parasuraman, A. (1995), ''Measuring and Monitoring Service Quality,'' in *Understanding Services Management: Integrating Marketing, Organisational Behaviour and Human Resource Management,* W. J. Glynn and J. G. Barnes, Eds., John Wiley & Sons, Chichester, pp. 143–177.

Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1985), ''A Conceptual Model of Service Quality and Its Implications for Further Research,'' *Journal of Marketing,* Vol. 49, Fall, pp. 41–50.

Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1988), ''SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality,'' *Journal of Retailing,* Vol. 64, Spring, pp. 12–40.

Parasuraman, A., Berry, L. L., and Zeithaml, V. A. (1991), ''Refinement and Reassessment of the SERVQUAL Scale,'' *Journal of Retailing,* Vol. 67, Winter, pp. 420–450.

Paulin, M., and Perrien, J. (1996), ''Measurement of Service Quality: The Effect of Contextuality,'' in *Managing Service Quality,* P. Kunst and J. Lemmink, Eds., Vol. 2, pp. 79–96.

Peter, J. P., Churchill, G. A., Jr., and Brown, T. J. (1993), ''Caution in the Use of Difference Scores in Consumer Research,'' *Journal of Consumer Research,* Vol. 19, March, pp. 655–662.

Peterson, R. A., and Wilson, W. R. (1992), ''Measuring Customer Satisfaction: Fact and Artifact,'' *Journal of the Academy of Marketing Science,* Vol. 20, No. 1, pp. 61–71.

Reeves, C. A., and Bednar, D. A. (1994), ''Defining Quality: Alternatives and Implications,'' *Academy of Management Review,* Vol. 19, No. 3, pp. 419–445.

Reichheld, F. F., and Sasser, W. E., Jr. (1990), ''Zero Defection: Quality Comes to Services,'' *Harvard Business Review,* Vol. 68, September–October, pp. 105–111.

Rust, R. T., and Oliver, R. L. (1994), ''Service Quality: Insights and Managerial Implications from the Frontier,'' in *Service Quality: New Directions in Theory and Practice,* R. T. Rust and R. L. Oliver Eds., Sage, Thousand Oaks, CA, pp. 1–20.

Rust, R. T., Zahorik, A. J., and Keiningham, T. L. (1995), ''Return on Quality (ROQ): Making Service Quality Financially Accountable,'' *Journal of Marketing,* Vol. 59, April, pp. 58–70.

Rust, R. T., Zahorik, A. J., and Keiningham, T. L. (1996), *Service Marketing,* HarperCollins, New York.

Schneider, B. (1980), ‘‘The Service Organization: Climate Is Crucial,’’ *Organizational Dynamics,* Vol. 9, Autumn, pp. 52–65.

Schneider, B., and Bowen, D. E. (1985), ‘‘Employee and Customer Perceptions of Service in Banks: Replication and Extension,’’ *Journal of Applied Psychology,* Vol. 70, pp. 423–433.

Shostack, G. L. (1977a), ‘‘Breaking Free from Product Marketing,’’ *Journal of Marketing,* Vol. 41, April, pp. 73–80.

Shostack, G. L. (1977b), ‘‘Banks Sell Services—Not Things,’’ *Bankers Magazine,* Vol. 160, Winter, p. 40.

Shostack, G. L. (1985), ‘‘Planning the Service Encounter,’’ in *The Service Encounter: Managing Employee/Customer Interaction in the Service Businesses,* J. A. Czepiel, M. R. Solomon, and C. F. Surprenant, Eds., Lexington Books, Lexington, MA, pp. 243–263.

Solomon, M. R., Surprenant, C. F., Czepiel, J. A., and Gutman, E. G. (1985), ‘‘A Role Theory Perspective on Dyadic Interactions: The Service Encounter,’’ *Journal of Marketing,* Vol. 49, Winter, pp. 99–111.

Steenkamp, J. B. E. M. (1989), *Product Quality: An Investigation into the Concept and How It Is Perceived by Consumers,* Van Gorcum, Assen.

Swartz, T. A., and Iacobucci, D. (2000), ‘‘Introduction,’’ in *Handbook of Services Marketing and Management,* Sage, Thousand Oaks, CA.

Swartz, T. A., Bowen, D. E., and Brown, S. W. (1992), ‘‘Fifteen Years after Breaking Free: Services Then, Now and Beyond,’’ in *Advances in Services Marketing and Management,* T. A. Swartz, D. E. Bowen, and S. W. Brown, Eds., Vol. 1, JAI Press, Greenwich, pp. 1–21.

Zeithaml, V. A. (1981), ‘‘How Consumer Evaluation Processes Differ between Goods and Services,’’ in *Marketing of Services,* J. H. Donnelly and W. R. George, Eds., American Marketing Assn., Chicago, pp. 186–190.

Zeithaml, V. A. (1988), ‘‘Consumer Perceptions of Price, Quality, and Value: A Means–End Model and Synthesis of Evidence,’’ *Journal of Marketing,* Vol. 52, July, pp. 2–22.

Zeithaml, V. A., and Bitner, M. J. (2000), *Services Marketing: Integrating Customer Focus Across the Firm,* Irwin/McGraw-Hill, Boston.

Zeithaml, V. A., Parasuraman, A., and Berry, L. L. (1985), ‘‘Problems and Strategies in Services Marketing,’’ *Journal of Marketing,* Vol. 49, Spring, pp. 33–46.

Zeithaml, V. A., Parasuraman, A., and Berry, L. L. (1988), ‘‘Communication and Control Process in the Delivery of Service Quality,’’ *Journal of Marketing,* Vol. 52, April, pp. 35–48.

Zeithaml, V. A., Berry, L. L., and Parasuraman, A. (1993), ‘‘The Nature and Determinants of Customer Expectations of Service,’’ *Journal of the Academy of Marketing Science,* Vol. 21, No. 1, pp. 1–12.

CHAPTER 22

Assessment and Design of Service Systems

MICHAEL HAISCHER
HANS-JÖRG BULLINGER
KLAUS-PETER FÄHNRICH
Fraunhofer Institute of Industrial Engineering

1. INTRODUCTION **634**
- 1.1. Customer Service as a Key Success Factor in Competition 634
- 1.2. The Need for Systematic Engineering of Services 635

2. FUNDAMENTALS OF SERVICE MANAGEMENT **636**
- 2.1. Differences between Services and Material Goods 636
- 2.2. Definitions and Terminology 637
- 2.3. Service Typologies 637

3. MANAGEMENT OF SERVICE QUALITY **638**
- 3.1. Service Quality Models 638
- 3.2. Measuring Service Quality 640
- 3.3. Design of Service Processes 641
- 3.4. Resource-Related Quality Concepts 641

4. THE STRUCTURE OF SERVICE SYSTEMS **642**

5. CONCEPTUAL FRAMEWORK FOR THE ASSESSMENT OF A SERVICE SYSTEM **645**
- 5.1. Quality Assessments 645
- 5.2. Areas for Quality Assessment in Service Organizations 645
- 5.3. A Maturity Model for Quality Management in Service Organizations 648
- 5.4. The Assessment and Design Procedure 648

6. CONCLUSION **649**

REFERENCES **649**

1. INTRODUCTION

1.1. Customer Service as a Key Success Factor in Competition

Services continuously gain relevance in all developed economies (Bullinger 1997). They are increasingly becoming a key issue in the discussion about growth and employment. Many firms now recognize the need, on the one hand, to strengthen ties with existing customers by offering innovative services and, on the other hand, to win over completely new customer groups and markets. Therefore, concepts for effective service management are relevant for both service companies entering into tougher and more global competition and manufacturing companies setting up a service business to support their core products by enhancing customer loyalty. Not only traditional service providers, but also companies for whom in the past services have only represented a small portion of total business, are nowadays required to satisfy ever-more complex needs by offering a whole series of new services to their customers. In particular, many companies are now awakening to the opportunities not merely for safeguarding and expanding their own competitive position with the aid of innovative services,

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

but also for acquiring entirely new business segments. At the same time, they are confronted with the problem of not being able to avail themselves of the systematic procedures and methods that are essential above all to develop and manage complex services. All too often, many successful services are still no more than the ad hoc outcome of individual projects and the personal efforts of employees or management.

Only a minority of firms have appreciated from the outset that crucial competitive advantages will no longer be secured by advanced technology, cost leadership, or product quality alone. On the contrary, innovative, subtly differentiated services are turning into unique selling features that set a company apart from its competitors, representing a promising strategy for success when it comes to tapping new market potentials. In Europe in particular, the discussion about services has been reduced for far too long to the simple formula ‘‘services = outsourcing = job cuts.’’ The opportunities that lie in the exploitation of new business segments and the accompanying creation of new jobs—especially when high-quality services are exported—have not been fully recognized.

The main objective of this chapter is therefore to provide guidelines for systematic design, management, and assessment of service systems with particular attention to the management of service quality. To this end, fundamental approaches for defining *service* and *service quality* are described, leading to a conceptual framework for the structure of service systems. From this an assessment method is derived that supports continuous improvement in service organizations. Additionally, an emerging research area, service engineering, is outlined.

1.2. The Need for Systematic Engineering of Services

While there exists a broad range of methodologies, and tools are available on the development of goods, the development and engineering of services have not been a common topic in the scientific literature. The service discussion has focused primarily on such issues as service marketing, management of service processes, and human resource management in service industries. Service development and design have been largely ignored. Of the small number of authors who have discussed the issue (e.g., Bowers 1986; Easingwood 1986), the majority belong to the service marketing discipline. Approaches based on engineering science are still rare in the service sector to this day. One exception is Ramaswamy (1996). This deficit can be attributed to a single basic cause: the lack of tangibility of services as a research and development goal. That is to say, the development of services is a much more abstract process than the development of material goods or software. So far only a few, very generic attempts have been made to reduce this high level of abstraction and try to capture services operationally as the goal of a systematic development procedure. However, many companies have recently begun (not least as a result of the growing pressures of competition) to rethink their strategies for service provision. They want their services to be ‘‘regular products,’’ that is, reproducible, exportable, even tangible and therefore developable. Services must undergo a systematic design process like any other product.

During the second half of the 1990s, service engineering has emerged as a new research discipline (Fähnrich 1998; Meiren 1999). It is a technical discipline that is concerned with the systematic development and design of service products using suitable methods and procedures. What distinguishes the development of services from the development of conventional, material products is that with services, the interaction between customers and employees plays a crucial role. Thus, several questions require special attention (Hofmann et al. 1998):

- The design of the customer interface and customer interaction
- The design of the service processes
- The selection and training of personnel
- Optimized support for front-office staff (i.e., those employees in direct contact with the customers).

An interdisciplinary approach is essential for solving development tasks of this nature. Service development must therefore integrate knowhow from a variety of scientific disciplines, notably engineering sciences, business studies, industrial engineering, and design of sociotechnical systems. Topics receiving particular attention within the field of service engineering are:

- *Definitions, classifications and standardization of services:* Until now there has been only a diffuse understanding of the topic of service. In particular, international definitions, classifications, and standards, which are necessary in order to develop, bundle, and trade services in the future in the same way as material products, are currently not available. Research questions to be solved include which structural elements of services can be identified, how they should be considered in the development, and how they can be made operational, as well as which methods and instruments could be used for the description and communication of internal and external services.

- *Development of service products:* The development and design of services requires reference models, methods, and tools. Research topics include reference models for different cases (e.g., development of new services, development of hybrid products, bundling of services, reengineering or redesigning of services) and different service types, analysis of the transferability of existing methods (e.g., from classic product development and software engineering) to the development of services, development of new service-specific engineering methods, and development of tools (e.g., ‘‘computer-aided service engineering’’).
- *Coengineering of products and services.* Offers of successful, high-quality products are often accompanied by service activities. Particularly for companies willing to make the move towards an integrated product/service package methods for combined product/service development are presently not available. Concepts for simultaneous engineering of products and services are necessary, for instance.
- *R&D management of services.* The development of services must be integrated into the organizational structure of companies. In practice, there are very few conclusive concepts. The consequences of this include a lack of allocation of resources and responsibilities, abstract development results, and insufficient registration of development costs. Research topics include which parts of a company must be integrated in the service development process; which organizational concepts (e.g., R&D departments for service development, virtual structures) are suitable; how information flow, communication, and knowledge can be managed; how the development process can be controlled; and how a company can continuously bring competitive services to market.

2. FUNDAMENTALS OF SERVICE MANAGEMENT

2.1. Differences between Services and Material Goods

Are services essentially different from material goods? Once service management had been addressed by several scientific disciplines, a considerable amount of research effort was spent on this question. The service marketing literature, in particular, provides an ample variety of theoretical arguments and examples from various industries that underscore the distinction between services and material goods (for an overview, see Fisk et al. 1993). Meanwhile, this distinction becomes more and more blurred. Almost any product can be seen as a combination of material goods and services, and thus the notion of hybrid products (i.e., bundles of material goods and services) has become widely accepted (Stanke and Ganz 1996).

However, three properties of services are assumed to be fundamental:

- *Intangibility:* Unlike material goods, services normally do not exist physically. They consist of concepts and activities fulfilling a value proposition given to the customer.
- *Simultaneity of production and consumption:* Generally speaking, services are delivered at the time the customer needs them and to the location where they are needed. Therefore they can hardly be stored and shipped. This property is usually labeled the ‘‘uno-actu-principle.’’
- *Customer integration:* In most cases, the customer is integrated in a service delivery process, either personally as an object of the service or by providing his or her property or information as input.

Counterexamples can be easily found at least for the first two properties, intangibility and uno-actu-principle. The last distinction, the involvement of the customer in the service delivery process, appears to be more fundamental. Some authors even label it the only valid difference between services and goods: ‘‘With services, the customer provides significant inputs into the production process. With manufacturing, groups of customers may contribute ideas to the design of the product, however, individual customers’ only part in the actual process is to select and consume the output’’ (Sampson 1999).

Although the relevance of each of these basic properties varies considerably between different types of services, they indicate some basic challenges that must be met in managing almost any service:

- *Definition and description of an intangible product:* Because product properties are mostly immaterial, describing and—even more important—demonstrating them to the customer is much more difficult than with material goods. In most cases, the customer does not have an opportunity to look at or to test the service prior to purchase, as can be done with a car, for example. Therefore, buying a service creates a certain extent of risk for the customer. The challenge for a marketing department consists at this point of communicating the service’s characteristics to the customer and reducing this feeling of risk as much as possible. For a development or an operations department the challenge lies in defining and describing the service product in a way that ensures its delivery without failure.

- *Managing resources* in terms of quantity and quality: Most services cannot be produced "on stock," which requires that the service provider keep capacity available, whether the service is purchased or not. For this reason, concepts such as yield management are extremely important in numerous service industries, such as the airline industry.
- *Managing the customer* as a part of the service that can be controlled in an indirect manner at most. In many cases, proper service delivery depends on the customer behaving and cooperating in a certain way. Therefore, a service provider has to find ways to ensure that the customer assumes his or her role in the process.

2.2. Definitions and Terminology

According to ISO 9004/2, services are "the results generated by activities at the interface between the supplier and the customer and by supplier internal activities to meet customer needs." The following notes are added to this definition:

- "The supplier or the customer may be represented at the interface by personnel or equipment."
- "Customer activities at the interface with the supplier may be essential to the service delivery."
- "Delivery or use of tangible product may form part of the service delivery."
- "A service may be linked with the manufacture and supply of tangible product."

The standard provides a rather general definition that contains various essential aspects. For purposes of service management, the definition needs to be refined. This can be done by analyzing the most common definitions of the services concept, the most widely accepted of which was originally established by Donabedian (1980). It states that a service bears three distinct dimensions:

1. A structure dimension (the structure or potential determines the ability and willingness to deliver the service in question)
2. A process dimension (the service is performed on or with the external factors integrated in the processes)
3. An outcome dimension (the outcome of the service has certain material and immaterial consequences for the external factors)

Combining the aspects that are considered in the standard with Donabedian's dimensions of service yields Figure 1.

2.3. Service Typologies

Most approaches to defining services aim at giving an general explanation of service characteristics that applies to all types of services. While this approach is valuable from a theoretical point of view, due to the broad variety of different services, it hardly provides very detailed guidelines for the design and management of a specific service. Classification schemes and typologies that classify services and provide distinctive classes calling for similar management tools are needed. Classifying services with respect to the industry that the company belongs to hardly provides further insight, since a particular service may be offered by companies from very different industries. Industries in the service sector are merging, generating competition among companies that were acting in separate marketplaces before (e.g., in the media, telecommunications, or banking industries). On the other hand, services within one industry may require completely different management approaches. Thus, services have to be classified and clustered according to criteria that relate to the particular service product rather than to the industry. Such typologies can give deeper insight into the special characteristics of a service as well as implications concerning its management.

Several attempts have been made in order to design such typologies. Some of them are based on empirical studies. The study presented in Eversheim et al. (1993) evaluated questionnaires from 249 German companies representing different service industries. By the application of a clustering algorithm, seven types of services were finally identified, based on 10 criteria.

A recent study on service engineering (Fähnrich and Meiren 1999) used 282 answered questionnaires for the determination of service types. It derived four separate types of services that differ significantly in two ways:

1. The intensity of interaction between service provider and customer
2. The number of different variants of the service that the customer may obtain

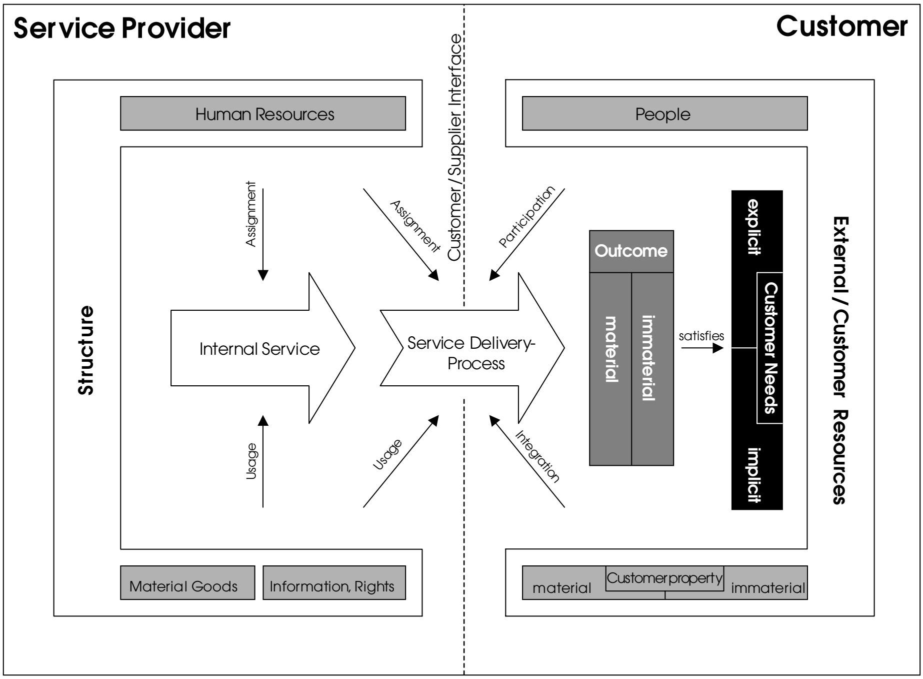

Service Provider
Customer
Customer / Supplier Interface
Human Resources
People
Structure
Assignment
Assignment
Participation
Internal Service
Service Delivery-Process
Usage
Usage
Integration
Outcome
material
immaterial
satisfies
explicit
Customer Needs
implicit
External / Customer Resources
Material Goods
Information, Rights
material
Customer property
immaterial

Figure 1 Elements of Service Definitions.

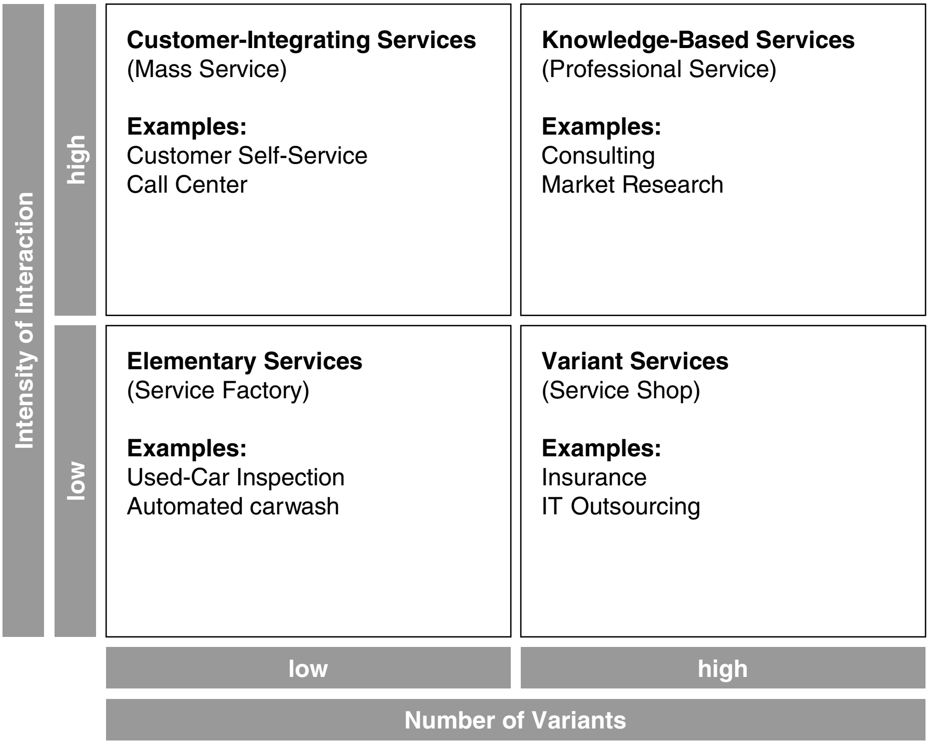
Strategies for service development and service operations can be assigned for the four types of services. Interestingly, these results are very similar to a typology of service operations introduced by Schmenner (1995), identifying four different types of service processes: service factory, service shop, mass service, and professional services. Figure 2 gives an overview of the typology, relating it to Schmenner's types.

3. MANAGEMENT OF SERVICE QUALITY

3.1. Service Quality Models

ISO 8402 defines quality as "the totality of characteristics of an entity that bear on its ability to satisfy stated or implied needs." The standard explicitly applies to services, too, but it leaves space for interpretation. Therefore, some different notions of quality have emerged, depending on the scientific discipline or practical objective under consideration. Those notions can be grouped into five major approaches (Garvin 1984) and can be applied to goods as well as to services:

- The *transcendent approach* reflects a "common sense" of quality, defining it as something that is "both absolute and universally recognizable, a mark of uncompromising standards and high achievement" (Garvin 1984). Therefore, quality cannot be defined exactly, but will be recognized if one experiences it.
- The *product-based approach* links the quality of an item to well-defined, measurable properties or attributes. Therefore, it can be assessed and compared objectively by comparing the values of those attributes.
- The *user-based approach* focuses exclusively on the customer's expectations and therefore defines quality as the degree to which a product or service satisfies the expectations and needs of an individual or a group.
- The *manufacturing-based approach* derives quality from the engineering and manufacturing processes that deliver the product. Quality equals the degree to which a product meets its specifications.



		low	high
Intensity of Interaction	high	**Customer-Integrating Services** (Mass Service) **Examples:** Customer Self-Service Call Center	**Knowledge-Based Services** (Professional Service) **Examples:** Consulting Market Research
	low	**Elementary Services** (Service Factory) **Examples:** Used-Car Inspection Automated carwash	**Variant Services** (Service Shop) **Examples:** Insurance IT Outsourcing
		Number of Variants	

Figure 2 Four Basic Types of Service Products. (From Fähnrich and Meiren 1999)

- The *value-based approach* relates the performance of a product or service to its price or the cost associated with its production. According to this notion, those products that offer a certain performance at a reasonable price are products with high quality.

While the relevance of the transcendent approach is only theoretical or philosophical, the other approaches do influence quality management in most companies. Normally, they exist simultaneously in one organization: while the marketing department applies a user-based approach to quality, the manufacturing and engineering departments think of quality in a product- or manufacturing-based manner. The coexistence of the two views carries some risk because it might lead to diverging efforts in assuring product quality. On the other hand, today's competitive environment requires a combination of the different definitions (Garvin 1984) because each of them addresses an important phase in the development, production, and sale of a product. First, expectations and needs of the targeted customers have to be analyzed through market research, which requires a customer-based approach to quality. From those requirements, product features have to be derived such that the product meets the customers' expectations (product-based approach). Then the production process must be set up and carried out in a way that ensures that the resulting product in fact bears the desired characteristics. This implies that the production process should be controlled with respect to quality goals and therefore a manufacturing-based approach. Finally, the product is sold to its customers, who assess it against their expectations and express their satisfaction or dissatisfaction. The product specifications will eventually be modified, which again calls for a customer-based approach.

Because most definitions of service quality are derived from marketing problems, they reflect more or less the customer-based approach or value-based approach. There are several reasons for this one-sided view of service quality. In general, services do not bear attributes that can be measured physically. Furthermore, the customer is involved in the service-delivery process, so process standards are difficult to define. Additionally, the understanding of quality has moved from the product- and process-based view towards the customer- and value-based view in almost all industries, equaling "quality" and "customer satisfaction." Therefore, the customer- and value-based approach has been used predominantly in recent research on service quality.

The most fundamental model for service quality that influences many other approaches is derived from Donabedian's dimensions, namely structure, process, and outcome (Donabedian 1980). Quality of potentials includes material resources and people, for example, the capability of customer agents. Process quality includes subjective experiences of customers (e.g., the friendliness of employees) during the service-delivery process, as well as criteria that can be measured exactly (e.g., the time needed for answering a phone call). In comparison to manufacturing, process quality has an even greater impact for services. A service-delivery process that is designed and performed extremely well ensures, as in manufacturing, the quality of its final outcome. Additionally, to the customer, the process is part of the service because the customer may observe it or even participate in it. The quality of the outcome of a service is the third component of service quality. It reflects the degree to which the service solves the customer's problems and therefore satisfies his or her needs and expectations.

Many models for service quality are derived from the concept of customer satisfaction. According to this notion, service quality results from the difference between the customers' expectations and their experiences with the actual performance of the service provider. If the expectations are met or even surpassed, the quality perceived by customers is good or very good, otherwise the customers will remain dissatisfied and rate the service quality as poor.

Based on this general assumption, several approaches to service quality aim at explaining the reasons for customer satisfaction (and therefore for service quality). This provides a basis for measuring the outcome quality of a service. Unfortunately, there are hardly any concepts that link the measurement of customer satisfaction or service quality to methods that influence it during service design or delivery. A first step towards this connection is the gap model, developed by Parasuraman et al. (1985) based on results from empirical research. The gap model identifies five organizational gaps within the process of service design and delivery that cause deficits in quality, leading to dissatisfied customers. The gaps occur in the following phases of the process:

- Gap 1 results from the fact that the management of the service provider fails to understand the customers' expectations correctly.
- Gap 2 denotes the discrepancy between management's conceptions of customers' expectations and the service specifications that are derived from it.
- Gap 3 is caused by a discrepancy between service specifications and the performance that is delivered by the service provider.
- Gap 4 consists of the discrepancy between the delivered perfomance and the performance that is communicated to the customer.
- Gap 5 is the sum of gap 1 through 4. It describes the discrepancy between the service the customer expected and the service that he or she actually experienced.

According to Parasuraman et al., customer satisfaction (i.e., gap 5) can be expressed in terms of five dimensions: tangibles, reliability, responsiveness, assurance, and empathy (Parasuraman et al. 1988). These dimensions are the basis for the SERVQUAL method, which is an well-known method for measuring the quality of services through assessing customer satisfaction. The following section addresses different approaches for measuring service quality.

3.2. Measuring Service Quality

Methods for the measurement of service quality are used during and after the delivery of a service in order to assess the customer's satisfaction and quality perception. They can be compared to quality tests in manufacturing, but there are some significant differences: service quality cannot be assessed by measuring physical attributes, and therefore some immaterial criteria that are related to quality have to be defined. Furthermore, if defects are detected, they generally cannot be "repaired" as can be done in case of material goods, since the customer normally participates in the service delivery process and therefore witnesses defects as they occur. Therefore, the main objective of measuring service quality is improving it on a mid- or long-term base rather than detecting and repairing defective units.

Depending on the chosen approach (see Section 3.1) service quality can be measured from the customer's and the service provider's point of view. Measuring from the service provider's point of view involves gathering data that are internally available, such as performance measures or quality cost (Eversheim 1997). They can be analyzed using well-known methods from quality management in manufacturing processes, such as statistical process control (Gogoll 1996). In addition, service quality can be assessed indirectly by an overall analysis of the quality system, which is done by a quality audit or quality assessment. Those provide information regarding the capability of a service provider to deliver quality services rather than information about the quality of a specific service. Therefore, they are based on the assumption that there is a correlation between the quality of the structure and processes, on the one hand, and the outcome, on the other hand.

There are two basic approaches for measuring service quality from the customer's point of view. Both of them are based on assessing customer satisfaction.

Multiattribute methods are based on the assumption that the quality perception of customers is determined by assessing distinctive attributes of the particular service. With regard to each attribute, the customer compares the expected and the received quality. The overall judgment then results from a weighted addition of those comparisons. The most prominent example of this type of method is SERVQUAL (Parasuraman et al. 1988), a multiple-item scale consisting of 22 items grouped into five dimensions.

Another group of methods uses the assessment of *service encounters* or *moments of truth* (i.e., the contact between customers and the service provider) for measuring service quality. An example of those methods is the critical incident technique (Bitner et al. 1990). This method uses structured interviews to gather information about customers' experiences that have roused either very negative or very positive emotions. From those interviews, the most relevant problem areas are determined. This type of method allows the customer to describe the service encounter from his or her point of view instead of assessing it by predefined criteria. The method normally leads to the most significant causes of service failures.

Besides those activities that are initiated by the service provider, systematic analysis of customer complaints yields valuable information about quality problems. Typically, only a very small percentage of dissatisfied customers do complain, while most of them just buy from a competitor without further notice (Heskett et al. 1997). Effective complaint management therefore requires that the customer be encouraged to complain and that communication between customer and service provider be facilitated as much as possible. Customers whose complaints are treated in a satisfying manner often reward the service provider with increased loyalty (Reichheld 1997). Therefore, defective services can be "repaired" to a certain degree, which is normally described by the concept of service recovery (Heskett 1997)

3.3. Design of Service Processes

In addition to customer satisfaction, the process dimension of service quality has been addressed by numerous researchers and practitioners.

An important tool for service process design and management is the service blueprinting method, originally described in Shostack (1984). Service blueprinting subdivides each process into process steps, which are perceived by the customer directly, and supporting activities, which are necessary to deliver the service but are only perceived indirectly. The two areas are separated by a "line of visibility." This differentiation is extremely important for customer-oriented service development processes because it permits concentration on the design of the customer interface at a very early stage. Service blueprinting was subsequently modified and named "service mapping" (Kingman-Brundage 1995). This method most notably includes two new lines of interaction. The first of these, the line of external interaction, subdivides the process as perceived by the customer into activities that are performed by the customer personally and activities in which the customer participates but which are in fact performed by the employees of the company offering the service. The second line, the line of internal interaction, differentiates between processes delivered by the supporting "back office" and those provided either by other companies or by other departments in the same company. Service blueprinting and service mapping are especially useful for service planning because they help identify potential errors very early on. Furthermore, they can be used as training material for employees and customers. Figure 3 shows an example of a service blueprint.

3.4. Resource-Related Quality Concepts

Approaches to service quality that apply to the *structure* dimension can be subsumed under the heading "resource concepts." These include, most importantly, human resources concepts (especially qualification concepts), as well as the infrastructure necessary to deliver the service and service-support tools in the form of suitable information and communication technologies. The analysis of customer–employee interaction is crucial to enable appropriate recruiting and qualification of the employees who are to deliver the service.

Two questions arise immediately in regard to human resource management in services.

- There is an ongoing debate on whether the knowhow required for running a service system should be put into people or into processes. The job enlargement approach requires significant effort in recruiting, training, and retaining the kind of employees who are able to ensure high

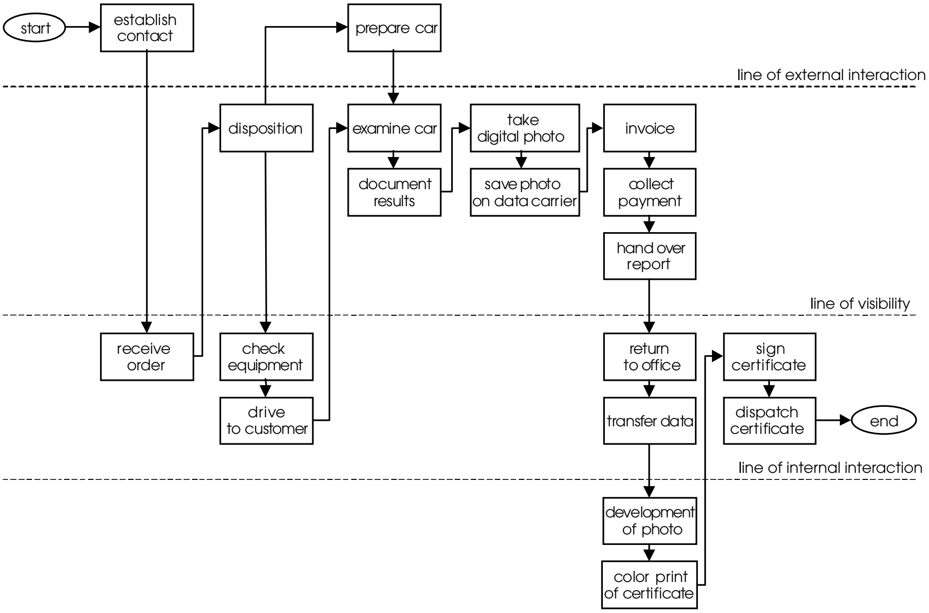

start
establish contact
prepare car
line of external interaction
disposition
examine car
take digital photo
invoice
document results
save photo on data carrier
collect payment
hand over report
line of visibility
receive order
check equipment
drive to customer
return to office
transfer data
sign certificate
dispatch certificate
end
line of internal interaction
development of photo
color print of certificate

Figure 3 Service Blueprint of a Car Certification Service. (From Meiren 1999)

quality through personal effort and skills. The production line approach tries to ensure service quality by elaborated process design, therefore calling for lower requirements from the workforce involved.

- The management of capacity is crucial in services because they cannot be produced ‘‘on stock.’’ This calls for concepts that enable a service provider to assign the employees in a flexible way.

Role concepts are one way to deploy human resources during the service-delivery phase (Frings and Weisbecker 1998; Hofmann et al. 1998) that addresses both questions described above. ‘‘Roles’’ are defined groups of activities within the framework of a role concept. The roles are then linked to employees and customers. It is possible for several employees or several customers to perform one and the same role and for one employee or one customer to perform several different roles. Role concepts are a particularly useful tool for simplifying personnel planning, for instance in connection with the selection and qualification of the employees who will later be required to deliver the service. Most importantly, potential bottlenecks in the service-delivery phase can be identified extremely early on and suitable action taken to avoid them. Moreover, role concepts provide a starting point for formulating tasks in the area of work planning and enable customer behavior to be analyzed and planned prior to the service-delivery process. Finally, the use of roles does not imply any direct relationship to fixed posts or organizational units, and the concept is thus extraordinarily flexible.

4. THE STRUCTURE OF SERVICE SYSTEMS

Most service-management concepts discussed so far clearly show an affiliation to either the production line view or the human resource or customer and marketing-oriented view on services. These different views have not been integrated much, mainly because very different scientific disciplines are involved. Obviously, both approaches are needed. Nowadays service processes are highly complex and call for the application of sophisticated methods from industrial engineering, while on the other hand the outcome depends heavily on the performance of employees and customers.

One way to handle this complexity is to apply a system view to services, as proposed by several authors (e.g., Lovelock 1995; Kingman-Brundage 1995). Systems are made up of elements bearing attributes and the relationships between those elements. Elements may be (sub)systems themselves. If the elements of a service system are defined as those different components that have to be designed and managed, then the most elementary system model for services is made up of three basic elements:

1. *Customer:* The customer normally initiates a service delivery process and performs certain tasks within it. The customer partly reflects the outcome dimension of service quality by being either satisfied or dissatisfied with the service.
2. *Resources:* Human and material resources perform defined functions within a service process and interact with the customer. They reflect the structure dimension of service quality.
3. *Process:* Processes can be viewed as the definition of interactions between the resources and the customers in a service system. However, there are reasons for defining them as the third basic element within the system. Assuming the production line approach, processes are objects that carry information regarding the service, especially in IT-based services. They exchange information with other elements in the system or even control them to a certain degree.

For several reasons, it seems useful to define a fourth basic element within the system, namely service products. So far, the notion of a service product has not been used in a clearly defined way. However, in recent years the term *product* has been used more frequently to express the idea that services can be developed and produced in a defined and repeatable way. Defining the concept of products within a service system therefore yields several benefits:

- So far, the outcome dimension of a service has been considered mainly from the customer's point of view, taking outcome quality as equivalent to customer satisfaction. However, aiming only at customer satisfaction means that service quality easily becomes a moving target. This can be avoided by integrating all relevant information regarding the outcome of a service into a product model that is communicated internally and (to an appropriate extent) externally.
- The definition of a product clearly separates the outcome of a service from the way it is delivered. The product therefore acts as an interface between the market and customer view on a service (''what is delivered'') and a production or process view (''How it is delivered'').
- Product models support the modular set-up of services. Services that are offered with a wide range of variants are difficult to handle, particularly if the service depends heavily on information technology. A potential solution is the division of a service in separate modules that can be bundled according to the customer's needs. This also greatly supports the development of new products or variants because they (and the corresponding software systems) only have to be assembled from existing parts.

A product model for services cannot simply consist of a one-to-one application of traditional product concepts from manufacturing. Instead, the product model must be derived from the essential characteristics of service processes. This can be achieved by an analysis of generic models of service delivery (e.g., the universal service map [Kingman-Brundage 1995] or the process model [Eversheim et al. 1997]) that yields those points within a service process that require information from a product model. Thus, a typical service process can be described in the following way:

1. A service is initiated by the definition of a service concept (Heskett et al. 1997) that is communicated to the market and its targeted customers. The required resources are assigned in parallel.
2. Potential customers decide to contact the service provider in order to obtain additional information or to order the service instantaneously. For this purpose, the ways in which customers gain access to the service have to be defined and communicated. Traditionally, customers enter the service facility in person (e.g. a restaurant, a car repair shop). Depending on the particular service concept, this may be replaced by telephone or Internet access or a combination of those. In any case, an access system has to be defined that is located directly at the line of interaction between customer and service provider.
3. If the customer decides to order the service, some specifications regarding the desired variant have to be recorded by the access system.
4. The service delivery process must then be activated, using the specification that has been collected by the access system. First, however, a consistency check must be performed in order to verify whether the desired variant of the service is obtainable.
5. If the variant is valid, the configuration of the service will be set up, that is, those delivery processes that are required for the particular variant are activated.
6. If a process includes participation of the customer, the customer must be informed about the expected outcome and form of the cooperation. Results produced by the customer must be fed back into the delivery process.

7. As soon as all threads of the delivery process (carried out by the service provider or the customer) have been terminated, the outcome will be delivered and an invoice sent to the customer.
8. If the results are satisfactory, the contact between customer and service provider comes to an end and the overall process is finished. The customer may order some additional service or may return later to issue a new order. If the results leave the customer dissatisfied, a service-recovery procedure may be initiated or the customer may simply terminate the business relation.

The information needed to control such a process is normally distributed between the people involved (employees and customers) and the documents that define the processes. However, the reasoning outlined above shows that it is useful to combine some elements within a product model. These considerations lead to the product model depicted in Figure 4, which uses an object-oriented notation according to the Unified Modeling Language (UML).

Using the object-oriented notion of a system, the individual elements of the product model are displayed as classes that have a unique name (e.g., ‘‘Service_Product’’), bear attributes (e.g., ‘‘External_Product_Information’’), and use methods for interaction with other classes.

The product model is centered around the class Service_Product. This class identifies the product, carries information for internal use, and provides all information regarding valid configurations of the product. Service_Product may be composed of other, less complex products, or it may be an elementary product. It relates to a class Service_Concept that defines the customer value provided by the product, the customers and market it aims at, and the positioning in comparison to its competitors. The class Service_Product is activated by whatever provides the customer's access to the service system. Access_Module carries information on the product for external distribution and records the specification of the product variant as requested by the particular customer. Access_Module eventually activates Service_Product, which subsequently checks the consistency and validity of the requested product configuration and sets up the configuration by ‘‘assembling’’ individual modules of the service. Then Deliver_Product is carried out by activating one or several Service_Functions that make up the particular product configuration. During and after the service delivery process Service_Product may return statistics to the management system, such as cost and performance data. Also, Product_Features may be attached to a Service_Product, potentially including Material_Components.

Each Service_Product contains one or several Service_Functions that represent the Outcome of the service. A Service_Function carries information about the expected Quality Level (which may

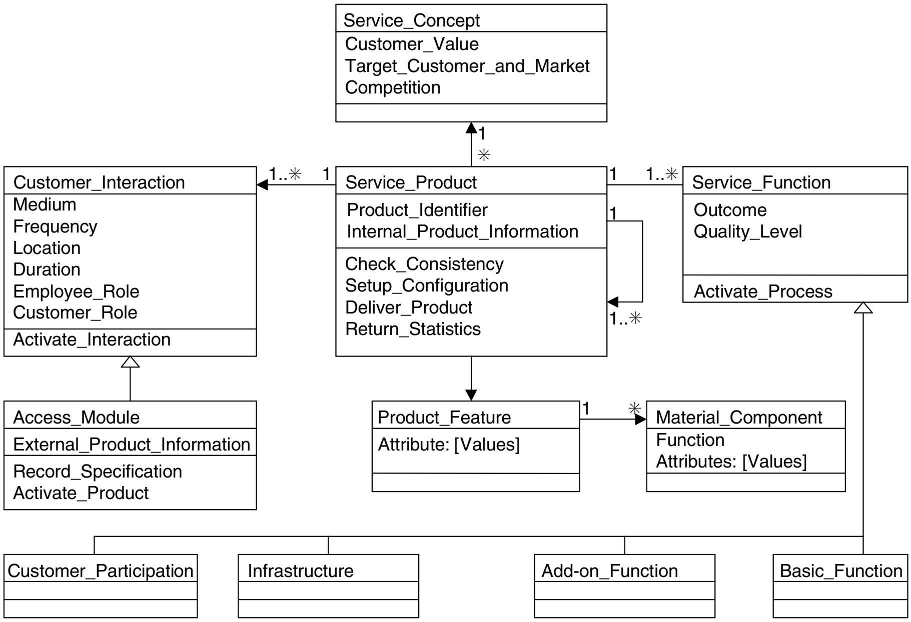

Service_Concept
Customer_Value
Target_Customer_and_Market
Competition
1
*
Customer_Interaction
Medium
Frequency
Location
Duration
Employee_Role
Customer_Role
Activate_Interaction
1..*
1
Service_Product
Product_Identifier
Internal_Product_Information
Check_Consistency
Setup_Configuration
Deliver_Product
Return_Statistics
1
1..*
1
1..*
Service_Function
Outcome
Quality_Level
Activate_Process
Access_Module
External_Product_Information
Record_Specification
Activate_Product
Product_Feature
Attribute: [Values]
1
*
Material_Component
Function
Attributes: [Values]
Customer_Participation
Infrastructure
Add-on_Function
Basic_Function

Figure 4 Service Product Model.

be expressed by criteria like the ones used in SERVQUAL and a relation to the internal or external service delivery processes that realize this function. Jaschinski (1998) names six different types of service functions, four of which can be expressed within the product model:

1. A Basic_Function contains the basic outcome a customer expects from the service, such as the transportation from location A to location B in the case of an airline.
2. An Add-on_Function realizes some additional features that enhance the service compared to its competitors. Staying with the airline example, the large variety of meals or the lounges in the airport are such Add-on_Functions.
3. Infrastructure is needed in many cases in order to enable the customer to use the service at all. An example is the check-in counters in the airport that provide the necessary preparation for boarding an airplane.
4. Customer_Participation denotes the part of the service's outcome that is produced by the customer. An example is the automated self-check-in that is used by several airlines.

According to Jaschinski (1998), interaction between customers and service provider and access to the service system represent two more functions. Within this product model, interaction is defined by the element Customer_Interaction, of which Access_Module must be regarded as a special instance.

The product model therefore acts as a central hub of a service system, coordinating the interactions of customers, employees, material resources, and processes. The product is activated by a customer, then initiates the required processes, which in turn call for necessary resources. This model explicitly shows the correlation between the elements of a service system and may serve as a template for designing a service from a product designer's point of view.

Based on this concept, a framework for the assessment of a service system will be introduced in the next section that serves as a methodology for continuous improvement.

5. CONCEPTUAL FRAMEWORK FOR THE ASSESSMENT OF A SERVICE SYSTEM

5.1. Quality Assessments

Usually the performance of a company is assessed by financial criteria. This may seem perfectly appropriate because the survival of a company depends on adequate profits. All other objectives (e.g., the quality of products and services) are pursued not for their own sake but rather in support of financial long-term success. Within the last decade, however, researchers and practitioners have come to the conclusion that because financial criteria relate to the past, they reflect only part of a company's performance. The future performance and growth of a company depend equally on soft factors that are difficult to quantify, such as the quality of leadership or the qualification of the employees. The idea of integrating soft and hard factors is realized in the balanced scorecard concept (Kaplan and Norton 1996), which is gaining increasing acceptance worldwide.

Assessment is another management concept that aims at integrating *all* relevant factors. Especially in the field of quality management, assessments have become very popular in recent years. There are numerous examples of successful application. Here the term *assessment* denotes several different but similar methods that serve mainly as a tool for analyzing and documenting the actual status of an organization. An assessment can also be used as the basis for a process of continuous improvement. This leads to three main objectives of an assessment:

1. Analysis of the status quo and determination of potential improvements
2. Support and acceleration of change processes
3. Measurement of whether objectives have been achieved

An assessment can be performed in different ways, depending on the particular objectives and the effort that is spent. Table 1 gives an overview of the most relevant variants.

Most companies use an established assessment model when they carry out a quality assessment. Well-known examples are the Malcolm Baldrige National Quality Award (MBNQA) in the United States and the European Quality Award in Europe. Several derivations of these models exist in different countries. All assessment models define criteria that address the various fields of improvement in an organization. Those criteria are grouped in categories and a procedure used both for gathering and evaluating information and for the calculation of the score is defined.

5.2. Areas for Quality Assessment in Service Organizations

While MBNQA and EQA address arbitrary organizations, this section presents ServAs (service assessment), an assessment model that has been specifically developed for service organizations. ServAs

TABLE 1 Various Types of Quality Assessment

Type	Objectives	Effort	Participants	Questionnaire	Shortcomings
Quick check	Getting started fast, overview of status quo	○	1 person, e.g., the quality manager	Complete questionnaire	The result shows the view of just one person
Assessment workshop	Awareness building for the management team, action plan	◑	Ca. 5–7 persons	Complete questionnaire	Only the view of management is integrated
Survey	Appraisal of practices or processes by all relevant employees	◕	All employees of the unit	Select and customized items	Broad understanding of the method and its topics has to be disseminated
Documented in-depth analysis	Profound analysis of all relevant processes and functions	●	Team of specialists	Specific questionnaire	Very high effort

contains 12 categories that describe those fields within a service organization that require attention from the management. Each category contains relevant criteria that describe the actual status of the organization with respect to the particular management field. The criteria are grouped within key areas. They have been derived from the relevant characteristics of a service system, combined with applicable elements from established assessment models (see Haischer 1996; Eversheim 1997). Figure 5 gives an overview of the categories of ServAs and the criteria.

According to the fundamental dimensions of service quality, the categories have been assigned to either the structure, the processes, or the outcome of a service. Five categories belong to the structure dimension:

1. *Customer focus:* How does the service organization ensure that the offered services meet the expectations and needs of customers?
2. *Leadership:* How is a service culture being established and maintained by senior management?
3. *Employees:* How are employees recruited, qualified and motivated in accordance with the company's overall objectives?
4. *Resources:* Is the usage of material resources effective and efficient?
5. *Quality system:* Does the service organization maintain a quality system that transfers quality objectives to products and processes?

Four categories assess process quality in a service organization:

1. *Service development:* Are service products systematically developed within a defined development process?
2. *Service delivery:* How does the service organization organize the core process of service delivery in an effective and efficient way?
3. *Cooperation management:* How does the organization build, use, and maintain partnerships along the value chain?
4. *Customer relationship and communication:* How does the service organization manage communication with customers and other external partners?

The three last categories address the fundamental dimensions of outcome any company strives for:

1. *Employee satisfaction:* How does the service organization measure and manage employee satisfaction?

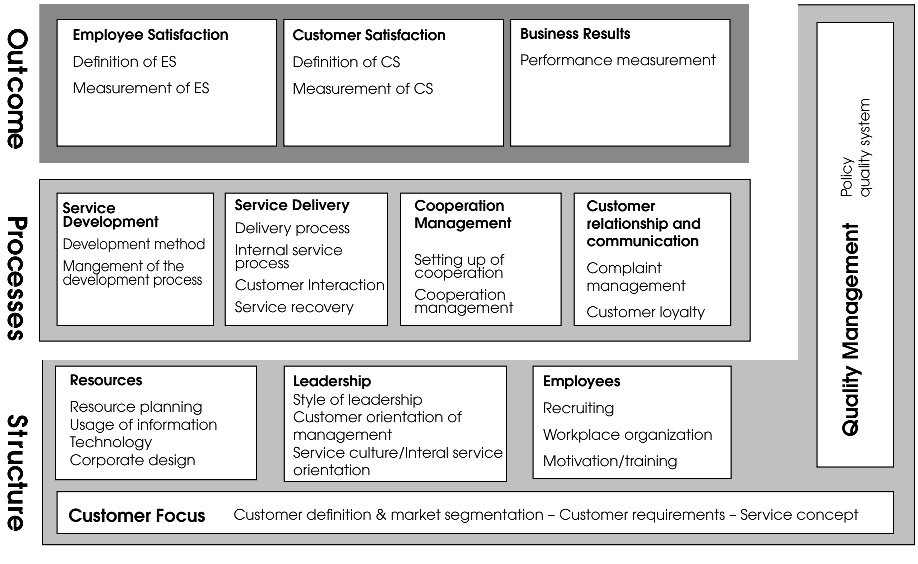

Outcome
Employee Satisfaction
Definition of ES
Measurement of ES
Customer Satisfaction
Definition of CS
Measurement of CS
Business Results
Performance measurement
Processes
Service Development
Development method
Mangement of the development process
Service Delivery
Delivery process
Internal service process
Customer Interaction
Service recovery
Cooperation Management
Setting up of cooperation
Cooperation management
Customer relationship and communication
Complaint management
Customer loyalty
Structure
Resources
Resource planning
Usage of information
Technology
Corporate design
Leadership
Style of leadership
Customer orientation of management
Service culture/Interal service orientation
Employees
Recruiting
Workplace organization
Motivation/training
Customer Focus
Customer definition & market segmentation – Customer requirements – Service concept
Quality Management
Policy
quality system

Figure 5 Categories and Key Areas for the Assessment of Service Systems.

2. *Customer satisfaction:* How does the service organization define and measure customer satisfaction?
3. *Business results:* How does the service organization establish performance measurements using financial and nonfinancial criteria?

A questionnaire has been derived from these categories and the related key areas. It is the most important tool in the assessment method. Each key area contains several criteria related to one or more items of the questionnaire. These items indicate the "maturity level" (see Section 5.3) of the service organization with respect to the criterion under consideration.

The questionnaire may be utilized for analyses at different depths, from an initial quick self-assessment by senior management to an in-depth discussion of the different key areas in workshops and quality circles. The results serve as a basis for planning improvement measures.

5.3. A Maturity Model for Quality Management in Service Organizations

The overall result of an assessment using ServAs is expressed in terms of the maturity level of the service organization. In recent years, models for maturity levels have been recognized as an appropriate base for improvement processes. They are based on the assumption that the development of an organization follows some distinctive steps or maturity levels. Several different maturity level models have recently been developed for particular industries (e.g., Humphrey 1989; Rommel 1995; Malorny 1996). They have been integrated into the development of ServAs.

Using a number between 1 and 5, the maturity levels indicate how far the service organization has proceeded towards total quality management (TQM). Each level describes a typical state that an organization assumes in developing quality management. Thus, the maturity level can be easily interpreted by management and employees. It can be used internally for motivating and communicating further efforts in an improvement program because it provides clear objectives: "We have reached level 3—within a year we want to get to level 4!"

Table 2 displays the five maturity levels of ServAs with their underlying principles, their most important characteristics, and the key tasks at each level.

After the questionnaire is filled in, the overall maturity level of an organization is calculated using a two-step procedure. First, the maturity level has to be calculated for each individual category. For each category and each maturity level a certain threshold has to be passed (i.e., a given number of positive answers in the questionnaire must be met or exceeded) for the maturity level to be reached. Finally, the overall maturity level is calculated from the category levels according to several scoring rules.

5.4. The Assessment and Design Procedure

Usually, the conduction of a ServAs assessment is done within the organizational frame of a project. Figure 6 shows an overview of the main phases and activities of such a project.

TABLE 2 Maturity Levels of Service Organizations

Level	Principle	Characteristics	Key Tasks
1	Ad hoc management	Service quality is attained by coincidence or by repairing mistakes.	Quality and service policy, dissemination of a common understanding of quality
2	Repeatable	Given identical circumstances, a certain quality level can be reached repeatedly.	Structured documentation of processes and products
3	Process definition	ISO 9000 type of quality management.	Ensuring effectiveness and efficiency of processes by use of performance measures
4	Quantitative management	Feedback loops and performance measures are established.	Refining of feedback loops, ensuring the participation of each employee
5	Continuous improvement	All members of the company are involved in improvement actions.	Continuous review of measures and feedback loops

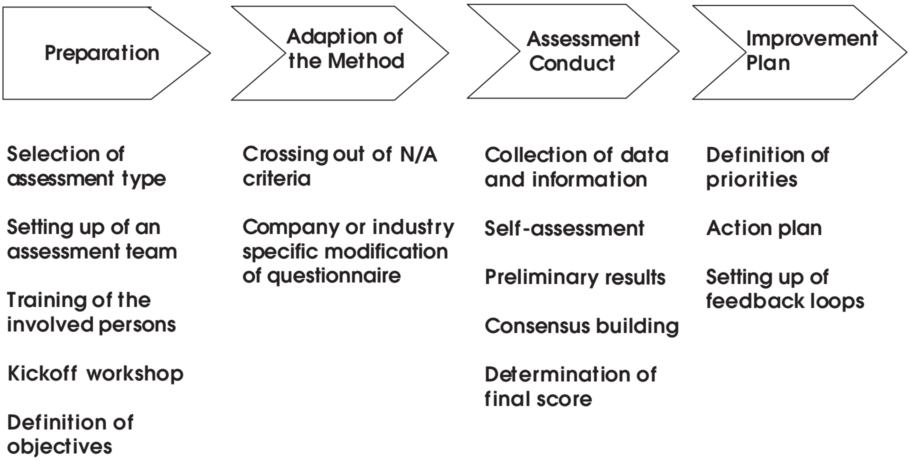

Preparation
Adaption of the Method
Assessment Conduct
Improvement Plan
Selection of assessment type
Setting up of an assessment team
Training of the involved persons
Kickoff workshop
Definition of objectives
Crossing out of N/A criteria
Company or industry specific modification of questionnaire
Collection of data and information
Self-assessment
Preliminary results
Consensus building
Determination of final score
Definition of priorities
Action plan
Setting up of feedback loops

Figure 6 Procedure for Conducting an Assessment.

After the appropriate type of assessment is selected and the project team is set up and trained, the project should be initiated with a kickoff workshop that helps to communicate the project's background and objectives to all employees. During this workshop the most important and urgent problems can be gathered in order to fine-tune the project's objectives. In a second step the questionnaire has to be adjusted to company- and industry-specific requirements, which is done by eliminating criteria that do not apply or by adding company- and industry-specific criteria. Then information concerning the considered criteria is gathered within the organization, depending on the chosen mode of assessment. This information yields a preliminary result that has to be consolidated by eliminating contradictory statements, using different consensus-building methods. After the final score is determined, an action plan for improvement measures has to be decided on and its execution measured on a regular basis.

6. CONCLUSION

Managing customer service and service quality is a complex task that requires a holistic approach that takes into account people (customers and employees), material resources, and abstract entities such as products and processes. A system-oriented approach is useful for mastering this complexity because it supports the understanding of the manifold dependencies within a service organization.

The concept of a service product model that has been presented in this chapter helps to organize the variety of information that a service provider has to handle. It facilitates modularization of services and therefore the development of product bundles, variants, and entirely new products in service organizations. Defining service products enables service providers to distinguish the outcome of a service clearly from the processes and resources that lead to it, opening a variety of strategic options. However, one must keep in mind that service delivery can be never automated completely. The right balance between skilled people and intelligent products and processes is still required. The ServAs method helps to achieve this balance by supporting continuous improvement of the entire service system.

REFERENCES

Bitner, M. J., Booms, B. H., and Tetreault, M. S. (1990), "The Service Encounter: Diagnosing Favorable and Unfavorable Incidents," *Journal of Marketing,* Vol. 54, January, pp. 71–84.

Bowers, M. R. (1986), "New Product Development in Service Industries," Texas A&M University, College Station, TX.

Bullinger, H.-J. (1997), "Dienstleistungen für das 21: Jahrhundert—Trends, Visionen und Perspektiven," in Dienstleistungen für das 21. Jahrhundert, Bullinger, H.-J., Ed., Schäffer-Poeschel, Stuttgart.

Donabedian, A. (1980), *The Definition of Quality and Approaches to Its Assessment,* Vol. 1 of *Explorations in Quality Assessment and Monitoring,* Health Administration Press, Ann Arbor, MI.

Easingwood, C. (1986), "New Product Development for Service Companies," *Journal of Product Innovation Management,* Vol. 3, No. 4, pp. 264–275.

Eversheim, W. (1997), Qualitätsmanagement für Dienstleister: Grundlagen—Selbstanalyse—Umsetzungshilfen, Springer, Berlin.

Eversheim, W., Jaschinski, C., and Roy, K.-P. (1993), *Typologie Dienstleistungen,* Forschungsinstitut für Rationalisierung, Aachen.

Fähnrich, K.-P. (1998), "Service Engineering—Perspektiven einer noch jungen Fachdisziplin," *IM Information Management and Consulting,* special edition on service engineering, pp. 37–39.

Fähnrich, K.-P., and Meiren, T. (1998), "Service Engineering," *Offene Systeme,* Vol. 7, No. 3, pp. 145–151.

Fähnrich, K.-P., and Meiren, T. (1999), *Service Engineering—Ergebnisse einer empirischen Studie zum Stand der Dienstleistungsentwicklung in Deutschland,* IRB, Stuttgart.

Fisk, R. P., Brown, S. W., and Bitner, M. J. (1993), "Tracking the Evolution of the Service Marketing Literature," *Journal of Retailing,* Vol. 69, No. 1, pp. 61–103.

Frings, S., and Weisbecker, A. (1998), "Für jeden die passende Rolle," *it Management,* Vol. 5, No. 7, pp. 18–25.

Garvin, D. A. (1984), "What Does Product Quality Really Mean?," Sloan Management Review, Fall, pp. 25–43.

Gogoll, A. (1996), *Untersuchung der Einsatzmöglichkeiten industrieller Qualitätstechniken im Dienstleistungsbereich,* IPK, Berlin.

Haischer, M. (1996), "Dienstleistungsqualität—Herausforderung im Service Management," HMD Theorie und Praxis der Wirtschaftsinformatik, No. 187, pp. 35–48.

Heskett, J. L., Sasser, W. E., and Schlesinger, L. A. (1997), *The Service Profit Chain: How Leading Companies Link Profit and Growth to Loyalty, Satisfaction, and Value,* Free Press, New York.

Hofmann, H., Klein, L., and Meiren, T. (1998), "Vorgehensmodelle für das Service Engineering," *IM Information Management and Consulting,* special edition on service engineering, pp. 20–25.

Humphrey, W. S. (1989), *Managing the Software Process,* Addison-Wesley, Reading, MA.

Jaschinski, C. M. (1998), *Qualitätsorientiertes Redesign von Dienstleistungen,* Shaker, Aachen.

Kaplan, R. S., and Norton, D. P. (1996), *The Balanced Scorecard,* Harvard Business School Press, Boston.

Kingman-Brundage, J. (1995), "Service Mapping: Back to Basics," in *Understanding Services Management,* W. J. Glynn and J. G. Barnes, Eds, Wiley, Chichester, pp. 119–142.

Lovelock, C. H. (1995), "Managing Services: The Human Factor," in *Understanding Services Management,* W. J. Glynn and J. G. Barnes, Eds, Wiley, Chichester, pp. 203–243.

Malorny, C. (1996), *Einführen und Umsetzen von Total Quality Management,* IPK, Berlin.

Meiren, T. (1999), "Service Engineering: Systematic Development of New Services," in *Productivity and Quality Management,* W. Werter, J. Takala, and D. J. Sumanth, Eds., MCB University Press, Bradford, pp. 329–343.

Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1985), "A Conceptual Model of Service Quality and Its Implications for Future Research," *Journal of Marketing,* Vol. 49, No. 3, pp. 41–50.

Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1988), "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality," *Journal of Retailing,* Vol. 64, No. 1, pp. 12–37.

Ramaswamy, R. (1996), *Design and Management of Service Processes,* Addison-Wesley, Reading, MA.

Reichheld, F. F. (1996), *The Loyalty Effect,* Harvard Business School Press, Boston.

Rommel, G., Kempis, R.-D., and Kaas, H.-W. (1994), "Does Quality Pay?," McKinsey Quarterly, No. 1, pp. 51–63.

Sampson, S. E. (1999), *The Unified Service Theory,* Brigham Young University, Provo, UT.

Schmenner, R. W. (1995), *Service Operations Management,* Prentice Hall, Englewood Cliffs, NJ.

Shostack, G. L. (1984), "Designing Services That Deliver," *Harvard Business Review,* Vol. 62, No. 1, pp. 133–139.

Stanke, A., and Ganz, W. (1996), "Design hybrider Produkte," in *Dienstleistungen für das 21. Jahrhundert: Eine öffentliche Diskussion,* V. Volkholz and G. Schrick, Eds., RKW, Eschborn, pp. 85–92.

CHAPTER 23

Customer Service and Service Quality

Richard A. Feinberg
Purdue University

1. **INTRODUCTION TO THE CUSTOMER SERVICE PARADIGM** **651**
2. **IT'S THE CUSTOMER STUPID . . . NOT SERVICE QUALITY, PRODUCT QUALITY, OR CUSTOMER SERVICE!** **654**
    2.1. Return on Investment for Service-Quality Improvements 654
3. **HOW TO CREATE A CUSTOMER-FOCUSED BUSINESS** **654**
    3.1. Step 1: Mission, Values, and Purpose 654
    3.2. Step 2: Proactive Policies and Procedures 656
        3.2.1. One Very Important Policy: The Guarantee 656
    3.3. Step 3: Determining What Is Important to the Customer 657
4. **THE CUSTOMER SERVICE DEPARTMENT** **657**
    4.1. Organization 657
    4.2. Centralization 657
    4.3. Setting Service-Quality Standards 657
    4.4. Managing Complaints 658
    4.5. Call Centers 658
        4.5.1. Call Center Operations and Logistics 658
    4.6. Hiring, Training, and Keeping Customer-Oriented Professionals 659
        4.6.1. Training for Exceptional Customer Service 659
    4.7. Serving the Internal Customer 659
5. **THE FUTURE OF CUSTOMER SERVICE AND SERVICE TECHNOLOGY** **660**
    5.1. Customer Assess Is the New Marketing Paradigm 660
6. **THE CUSTOMER SERVICE AUDIT** **662**
7. **A FINAL WORD** **662**

REFERENCES **663**

1. INTRODUCTION TO THE CUSTOMER SERVICE PARADIGM

A fundamental change in understanding business success has occurred. The new paradigm stresses that sales and profit are the result of customer satisfaction (Anton 1996) (see Figure 1).

Competitive advantage is most likely to come from innovation and creativity, flexibility to shift quickly as markets and customers change, marketing and tailoring value to the specific needs of profitable customers, and developing and creating long-term customer relationships. Today (and even more so in the future) competitive advantage will come from doing things sooner and better than the competition to a customer who will want to do business with you for an extended period of time.

The essence of business is to create satisfied customers who purchase your products and services and come back for more, and products that do not come back. The lifetime value of a customer is often overlooked in the day-to-day strategic and operational decisions. But this lifetime value is the critical issue for understanding why investments in service quality and customer satisfaction are not just expense lines in some budget but investments in bottom-line profit and the future of the company

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

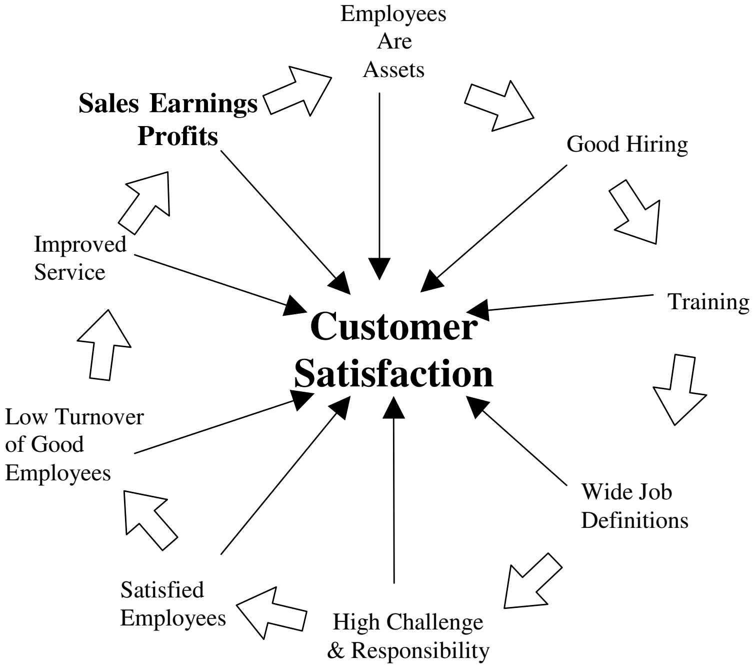

Employees Are Assets
Good Hiring
Training
Wide Job Definitions
High Challenge & Responsibility
Satisfied Employees
Low Turnover of Good Employees
Improved Service
Sales Earnings Profits
Customer Satisfaction

Figure 1 Customer Satisfaction Wheel of Success.

(Rust et al. 2000). Service quality and customer satisfaction must be calculated as a function of the sales and profitability related to the length of time a customer stays with your company as well as the costs involved in losing dissatisfied customers (see Tables 1 and 2). Research clearly shows that the longer a customer stays with a company the greater the sales and profits to the company because that consumer will buy more, buy more profitable items/services, requires smaller sales expenses, and be responsible for new customers through positive recommendations (see Figure 2 for a hypothetical look at this relationship).

TABLE 1 Increased Revenues That Can Result over Time from Improved Customer Service

Year	Revenues at 70% Retention Rate	Revenues at 80% Retention Rate	Revenues at 90% Retention Rate	Revenues at 100% Retention Rate
1	$1,000,000	$1,000,000	$1,000,000	$1,000,000
2	770,000	880,000	990,000	1,100,000
3	593,000	774,000	980,000	1,210,000
4	466,000	681,000	970,000	1,331,000
5	352,000	600,000	961,000	1,464,100
6	270,000	528,000	951,000	1,610,510
7	208,000	464,000	941,000	1,771,561
8	160,000	409,000	932,000	1,948,717
9	124,000	360,000	923,000	2,143,589
10	95,000	316,000	914,000	2,357,948
Totals	$4,038,000	$6,012,000	$9,562,000	$15,937,425

Reprinted from *Customer Service Operations: The Complete Guide* by Warren Blanding. Copyright © 1991 AMACOM, a division of the American Management Association International. Reprinted by permission of AMACOM, a division of American Management Association International, New York, NY. All rights reserved. http://www.amacombooks.org.

Figures are based on 10% account growth annually.

TABLE 2 Annual Revenue Loss from Customer Defection

If you Lose:	*SPENDING* **$5 Weekly**	*SPENDING* **$10 Weekly**	*SPENDING* **$50 Weekly**	*SPENDING* **$100 Weekly**	*SPENDING* **$200 Weekly**	*SPENDING* **$300 Weekly**
1 *customer a day*	$94,900	$189,800	$949,000	$1,898,000	$3,796,000	$5,694,000
2 *customers a day*	189,800	379,600	1,898,000	3,796,000	7,592,000	11,388,000
5 *customers a day*	474,500	949,000	4,745,000	9,490,000	18,980,000	28,470,000
10 *customers a day*	949,000	1,898,000	9,490,000	18,980,000	37,960,000	56,940,000
20 *customers a day*	1,898,000	3,796,000	18,980,000	37,960,000	75,920,000	113,880,000
50 *customers a day*	4,745,000	9,490,000	47,450,000	94,900,000	189,800,000	284,700,000
100 *customers a day*	9,490,000	18,980,000	94,900,000	189,800,000	379,600,000	569,400,000

Reprinted from *Customer Service Operations: The Complete Guide* by Warren Blanding. Copyright © 1991 AMACOM, a division of the American Management Association International. Reprinted by permission of AMACOM, a division of American Management Association International, New York, NY. All rights reserved. http://www.amacombooks.org.

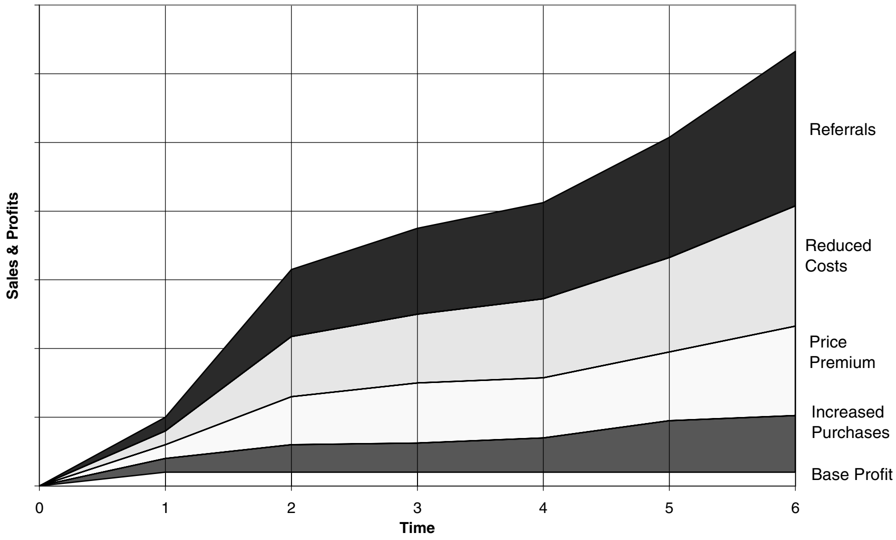


Figure 2 The Lifetime Value of One Customer.

Service quality and customer satisfaction are not new constructs. As early as the mid-1960s, Peter Drucker, the original management guru, prescribed the following questions for all businesses: Who is your customer? What is the essence of your business? What does your customer value? And how can your business serve the customer better than the competition?

It is clear that although the 1990s were an era of talking about service quality and customer satisfaction, business has a long way to go in delivering on a customer satisfaction promise. It would certainly be depressing if all we have accomplished since the 1982 publication of Peters and Waterman's *In Search of Excellence* is to add a customer service customer satisfaction statement to the mission statements of our Fortune 500 companies. Five years ago, we would have been pleased to state that customer satisfaction was being discussed in the boardrooms and had a chance to become a part of business strategy. Now this is not enough. Delighting customers must become a focal strategy. Every business, every executive, and every manager must assess how their "function/work" contributes to customer satisfaction.

2. IT'S THE CUSTOMER, STUPID . . . NOT SERVICE QUALITY, PRODUCT QUALITY, OR CUSTOMER SERVICE!

The issue is satisfaction. Consumers will pay for two things: solutions to problems and good feelings. When the consumer can purchase almost any product from multiple channels (TV, catalog, Internet, multiple stores) the only thing that can differentiate one business from another is the way it makes the customer feel. Businesses must decide whether they will simply be a vending machine and wait for the customer to deposit money in the slot and pick up their product/service, or whether they will add value. If businesses choose to be a vending machine, consumers will choose the cheapest, most convenient vending machine. But if businesses add value, consumers will expend effort, remain loyal, and purchase at greater margin for longer periods of time.

The issue is not simply service and product quality. Businesses can make their "stuff" better and better, but if it is not the stuff that consumers want, then consumers will not buy. In the 1960s, the big names in slide rules were Post, Pickett, and K&E. Each made considerable profit by selling better-quality and better-functioning slide rules—year after year after year. These three companies did exactly what the experts say is important—make quality products. Yet in the late 1970s slide rule sales disappeared, and so did these companies, not because they made slide rules of lower quality but because new companies by the name of Hewlett-Packard and Texas Instruments began making electronic calculators. Not a single slide rule company exists today. None of the three dominant companies noticed the shift. All three continued to make quality slide rules for customers who were now buying electronic calculators. Making the best-functioning product does not guarantee survival. The customer passed them by. Satisfaction results not simply from quality products and services but from products and services that consumers want in the manner they want. The side of the business highway is littered with companies that continued to make their stuff better without watching where the consumer was headed. In 1957, the top 10 businesses in Chicago were Swift, Standard Oil, Armour, International Harvester, Inland Steel, Sears, Montgomery Wards, Prudential, and the First Bank of Chicago. Thirty-five years later, the list includes only Sears and First National Bank from the original list, plus Ameritech, Abbott Labs, McDonald's, Motorola, Waste Management, Baxter, CAN Financial, and Commonwealth Edison. These two lists dramatically show the evolution that business goes through.

The discussion of a customer-satisfaction orientation is not limited to a few select businesses and industries. A satisfaction orientation exists in profit, not-for-profit, public and private, and business-to-business, and business to consumer (see Table 3).

2.1. Return on Investment for Service-Quality Improvements

Increases in service quality and customer satisfaction have direct bottom-line benefits. According to Robert LaBant of IBM, a 1% increase in customer satisfaction is worth $275,000,000 to IBM. Reichheld and Sasser (1990) estimated that a 5% increase in customer retention yields a 20% profit increase for catalog companies, 30% for auto service chains, 35% for software companies, 50% for insurance companies, and 125% for credit card companies. Costs associated with satisfaction and service-quality improvements are not costs but investments in sales, profit, and long-term viability. Assistance in figuring the return on investment for service quality and satisfaction improvements is available in books (Rust et al. 1994) and ROI calculators available on the Web (www.1to1.com—click on tools and ideas—look for downloads; www.e-interactions.com/download.html); (www.cfs.purove.edu/conscirt/quality.html).

3. HOW TO CREATE A CUSTOMER-FOCUSED BUSINESS

3.1. Step 1: Mission, Values, and Purpose

A customer-service focus emanates from the core values implicitly and explicitly expressed throughout the organization. It is represented in the mission statement of the company and the development

TABLE 3 Examples of Customer Service Applications

Consumer		Business		Public/Institutional/Associations/Not for Profit	
Products	Services	Products	Services	Products	Services
Retail, general	Insurance	Manufacturing to inventory	Transportation	Infrastructure	Defense
Retail, special	Banking, financial	Manufacturing to order	Warehousing	Currency/stamps	Education
Mail order	Travel	Commodities	Insurance	Publications	Licensing, regulation
Do-it-yourself	Health care	Finished goods	Financial	Power/light	Police protection
Home delivery	Real estate	Hi-tech/low-tech	Factoring	Museums, parks	Health
Party sales	Domestic help	Bulk/packaged	Engineering	Sheltered workshops	Information
Door-to-door	Lawn, yard care	Consumer/industrial	Environmental	Research spinoff	Postal service
In-home demos	Bridal counseling	Original equipment manufacturers/distributors	Computer services	Surplus goods	Subsidies
Auctions	Catering		Security services	Timberlands	Taxes
Estate/yard sales	Riding, sports	Direct/indirect	Consulting	Oil, minerals	Lotteries
Equipment rental	Recreation	Consignment	Leasing		Agriculture department
Subscriptions	Entertainment	Site delivery	Waste management		Disaster relief
Negative options	Beauty	On-site construction			Standards
	Diet plans				Export/import
	Child care				Social services
	Education				Fire department
	Consulting				Waste management

Reprinted from *Customer Service Operations: The Complete Guide* by Warren Blanding. Copyright © 1991 AMACOM, a division of the American Management Association International. Reprinted by permission of AMACOM, a division of American Management Association International, New York, NY. All rights reserved. http://www.amacombooks.org.

of a functional department to oversee customer relationship management and an executive responsible for the voice of the consumer.

A customer-focused mission statement should make it clear to all who work in the company and do business with the company that the customer is not to be ignored and should be the focal point of all components of the business. Every department and every individual should understand how they are serving the customer or serving someone who is serving the customer. The mission can be something as simple as Lands' End's "guaranteed period" or the engraving on a two-ton granite rock that marks the entrance to Stew Leonard's grocery stores in Connecticut:" Stew's Rules—Rule 1. The customer is always right. Rule 2. If you think the customer is wrong reread rule number 1." Then there is the elegant mission statement on the walls of the Ritz-Carlton: "We are Ladies and Gentleman Serving Ladies and Gentleman. . . . The Ritz-Carlton experience enlivens the senses, instills well-being, and fulfills even the unexpressed wishes and needs of our guests." And the mission statement that helps explain why Wal-Mart is the top retailer in the world: "We exist to provide value to our customers. To make their lives better via lower prices and greater selection; all else is secondary."

The foundation for a visionary company is the articulation of a core ideology (values and guiding principles) and essence (the organizations fundamental reason for existence beyond just making money) (Collins and Porras 1994).

3.2. Step 2: Proactive Policies and Procedure

Once an organization has established the principles of customer orientation and service quality as part of its reason for being, it must implement that vision throughout in specific policies and procedures that define doing business as and at that company.

These policies can be taught as simply as they do at Nordstrom's, where customer service training consists of the prescription "Use your good judgment always," or as complex as a 250-page policies and procedures manual covering acceptance or orders, handling major and minor accounts, order changes, new customers, phone inquiries, return policies, ship dates, orders for future delivery, and hundreds of other specific policies and procedures.

In a business world that no longer has time for a five-year apprenticeship in which the policies and procedures are passed through in one-on-one learning, how things are done and their standards must be available for all. High levels of turnover means that teaching policies and procedures is a 7-day a week, 24-hour-a-day requirement.

3.2.1. One Very Important Policy: The Guarantee

One of the more important policies and procedures for creating service quality and satisfaction is the guarantee. In a world in which products and services are commodities (interchangeable products that can be purchased at a number of channels), customers want to know that the company stands behind the purchase. Companies may not be able to guarantee that they won't make mistakes or that the product will never fail, but they can guarantee that they will stand behind it. Unfortunately, many companies have policies that state what they will not do rather than reassuring the customer with what they will do. Lands' End tells us that everything is "Guaranteed period." Restoration Hardware tells us, "Your satisfaction is not only our guarantee, it's why we're here. To be perfectly clear, we insist you're delighted with your purchase. If for any reason, a selection doesn't meet your expectations, we stand ready with a full refund or exchange." Nordstrom's has been built on its legendary policy of taking back *anything* (it is part of the Nordstrom's legend that they once accepted a set of tires despite never having sold tires). Same-day delivery, "If something goes wrong in 30 days after we fix it we will repair it free," "We accept merchandise for 30 days after purchase but only with a sales slip" are examples of policies affecting guarantees.

In establishing guarantees, there are a number of factors that should be considered. First, the costs of administering return/guarantee policies may outweigh their benefits. For example, if a manager must approve all returns and the approval rate exceeds 90% anyway, the costs of the manager's time and effort probably exceed the benefit. Second, uncertain and restrictive policies that increase the probability that irate consumers will confront poorly prepared front-line people will decrease morale and increase turnover. Third, guarantees can be a significant source of profit when companies sell extended guarantees for products and services that have naturally low levels of failure. Electronics companies have found these extended policies to be very profitable. Fourth, guarantees, which lay out an organization to outperform historical or market standards, force companies to become extraordinary. Making extraordinary goals public may force companies to aim at and reach high levels of customer service and satisfaction. Finally, return policies and product service guarantees must be developed in light of the lifetime value of the customer. The probability of losing a customer for life because of a restrictive policy must be balanced against the value of the potential purchasing power and loyalty of the consumer in the formation of policies. Sewell Cadillac in Texas estimates the value of each customer to be over $250,000 (the value of car sales, leasing, repairs to customers buying cars over their lifetime). To create policies that drive these customers away for what really are insignificant dollar amounts given this lifetime value is tantamount to stealing future income.

On the other hand, very liberal guarantees and return policies may be inappropriate and silly. A contractor who has a satisfaction-guaranteed policy might be giving away the business when the customer is dissatisfied with the room design they selected and approved although the workmanship was outstanding. Although unlimited guarantees are an inducement to consumers and may be a competitive advantage, a profitable business must be run. Unlimited and liberal policies may require higher prices, which will allow competitors to offer comparable products and services at lower prices.

For most businesses, a guarantee in the form of a limited warranty probably makes sense. These limited warranties spell out the conditions that define replacement or exchange. Customer abuses are less likely, but reasonable customers will still find these guarantees an inducement.

3.3. Step 3: Determining What Is Important to the Customer

Perhaps the biggest drawback to developing exceptional customer satisfaction is that information about what is important to customers comes from talking to ourselves (the executives of the company), not the customer. Understanding the customer experience from representative samples of customers using qualitative measurement and analysis is generally possible, and it guarantees that policies and procedures instituted by a company are not based on idiosyncratic and/or nonrepresentative (of all a company's customers) pieces of data or belief.

The customer satisfaction measurement system must connect to the internal measures of a company and to external customer evaluations. To provide evidence, surveys can be performed to make possible a gap analysis (the gap is the difference between what the customer should experience and what the customer actually experiences). The data of actual-to-expected performance allow the application of advanced statistical techniques such as regression analysis to determine empirically the relative impact and/or importance of attributes and processes to customer satisfaction (the power of this technique and an example of its application are illustrated in Anton [1996]). To improve customer satisfaction, identifying the specific attributes and processing the most predictive of customer satisfaction decisions about investment of resources will yield the greatest benefit.

4. THE CUSTOMER SERVICE DEPARTMENT

Companies at the forefront of customer satisfaction have found that positioning, organizing, and staffing a department for consumer affairs/customer satisfaction is essential. The growth in importance of customer service departments is dramatically shown by the proliferation of professional associations and professional newsletters. The Society of Consumer Affairs Professionals (SOCAP, at www.socap.org) serves to professionalize the industry and provides a range of training, educational material, and meetings. The *Customer Service Newsletter,* offered by the Customer Service Group (www.alexcommgrp.com), is part of a set of useful newsletters and practical materials for improving and understanding a company's customer service offerings.

4.1. Organization

Organizing a department whose top executive reports to the CEO is likely to send a signal to the organization that customer service/satisfaction is important. Direct reporting to the CEO allows policies about customers to be made at the highest levels and reduces the probability that these policies will get lost in the history and bureaucracy of the business.

4.2. Centralization

There is much to be gained from having a centralized customer service initiative. These advantages include economies of scale, greater practicality for using state of the art computers and telephone strategies, adaptability, easy adherence to standards, better access to top management, opportunities to create career paths, and being well suited for manufacturing and consumer products businesses. Probably the greatest advantage of centralization is the centralization of information. Information is easily aggregated and analyzed for use by the strategic areas in the business.

4.3. Setting Service-Quality Standards

Service standards are the measuring stick that guides a customer service report card. Standards must be carefully engineered to be consistent with the operational and strategic goals of the business. One of the more common standards in customer call centers may actually be antithetical to the customer satisfaction goals of the organizations. Forcing customer service agents to take too many calls per hour or make each call as short as possible (both metrics logical under a cost cutting/control mission) may interfere with a mission to maximize customer satisfaction (which might require longer calls and therefore fewer calls per hour).

Mystery shoppers, mystery callers, and quality monitoring are all techniques used to monitor service quality standards. Unfortunately, more attention is paid to how those standards are measured than to the establishment of standards that are really related to outcome measures that matter.

Simply stated, what gets measured gets done. What gets rewarded gets repeated. Telling front-line people that they must greet the customer within 25 seconds can easily be monitored and will increase the chance that the front-line person will greet the customer quickly. If a standard is established, it should be related to satisfaction, purchase, or loyalty in some way. Standards should be related to issues of bottom-line significance. Unfortunately, many customer satisfaction and service quality standards are set because they can be easily measured and monitored (greet customers within 25 seconds when they enter the store, answer the phone by the fourth ring, respond to the e-mail with 24 hours). Or they are set because they have been historically used in the organization.

Creating meaningful customer service and service quality standards plays a role in establishing the company as an outstanding customer-oriented organization. Few companies and organizations have standards, and those who have them do not tie them to strategy and mission. But a few have standards that they have found to be causal determinants of customer satisfaction and profitability. These companies are leaders in their fields.

4.4. Managing Complaints

At any time, almost 25% of your customers are dissatisfied with your products or service. Yet fewer than 5% of these consumers ever complain. Complaints are a fertile source of consumer intelligence. Businesses should do everything to maximize the number of complaints from dissatisfied customers. Complaints define what companies are doing wrong so that systemic changes can be made if needed. Second, research is clear in showing that a dissatisfied consumer who complains and is taken care of is significantly more loyal than a consumer who does not complain. Complaints are strategic opportunities. Most consumers who complain are not irate. Systems and employees who are not responsive create irate consumers. Training and empowerment allow front-line employees to reduce anger and create loyal customers.

Companies that understand the strategic value of complaints have instituted systems and access points that literally encourage consumers to complain. Internet access sites and e-mail addresses are the wave of the future, and companies will need to be prepared for the volume of contacts received in this manner. More likely today's companies have a call center at the center of their complaint-management system.

With simple training and sound procedures and policies, most consumer complaints can be resolved quickly and effectively at the lowest levels of contact. It costs five times as much to get new customers as it does to keep customers. Call centers, and in the future e-mail and Web access, provide companies with the cost-effective ability to manage complaints, turning a dissatisfied customer into a loyal one. But maybe more important, a company that recognizes a complaint as a strategic opportunity encourages complaints and is more likely to use the information to make strategic development and marketing decisions. What you do not hear can and will hurt you.

4.5. Call Centers

The call center has emerged as a vital link between customers and businesses after the consumer has purchased the products and/or services. These centers range from small operations to massive operations employing thousands of telephone service representatives.

The birth of the 800 toll-free number made access to companies cheap and easy for consumers. Subsequent advances in telecommunications technology have enabled businesses to handle volumes of calls unimaginable five years ago at affordable costs.

4.5.1. Call Center Operations and Logistics

Inbound and outbound communications form the thrust of call center operations. The Internet is forming the basis of low-cost communication for the business-to-business and consumer-to-business enterprise. EDI (electronic data interchange) was a novelty five years ago. Now consumers (and businesses who are consumers of other businesses) would probably prefer to be able to order, check order, check inventory, locate where the products are en route, pay, and follow up without having to touch or talk to a live person.

Sophisticated natural language recognition voice recognition programs (interactive voice response technology [IVR]) are replacing the boring and ineffective first-generation IVR (''press or say 1 if you . . . press or say 2 if you . . . ''). IVR can become a cost-effective means of handling 50–75% of all incoming phone conversations. Telephonic advances allow a consumer to speak in his or her natural voice about the problem he or she is experiencing and the call to be routed so that the most qualified customer service agents will get the call before it has even been picked up.

More importantly, switch technology allows routing of customers based on their value to the company. High-priority customers can get through quickly, while lower-valued customers can wait in the queue.

4.6. Hiring, Training, and Keeping Customer-Oriented Professionals

The customer is the most important part of a business. No customer means no sales means no reason to exist. If the customer is so important, do we see few executive-level positions with customer satisfaction in the title? The financial side of the business is important, so we have a vice president of finance. The marketing side of the business is important, so we have a vice president of marketing. The consumer is important, but we do not have a vice president for consumer satisfaction.

Where to begin:

1. Make a commitment to exceptional customer satisfaction by making certain that job descriptions have customer satisfaction accountabilities. Have a person whose only responsibility is to think about how the operations and marketing and recruitment affects the customer. Because we believe that hiring minority individuals is important, we put into place an executive who audits all aspects of this part of the company. Without a key individual whose focus is customer satisfaction, customer satisfaction may be lost in the day-to-day pressures.
2. The expression is, if you are not serving the customer, you'd better be serving someone who is. Thus, positions should include some accountability for customer satisfaction. What gets measured and rewarded gets done. If customer satisfaction is important, then monitoring, measuring, and rewarding based on customer satisfaction are critical.
3. Customer satisfaction must be lived by the top executives. Customer-oriented cultures wither when senior executives only talk the talk. Every Saturday morning, executives at Wal-Mart's Bentonville, Arkansas, headquarters gather for a meeting that is satellite linked to each of their 2600+ stores. They do a number of things at this meeting (point out and discuss hot items, cost savings, the Wal-Mart cheer, etc.), but nothing is more important than the senior executive at the meeting getting up and asking all there, "Who is the most important person in Wal-Mart's life?" and then hearing all respond, "The customer."
4. Hire for attitude, train for skill. There is simply too much looking for skilled people and then hoping they can learn customer service skills. In studying at the best-in-class companies, we have observed that selecting people at all levels who are "eagles" (show evidence of the ability to soar) and then teaching them the skill set for the job is better than getting people who have some skills and hoping they will become eagles.
5. There are any number of commercially available screening and selection tools focusing on customer satisfaction.

Selecting, training, and developing a fanatical devotion to the customer in employees is the critical piece of the puzzle.

4.6.1. Training for Exceptional Customer Service

Most companies expect a great deal from their customer service professionals. Yet formal training is sketchy and infrequent and lacks specificity and impact. Training works well because it sets expectations, teaches skills, and allows employees to be successful in the job.

Instituting customer service training in companies with established training programs requires a customer-orientation module in existing training programs and the development of a complete customer service representative training program for those individuals with total customer service functionality.

Some companies find that a call center is the best place for any new hire to gain experience. Few engineers at Ford will forget the customer in their later jobs after spending some time in the Ford Customer Call Center listening to problems and complaints. In addition, after working in a call center, these future leaders of the company (whether they come from sales, finance, marketing, legal, or administration) become sensitive to customer satisfaction issues and see more clearly how customer service fits the big picture.

Stories of companies with exceptional customer service orientation and excellent training obscures the fact that training alone will never work. The organization, its people, and its processes must support the jobs of people who have direct customer service functionality. Customer service training is a continuous process. Measuring performance, holding people accountable, providing feedback, cross-training, refresher training, and customer visits all allow the growth and development of a customer service culture and mission.

4.7. Serving the Internal Customer

One of the greater gaps found in organizations attempting to become more customer focused is the lack of attention to the internal customer. A good example of an internal customer orientation is provided by the Haworth Company, an office systems manufacturer. Their quality-improvement program revolved around the internal customer. Each work unit identified all its internal customers. They prioritized these customers, identified key work output in measurable terms, set standards, and measured their performance with their internal customers. Many companies have instituted internal help desks to which employees can call to get answers and solutions to their work problems much the same way that external customers use call centers for answers to questions and solutions to problems.

5. THE FUTURE OF CUSTOMER SERVICE AND SERVICE TECHNOLOGY

The 21st century will be an era when customer satisfaction and service quality will be defined by providing customers (business to business and consumer to business) consistent accessibility. Total customer touch, at anytime, from anywhere, in any form. However, few businesses use customer access as a key component of a business strategy. Access seems to be developing piecemeal, not as the central core of an overall strategy to allow customer access and deliver customer value anytime and anywhere.

5.1. Customer Access Is the New Marketing Paradigm

It really is very simple. If you want people to solve problems right now, give them right-now information. And magically, if you want employees to create an enhanced customer experience (internal and/or external), give them the right information exactly when they need it. Access will be realized in a combination of bricks-and-mortar storefronts, the call center, and the Internet. If you don't deliver consistent accessibility, someone else will.

Dell Computers understands the new paradigm probably better than any company as we enter the new millennium. It began selling products online in 1996. In 1999, it received 2 million hits per week and did 40% of its business on the Web. That is 20 million dollars of computers and "stuff" each day. No computer is built until it is ordered, making inventory expenses minimal. Michael Dell believes that three words control the future of business. Businesses that understand the power of these words and can implement them will win. What are these three words? "The consumer experience." That expression placed on or in view of everyone's desk reminds all Dell employees that their salaries, cars, desks, and retirement accounts and the presents they give all depend on the experience they provide to the customer at all customer touch points. To make certain they are on top of that "experience," Dell created the Consumer Experience Council, a group that scrutinizes every aspect of how Dell interacts with the customer.

Technological advances have made the ability to integrate telephone and computer technologies with front- and back-office functions a realistic weapon in creating and maintaining long-term customer relations. The business that treats the telephone, the Internet, e-mail, and storefronts as complementary channels will just create more opportunities for capturing market share and the share of the consumer.

There are some very strong reasons why accessibility is the central issue, but none more important than the simple fact that lack of accessibility turns out to be the prime customer dissatisfier and reason why consumers desert a company. Research at our Center for Customer Driven Quality at Purdue University has affirmed that over 50% of consumers who desert a company because of bad service experience (Figure 3) and that bad service is primarily defined as an accessibility issue (Figure 4).

In order to support the total customer experience with information and intelligence at all points of contact, businesses must develop systems that can pull information from many differing databases. This requires a technology platform that allows for real-time reporting to the business and immediate updating to the customer experience file. Our points of contact must be aligned with all databases. Finally, the people who work in all of the customer access channels (contact centers, storefronts, and webcenters) need to understand the value that is added at each and every point of customer contact. This requires a redefinition of the role of each person who is acting in one of these contact channels. As an example, storefront people need to be comfortable with the Web and the call center. They need to be aware of how each of these access channels operates. This will require them to access to these other channels within the store. The company face to the customer should be such that whatever channel the customer selects, he or she will be able to conduct the business he or she wants and be recognized by his or her contact point as a person, not a transaction. No story will need to be told twice. No problem will go unsolved. No question will go unanswered. The customer will be proactively informed of any changes.

The organization must realize that this is not simply a technology purchase. This is not just pasting the latest electronic gizmo or switch or software onto your call center or website. Technology only enables the organization to enhance the customer experience. Customer relationship management (CRM) is not a toolbox; it is a new way of seeing the customer and then using a set of tools to enhance the total customer experience.

In a similar manner, the director, manager, or senior executive will have access to a set of cockpit-like analytical tools that will provide her or him enterprise-wide access to aggregate real-time information. So while the individual performance of a telephone service representative may not be

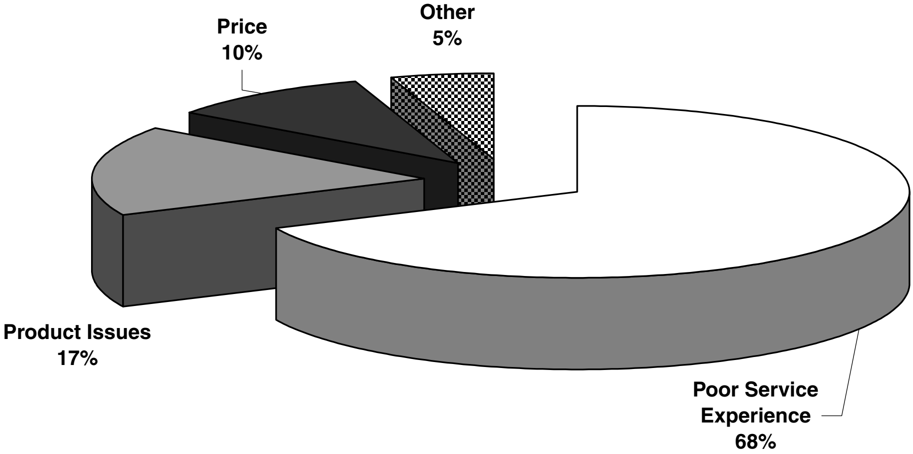


Figure 3 Why Do Customers Desert?

important to her/him, the enterprise-wide customer issues being probed in the contact centers are an issue—and this cockpit-type software provides that information when requested in real time.

So companies must remember to keep the customer at the focus of the decisions and solutions. Systems must be based on processes and technology that allow for simplified customer relationship management. General Electric has over 200,000 products across 80+ industries but only one number to call with a question, problem, concern whether it is about a jet engine, MRI machine, or a lightbulb . . . and they do this 24 hours a day, 7 days a week, 365 days a year. The systems must be open and easily integrated. The news is that this version of the future is available with technology today (and it will only get better and cheaper and easier).

Perhaps the most critical but undervalued aspect in creating total enterprise access will be the development of a layer of technology infrastructure called middleware. This middleware is critical for integrating the information now housed in separate and disparate databases. This "plumbing" does not get the attention it deserves. It is not glamorous or fashionable, partly because it is difficult to get a handle on, and it represents separate and distinct territories, each protected by herds of executives and workers who have vested interests in status quo systems. Middleware is the bridge between the databases. It is the piece that prevents the customer from hearing, "I am sorry, but the system does not allow me to do that" (heard often at airline counters) or "I don't have access to that database here" (heard often at banks)—the kinds of things that frustrate consumers and drive them elsewhere. No more "Why is it so hard for you people to get the order right?"

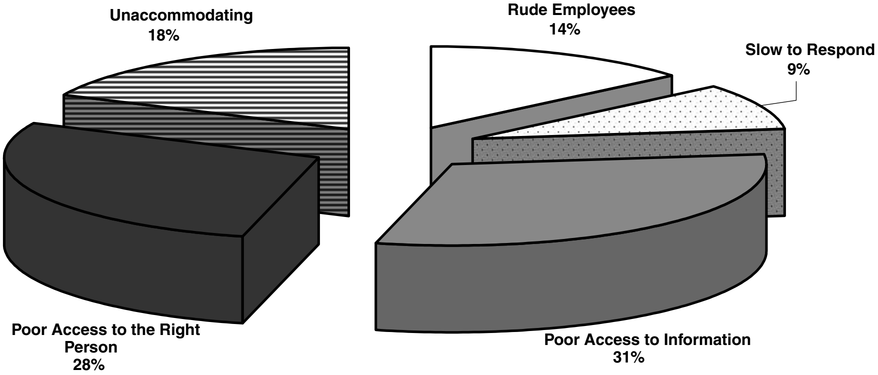


Figure 4 What Does the Customer Mean by Poor Service Experience?

Accessibility is clearly a business issue. Dell has recently moved past Compaq computer as the leading PC in the United States, yet Dell doesn't sell a single computer in the store. Intel and Cisco will book more from direct Internet orders than all the business-to-consumer sales that have taken place up until now. We can check inventory levels, place orders, and track these orders and delivery dates from these companies anytime we want to from our desk (as well as answer or have answered any of the questions we might have).

Accessibility is a control issue. Our ability to track our FedEx and UPS packages gives us power (and satisfaction). Imagine my smile when, after the recipient of a report I sent by FedEx claimed not to have received it, meaning my bonus for doing the work at that date would not be forthcoming, I was able to access (while we were speaking) the very signature that signed for the FedEx package exactly one hour before the due time (signed by this guy's secretary).

Accessibility is not just a marketing issue but an enterprise one. The story of how Avis trains its people better and more efficiently is the story of accessibility. Avis had the goal of consistent training across all employees anywhere anytime in the organization. The company developed a system called Spotlight, a virtual multimedia learning center that can be accessed in any of the 15,000 offices across 1,210 countries. After hearing a greeting and motivational talk by the CEO and learning the basic Avis skill set, the new employee meets a customer who takes him or her through the most common problems (representing 80% of all the escalated customer dissatisfaction issues). There are multiple lessons of accessibility here. First, anyone in the organization has immediate access to training. Classes do not have to be scheduled. Trainers do not have to be trained. The associate in Asia can access the training as well as an associate in New York City. This training is infinitely repeatable. But less obvious is the accumulation of customer information so that employees can know and learn from the top consumer problems, questions, and difficulties in an almost real-time process. While Avis is training to the specific situation, the competition is talking about general issues of customer dissatisfaction/satisfaction. Accessibility of information has led to specificity of attack.

Accessibility is an inventory issue. Accessibility among vendors, manufacturers, and retailers will lead to automatic replenishment. The complex accessibility among Dell, FedEx, and the many manufacturers who make Dell's parts results in FedEx managing just-in-time delivery of all the parts needed to build that special PC that the consumer just ordered today. At the end of the day, the system will tell us what we have on hand compared with what we want. If there is a difference, the difference will be ordered and the system will learn about proper inventory levels so that differences will be less likely in the future. In other words, the system learns from its experiences today and adjusts for tomorrow.

Accessibility is a retooling issue. Retooling a company for customer access means reengineering the technology and the people.

Most importantly, whatever the bottom-line impact of accessibility is on running a better business, the bottom-line impact on consumers is key. Accessibility enhances the customer's total experience. Accessibility builds customer and employee relationships with the company that empower them to change the enterprise and the enterprise to change them.

Accessibility builds brand meaning and value. Consumers are finding accessibility as a differentiating brand value.

Accessibility can create "delight"—the difference between the just satisfied and the WOWed.

In addition to accessibility, the future of service quality and customer satisfaction has to do with how information about the consumer will lead to extraordinary relationship building. Many names are emerging for this, such as database marketing, relationship marketing, and one to one marketing. This is not a new name for direct mail, or an order-taking system, or a substitute for a solid marketing strategy, or a solution to a bad image, or quick fix for a bad year. This is a new paradigm that Peppers and Rogers (1997, 1999) call "one to one marketing." The goal is to identify your best customers ("best" can mean more profitable, most frequent purchasers, highest proportions of business, loyalty) and then spend the money to get them, grow them, and keep them loyal. You need to fence your best customers off from competition. How easy would it be for another company to come in and steal these clients? How easy would it be for these best customers to form a relationship similar to what they have with you with another company?

6. THE CUSTOMER SERVICE AUDIT

The list of questions in Table 4 is a useful checklist to monitor performance of your company for customer service and satisfaction.

7. A FINAL WORD

Customer satisfaction is everything. (For a free downloadable collection of customer service and customer satisfaction quotes, go to customer service graffiti at www.cfs.purdue.edu\conscirt\quality.html—click on customer service graffiti). In the 21st century, satisfaction will be driven by

TABLE 4 Customer Service Audit Checklist

OBJECTIVE: The following list is an aid to measure and monitor customer orientation.

PROCEDURE
- Do you have a written customer service policy?
- Is this given a wide circulation within the company?
- Do customers receive a copy of this policy?
- Is customer service included in the marketing plan?
- What elements of customer service do you regularly monitor?
- Do you think other aspects of service should be monitored?
- Do you monitor competitive service performance?
- Do you know the true costs of providing customer service?
- Do customers have direct access to information on stock availability and delivery?
- How do you report to customers on order status?
- Is there a single point of contact for customers in your company?
- Do customers know who this individual is?
- Is any attempt made to estimate the cost of customer service failures (for example, a part delivery, late delivery, etc.)?
- Do you seek to measure the costs of providing different levels of service?
- Do you have internal service measures as well as external measures?
- How do you communicate service policies to customers?
- What is your average order cycle time?
- How does this compare with that of your major competitors?
- Do you monitor actual order-to-delivery lead-time performance?
- Do you have a system for accounting for customer profitability?
- Does the chief executive regularly receive a report on customer service performance?
- Do you consciously seek to hire individuals with a positive attitude towards customer service?
- Does customer service feature in the criteria for staff promotion?
- Do you use quality control concepts in managing customer service?
- Do you differentiate service levels by product?
- Do you differentiate customer service levels by customer type?
- Do you have a standard cost for an out of stock situation (for example, cost of lost sales, cost of back orders, etc.)?
- Do you provide customers with a customer service manual?
- Do you monitor the internal customer service “climate” on a regular basis?
- Does your customer service organization effectively manage the client relationship from order to delivery and beyond?
- How do you monitor and respond to complaints?
- How responsive are you to claims from customers?
- Do you allocate adequate resources to the development of customer service?
- How do you seek to maintain a customer focus?
- Does customer service regularly feature at management meetings and in training programs?
- What specific actions do you take to ensure that staff motivation re customer service is maintained at a high level?
- Is the company image re customer service adequate for the markets in which it operates?

Adapted from Leppard and Molyneux 1994. Used with permission of International Thomson Publishings.

customer access. But first, senior executives must agree that providing customers with a consistent, thoughtful, and value-added total customer experience at any and all touch points is vital to their retention and loyalty and future acquisition. This will allow their organizations to be moving towards a yet-to-be-defined level of enhanced total enterprise access for employees and customers . . . which will enhance the employee and customer experience . . . which will create loyal and long-lasting employee and consumer relationships with your company . . . which means happy customers, happy employees, happy senior executives, happy shareholders, happy bankers, and of course happy consumers.

REFERENCES

Anton, J. (1996), *Customer Relationship Management,* Prentice Hall, Upper Saddle River, NJ.

Blanding, W. (1991), *Customer Service Operations:* The Complete Guide, Amacon, New York.

Collins, J., and Porras, J. (1994), *Built to Last: Successful Habits of Visionary Companies,* HarperBusiness, New York.

Leppard, J., and Molyneux, L. (1994), *Auditing Your Customer Service,* Rutledge, London.

Peppers, D., and Rogers, M. (1997), *The One to One Future: Building Relationships One Customer at a Time,* Currency Doubleday Dell, New York.

Peppers, D., and Rogers, M. (1999), *Enterprise One to One: Tools for Competing in the Interactive Age,* Doubleday, New York.

Peters, T., and Waterman, R. (1988), *In Search of Excellence,* Warner, New York.

Reichheld, F., and Sasser, W. (1990), ‘‘Zero Defections: Quality Comes to Services,’’ *Harvard Business Review,* Vol. 68, September–October, pp. 105–111.

Rust, R., Zahorik, A., and Keningham, T. (1994), *Return on Quality: Measuring the Financial Impact of Your Company's Quest for Quality,* Irwin, Homewood, IL.

Rust, R., Zeithaml, V., and Lemon, K. (2000), *Driving Customer Equity: How Customer Lifetime Value Is Reshaping Corporate Strategy,* Free Press, New York.

CHAPTER 24

Pricing and Sales Promotion

Kent B. Monroe
University of Illinois at Urbana-Champaign

1. **INTRODUCTION TO PRICING MANAGEMENT** **666**
   - 1.1. The Definition of Price 666
   - 1.2. Proactive Pricing 667
   - 1.3. Factors to Consider When Setting Price 667
2. **PRICING OBJECTIVES** **667**
   - 2.1. Profit Objectives 667
   - 2.2. Volume-Based Objectives 667
   - 2.3. Competitive Objectives 668
3. **DEMAND CONSIDERATIONS** **668**
   - 3.1. Influence of Price on Buyer Behavior 668
   - 3.2. Useful Economic Concepts 668
     - 3.2.1. Demand Elasticity 668
     - 3.2.2. Revenue Concepts 669
     - 3.2.3. Consumers' Surplus 669
   - 3.3. Understanding How Buyers Respond to Prices 669
     - 3.3.1. Price, Perceived Quality, and Perceived Value 669
     - 3.3.2. Price Thresholds 669
     - 3.3.3. Effects of Reference Prices on Perceived Value 671
   - 3.4. The Effect of E-commerce on Pricing 671
     - 3.4.1. The Role of Information 671
     - 3.4.2. Pressures on E-commerce Prices 672
     - 3.4.3. The Feasibility of Sustainable Low Internet Prices 672
4. **THE ROLE OF COSTS IN PRICING** **672**
   - 4.1. Cost Concepts 672
   - 4.2. Profitability Analysis 673
   - 4.3. Profit Analysis for Multiple Products 673
5. **TYPES OF PRICING DECISIONS** **674**
   - 5.1. Pricing New Products 675
     - 5.1.1. Skimming Pricing 675
     - 5.1.2. Penetration Pricing 675
     - 5.2. Pricing During Growth 675
   - 5.3. Pricing During Maturity 675
   - 5.4. Pricing a Declining Product 675
   - 5.5. Price Bundling 676
     - 5.5.1. Rationale for Price Bundling 676
     - 5.5.2. Principles of Price Bundling 676
   - 5.6. Yield Management 676
6. **ADMINISTERING THE PRICING FUNCTION** **677**
7. **PRICE AND SALES PROMOTIONS** **678**
   - 7.1. Price Promotion as Price Segmentation 678
   - 7.2. Price Promotion Decision Variables 678
   - 7.3. Some Perspectives on Price and Sales Promotions 679
     - 7.3.1. Immediate Effects of Price and Sales Promotions 679
     - 7.3.2. Intermediate Effects of Price and Sales Promotions 679
     - 7.3.3. Long-Term Effects of Price and Sales Promotions 680
8. **LEGAL ISSUES IN PRICING** **680**
   - 8.1. Price Fixing 680
   - 8.2. Exchanging Price Information 681

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

8.3. Parallel Pricing and Price Signaling 681
8.4. Predatory Pricing 681
8.5. Illegal Price Discrimination 681
9. FOUR BASIC RULES FOR PRICING 682
9.1. Know Your Objectives 682
9.2. Know Your Demand 682
9.3. Know Your Competition and Your Market 682
9.4. Know Your Costs 683
10. SUMMARY 683
REFERENCES 683

1. INTRODUCTION TO PRICING MANAGEMENT

A husband and wife were interested in purchasing a new full-sized automobile. They found just what they were looking for, but its price was more than they could afford—$28,000. They then found a smaller model for $5,000 less that seemed to offer the features they wanted. However, they decided to look at some used cars and found a one-year-old full-size model with only 5,000 miles on it. The used car included almost a full warranty and was priced at $20,000. They could not overlook the $8,000 difference between their first choice and the similar one-year-old car and decided to purchase the used car.

This simple example illustrates an important aspect of pricing that often is not recognized: buyers respond to price differences rather than to specific prices. While this basic point about how prices influence buyers' decisions may seem complex and not intuitive, it is this very point that drives how businesses are learning to set prices. You may reply that this couple was simply responding to the lower price of the used car. And you would be correct—up to a point. These buyers were responding to a *relatively* lower price, and it was the *difference* of $8,000 that eventually led them to buy a used car.

Now consider the automobile maker who has to set the price of new cars. This decision maker needs to consider how the price will compare (1) to prices for similar cars by other car makers; (2) with other models in the seller's line of cars; and (3) with used car prices. The car maker also must consider whether the car dealers will be able to make a sufficient profit from selling the car to be motivated to promote the car in the local markets. Finally, if the number of new cars sold at the price set is insufficient to reach the profitability goals of the maker, price reductions in the form of cash rebates or special financing arrangements for buyers might have to be used. Besides these pricing decisions, the car maker must decide on the discount in the price to give to fleet buyers like car rental companies. Within one year, these new rental cars will be sold at special sales to dealers and to individuals like the buyer above.

Pricing a product or service is one of the most important decisions made by management. Pricing is the only marketing strategy variable that directly generates income. All other variables in the marketing mix—advertising and promotion, product development, selling effort, distribution—involve expenditures. The purpose of this chapter is to introduce this strategically important marketing decision variable, define it, and discuss some of the important ways that buyers may respond to prices. We will also discuss the major factors that must be considered when setting prices, as well as problems managers face when setting and managing prices.

1.1. The Definition of Price

It is usual to think of price as the amount of money we must give up to acquire something we desire. That is, we consider price as a formal ratio indicating the quantities of money (or goods and services) needed to acquire a given quantity of goods or services. However, it is useful to think of price as a ratio of what buyers receive in the way of goods and services relative to what they give up in the way of money or goods and services. In other words, price is the ratio of what is *received* relative to what is *given up*.

Thus, when the price of a pair of shoes is quoted as $85, the interpretation is that the seller receives $85 from the buyer and the buyer receives one pair of shoes. Similarly, the quotation of two shirts for $55 indicates the seller receives $55 and the buyer receives two shirts. Over time, a lengthy list of terms that are used instead of the term *price* has evolved. For example, we pay a postage *rate* to the Postal Service. *Fees* are paid to doctors and dentists. We pay *premiums* for insurance coverage, *rent* for apartments, *tuition* for education, and *fares* for taxis, buses, and airlines. Also, we pay *tolls* to cross a bridge, *admission* to go to a sporting event, concert, movie, or museum. Banks may have *user fees* for credit charges, *minimum required balances* for a checking account service, *rents* for safety deposit boxes, and fees or *interest charges* for automatic teller machine (ATM) use or cash advances. Moreover, in international marketing, *tariffs* and *duties* are paid to import goods into another country.

The problem that this variety of terms creates is that we often fail to recognize that the setting of a rent, interest rate, premium, fee, admission charge, or toll is a pricing decision exactly like that for the price of a product purchased in a store. Moreover, most organizations that must set these fees, rates, and so on must also make similar pricing decisions to that made by the car maker discussed above.

1.2. Proactive Pricing

The need for correct pricing decisions has become even more important as global competition has become more intense. Technological progress has widened the alternative uses of buyers' money and time and has led to more substitute products and services. Organizations that have been successful in making profitable pricing decisions have been able to raise prices successfully or reduce prices without competitive retaliation. Through careful analysis of pertinent information and deliberate acquisition of relevant information, they have become successful pricing strategists and tacticians (Cressman 1997).

There are two essential prerequisites to becoming a successful proactive pricer. First, it is necessary to understand how pricing works. Because of the complexities of pricing in terms of its impact on suppliers, salespeople, distributors, competitors, and customers, companies that focus primarily on their internal costs often make serious pricing errors.

Second, it is essential for any pricer to understand the pricing environment. It is important to know how customers perceive prices and price changes. Most buyers do not have complete information about alternative choices and most buyers are not capable of perfectly processing the available information to arrive at the ''optimum'' choice. Often, price is used as an indicator not only of how much money the buyer must give up, but also of product or service quality. Moreover, differences between the prices of alternative choices also affect buyers' perceptions. Thus, the price setter must know who makes the purchase decision for the products being priced and how these buyers perceive price information.

1.3. Factors to Consider When Setting Price

There are five essential factors to consider when setting price. *Demand* considerations provide a ceiling or maximum price that may be charged. This maximum price depends on the customers' perceptions of value in the seller's product or service offering. On the other hand, *costs* provide a floor or minimum possible price. For existing products or services, the relevant costs are those costs that are directly associated with the production, marketing, and distribution of these products or services. For a new product or service, the relevant costs are the *future costs* over that offering's life. The difference between the maximum price that some buyers are willing to pay (value) and the minimum cost-based price represents an initial pricing discretion. However, this range of pricing discretion is narrowed by *competition, corporate profit and market objectives,* and *regulatory constraints*.

2. PRICING OBJECTIVES

2.1. Profit Objectives

Pricing objectives need to be measured precisely. Performance can then be compared with objectives to assess results. In practice, the objective of profit maximization may be realized in multiple ways. In some markets, relatively low prices result in greater sales and higher profits. But in other markets, relatively high prices result in slightly decreased unit sales and also higher profits. Thus, the profits of some firms may be based on low prices and high sales volume, while for other firms high prices and low sales volume may be more profitable. Another common pricing objective is some form of target return on investment, that is, regaining a specified percentage of investment as income. Return on investment (ROI) is expressed as the ratio of profits to investments. For manufacturers, investments include capital, machinery, buildings, and land, as well as inventory. For wholesalers and retailers, inventory and buildings constitute the bulk of investments.

2.2. Volume-Based Objectives

Some organizations set pricing objectives in terms of sales volume. A common goal is sales growth, in which case the firm sets prices to increase demand. Other firms may seek sales maintenance, knowing that growth does not ensure higher profits and that they may not have the resources needed to pursue sales growth.

If capturing a high market share is a marketing objective, pricing objectives should reflect this goal. In general, a high market share is achieved by setting prices relatively low to increase sales. From a profitability perspective, the organization must be willing to accept lower initial profits in exchange for the profits that may be produced over time by increased volume and high market share. However, other companies achieve a strong position in selected markets by setting high prices and offering high-quality products and service.

2.3. Competitive Objectives

At times, firms base their pricing objectives on competitive strategies. Sometimes, the goal is to achieve price stability and engage in nonprice competition, while at other times, they price aggressively. When marketing a mature product and when the firm is the market leader, it may seek to stabilize prices. Price stability often leads to nonprice competition in which a firm's strategy is advanced by other components of the marketing mix: the product itself, the distribution system, or the promotional efforts.

In some markets, a firm may choose to price aggressively, that is, price below competition, to take advantage of market changes, for example, when products are in early stages of the life cycle, when markets are still growing, and when there are opportunities to establish or gain a large market share. As with a market share or volume objective, this aggressiveness must be considered within the context of a longer term perspective.

3. DEMAND CONSIDERATIONS

One of the most important cornerstones of price determination is demand. In particular, the volume of a product that buyers are willing to buy at a specific price is that product's demand. In this section we will review some important analytical concepts for practical pricing decisions.

3.1. Influence of Price on Buyer Behavior

In economic theory, price influences buyer choice because price serves as an indicator of product or service cost. Assuming the buyer has perfect information concerning prices and wants satisfaction of comparable product alternatives, he or she can determine a product/service mix that maximizes satisfaction within a given budget constraint. However, lacking complete and accurate information about the satisfaction associated with the alternative choices, the buyer assesses them on the basis of known information. Generally, one piece of information available to the buyer is a product's price. Other pieces of information about anticipated purchases are not always known, and buyers cannot be sure how reliable and complete this other information is. And because this other information is not always available, buyers may be uncertain about their ability to predict how much they will be satisfied if they purchase the product. For example, if you buy a new car, you do not know what the relative incidence of car repairs will be for the new car until after some months or years of use. *As a result of this imperfect information, buyers may use price both as an indicator of product cost as well as an indicator of quality* (*want satisfaction attributes*).

3.2. Useful Economic Concepts

This brief outline of how price influences demand does not tell us about the extent to which price and demand are related for each product/service choice, nor does it help us to compare, for example, engineering services per dollar to accounting services per dollar. The concept of elasticity provides a quantitative way of making comparisons across product and service choices.

3.2.1. Demand Elasticity

Price elasticity of demand measures how the quantity demanded for a product or service changes due to a change in the price of that product or service. Specifically, price elasticity of demand is defined as the percentage change in quantity demanded relative to the percentage change in price. Normally, it is assumed that quantity demanded falls as price increases; hence, price elasticity of demand is a negative value ranging between 0 and $-\infty$.

Because demand elasticity is relative, various goods and services show a range of price sensitivity. Elastic demand exists when a given percentage change in price results in a greater percentage change in the quantity demanded. That is, price elasticity ranges between -1.0 and $-\infty$. When demand is inelastic, a given percentage change in price results in a smaller percentage change in the quantity demanded. In markets characterized by inelastic demand, price elasticity ranges between 0 and -1. Another important point about price elasticity: it does change and is different over time for different types of products and differs whether price is increasing or decreasing.

A second measure of demand sensitivity is cross price elasticity of demand, which measures the responsiveness of demand for a product or service relative to a change in price of another product or service. Cross price elasticity of demand is the degree to which the quantity of one product demanded will increase or decrease in response to changes in the price of another product. If this relation is negative, then, in general, the two products are complementary; if the relation is positive, then, in general, the two products are substitutes.

Products that can be readily substituted for each other are said to have high cross price elasticity of demand. This point applies not only to brands within one product class but also to different product classes. For example, as the price of ice cream goes up, consumers may switch to cakes for dessert, thereby increasing the sales of cake mixes.

3.2.2. Revenue Concepts

There is a relationship between sellers' revenues and the elasticity of demand for their products and services. To establish this relationship we need to define the concepts of total revenue, average revenue, and marginal revenue. Total revenue is the total amount spent by buyers for the product ($\text{TR} = P \times Q$). Average revenue is the total outlay by buyers divided by the number of units sold, or the price of the product ($\text{AR} = \text{TR}/Q$). Marginal revenue refers to the change in total revenue resulting from a change in sales volume.

The normal, downward-sloping demand curve reveals that to sell an additional unit of output, price must fall. The change in total revenue (marginal revenue) is the result of two forces: (1) the revenue derived from the additional unit sold, which is equal to the new price; and (2) the loss in revenue which results from marking down all prior saleable units to the new price. If force (1) is greater than force (2), total revenue will increase, and total revenue will increase only if marginal revenue is positive. Marginal revenue is positive if demand is price elastic *and* price is decreased, or if demand is price inelastic *and* price is increased.

3.2.3. Consumers' Surplus

At any particular price, there are usually some consumers willing to pay more than that price in order to acquire the product. Essentially, this willingness to pay more means that the price charged for the product may be lower than some buyers' perceived value for the product. The difference between the maximum amount consumers are willing to pay for a product or service and the lessor amount they actually pay is called consumers' surplus. In essence, it is the money value of the willingness of consumers to pay in excess of what the price requires them to pay. This difference represents what the consumers gain from the exchange and is the money amounts of value-in-use (what is gained) minus value-in-exchange (what is given up). Value-in-use always exceeds value-in-exchange simply because the most anyone would pay must be greater than what they actually pay, otherwise they would not enter into the trade.

3.3. Understanding How Buyers Respond to Prices

As suggested above, a successful pricer sets price consistent with customers' perceived value (Leszinski and Marn 1997). To understand how customers form value perceptions, it is important to recognize the relative role of price in this process. Because of the difficulty of evaluating the quality of products before and even after the product has been acquired, how customers form their perceptions of the product becomes an important consideration when setting prices. During this perceptual process, buyers make heavy use of information cues, or clues. Some of these cues are price cues and influence buyers' judgments of whether the price differences are significant. For example, buyers may use the prices of products or services as indicators of actual product quality.

3.3.1. Price, Perceived Quality, and Perceived Value

Would you buy a package of 25 aspirin that costs only 50 cents? Would you be happy to find this bargain, or would you be suspicious that this product is inferior to other brands priced at 12 for \$1.29? In fact, many consumers would be cautious about paying such a low relative price. Thus, the manufacturers of Bayer and Excedrin know that some people tend to use price as an indicator of quality to help them assess the relative value of products and services.

Since buyers generally are not able to assess product quality perfectly (i.e., the ability of the product to satisfy them), it is their *perceived quality* that becomes important. Under certain conditions, the perceived quality in a product is positively related to price. Perceptions of value are directly related to buyers' preferences or choices; that is, the larger a buyer's perception of value, the more likely would the buyer express a willingness to buy or preference for the product. Perceived value represents a trade off between buyers' perceptions of quality and sacrifice and is positive when perceptions of quality are greater than the perceptions of sacrifice. Figure 1 illustrates this role of price on buyers' perceptions of product quality, sacrifice, and value. Buyers may also use other cues, such as brand name, and store name as indicators of product quality.

3.3.2. Price Thresholds

Those of us who have taken hearing tests are aware that some sounds are either too low or too high for us to hear. The low and high sounds that we can just barely hear are called our lower and upper absolute hearing thresholds. From psychology, we learn that small, equally perceptible changes in a response correspond to proportional changes in the stimulus. For example, if a product's price being raised from \$10 to \$12 is sufficient to deter you from buying the product, then another product originally priced at \$20 would have to be repriced at \$24 before you would become similarly disinterested.

Our aspirin example above implies that consumers have lower and upper price thresholds; that is, buyers have a range of acceptable prices for products or services. Furthermore, the existence of

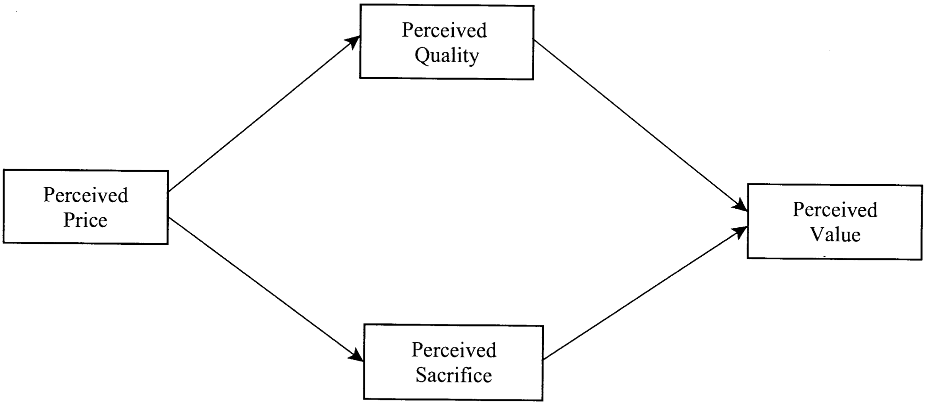


Figure 1 Perceived Price–Perceived Value Relationship

a lower price threshold implies that there are prices greater than $0 that are unacceptable because they are considered to be too low, perhaps because buyers are suspicious of the product's quality. Practically, this concept means that rather than a single price for a product being acceptable to pay, buyers have some range of acceptable prices. Thus, people may refrain from purchasing a product not only when the price is considered to be too high, but also when the price is considered to be too low. The important lesson to learn is that there are limits or absolute thresholds to the relationship between price and perceived quality and perceived value (Monroe 1993).

When buyers perceive that prices are lower than they expect, they may become suspicious of quality. At such low prices, this low perceived quality may be perceived to be less than the perceived sacrifice of the low price. Hence, the mental comparison or trade-off between perceived quality and perceived sacrifice may lead to an unacceptable perceived value. Thus, a low price may actually reduce buyers' perceptions of value. At the other extreme, a perceived high price may lead to a perception of sacrifice that is greater than the perceived quality, also leading to a reduction in buyers' perceptions of value. Thus, not only is it important for price setters to consider the relationship among price, perceived quality, and perceived value, but to recognize that there are limits to these relationships.

Usually a buyer has alternative choices available for a purchase. However, even if the numerical prices for these alternatives are different, it cannot be assumed that the prices are perceived to be different. Hence, the price setter must determine the effect of perceived price differences on buyers' choices. As suggested above, the perception of a price change depends on the magnitude of the change.

Generally, it is the perceived relative differences between prices that influence buyers' use of price as an indicator of quality. In a similar way, relative price differences between competing brands, different offerings in a product line, or price levels at different points in time affect buyers' purchase decisions. A recent experience of a major snack food producer illustrates this point. At one time, the price of a specific size of this brand's potato chips was $1.39 while a comparable size of the local brand was $1.09, a difference of 30 cents. Over a period of time, the price of the national brand increased several times until it was being retailed at $1.69. In like manner, the local brand's price also increased to $1.39. However, while the local brand was maintaining a 30-cent price differential, the national brand obtained a significant gain in market share. The problem was buyers perceived a 30-cent price difference relative to $1.69 as less than a 30-cent price difference relative to $1.39. This example illustrates the notion of differential price thresholds, or the degree that buyers are sensitive to relative price differences.

From behavioral price research, a number of important points about price elasticity have emerged:

1. Buyers, in general, are more sensitive to perceived price increases than to perceived price decreases. In practical terms, this difference in relative price elasticity between price increases vs. price decreases means it is easier to lose sales by increasing price than it is to gain sales by reducing price.
2. Sometimes a product may provide a unique benefit or have a unique attribute that buyers value. These unique benefits or attributes serve to make the product less price sensitive.

3. The frequency of past price changes can influence buyers' sensitivity to price changes. If prices have been changing relatively frequently, buyers may not have adjusted to the previous price change when a new change occurs. If buyers have not adjusted to the last price increase, then another price increase will be perceived as a larger increase than it actually is, making them more sensitive to the increase. The issue this point raises is the concept of reference prices and is discussed next.

3.3.3. Effects of Reference Prices on Perceived Value

In the past few years, evidence has emerged confirming the existence of a reference price serving as an anchor for price judgments. One of the most important points stemming from the concept of reference prices is that buyers do judge or evaluate prices comparatively. That is, for a price to be judged acceptable, too high, or too low, it has to be compared to another price. This other comparative price is the buyer's reference price for that particular judgment. The reference price serves as an anchor for buyers' price judgments. Buyers may use as a reference point the range of prices last paid, the current market price or perceived average market price, a belief of a fair price to pay, or an expected price to pay to judge actual prices.

Price perceptions are relative. That is, a specific price is compared to another price or a reference price. The illustration for relating this important point to pricing strategy and tactics comes from a firm introducing a new product with an introductory low price. Initially, the product was targeted to sell at a price of $17.50. However, the firm used the tactic of introducing the product at a low price of $14.95. Later, when it was time to remove the introduction price because of increased costs, the regular price was set at $20.00. The product failed to sustain sufficient sales volume to warrant its continued existence. The error in this situation was that the pricing tactic of a low introductory price established a baseline or reference price of $14.95 rather than $17.50. Hence, the $20.00 price, when compared to $14.95, was perceived to be too expensive and buyers stopped buying.

Recent behavioral research has provided additional explanations of how people form value judgments and make decisions when they do not have perfect information about alternatives. Moreover, these explanations further our understanding of why buyers are more sensitive to price increases than to price decreases and how they respond to comparative price advertisements, coupons, rebates, and other special price promotions. The common element in these explanations is that buyers judge prices comparatively, that is, a reference price serves to anchor their judgments (Monroe 1993).

3.4. The Effect of E-commerce on Pricing

There is a clear emerging trend toward electronic commerce becoming a significant aspect of the economy. Estimates of online sales indicate that the $7.8 billion in sales in the United States recorded in 1998 grew to approximately $15 billion in 1999 and is expected to grow to $23–28 billion in 2000 (Schmeltzer 2000; Sparks 2000). Accompanying this rapid growth of electronic sales is an emphasis on exchanges occurring at lower prices than in conventional exchanges. Several questions naturally arise with this knowledge of increasing electronic sales at relatively lower prices:

1. How does information play a role in this phenomenon?
2. What forces exist to pressure or enable online sellers to sell at lower prices?
3. Can these relatively lower prices (and accompanying lower profit margins) be sustained?

3.4.1. The Role of Information

In Section 3.1, we pointed out that traditional economic theory assumes that buyers and sellers have perfect information about prices, their own tastes and preferences, and their budget or income available for purchasing goods and services. However, when buyers are faced with imperfect information and the inability to assess quality, and therefore the ability to determine their degree of satisfaction prior to purchase, they may use price to infer quality and their expected degree of satisfaction. However, the quality of the attributes of some goods can be assessed prior to purchase. We call these goods search products. Examples of search products would include books, videotapes and music CDs, brand-name hard goods, airline tickets, toys, pet supplies, and standard industrial supplies. Indeed, a recent survey indicated that online shoppers purchased more books and videotapes online than in stores (*Chicago Tribune* 2000). Further, the CPI inflation rate for November 1999 for recreational products was 0.7% compared to the overall CPI inflation rate of 2.2% (Cooper and Madigan 1999). Products in the recreational category include toys, music, books, and pet supplies. Buyers perceive less risk in buying products that they believe vary little in quality across alternative sellers. Thus, searching online for the lowest prices for these types of products is convenient, quick, and virtually costless. For example, CompareNet offers detailed information on more than 100,000 products. Other sites provide software agents to find products (Green 1998). Moreover, the online shopper can search in a wider geographical area than in the local community. Finally, consumers and corporate buyers can pool their purchasing power and get volume discounts. General Electric Co. was able to reduce its purchase cost by 20% on more than $1 billion in purchases of operating materials by pooling orders from divisions on a worldwide basis (Hof 1999).

3.4.2. Pressures on E-commerce Prices

As indicated above, the online buyer may be able to reduce search costs while not incurring any significant increase in risk. At the same time, the direct supplier–buyer relationship in e-commerce transactions reduces or eliminates the various intermediaries in traditional distribution systems. Further, the information intermediary may take a smaller percentage of the final selling price for the information service provided (perhaps 10% as opposed to 40–50% traditional intermediary margins) (Hof 1999). Also, a large and growing number of information intermediaries, manufacturers, distributors, and retailers are providing buyers access to buying on the internet. This increasing competitive pressure will keep prices relatively low, as expected from traditional economic theory. But the question is, how long will we see both this intensive competition and these relatively very low prices?

3.4.3. The Feasibility of Sustainable Low Internet Prices

Currently almost all online sellers are losing money because they do not have sufficient sales at the relatively low profit margins to recover their fixed costs (Sparks 1999). Further, to get the sales growth necessary to become profitable will require increasing amounts of investment in marketing, that is, increasing fixed costs. Without infusion of large amounts of venture capital, many of these enterprises will not succeed and will either fail outright or will be acquired by more successful online sellers or manufacturers and retailers with the financial and marketing capital to generate the necessary sales growth.

Further, as more manufacturers, distributors, and retailers enter electronic selling, it is likely that their online sales will be at lower unit profit margins, given the current pressure on prices. Initially, their online sales will be at the expense of their higher-margin conventional sales, reducing their overall profit margins. Overcoming this negative effect on profitability will require new sales growth from new buyers who may shift from other electronic sellers. At some point there will be a shakeout between the e-commerce sellers and prices will likely stabilize at relatively higher levels than we encounter in this early stage of electronic commerce. This expectation is not different from the various revolutions in selling that have occurred over the years. The real problem will be predicting when the shakeout will occur leading to prices stabilizing profits at sustainable levels from electronic sellers. And if there is little perceived quality variation across sellers, then buyers are more likely to minimize the price paid for these items.

However, searching for the lowest price from alternative sellers can be time consuming in traditional shopping. Most busy people have neither the time nor the willingness to visit multiple outlets seeking the lowest price for a purchase they are considering.

4. THE ROLE OF COSTS IN PRICING

As indicated earlier, demand provides an upper limit on the pricing discretion the firm has. This limit is the willingness of buyers to purchase at a stated price. On the other hand, the other variable directly affecting profits—costs—sets a floor to a firm's pricing discretion. If prices are too low in comparison with costs, volume may be high but profitless. By determining the difference between costs and the price under consideration and then balancing that margin against the estimated sales volume, the seller can determine whether the product or service will contribute sufficient money to enable the firm to recover its initial investment. In considering the cost aspect of a pricing decision, a crucial question is what costs are relevant to the decision.

It is important for the seller to know the causes and behavior of product costs in order to know when to accelerate cost recovery, how to evaluate a change in selling price, how to profitably segment a market, and when to add or eliminate products. Even so, costs play a limited role in pricing. They indicate whether the product or service can be provided and sold profitably at any price, but they do not indicate the actual prices that buyers will accept. Proper costs serve to guide management in the selection of a profitable product mix and determine how much cost can be incurred without sacrificing profits.

4.1. Cost Concepts

To determine profit at any volume, price level, product mix, or time, proper cost classification is required. Some costs vary directly with the rate of activity, while others do not. If the cost data are classified into their fixed and variable components and properly attributed to the activity causing the cost, the effect of volume becomes apparent and sources of profit are revealed.

In addition to classifying costs according to ability to attribute a cost to a product or service, it is also important to classify costs according to variation with the rate of activity (Ness and Cucuzza

1995). Some costs vary directly with the activity level, while other costs, although fixed, are directly attributable to the activity level. Hence, it is important to clarify specifically what is meant by the terms *direct* and *indirect*. The directly traceable or attributable costs are those costs that we can readily determine as contributing to the product or service's cost. However, whether a direct cost is variable, fixed, or semivariable depends on properly determining the cause of that cost. Perhaps more than anything else, managers need to exert the will to understand how costs are incurred and how they behave as activity levels change in their organizations.

4.2. Profitability Analysis

Virtually every planned action or decision in an organization affects costs and therefore profits. Profit analysis attempts to determine the effect of costs, prices, and volume on profits. The goal of this analysis is to provide accurate information about the profit contributions made by each product, thereby giving management a sound basis for managing its product lines.

One of the most important pieces of data resulting from a profit analysis is the contribution ratio, which is usually referred to as the profit-volume ratio (PV). The PV ratio is the proportion of sales dollars available to cover fixed costs and profits after deducting variable costs:

$$\mathrm{PV} = \frac{\text{unit price} - \text{unit variable cost}}{\text{unit price}}$$

This PV ratio is an important piece of information for analyzing the profit impact of changes in sales volume, changes in the cost structure of the firm (i.e., the relative amount of fixed costs to total costs), as well as changes in price.

For a single product situation, the elements of profitability that the firm must consider are price per unit, sales volume per period, variable costs per unit, and the direct and objectively assignable fixed costs per period. However, typically firms sell multiple products or offer multiple services, and the cost-classification requirements noted above become more difficult. Further, many companies now recognize that it is important to avoid arbitrary formulas for allocating overhead (common costs) so that each product carries its fair share of the burden. In part, these companies understand that the concept of variable costs extends beyond the production part of the business. That is why we have stressed the notion of activity levels when defining variable costs. These companies have found out that using the old formula approach to cost allocation has meant that some products that were believed to be losing money were actually profitable and that other so-called profitable products were actually losing money.

A key lesson from the above discussion and the experiences of companies that use this approach to developing the relevant costs for pricing is that profits must be expressed in monetary units, not units of volume or percentages of market share. Profitability is affected by monetary price, unit volume, and monetary costs.

4.3. Profit Analysis for Multiple Products

In multiproduct firms, it is important to place emphasis on achieving the maximum amount of contribution revenue for each product instead of attempting to maximize sales revenues. Each product offering faces different competition, has a different demand elasticity, and perhaps depends for its sales, at least in part, on the sales of the other products in the line.

Within a multiproduct firm, each offering generates a different amount of volume, a different cost structure, including variable and fixed costs, different unit prices, and, of course, different revenues. Not only are these important factors different, but they are changing. The PV ratio can be used to analyze the relative profit contributions of each product in the line. Each product has a different PV value and different expected dollar sales volume as a percentage of the line's total dollar volume. In multiple-product situations, the PV is determined by weighting the PV of each product by the percentage of the total dollar volume for all products in the line.

This issue of managing the prices and profitability of the firm's product line is extremely important. For example, consider the prices of a major hotel chain. For the same room in a given hotel, there will be multiple rates: the regular rate (commonly referred to as the rack rate), a corporate rate that represents a discount for business travelers, a senior citizen rate, a weekend rate, single vs. double rates, group rate, and conference rate, to name just a few. Over time, the hotel's occupancy rate expressed as a percentage of rooms sold per night had increased from about 68% to 75%. Yet despite increasing prices that more than covered increases in costs, and increasing sales volume, profitability was declining.

After a careful examination of this problem, it was discovered that they were selling fewer and fewer rooms at the full rack rate while increasing sales, and therefore occupancy, of rooms at discounted rates that were as much as 50% off the full rack rate. As a result, the composite weighted PV ratio had significantly declined, indicating this demise in relative profitability. Further, to illustrate the importance of considering buyers' perceptions of prices, the price increases had primarily been at the full rack rate, creating a greater and more perceptible difference between, for example, the full rate and the corporate rate. More and more guests were noticing this widening price difference and were requesting and receiving the lower rates.

When there are differences in the PVs among products in a line, a revision in the product selling mix may be more effective than an increase in prices. That is, a firm, by shifting emphasis to those products with relatively higher PVs, has a good opportunity to recover some or all of its profit position. Hence, profit at any sales level is a function of prices, volume, costs and the product dollar sales mix (Monroe and Mentzer 1994).

5. TYPES OF PRICING DECISIONS

As shown in Table 1, there are many kinds of pricing decisions that a firm must make. There is the decision on the specific price to set for each product and service marketed. But this specific price depends on the type of customer to whom the product is sold. For example, if different customers purchase in varying quantities, should the seller offer volume discounts?

The firm must also decide whether to offer discounts for early payment and, if so, when a customer is eligible for a cash discount and how much to allow for early payment. Should the firm attempt to suggest retail or resale prices, or should it only set prices for its immediate customers? When a firm uses other business firms as intermediaries between itself and its customers, it does not have direct contact with its ultimate customers. Yet the way customers respond to prices depends on how they perceive the prices and the relationships between prices. Hence, the manufacturer is very interested in having the prices that the final customer responds to correspond to its strategic objectives.

Normally, the firm sells multiple products and these questions must be answered for each product or service offered. Additionally, the need to determine the number of price offerings per type of product and the price relationships between the products offered make the pricing problem more complex. Further, different types of market segments respond differently to prices, price differences, and price changes.

The firm must also decide on whether it will charge customers for the transportation costs incurred when shipping the products to them. Customers located at different distances from the manufacturer will pay a different total price if they are charged for the transportation costs. Yet if the seller quotes a uniform price that includes the transportation costs regardless of distance from the manufacturer, some buyers may receive the products at a total price that is less than the costs incurred to the manufacturer while other customers will pay a price that exceeds the total costs incurred by the manufacturer. Such differences in prices to similar types of customers may lead to some concerns about the legality of the pricing policy.

5.1. Pricing New Products

One of the most interesting and challenging decision problems is that of determining the price of a new product or service. Such pricing decisions are usually made with very little information on demand, costs, competition, and other variables that may affect the chances of success. Many new products fail because they do not provide the benefits desired by buyers, or because they are not available at the right time and place. Others fail because they have been incorrectly priced, and the error can as easily be in pricing too low as in pricing too high. Pricing decisions usually have to be made with little knowledge and with wide margins of error in the forecasts of demand, cost, and competitors' capabilities.

The core of new product pricing takes into account the price sensitivity of demand and the incremental promotional and production costs of the seller. What the product is worth to the buyer, not what it costs the seller, is the controlling consideration. What is important when developing a

TABLE 1 Basic Pricing Decisions

1. What to charge for the different products and services marketed by the firm
2. What to charge different types of customers
3. Whether to charge different types of distributors the same price
4. Whether to give discounts for cash and how quickly payment should be required to earn them
5. Whether to suggest resale prices or only set prices charged one's own customers
6. Whether to price all items in the product line as if they were separate or to price them as a "team"
7. How many different price offerings to have of one item
8. Whether to base prices on the geographical location of buyers (i.e., whether to charge for transportation)

Source: Monroe 1990, p. 278.

new product's price is the relationship between the buyers' perceived benefits in the new product relative to the total acquisition cost (i.e., financial value), and relative to alternative offerings available to buyers.

It has been generally presumed that there are two alternatives in pricing a new product: *skimming* pricing, calling for a relatively high price, and *penetration* pricing, calling for a relatively low price. There are intermediate positions, but the issues are made clearer by comparing the two extremes.

5.1.1. Skimming Pricing

Some products represent drastic improvements upon accepted ways of performing a function or filling a demand. For these products, a strategy of high prices during market introduction (and lower prices at later stages) may be appropriate. Skimming is not always appropriate and does have drawbacks. A skimming strategy is less likely to induce buyers into the market and does not encourage rapid adoption or diffusion of the product. Moreover, if skimming results in relatively high profit margins, competitors may be attracted into the market.

5.1.2. Penetration Pricing

A penetration strategy encourages both rapid adoption and diffusion of new products. An innovative firm may thus be able to capture a large market share before its competitors can respond. One disadvantage of penetration, however, is that relatively low prices and low profit margins must be offset by high sales volumes. One important consideration in the choice between skimming and penetration pricing at the time a new product is introduced is the ease and speed with which competitors can bring out substitute offerings. If the initial price is set low enough, large competitors may not feel it worthwhile to make a big investment for small profit margins. One study has indicated that a skimming pricing strategy leads to more competitors in the market during the product's growth stage than does a penetration pricing strategy (Redmond 1989).

5.2. Pricing During Growth

If the new product survives the introductory period, as demand grows, usually a number of competitors are producing and selling a similar product and an average market price begins to emerge. Normally there is a relatively wide range of market prices early in the growth stage, but this market price range narrows as the product approaches maturity.

In regard to pricing products during the growth stage, three essential points should be noted: (1) the range of feasible prices has narrowed since the introductory stage; (2) unit variable costs may have decreased due to the experience factor; and (3) fixed expenses have increased because of increased capitalization and period marketing costs. The pricing decision during the growth stage is to select a price that, subject to competitive conditions, will help generate a sales dollar volume that enables the firm to realize its target profit contribution.

5.3. Pricing During Maturity

As a product moves into the maturity and saturation stages, it is necessary to review past pricing decisions and determine the desirability of a price change. Replacement sales now constitute the major demand, and manufacturers may also incur regional competition from local brands. Market conditions do not appear to warrant a price increase, hence the pricing decision usually is to reduce price or stand pat.

When is a price reduction profitable? We know that when demand is price elastic it is profitable to reduce prices if costs do not rise above the increase in revenues. But since it can be expected that any price decrease will be followed by competitors, it is also necessary that the market demand curve be elastic within the range of the price reduction. Moreover, the requirements for a profitable price reduction strategy include beginning with a relatively high contribution margin (i.e., relatively high PV ratio), opportunity for accelerating sales growth and a price-elastic demand (Monroe and Mentzer 1994). When a product has reached the maturity stage of its life cycle, it is most likely that these conditions will not exist.

At the maturity stage of the life cycle, the firm probably should attempt to maximize short-run direct product contribution to profits. Hence, the pricing objective is to choose the price alternative leading to maximum contribution to profits. If competition reduces prices, the firm may, however reluctantly, match the price reduction. On the other hand, it may try to reduce costs by using cheaper materials, eliminating several labor operations, or reducing period marketing costs. All or any of these actions may allow the firm to match competitively lower prices and still maintain target contributions to profit.

5.4. Pricing a Declining Product

During the declining phase of a product's life, direct costs are very important to the pricing decision. Normally, competition has driven the price down close to direct costs. Only those sellers who were able to maintain or reduce direct costs during the maturity stage are likely to have remained. If the decline is not due to an overall cyclical decline in business but to shifts in buyer preferences, then the primary objective is to obtain as much contribution to profits as possible.

So long as the firm has excess capacity and revenues exceed all direct costs, the firm probably should consider remaining in the market. Generally, most firms eliminate all period marketing costs (or as many of these costs as possible) and remain in the market as long as price exceeds direct variable costs. These direct variable costs are the minimally acceptable prices to the seller. Thus, with excess capacity, any market price above direct variable costs would generate contributions to profit. Indeed, the relevant decision criterion is to maximize contributions per sales dollar generated. In fact, it might be beneficial to raise the price of a declining product to increase the contributions per sales dollar. In one case, a firm raised the price of an old product to increase contributions while phasing it out of the product line. To its surprise, sales actually grew. There was a small but profitable market segment for the product after all!

5.5. Price Bundling

In marketing, one widespread type of multiple products pricing is the practice of selling products or services in packages or bundles. Such bundles can be as simple as pricing a restaurant menu for either dinners or the items à la carte, or as complex as offering a ski package that includes travel, lodging, lift and ski rentals, and lessons. In either situation, there are some important principles that need to be considered when bundling products at a special price.

5.5.1. Rationale for Price Bundling

Many businesses are characterized by a relatively high ratio of fixed to variable costs. Moreover, several products or services usually can be offered using the same facilities, equipment, and personnel. Thus, the direct variable cost of a particular product or service is usually quite low, meaning that the product or service has a relatively high profit–volume (PV) ratio. As a result, the incremental costs of selling additional units are generally low relative to the firm's total costs.

In addition, many of the products or services offered by most organizations are interdependent in terms of demand, either being substitutes for each other or complementing the sales of another offering. Thus, it is appropriate to think in terms of relationship pricing, or pricing in terms of the inherent demand relationships among the products or services. The objective of price bundling is to stimulate demand for the firm's product line in a way that achieves cost economies for the operations as a whole, while increasing net contributions to profits.

5.5.2. Principles of Price Bundling

Underlying the notion of bundling is the recognition that different customers have different perceived values for the various products and services offered (Simon et al. 1995). In practical terms, these customers have different maximum amounts they would be willing to pay for the products. For some customers, the price of the product is less than this maximum acceptable price (upper price threshold), resulting in some consumer surplus. However, for these customers, the price of a second product or service may be greater than they are willing to pay and they do not acquire it. If the firm, by price bundling, can shift some of the consumer surplus from the highly valued product to the less valued product, then there is an opportunity to increase the total contributions these products make to the firm's profitability.

The ability to transfer consumer surplus from one product to another depends on the complementarity of demand for these products. Products or services may complement each other because purchasing them together reduces the search costs of acquiring them separately. It is economical to have both savings and checking accounts in the same bank to reduce the costs of having to go to more than one bank for such services. Products may complement each other because acquiring one may increase the satisfaction of acquiring the other. For the novice skier, ski lessons will enhance the satisfaction of skiing and increase the demand to rent skis. Finally, a full product-line seller may be perceived to be a better organization than a limited product-line seller, thereby enhancing the perceived value of all products offered.

5.6. Yield Management

A form of segmentation pricing that was developed by the airlines has been called yield management. Yield management operates on the principle that different segments of the market for airline travel have different degrees of price sensitivity. Therefore, seats on flights are priced differently depending on the time of day, day of the week, length of stay in the destination city before return, when the ticket is purchased, and willingness to accept certain conditions or restrictions on when to travel. Besides the airlines, hotels, telephone companies, rental car companies, and banks and savings and loans have used yield management to increase sales revenues through segmentation pricing. It seems likely that retail firms could use yield management to determine when to mark down slow-moving merchandise and when to schedule sales, or how to set a pricing strategy for products with short selling seasons.

The unique benefits of the yield-management pricing program is it forces management to continuously monitor demand for its products. Further, changes in demand lead to pricing changes to reflect these changes. If the product is not selling fast enough, then price reductions can be initiated to stimulate sales. Because of a relatively high fixed costs to total variable costs cost structure, these focused price reductions can provide leverage for increasing operating profits, by the effect on sales volume. With relatively high contribution margins (high PV ratios), small price reductions do not require large increases in volume to be profitable.

6. ADMINISTERING THE PRICING FUNCTION

Perhaps the most difficult aspect of the pricing decision is to develop the procedures and policies for administering prices. Up to this point, the issue has been on the setting of base or list prices. However, the list price is rarely the actual price paid by the buyer. The decisions to discount from list price for volume purchases or early payment, to extend credit, or to charge for transportation effectively change the price actually paid. In this section, we consider the problems of administering prices.

An important issue relative to pricing is the effect that pricing decisions and their implementation have on dealer or distributor cooperation and motivation as well as on salespeople's morale and effort. While it is difficult to control prices legally through the distribution channel, nevertheless, it is possible to elicit cooperation and provide motivation to adhere to company-determined pricing policies. Also, because price directly affects revenues of the trade and commissions of salespeople, it can be used to foster desired behaviors by channel members and salespeople.

Price administration deals with price adjustments or price differentials for sales made under different conditions, such as:

1. Sales made in different quantities
2. Sales made to different types of middlemen performing different functions
3. Sales made to buyers in different geographic locations
4. Sales made with different credit and collection policies
5. Sales made at different times of the day, month, season, or year

Essentially, price structure decisions define how differential characteristics of the product and/or service will be priced. These price structure decisions are of strategic importance to manufacturers, distributors or dealers, and retailers (Marn and Rosiello 1992). In establishing a price structure, there are many possibilities for antagonizing distributors and even incurring legal liability. Thus, it is necessary to avoid these dangers while at the same time using the price structure to achieve the desired profit objectives.

We have noted that the definition of price includes both the required monetary outlay by the buyer, plus a number of complexities including terms of agreement, terms of payment, freight charges, offering a warranty, timing of delivery, or volume of the order. Moreover, offering different products or services in the line with different features or benefits at different prices permits the opportunity to develop prices for buyers who have different degrees of sensitivity to price levels and price differences. Moving from a simple one-price-for-all-buyers structure to a more complex pricing structure provides for pricing flexibility because the complexity permits price variations based on specific product and service characteristics as well as buyer or market differences. Moreover, a more complex price structure enhances the ability of firms to (Stern 1986):

1. Respond to specific competitive threats or opportunities
2. Enhance revenues while minimizing revenue loss due to price changes
3. Manage the costs of delivering the product or service
4. Develop focused price changes and promotions
5. Be more effective in gaining distributors' cooperation

To accomplish this goal of differential pricing requires identifying the key factors that differentiate price-market segments. Then the elements of the price structure may be developed that reflect these factors.

A product's list price is the product's price to final buyers. Throughout the distribution system, manufacturers grant intermediaries discounts, or deductions from the list price. These price concessions from producers may be seen as payment to intermediaries for performing the distribution function and for providing time and place utilities. The difference between the list price and the amount that the original producer receives represents the total discounts provided to channel members.

Channel members themselves employ discounts in various ways. Wholesalers pass on discounts to retailers just as manufacturers pass along discounts to wholesalers. Retailers may offer promotional discounts to consumers in the form of sweepstakes, contests, and free samples. Some stores offer quantity and cash discounts to regular customers. Even seasonal discounts may be passed along—for example, to reduce inventory of Halloween candy or Christmas cards.

7. PRICE AND SALES PROMOTIONS

Price promotions consist of tactics designed to achieve specific objectives in a limited time in a target market. Price promotions are directed at two primary audiences: consumers and the trade. Some commonly used promotion tactics are summarized next.

Aimed at both consumers and distributors, price-off promotions involve temporary price reductions to retailers with the intent that savings will be passed along to consumers. In an immediate price-off promotion, the marketer usually affixes a special label to the product's package to indicate the percentage or cash savings, which the retailer then honors. In a delayed price-off promotion, the consumer has to pay the normal price but can send the marketer a proof-of-purchase label to claim a cash rebate (refund).

Coupons offer buyers price reductions at the point of sale (most often at a retail checkout counter). Coupons are made available to consumers in a variety of ways. Some can be clipped from daily newspaper ads for supermarkets; some come as colorful freestanding inserts in Sunday newspapers; and some are distributed as fliers to homes within a store's immediate vicinity. Coupons are also mailed directly to consumers, often in booklets or packets assembled by distributors or companies specializing in such promotions. Finally, in-pack and on-pack coupons may be affixed to the product and allow savings on the current or a future purchase.

7.1. Price Promotion as Price Segmentation

As mentioned earlier, manufacturers and retailers have made increasing use of coupons, rebates and refunds, and short-term price reductions to stimulate short-term demand for their products and services. However, despite the popularity of these price deals, it is not at all clear that a majority of these deals are profitable. Simply offering a coupon, rebate, or price-off deal to current buyers in the hope of stimulating a substantial increase in unit sales likely will reduce profits unless mostly new buyers take advantage of the deal. Thus, the underlying objective of a price promotion to final customers should be to focus on a price-market segment that is more price sensitive or deal responsive than existing buyers. Essentially, then, price promotions to final customers simply becomes an aspect of segment pricing.

Not all buyers of a product on price promotion take advantage of the promotion. Redemption rates for coupons and rebates indicate that many buyers do not redeem coupons, nor do they follow through and take the rebate that is offered. One reason for this lack of redemption or follow-through is that the perceived cost of redeeming coupons or qualifying for the rebate is larger than the perceived value of the coupon or rebate. Thus, buyers who do not take advantage of the price promotion pay the regular price. Therefore, by offering a coupon or rebate, the seller has effectively segmented the market into two price segments.

For this type of segmented pricing to be profitable, some additional conditions must prevail. First, the segments must be separable to some degree. This separation must either be due to some natural separation, such as geographic region or location, time of purchase, or category of buyer, e.g., business buyer vs. consumer. Second, these segments must have different degrees of price sensitivity and/or different variable purchasing costs. Third, the variable costs of selling to these segments must not be so different that the effect of the lower price is not canceled by increased selling costs.

7.2. Price Promotion Decision Variables

To understand the complexity of developing a price promotion policy, it is useful to review the types of decisions that need to be made before a promotion can be implemented. The decisions outlined below indicate there is a need to plan this activity carefully.

1. Should a price promotion be offered? The first decision is simply whether the seller should offer a coupon, rebate, cents-off deal, or promotional discount.
2. To whom should the deal be offered: dealers or final customers? Offering a price promotion to dealers or distributors in anticipation that the deal will be offered to final customers is not the same as offering it directly to final customers. Dealers may choose to reduce their prices in the full amount or less, or not at all. Further, they may buy excess amounts of the deal merchandise, relative to actual demand, and sell some units at full price after the deal period is over.
3. When should the promotion be offered? This question refers not only to the specific time of the year but also to whether to offer the deal during peak or slack selling seasons. For example,

should ketchup manufacturers offer price promotions during the peak outdoor cooking season (May–September) or during the winter months?

4. How frequently should a promotion be run? If a product is offered frequently on some form of a price deal, customers may come to expect the deal and only buy on deal. The effect is that the product is rarely sold at full list price and the advantage of price segmentation is lost.
5. How long should the promotion be run? This decision is related to the issue of promotion frequency, peak vs. slack selling periods, and the length of buyers' repurchase periods. Some consumer coupons do not have an expiration date, thereby extending the promotion period indefinitely and making it difficult to measure the effectiveness of the coupon promotion.
6. How many units should be covered by the promotion? When deals are offered to the trade, there often is a restriction on the amount that a dealer can order on deal. This restriction may be related to the dealer's usual order size for the length of the promotion period.
7. What products and/or sizes should be promoted? Should a coffee producer offer a coupon on all package sizes of coffee, or only the 13 oz. size? Should a luggage manufacturer feature all of its luggage in the promotion, or only certain models?
8. How much should the price be reduced? What should be the amount of the rebate? What should be the value of the coupon? As discussed earlier, the degree of buyer sensitivity to the price reduction, or the degree that buyers perceive additional transaction value, will have an impact on the success of the promotion. The important issue here is how much of a price difference (reduction) is necessary to induce buyers to take advantage of the offer.

7.3. Some Perspectives on Price and Sales Promotions

Promotions comprise a significant portion of the marketing communications budget. With this increased managerial importance of price and sales promotions has come considerable research on how promotions affect sales. However, much of this new research information is still quite recent, and the evidence on how promotions affect sales is still emerging. In this section we will look at three different time frames to consider these effects (Blattberg and Neslin 1989): (1) the week or weeks in which the promotion occurs (immediate); (2) the weeks or months following the promotion (intermediate); and (3) the months or years following the implementation of several promotions (long term).

7.3.1. Immediate Effects of Price and Sales Promotions

Promotions seem to have a substantial immediate impact on brand sales. For example, price cuts for bathroom tissue are accompanied by immediate increases in brand sales. When such price promotions are coordinated with special point-of-purchase displays and local feature advertising, sales might increase as much as 10 times normal sales levels. Because of such observable immediate sales impact, many brands of packaged consumer goods are frequently promoted.

A large proportion of this immediate increase in sales is due to nonloyal buyers (brand switchers). For example, one study showed that 84% of the increase in coffee brand sales generated by promotions came from brand-switching consumers (Gupta 1988). However, not all brands have the same capability of inducing consumers to switch brands with a promotional activity.

The effect of brand A's promotions on brand B's sales likely is different than the effect of brand B's promotions on brand A's sales. This asymmetry of the promotion cross-elasticities is a very important managerial finding. That is, a strong brand, say brand A, may be more successful in inducing buyers of brand B to switch to brand A with a promotion than the weaker brand B would be in using a similar promotion to induce buyers of brand A to switch to brand B. This finding implies that when there are simultaneous promotions by both brands, brand A likely will experience a more positive sales impact than will brand B.

Another important finding is that different forms of price and sales promotions have separate effects on sales. Further, when several forms are used together, the total impact may be greater than the sum of the effects due to each form.

An important reason for the immediate sales effect is purchase acceleration, which occurs because loyal consumers buy larger quantities when the brand is promoted and/or purchase the brand sooner than normal. The issue of purchase acceleration is very important when we try to determine whether a promotion has been profitable. Indeed, if buyers do not change their rate of consumption of the product, then larger purchase quantities or earlier purchases means that there will be fewer sales later at the regular price.

7.3.2. Intermediate Effects of Price and Sales Promotions

Usually, promotions have been considered short-term tactics to provide for an immediate increase in sales. However, as in advertising, there are effects that occur even after a particular promotion campaign has expired. For example, if a consumer purchases brand A for the first time when A is being promoted, will that consumer buy brand A on the next purchase occasion (repeat purchase)? Or will the consumer develop a loyalty to the deal and look for another promoted brand on the next purchase occasion? The managerial implications of a promotion extend beyond the immediate sales effect of that promotion.

While there has been considerable research on whether brand purchasing enhances repeat brand purchases, such research provides conflicting results. There is some evidence that a prior brand purchase may increase the likelihood of buying that brand again. However, given the increasing use of promotions, other research shows no evidence that promotions enhance repeat brand purchases.

A reason why promotions may not enhance repeat brand purchases involves the question of why consumers may decide to buy the brand initially. It has been suggested that people may attribute their purchase to the deal that was available and not to the brand itself being attractive (Scott and Yalch 1980). If consumers indicate a negative reason for a purchase decision, there is a smaller likelihood that the experience will be a positive learning experience relative to the brand itself. If a brand is promoted quite frequently, then the learning to buy the brand occurs because of the reward of the deal, not the positive experience of using the brand itself.

7.3.3. Long-Term Effects of Price and Sales Promotions

In the long run, the relevant issue is the ability of the brand to develop a loyal following among its customers (Jones 1990). The important objective is to develop favorable brand attitudes of the customers such that they request that the retailer carry the brand and are unwilling to buy a substitute. As we observed above, many marketing practitioners fear that too-frequent promotional activity leads to loyalty to the deal, thereby undermining the brand's franchise. And if buyers come to expect that the brand will often be on some form of a deal, they will not be willing to pay a regular price and will be less likely to buy because they do not believe that the brand provides benefits beyond a lower price.

A second problem develops if consumers begin to associate a frequently promoted brand with lower product quality. That is, if consumers believe that higher-priced products are also of higher quality, then a brand that is frequently sold at a reduced price may become associated with a lower perceived quality level. Thus, the result may be that buyers form a lower reference price for the product. Products that are perceived to be of relatively lower quality have a more difficult time trying to achieve market acceptance.

8. LEGAL ISSUES IN PRICING

The development of a price structure to implement a pricing strategy is not only a difficult and complex task but is fraught with the potential for violating federal and state laws. In fact, the legal aspects of pricing strategy comprise one of the most difficult parts of marketing strategy and have left many business people not only frightened of making pricing decisions but often vulnerable to legal action because of their pricing activities.

As indicated earlier, the short-term effects of reducing price on profits likely will not be positive. Yet many firms attempting to gain volume or market share often resort to price reductions. Other competing firms often feel compelled to follow these price reductions and often to undercut the original price-reducing firm. Also, there is pressure, real or perceived, to provide certain buyers a favored status by giving them additional discounts for their business. To counteract these pressures to reduce prices and stabilize profits, some businesses have attempted by either overt or covert actions to stabilize prices and market share. In other situations, a larger firm has aggressively reduced prices in one market area or to a specific set of customers in order to reduce competition or drive a competitor out of business. Moreover, we have suggested that there are typically several price-market segments distinguished by different degrees of sensitivity to prices and price differences. Thus, the opportunity exists to set different prices or to sell through different channels to enhance profits.

Each of these various possible strategies or tactics is covered by some form of legislation and regulation. In fact, there are laws concerning price fixing amongst competitors, exchanging price information or price signalling to competitors, pricing similarly to competitors (parallel pricing), predatory pricing, and price discrimination. This section briefly overviews the laws that cover these types of activities.

8.1. Price Fixing

The Sherman Antitrust Act (1890) specifically addresses issues related to price fixing, exchanging price information, and price signaling. It also has an effect on the issue of predatory pricing. Section 1 of the Act prohibits all agreements in restraint of trade. Generally, violations of this section are divided into per se violations and rule of reason violations. Per se violations are automatic. That is, if a defendant has been found to have agreed to fix prices, restrict output, divide markets by agreement, or otherwise act to restrict the forces of competition, he or she is automatically in violation of the law and subject to criminal and civil penalties. There is no review of the substance of the situation, that is, whether there was in fact an effect on competition. In contrast, the rule-of-reason doctrine calls for an inquiry into the circumstances, intent, and results of the defendants' actions. That is, the courts will examine the substantive facts of the case, including the history, the reasons for the actions, and the effects on competition and the particular market. The current attitude of federal and state agencies and courts is that price fixing is a per se violation and that criminal sanctions should be applied to the guilty persons.

Section 2 of the Act prohibits the act of monopolizing, that is, the wrongful attempt to acquire monopoly power. Thus, having a monopoly is not illegal, but the deliberate attempt to become a monopoly is illegal. The issue here in recent cases has been not whether a firm had acquired a monopoly per se, but the methods of achieving such market power. Thus, the courts have been somewhat more aware of a firm's need to develop a strong competitive position in the markets it serves and that these strong competitive actions may lead to dominant market shares. However, if this dominant market share leads to above-market-average prices, some form of legal or regulatory action may take place.

8.2. Exchanging Price Information

Many trade associations collect and disseminate price information from and to their members. The legal issue arises when there is an apparent agreement to set prices based on the exchanged price information. Thus, if members discuss prices and production levels in meetings and prices tend to be uniform across sellers, it is likely that the exchange of information led to some form of price fixing. Again, the issue is whether the exchange of price information seems to have the effect of suppressing or limiting competition, which is a violation of section 1 of the Sherman Act. Care must be exercised about when and how price information is exchanged by competing sellers. The trend in recent years has been to make it more difficult to prove that such exchanges do not violate section 1.

8.3. Parallel Pricing and Price Signaling

In many industries and markets, one firm may emerge as a price leader. That is, one firm often is the first to announce price changes, and most rival sellers will soon follow the price changes made by the leader. At other times, another firm may initiate the price changes, but if the price leader does not introduce similar price changes, the other firms as well as the initial firm will adjust their prices to correspond to the price leader's prices. The legal question arises as to whether these somewhat concerted prices and price changes constitute a tacit, informal, and illegal agreement in violation of section 1.

Recently, some questions have been raised as to whether the public announcement of prices and price changes has been used by sellers to signal prices and achieve this common understanding about prices. That is, do sellers achieve this common understanding about prices through public announcements about their prices and price changes? If so, then a violation of section 1 may exist. For example, if one company announces a price increase effective 60 days later, some legal people have suggested that the announcement serves as a signal to competition. The suggestion seems to imply that if others follow with similar announcements, then the price increase will remain in effect; if others do not follow, the price increase will be rescinded. Announcing price increases ahead of the effective date provides time for customers, distributors, and the sales force to adjust their price frame of reference. However, what may be an effective managerial practice could be interpreted as a mechanism for attempting to achieve a common understanding among rival sellers. The ramifications of price signaling from a legal perspective remain to be determined by either legislative action, litigation, or further debate.

8.4. Predatory Pricing

Predatory pricing is the cutting of prices to unreasonably low and/or unprofitable levels so as to drive competitors from the market. If this price cutting is successful in driving out competitors, then the price cutter may have acquired a monopoly position via unfair means of competition—a violation of section 2 of the Sherman Act.

There is considerable controversy about predatory pricing, particularly how to measure the effect of a low-price strategy on the firm's profits and on competitors. Predatory pricing occurs whenever the price is so low that an equally efficient competitor with smaller resources is forced from the market or discouraged from entering it. Primarily, the effect on the smaller seller is one of a drain on cash resources, not profits per se. Much of the controversy surrounding measuring the effects of an alleged predatory price relates to the proper set of costs to be used to determine the relative profitability of the predator's actions. Recently, the courts seem to have adopted the rule that predatory pricing exists if the price does not cover the seller's average variable or marginal costs. However, the intent of the seller remains an important consideration in any case.

8.5. Illegal Price Discrimination

The Robinson-Patman Act was passed in 1936 to protect small independent wholesalers and retailers from being charged more for products than large retail chains were. Sellers can, however, defend themselves against discrimination charges by showing that their costs of selling vary from customer to customer. Cost savings can then be passed along to individual buyers in the form of lower prices. The Robinson-Patman Act therefore permits price discrimination in the following cases:

1. If the firm can demonstrate that it saves money by selling large quantities to certain customers, it can offer these buyers discounts equal to the amount saved.
2. Long-term sales agreements with customers also reduce costs; again, discounts may be granted equal to the amount saved.
3. In cases where competitors' prices in a particular market must be met, it is legal to match the competitors' lower prices in such a market while charging higher prices in other markets.

9. FOUR BASIC RULES FOR PRICING

The four rules given below are intended to capture the essence of the analysis necessary to determine and evaluate pricing strategies. The order in which the rules are presented does not imply a hierarchy of importance; each rule is equally important.

9.1. Know Your Objectives

As demonstrated earlier, many firms stress the profit objective of return on investment. Other firms stress the objective of maintaining specified profit margins, while still other firms seek to achieve market-share goals. It is not necessary for each product to maintain the same profit margin in order to achieve a particular return on investment. Similarly, different margins on products may still produce an overall desired corporate profit goal. Finally, firms stressing market share may utilize the experience curve factor and build profits by reducing prices.

The important point to remember is that differences in corporate profit objectives eventually will lead to differences in prices and the role of price in influencing actual profits. Ultimately, regardless of the financial goal, the pricing objective is behavioral in nature. That is, whether buyers buy more, whether nonbuyers decide to buy now, and whether buyers decide to purchase less frequently but in greater volume per order, or to pay earlier, are influenced by the prices and price structure of the seller. Further, the degree to which distributors and dealers are cooperative and motivated to sell the firm's products depends largely on the financial incentives provided by the suppliers' prices and price structure. Also, the sales force's motivation to help the firm achieve its financial objectives depends on its understanding and acceptance of the pricing strategy being followed. Price has an important role in developing incentives for distributors, salespeople, and buyers to perform in ways that will be beneficial to the firm. Thus, it is important that the seller develop a positive attitude toward pricing, leading to a proactive pricing approach.

9.2. Know Your Demand

This prescription suggests that the firm understand fully the factors influencing the demand for its products and services. The key question is the role of price in the purchaser's decision process. Price and price differentials influence buyer perceptions of value. Indeed, many companies have achieved positive results from differentially pricing their products and services.

Coupled with knowing how price influences buyers' perceptions of value, it is necessary to know how buyers use the product or service. Is the product used as an input in the buyer's production process? If so, does the product represent a significant or insignificant portion of the buyer's manufacturing costs? If the product is a major cost element in the buyer's production process, then small changes in the product's price may significantly affect the buyer's costs and the resulting price of the manufactured product. If the final market is sensitive to price increases, then a small price increase to the final manufacturer may significantly reduce demand to the initial seller of the input material. Thus, knowing your buyers also means understanding how they react to price changes and price differentials as well as knowing the relative role price plays in their purchase decisions.

Further, the seller should also know the different types of distributors and their functions in the distribution channel. This prescription is particularly important when the manufacturer sells both to distributors and to the distributors' customers.

9.3. Know Your Competition and Your Market

In addition to the influence of buyers, a number of other significant market factors influence demand. It is important to understand the operations of both domestic and foreign competitors, their rate of capacity utilization, and their products and services. In many markets, the dynamic interaction of supply and demand influences prices. Moreover, changes in capacity availability due to capital investment programs will influence supply and prices. A second important aspect of knowing the market is the need to determine price–volume relationships.

9.4. Know Your Costs

It is important to determine the basic cost data necessary for the pricing decision. As stated earlier, it is necessary to know which costs vary directly with changes in levels of activity and the underlying causes of the changes in costs. It is also necessary to identify the costs that are directly related to the product or service being costed but do not vary with activity levels-direct period or fixed costs. Furthermore, marketing and distribution costs should be objectively assigned to the products and not simply lumped into a general overhead category.

Valid cost data provide an objective basis for choosing between pricing alternatives, determining discounts, and establishing differential pricing alternatives. Furthermore, objective cost studies that are completed before the pricing decisions provide the firm with a valid legal justification for its price structure.

10. SUMMARY

It is important to realize there is no one right way to determine price. Pricing simply cannot be reduced to a formula—there are too many interacting factors. Successful pricing requires considering all internal and external factors and adapting to changes as they occur. Successful pricing is adaptive pricing. Pricing decisions should be logically made and should involve rigorous thinking, with minimum difficulty from human and organizational factors. Further, it should be recognized that judgment and prediction are needed about the future, not the past. Finally, pricing decisions should be made within a dynamic, long-run corporate and marketing strategy.

REFERENCES

Blattberg, R. C., and Neslin, S. (1989), "Sales Promotion: The Long and Short of It," *Marketing Letters,* Vol. 1, December, pp. 81–97.

Chicago Tribune (2000), "Survey: More Buying Books, Video On-line Than in Stores," January 11, Section 4, p. 2.

Cressman, G. E. (1997), "Snatching Defeat From the Jaws of Victory: Why Do Good Managers Make Bad Pricing Decisions?" *Marketing Management,* Vol. 6, Summer, pp. 9–19.

Cooper, J. C., and Madigan, K. (1999), "Happy New Year, But for How Long?" *Business Week,* December 27, pp. 49–50.

Green, H. (1998), "A Cybershopper's Best Friend," *Business Week,* May 4, p. 84.

Gupta, S. (1988), "Impact of Sales Promotions on When, What, and How Much to Buy," *Journal of Marketing Research,* Vol. 25, November, pp. 342–355.

Hof, R. D. (1999), "The Buyer Always Wins," *Business Week,* March 22, pp. EB 26–28.

Jones, J. P. (1990), "The Double Jeopardy of Sales Promotions," *Harvard Business Review,* Vol. 68, September–October, pp. 145–152.

Leszinski, R., and Marn, M. V. (1997), "Setting Value, Not Price," *McKinsey Quarterly,* No. 1, pp. 98–115.

Marn, M. V., and Rosiello, R. L. (1992), "Managing Price, Gaining Profit," *Harvard Business Review,* Vol. 70, September–October, pp. 84–94.

Monroe, K. B. (1990), *Pricing: Making Profitable Decisions,* 2nd Ed., McGraw-Hill, New York.

Monroe, K. B. (1993), "Pricing Practices which Endanger Profits," *Pricing Strategy & Practice,* Vol. 1, No. 1, pp. 4–10.

Monroe, K. B., and Mentzer, J. T. (1994), "Some Necessary Conditions for When Price Reduction Strategies May Be Profitable," *Pricing Strategy and Practice,* Vol. 2, No. 1, pp. 11–20.

Ness, J. A., and Cucuzza, T. G. (1995), "Tapping the Full Potential of ABC," *Harvard Business Review,* Vol. 73, July–August, pp. 130–138.

Redmond, W. H. (1989), "Effects of New Product Pricing on the Evolution of Market Structure," *Journal of Product Innovation Management,* Vol. 6, pp. 99–108.

Schmeltzer, J. (2000), "United Airlines Moving to Get Friendlier with Web Commerce," *Chicago Tribune,* January 11, Section 4, p. 2.

Scott, C. A., and Yalch, R. F. (1980), "Consumer Response to Initial Trial: A Bayesian Analysis," *Journal of Consumer Research,* Vol. 7, June, pp. 32–41.

Simon, H., Fassnacht, M., and Wubker, G. (1995), "Price Bundling," *Pricing Strategy and Practice,* Vol. 3, No. 1, pp. 34–44.

Sparks, D. (1999), "Who Will Survive the Internet Wars?" *Business Week,* December 27, 98–100.

Stern, A. A. (1986), "The Strategic Value of Price Structure," *Journal of Business Strategy,* Vol. 7, Fall, pp. 22–31.

CHAPTER 25

Mass Customization

MITCHELL M. TSENG
Hong Kong University of Science and Technology

JIANXIN JIAO
Nanyang Technological University

1. INTRODUCTION 684
  1.1. Concept Implication 685
  1.2. Technical Challenges 686
    1.2.1. Maximizing Reusability 686
    1.2.2. Product Platform 686
    1.2.3. Integrated Product Life Cycle 687
2. DESIGN FOR MASS CUSTOMIZATION 687
  2.1. Product Family 688
    2.1.1. Modularity and Commonality 688
    2.1.2. Product Variety 689
  2.2. Product Family Architecture 690
    2.2.1. Composition of PFA 690
    2.2.2. Synchronization of Multiple Views 691
  2.3. Product Family Design 692
3. MASS CUSTOMIZATION MANUFACTURING 694
  3.1. Managing Variety in Production Planning 694
  3.2. Coordination in Manufacturing Resource Allocation 697
  3.3. High-Variety Shop-Floor Control 699
4. SALES AND MARKETING FOR MASS CUSTOMIZATION 701
  4.1. Design by Customers 701
  4.2. Helping Customers Making Informed Choices: Conjoint Analysis 702
  4.3. Customer Decision-Making Process 703
    4.3.1. Phase I: Customer Needs Acquisition 703
    4.3.2. Phase II: Product Design 703
  4.4. One-to-One Marketing 704
5. MASS CUSTOMIZATION AND E-COMMERCE 705
6. SUMMARY 706
REFERENCES 706

1. INTRODUCTION

With the increasing competition in the global market, the manufacturing industry has been facing the challenge of increasing customer value. Much has been done to reduce costs and improve quality. Quality does not mean only conforming to specifications. More importantly, quality means ensuring customer satisfaction and enhancing customer value to the extent that customers are willing to pay for the goods and services. To this end, a well-accepted practice in both academia and industry is the exploration of flexibility in modern manufacturing systems to provide quick response to customers with new products catering to a particular spectrum of customer needs. Consequently, there is a growing trend toward increasing product variety, as evident in supermarkets. Various food and beverage companies are fighting for shelf space to display the explosive growth of product varieties. Rapidly changing design and product technologies further accentuate this trend. The key to success in the highly competitive manufacturing enterprise often is the company's ability to design, produce,

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

and market high-quality products within a short time frame and at a price that customers are willing to pay. These counterdemands for final products create enormous productivity challenges that threaten the very survival of manufacturing companies. In addition, increasingly high labor and land costs often put developed countries or regions at a disadvantage in attracting manufacturing plants comparing with neighboring developing countries. In order to meet these pragmatic and highly competitive needs of today's industries, it is imperative to promote high-value-added products and services (Ryan 1996). It was reported that 9 out of 10 bar code scanner vendors were planning to repackage their product offerings in 1997 to include a larger scope of value-added features and pursue application-specific solution opportunities (Rezendes 1997).

This chapter discusses the opportunities brought by mass customization for high-value-added products and services. Mass customization enhances profitability through a synergy of increasing customer-perceived values and reducing the costs of production and logistics. Therefore, mass customization inherently makes high-value-added products and services possible through premium profits derived from customized products. The chapter also introduces techniques of integrating product life-cycle concerns in terms of how to connect customer needs proactively with the capabilities of a manufacturer or service provider during the product-development process. Major technical challenges of mass customization are also summarized.

1.1. Concept Implication

Mass customization is defined here as "producing goods and services to meet individual customer's needs with near mass production efficiency" (Tseng and Jiao 1996). The concept of *mass customization* was anticipated by Toffler (1971) and the term was coined by Davis (1987). Pine (1993) documented its place in the continuum of industrial development and mapped out the management implications for firms that decide to adopt it. Mass customization is a new paradigm for industries to provide products and services that best serve customer needs while maintaining near-mass production efficiency. Figure 1 illustrates the economic implications of mass customization (Tseng and Jiao 1996). Traditionally, mass production demonstrates an advantage in high-volume production, where the actual volume can defray the costs of huge investments in equipment, tooling, engineering, and training. On the other hand, satisfying each individual customer's needs can often be translated into higher value, in which, however, low production volume is unavoidable and thus may lend itself to becoming economically not viable. Accommodating companies to garner economy of scale through repetitions, mass customization is therefore capable of reducing costs and lead time. As a result, mass customization can achieve higher margins and thus be more advantageous. With the increasing flexibility built into modern manufacturing systems and programmability in computing and communication technologies, companies with low to medium production volumes can gain an edge over competitors by implementing mass customization.

In reality, customers are often willing to pay premium price for their unique requirements being satisfied, thus giving companies bonus profits (Roberts and Meyer 1991). From an economic perspective, mass customization enables a better match between the producers' capabilities and customer needs. This is accomplished through either developing the company's portfolio, which includes prod-

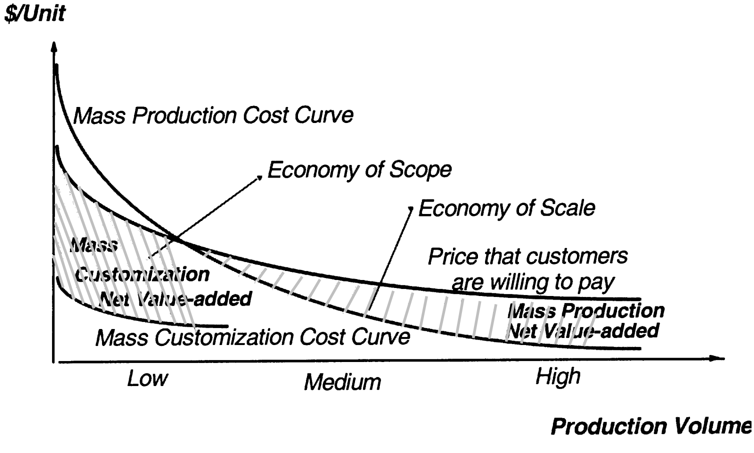


Figure 1 Economic Implications of Mass Customization.

ucts, services, equipment, and skills, in response to market demands, or leading customers to the total capability of the company so that customers are better served. The end results are conducive to improvement in resources utilization. Mass customization also has several significant ramifications in business. It can potentially develop customer loyalty, propel company growth, and increase market share by widening the product range (Pine 1993).

1.2. Technical Challenges

The essence of mass customization lies in the product and service providers' ability to perceive and capture latent market niches and subsequently develop technical capabilities to meet the diverse needs of target customers. Perceiving latent market niches requires the exploration of customer needs. To encapsulate the needs of target customer groups means to emulate existing or potential competitors in quality, cost, quick response. Keeping the manufacturing cost low necessitates economy of scale and development of appropriate production capabilities. Therefore, the requirements of mass customization depend on three aspects: time-to-market (quick responsiveness), variety (customization), and economy of scale (volume production efficiency). In other words, successful mass customization depends on a balance of three elements: features, cost, and schedule. In order to achieve this balance, three major technical challenges are identified as follows.

1.2.1. Maximizing Reusability

Maximal amounts of repetition are essential to achieve the efficiency of mass production, as well as efficiencies in sales, marketing, and logistics. This can be attained through maximizing commonality in design, which leads to reusable tools, equipment, and expertise in subsequent manufacturing. From a commercial viewpoint, mass customization provides diverse finished products that can be enjoyed uniquely by different customers. Customization emphasizes the differentiation among products. An important step towards this goal will be the development and proliferation of design repositories that are capable of creating various customized products. This product proliferation naturally results in the continuous accretion of varieties and thus engenders design variations and process changeovers, which seemingly contradict the pursuit of low cost and high efficiency of mass production. Such a setup presents manufacturers with a challenge of ensuring "dynamic stability" (Boynton and Bart 1991), which means that a firm can serve the widest range of customers and changing product demands while building upon existing process capabilities, experience, and knowledge. Due to similarity over product lines or among a group of customized products, reusability suggests itself as a natural technique to facilitate increasingly efficient and cost-effective product realization. Maximizing reusability across internal modules, tools, knowledge, processes, components, and so on means that the advantages of low costs and mass production efficiency can be expected to maintain the integrity of the product portfolio and the continuity of the infrastructure. This is particularly true in savings resulting from leveraging downstream investments in the product life cycle, such as existing design capabilities and manufacturing facilities.

Although commonality and modularity have been important design practices, these issues are usually emphasized for the purpose of physical design or manufacturing convenience. To achieve mass customization, the synergy of commonality and modularity needs to be tackled starting from the functional domain characterized by customer needs or functional requirements, and needs to encompass both the physical and process domains of design (Suh 1990). In that way, the reusability of both design and process capabilities can be explored with respect to repetitions in customer needs related to specific market niches.

1.2.2. Product Platform

The importance of product development for corporate success has been well recognized (Meyer and Utterback 1993; Roberts and Meyer 1991). The effectiveness of a firm's new product generation lies in (1) its ability to create a continuous stream of successful new products over an extended period of time and (2) the attractiveness of these products to the target market niches. Therefore, the essence of mass customization is to maximize such a match of internal capabilities with external market needs.

Towards this end, a product platform is impelled to provide the necessary taxonomy for positioning different products and the underpinning structure describing the interrelationships between various products with respect to customer requirements, competition information, and fulfillment processes. A product platform in a firm implicates two aspects: to represent the entire product portfolio, including both existing products and proactively anticipated ones, by characterizing various perceived customer needs, and to incorporate proven designs, materials, and process technologies.

In terms of mass customization, a product platform provides the technical basis for catering to customization, managing variety, and leveraging existing capabilities. Essentially, the product platform captures and utilizes reusability underlying product families and serves as a repertoire of knowledge bases for different products. It also prevents variant product proliferation for the same set of customer requirements. The formulation of product platform involves inputs from design concepts, process capabilities, skills, technological trends, and competitive directions, along with recognized customer requirements.

1.2.3. Integrated Product Life Cycle

Mass customization starts from understanding customers' individual requirements and ends with a fulfillment process targeting each particular customer. The achievement of time-to-market through telescoping lead times depends on the integration of the entire product-development process, from customer needs to product delivery. Boundary expansion and concurrency become the key to the integration of the product development life cycle from an organizational perspective. To this end, the scope of the design process has to be extended to include sales and service.

On the other hand, product realization should simultaneously satisfy various product life cycle concerns, including functionality, cost, schedule, reliability, manufacturability, marketability, and serviceability, to name but a few. The main challenge for today's design methodologies is to support these multiple viewpoints to accommodate different modeling paradigms within a single, coherent, and integrated framework (Subrahmanian et al. 1991).

In other words, the realization of mass customization requires not only integration across the product development horizon, but also the provision of a context-coherent integration of various viewpoints of product life cycle. It is necessary to employ suitable product platforms with unified product and product family structure models serving as integration mechanisms for the common understanding of general construction of products, thereby improving communication and consistency among different aspects of product life cycle.

2. DESIGN FOR MASS CUSTOMIZATION

Design has been considered as a critical factor to the final product form, cost, reliability, and market acceptance. The improvements made to product design may significantly reduce the product cost while causing only a minor increase in the design cost. As a result, it is believed that mass customization can best be approached from design, in particular, the up-front effort in the early stages of the product-development process.

Design for Mass Customization (DFMC) (Tseng and Jiao 1996) aims at considering economies of scope and scale at the early design stage of the product-realization process. The main emphasis of DFMC is on elevating the current practice of designing individual products to designing product families. In addition, DFMC advocates extending the traditional boundaries of product design to encompass a larger scope, from sales and marketing to distribution and services. To support customized product differentiation, a product family platform is required to characterize customer needs and subsequently to fulfill these needs by configuring and modifying well-established modules and components. Therefore, there are two basic concepts underpinning DFMC: product family architecture and product family design. Figure 2 summarizes the conceptual implications of DFMC in terms of the expansion of context from both a design scope perspective and a product-differentiation perspective.

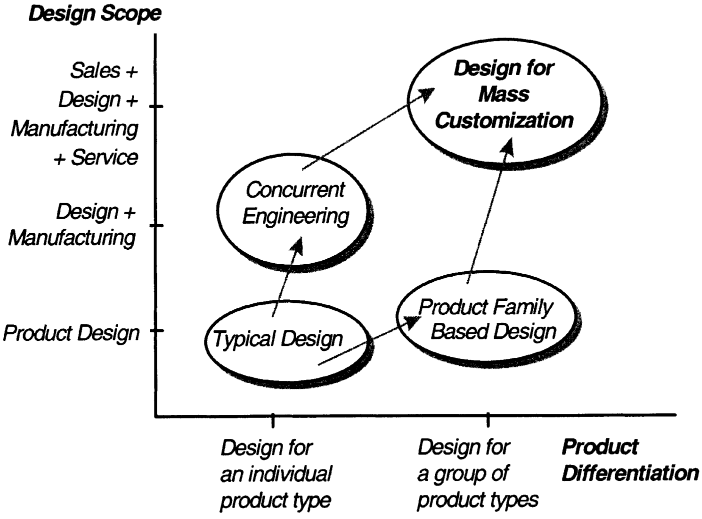

Design Scope
Sales + Design + Manufacturing + Service
Design + Manufacturing
Product Design
Design for Mass Customization
Concurrent Engineering
Typical Design
Product Family Based Design
Design for an individual product type
Design for a group of product types
Product Differentiation

Figure 2 Understanding DFMC.

2.1. Product Family

A product family is a set of products that are derived from a common platform (Meyer and Lehnerd 1997). Each individual product within the family (i.e., a product family member) is called a product variant. While possessing specific features/functionality to meet a particular set of customer requirements, all product variants are similar in the sense that they share some common customer-perceived value, common structures, and/or common product technologies that form the platform of the family. A product family targets a certain market segment, whereas each product variant is developed to address a specific set of customer needs of the market segment.

The interpretation of product families depends on different perspectives. From the marketing/sales perspective, the functional structure of product families exhibits a firm's product portfolio, and thus product families are characterized by various sets of functional features for different customer groups. The engineering view of product families embodies different product and process technologies, and thereby product families are characterized by different design parameters, components, and assembly structures.

2.1.1. Modularity and Commonality

There are two basic issues associated with product families: modularity and commonality. Table 1 highlights different implications of modularity and commonality, as well as the relationship between them.

The concepts of modules and modularity are central in constructing product architecture (Ulrich 1995). While a module is a physical or conceptual grouping of components that share some characteristics, modularity tries to separate a system into independent parts or modules that can be treated as logical units (Newcomb et al. 1996). Therefore, decomposition is a major concern in modularity analysis. In addition, to capture and represent product-structures across the entire product-development process, modularity is achieved from multiple viewpoints, including functionality, solution technologies, and physical structures. Correspondingly, there are three types of modularity involved in product realization: functional modularity, technical modularity, and physical modularity.

What is important in characterizing modularity is the *interaction* between modules. Modules are identified in such a way that between-module (intermodule) interactions are minimized, whereas within-module (inframodule) interactions may be high. Therefore, three types of modularity are characterized by specific measures of interaction in particular views. As for functional modularity, the interaction is exhibited by the relevance of functional features (FFs) across different customer groups. Each customer group is characterized by a particular set of FFs. Customer grouping lies only in the functional view and is independent of the engineering (including design and process) views. That is, it is solution neutral. In the design view, modularity is determined according to the technological feasibility of design solutions. The interaction is thus judged by the coupling of design parameters (DPs) to satisfy given FFs regardless of their physical realization in manufacturing. In the process view, physical interrelationships among components and assemblies (CAs) are mostly derived from manufacturability. For example, on a PCB (printed circuit board), physical routings of CAs determine the physical modularity related to product structures.

It is the commonality that reveals the difference of the architecture of product families from the architecture of a single product. While modularity resembles decomposition of product structures and is applicable to describing module (product) types, commonality characterizes the grouping of similar module (product) variants under specific module (product) types characterized by modularity. Corresponding to the three types of modularity, there are three types of commonality in accordance with functional, design, and process views. Functional commonality manifests itself through functional classification, that is, grouping similar customer requirements into one class, where similarity is measured by the Euclidean distance among FF instances. In the design view, each technical module, characterized by a set of DPs corresponding to a set of FFs, exhibits commonality through clustering similar DP instances by chunks (Ulrich and Eppinger 1995). Instead of measuring similarity among CA instances, physical instances (instances of CAs for a physical module type) are grouped mostly

TABLE 1 A Comparison of Modularity and Commonality

Issues	Modularity	Commonality
Focused Objects	Type (Class)	Instances (Members)
Characteristic of Measure	Interaction	Similarity
Analysis Method	Decomposition	Clustering
Product Differentiation	Product Structure	Product Variants
Integration/Relation	Class-Member Relationship	

according to appropriate categorization of engineering costs derived from assessing existing capabilities and estimated volume, that is, economic evaluation.

The correlation of modularity and commonality is embodied in the class-member relationships. A product structure is defined in terms of its modularity where module types are specified. Product variants derived from this product structure share the same module types and take on different instances of every module type. In other words, a class of products (product family) is described by modularity and product variants differentiate according to the commonality among module instances.

2.1.2. Product Variety

Product variety is defined as the diversity of products that a manufacturing enterprise provides to the marketplace (Ulrich 1995). Two types of variety can be observed: functional variety and technical variety. *Functional variety* is used broadly to mean any differentiation in the attributes related to a product's functionality from which the customer could derive certain benefits. On the other hand, *technical variety* refers to diverse technologies, design methods, manufacturing processes, components and/or assemblies, and so on that are necessary to achieve specific functionality of a product required by the customer. In other words, technical variety, though it may be invisible to customers, is required by engineering in order to accommodate certain customer-perceived functional variety. Technical variety can be further categorized into product variety and process variety. The technical variety of products is embodied in different components/modules/parameters, variations of structural relationships, and alternative configuration mechanisms, whilst process variety involves those changes related to process planning and production scheduling, such as various routings, fixtures/setups, and workstations. While functional variety is mostly related to customer satisfaction from the marketing/sales perspective, technical variety usually involves manufacturability and costs from the engineering perspective.

Even though these two types of variety have some correlation in product development, they result in two different variety design strategies. Since functional variety directly affects customer satisfaction, this type of variety should be encouraged in product development. Such a design for "functional" variety strategy aims at increasing functional variety and manifests itself through vast research in the business community, such as product line structuring (Page and Rosenbaum 1987; Sanderson and Uzumeri 1995), equilibrium pricing (Choi and DeSarbo 1994), and product positioning (Choi et al. 1990). In contrast, design for "technical" variety tries to reduce technical variety so as to gain cost advantages. Under this category, "research" includes variety reduction program (Suzue and Kohdate 1990), design for variety (Ishii et al. 1995a; Martin and Ishii 1996, 1997), design for postponement (Feitzinger and Lee 1997), design for technology life cycle (Ishii et al. 1995b), function sharing (Ulrich and Seering 1990), and design for modularity (Erixon 1996).

Figure 3 illustrates the implications of variety and its impact on variety fulfillment. While exploring functional variety in the functional view through customer requirement analysis, product

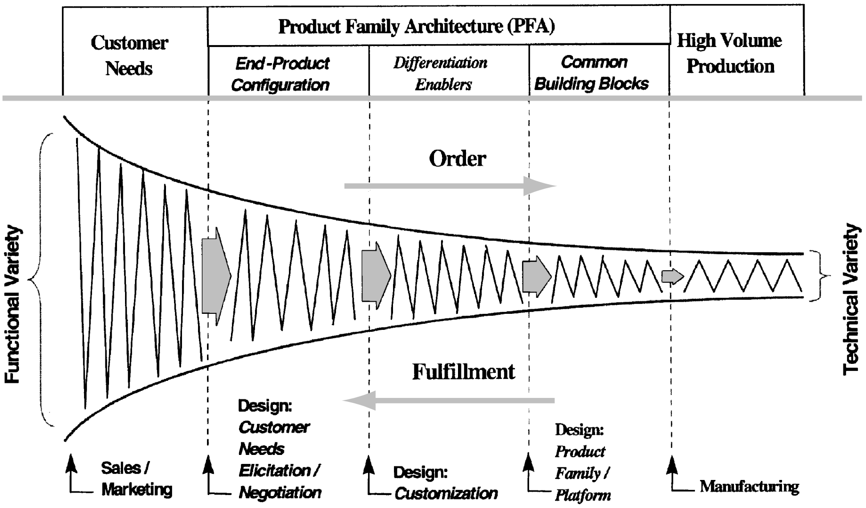


Figure 3 Variety Leverage: Handling Variety for Mass Customization.

family development should try to reduce technical variety in the design and process views by systematic planning of modularity and commonality so as to facilitate plugging in modules that deliver specific functionality, reusing proven designs and reducing design changes and process variations.

2.2. Product Family Architecture

As the backdrop of product families, a well-planned product family architecture (PFA)—the conceptual structure and overall logical organization of generating a family of products—provides a generic umbrella for capturing and utilizing commonality, within which each new product instantiated and extends so as to anchor future designs to a common product line structure. The rationale of such a PFA resides not only in unburdening the knowledge base from keeping variant forms of the same solution, but also in modeling the design process of a class of products that can widely variegate designs based on individual customization requirements within a coherent framework.

Figure 4 illustrates the principle of PFA with respect to product family development for mass customization. From the sales point of view, customers are characterized by combinations of functional features, $\{f\}$, and associated feature values, $\{f^*\}$. A product family, $\{\mathrm{V}_1, \mathrm{V}_2, \mathrm{V}_3, \ldots, \mathrm{V}_i, \ldots, \mathrm{V}_m\}$, is designed to address the requirements of a group of customers in the market segment, $\{\mathrm{Customer}_1, \mathrm{Customer}_2, \mathrm{Customer}_3, \ldots, \mathrm{Customer}_i, \ldots, \mathrm{Customer}_m\}$, in which customers share certain common requirements, f_0^* along with some similar and/or distinct requirements, $\{f_1^*, f_2^*, f_3^*, \ldots, f_n^*\}$. From the engineering perspective, product variants of the product family are derived from configuring common bases, {C}, and differentiation enablers, {E}, that are predefined for a product family. Configuration mechanisms determine the generative aspect of PFA. They guarantee that only both technically feasible and market-wanted product variants can be derived (Baldwin and Chung 1995).

2.2.1. *Composition of PFA*

The PFA consists of three elements: the common base, the differentiation enabler, and the configuration mechanism.

1. *Common base:* Common bases (CBs) are the shared elements among different products in a product family. These shared elements may be in the form of either common (functional) features from the customer or sales perspective or common product structures and common components from the engineering perspective. Common features indicate the similarity of

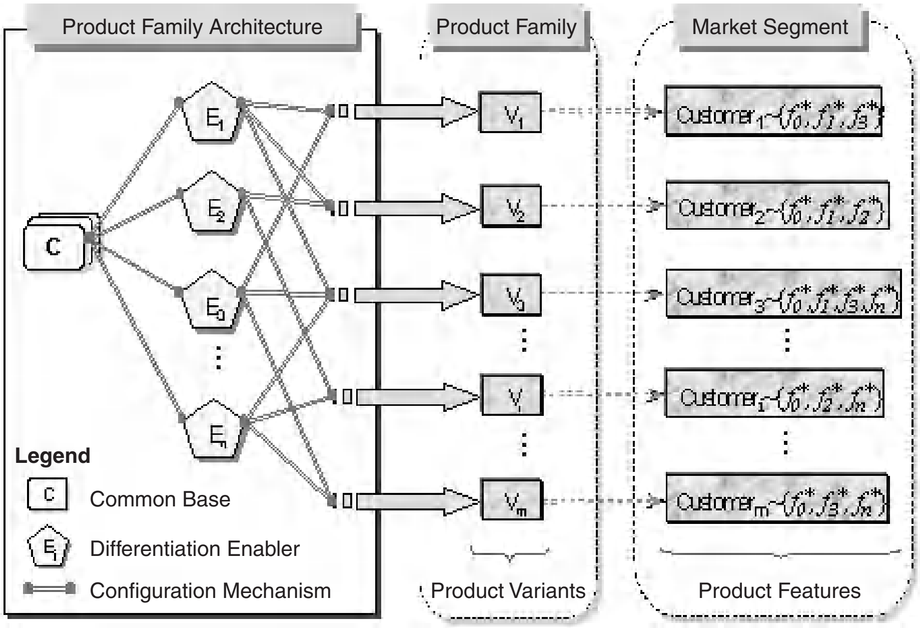

Product Family Architecture
Product Family
Market Segment
C
E1
E2
E3
⋮
En
V1
V2
V3
⋮
Vi
⋮
Vm
Customer1~{f0*, f1*, f3*}
Customer2~{f0*, f1*, f2*}
Customer3~{f0*, f1*, f3*, fn*}
⋮
Customeri~{f0*, f2*, fn*}
⋮
Customerm~{f0*, f3*, fn*}
Legend
C Common Base
E1 Differentiation Enabler
Configuration Mechanism
Product Variants
Product Features

Figure 4 PFA and Its Relationships with Market Segments.

customer requirements related to the market segment. Common product structures and components are determined by product technologies, manufacturing capabilities and economy of scale.

2. *Differentiation enabler:* Differentiation enablers (DEs) are basic elements making products different from one another. They are the source of variety within a product family. From the customer perspective, DEs may be in the form of optional features, accessories, or selectable feature values. In a computer, for example, while a CD drive is an optional feature (yes/no), the RAM must be one instance of a set of selectable feature values, such as 64K, 128K, and 256K bites. In the engineering view, DEs may be embodied in distinct structural relationship (structural DEs) and/or various modules with different performance (constituent DEs). Each engineering DE usually has more than one alternative applicable to product variant derivation for specific applications.
3. *Configuration mechanism:* Configuration mechanisms (CMs) define the rules and means of deriving product variants. Three types of configuration mechanisms can be identified: selection constraints, include conditions, and variety generation.

Selection constraints specify restrictions on optional features because certain combinations of options (i.e., alternative values of optional features) are not allowed or feasible or, on the contrary, are mandatory. An example of the selection constraint for a car might be: "If cylinder (feature) is 1.3 liter (option) and fuel (feature) is diesel (option), a five-speed (option) gearbox (feature) is mandatory." Selection constraints eliminate those technically infeasible or market-unwanted products from all possible combinations of the offered options (Baldwin and Chung 1995). The theoretical number of combination is the Cartesian product of possible feature values (options).

Include conditions are concerned with the determination of alternative variants for each differentiation enabler. The include condition of a variant defines the condition under which the variant should be used or not used with respect to achieving the required product characteristics. It may be in the form of a logic function with parameter values of the differentiation enabler or with its parent constituent as independent variables. For example, an office chair (a parent) consists of one supporting module (a child), which performs as a differentiation enabler. Supposed there are two variants for this supporting module: "using wheels" and "using pads." The include condition of "using wheels" is "the office chair is drivable," while the include condition of "using pads" is "the office chair is not drivable." This include condition is defined in the form of a logic function of the parent's (office chair) variable, "drivable or not." Essentially, include conditions involve the engineering definition stage of product development.

Variety generation refers to the way in which the distinctiveness of product features can be created. It focuses on the engineering realization of custom products in the form of product structures. Such variety fulfillment is related to each differentiation enabler. This chapter identifies three basic methods of variety generation (Figure 5): attaching, swapping, and scaling, in light of the rationale of modular product architecture (Ulrich 1995; Ulrich and Tung 1991). More complex variety-generation methods can be composed by employing these basic methods recursively with reference to the hierarchical decomposition of product structures.

2.2.2. *Synchronization of Multiple Views*

It has been a common practice for different departments in a company have different understandings of product families from their specific perspectives. Such incoherence in semantics and subsequent deployment of information represents a formidable hindrance to current engineering data management systems (EDBS) (Krause et al. 1993). It is necessary to maintain different perspectives of product family representation in a single context. In addition, variety originates from the functional domain and propagates across the entire product-development process. Therefore, the representation of product families should characterize various forms of variation at different stages of product development.

The strategy is to employ a generic, unified representation and to use its fragments for different purposes, rather than to maintain consistency among multiple representations through transformation of different product data models to standard ones. Figure 6 illustrates such a representation scheme of PFA, where functional, behavioral, and structural (noted as FBS) views are tailored for specific departments and design phases.

While corresponding to and supporting different phases of product development, the FBS view model integrates several business functions in a context-coherent framework. This is embodied by mappings between the three views (Figure 6). Various types of customer needs (customer groups) are mapped from the functional view to the behavioral view characterized by solution principles (TPs and modular structures). Such a mapping manifests design activities. The mapping between the behavioral view and the structural view reflects considerations of manufacturing and logistics, where the modular structure and technical modules in terms of TPs are realized by physical modules in terms of components and assemblies through incorporating assessments of available process capa-

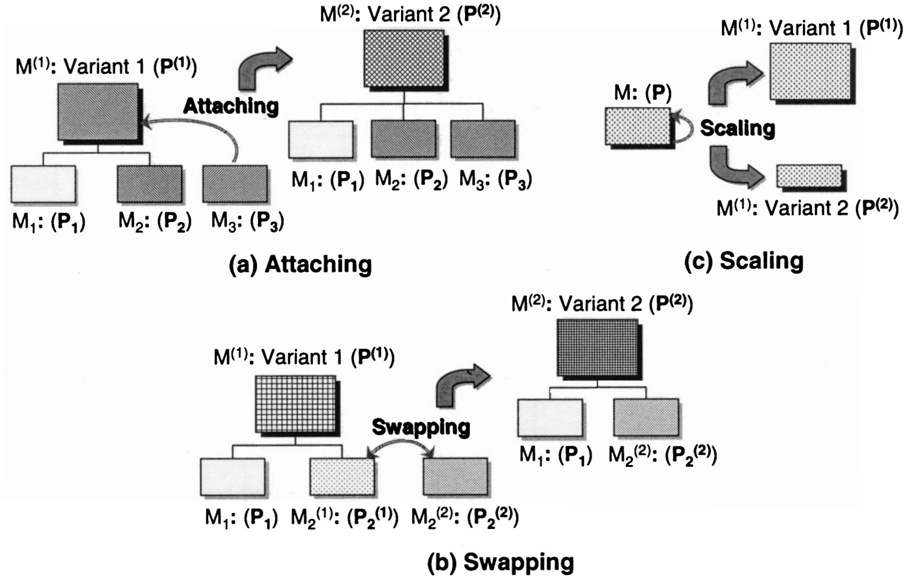

M(1): Variant 1 (P(1))
Attaching
M(2): Variant 2 (P(2))
M1: (P1) M2: (P2) M3: (P3)
M1: (P1) M2: (P2) M3: (P3)
(a) Attaching
M: (P)
Scaling
M(1): Variant 1 (P(1))
M(1): Variant 2 (P(2))
(c) Scaling
M(1): Variant 1 (P(1))
Swapping
M(2): Variant 2 (P(2))
M1: (P1) M2(2): (P2(2))
M1: (P1) M2(1): (P2(1)) M2(2): (P2(2))
(b) Swapping

Figure 5 Basic Methods of Variety Generation.

bilities and the economy of scale. The sales and marketing functions involve mapping between the structural and functional views, where the correspondence of a physical structure to its functionality provides necessary information to assist in negotiation among the customers, marketers, and engineers, such as facilitating the request for quotation (RFQ).

2.3. Product Family Design

Under the umbrella of PFA, product family design manifests itself through the derivation processes of product variants based on PFA constructs. Figure 7 illustrates the principle of PFA-based product

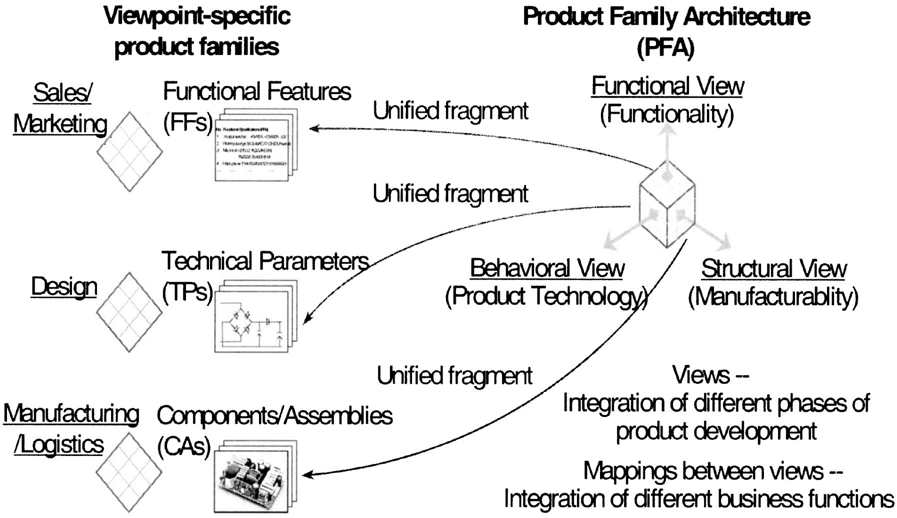

Viewpoint-specific product families
Product Family Architecture (PFA)
Sales/ Marketing
Functional Features (FFs)
Unified fragment
Functional View (Functionality)
Unified fragment
Design
Technical Parameters (TPs)
Behavioral View (Product Technology)
Structural View (Manufacturablity)
Unified fragment
Manufacturing /Logistics
Components/Assemblies (CAs)
Views -- Integration of different phases of product development
Mappings between views -- Integration of different business functions

Figure 6 Representing Multiple Views of Product Family within a Single Context.

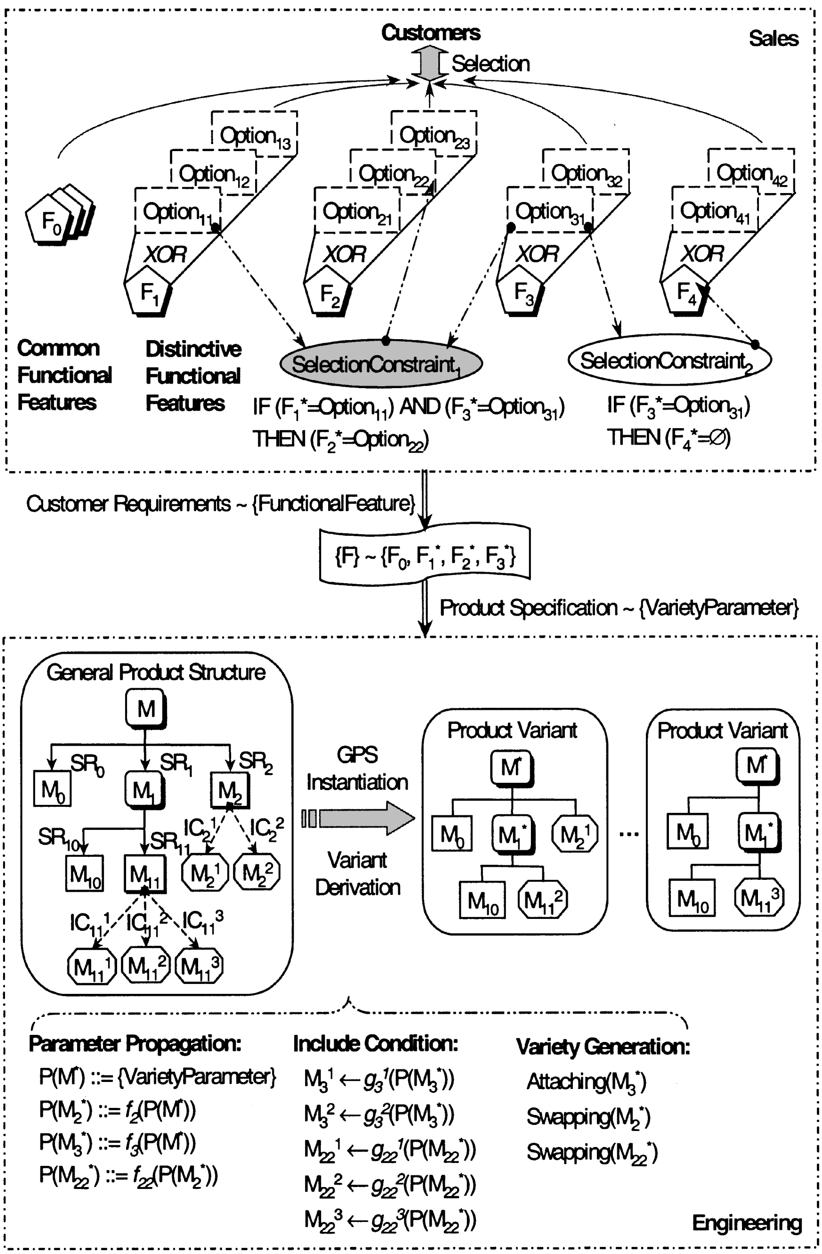

Customers
Sales
Selection
Option_{13}
Option_{12}
Option_{11}
Option_{23}
Option_{22}
Option_{21}
Option_{32}
Option_{31}
Option_{42}
Option_{41}
XOR
XOR
XOR
XOR
F_0
F_1
F_2
F_3
F_4
Common Functional Features
Distinctive Functional Features
$\text{SelectionConstraint}_1$
$\text{IF }(F_1{}^*=\text{Option}_{11})\text{ AND }(F_3{}^*=\text{Option}_{31})$
$\text{THEN }(F_2{}^*=\text{Option}_{22})$
$\text{SelectionConstraint}_2$
$\text{IF }(F_3{}^*=\text{Option}_{31})$
$\text{THEN }(F_4{}^*=\varnothing)$
Customer Requirements ~ {FunctionalFeature}
$\{F\} \sim \{F_0, F_1{}^*, F_2{}^*, F_3{}^*\}$
Product Specification ~ {VarietyParameter}
General Product Structure
M
SR_0
SR_1
SR_2
M_0
M_1
M_2
SR_{10}
SR_{11}
$IC_2{}^1$
$IC_2{}^2$
M_{10}
M_{11}
$M_2{}^1$
$M_2{}^2$
$IC_{11}{}^1$
$IC_{11}{}^2$
$IC_{11}{}^3$
$M_{11}{}^1$
$M_{11}{}^2$
$M_{11}{}^3$
GPS Instantiation
Variant Derivation
Product Variant
M^*
M_0
$M_1{}^*$
$M_2{}^1$
M_{10}
$M_{11}{}^2$
...
Product Variant
M^*
M_0
$M_1{}^*$
M_{10}
$M_{11}{}^3$
Parameter Propagation:
$P(M^*) ::= \{\text{VarietyParameter}\}$
$P(M_2{}^*) ::= f_2(P(M^*))$
$P(M_3{}^*) ::= f_3(P(M^*))$
$P(M_{22}{}^*) ::= f_{22}(P(M_2{}^*))$
Include Condition:
$M_3{}^1 \leftarrow g_3{}^1(P(M_3{}^*))$
$M_3{}^2 \leftarrow g_3{}^2(P(M_3{}^*))$
$M_{22}{}^1 \leftarrow g_{22}{}^1(P(M_{22}{}^*))$
$M_{22}{}^2 \leftarrow g_{22}{}^2(P(M_{22}{}^*))$
$M_{22}{}^3 \leftarrow g_{22}{}^3(P(M_{22}{}^*))$
Variety Generation:
$\text{Attaching}(M_3{}^*)$
$\text{Swapping}(M_2{}^*)$
$\text{Swapping}(M_{22}{}^*)$
Engineering

Figure 7 PFA-Based Product Family Design: Variant Derivation through GPS Instantiation.

family design. Customers make their selections among sets of options defined for certain distinctive functional features. These distinctive features are the differentiation enablers of PFA in the sales view. Selection constraints are defined for presenting customers only feasible options, that is, both technically affordable and market-wanted combinations. A set of selected distinctive features together with those common features required by all customers comprise the customer requirements of a customized product design. As shown in Figure 7, customized products are defined in the sales view in the form of functional features and their values (options), whereas in the engineering view, product family design starts with product specifications in the form of variety parameters. Within the PFA, variety parameters correspond to distinctive functional features and the values of each variety parameter correspond to the options of each functional feature.

To realize variety, a general product structure (GPS) is employed as a generic data structure of product family in the engineering view. The derivation of product variants becomes the instantiation of GPS. While the GPS characterizes a product family, each instance of GPS corresponds to a product variant of the family. Each item in the GPS (either a module or a structural relationship) is instantiated according to certain include conditions that are predefined in terms of variety parameters. Variety parameters originate from functional features specified by the customers and propagate along levels of abstraction of GPS. Variety generation methods, such as attaching, swapping, and scaling, are implemented through different instantiations of GPS items. While the GPS provides a common base for product family design, distinctive items of GPS, such as distinctive modules and structural relationships, perform as the differentiation enablers of the family. Distinctive items are embodied in different variants (instances) that are identified by associated conditions. Therefore, these include conditions and the variety generation capability constitutes configuration mechanisms of PFA in the engineering view.

3. MASS CUSTOMIZATION MANUFACTURING

Competition for mass customization manufacturing is focused on the flexibility and responsiveness in order to satisfy dynamic changes of global markets. The traditional metrics of cost and quality are still necessary conditions for companies to outpace their competitors, but they are no longer the deciding factors between winners and losers. Major trends are:

1. A major part of manufacturing will gradually shift from mass production to the manufacturing of semicustomized or customized products to meet increasingly diverse demands.
2. The ‘‘made-in-house’’ mindset will gradually shift to distributed locations, and various entities will team up with others to utilize special capabilities at different locations to speed up product development, reduce risk, and penetrate local markets.
3. Centralized control of various entities with different objectives, locations, and cultures is almost out of the question now. Control systems to enable effective coordination among distributed entities have become critical to modern manufacturing systems.

To achieve this, it is becoming increasingly important to develop production planning and control architectures that are modifiable, extensible, reconfigurable, adaptable, and fault tolerant. Flexible manufacturing focuses on batch production environments using multipurpose programmable work cells, automated transport, improved material handling, operation and resource scheduling, and computerized control to enhance throughput. Mass customization introduces multiple dimensions, including drastic increase of variety, multiple product types manufactured simultaneously in small batches, product mixes that change dynamically to accommodate random arrival of orders and wide spread of due dates, and throughput that is minimally affected by transient disruptions in manufacturing processes, such as breakdown of individual workstations.

3.1. Managing Variety in Production Planning

Major challenge of mass customization production planning results from the increase of variety. The consequence of variety may manifest itself through several ramifications, including increasing costs due to the exponential growth of complexity, inhibiting benefits from economy of scale, and exacerbating difficulties in coordinating product life cycles. Facing such a variety dilemma, many companies try to satisfy demands from their customers through engineering-to-order, produce-to-order, or assembly-to-order production systems (Erens and Hegge 1994).

At the back end of product realization, especially at the component level and on the fabrication aspect, today we have both flexibility and agility provided by advanced manufacturing machinery such as CNC machines. These facilities accommodate technical variety originating from diverse needs of customers. However, at the front end, from customer needs to product engineering and production planning, managing variety is still very ad hoc. For example, production control information systems, such as MRPII (manufacturing resource planning) and ERP (enterprise resource planning), are falling behind even though they are important ingredients in production management (Erens et al. 1994). The difficulties lie in the necessity to specify all the possible variants of each product and in the fact that current production management systems are often designed to support a production that is based on a limited number of product variants (van Veen 1992).

The traditional approach to variant handling is to treat every variant as a separate product by specifying a unique BOM for each variant. This works with a low number of variants, but not when customers are granted a high degree of freedom in specifying products. The problem is that a large number of BOM structures will occur in mass customization production, in which a wide range of combinations of product features may result in millions of variants for a single product. Design and maintenance of such a large number of complex data structures are difficult, if not impossible. To overcome these limitations, a generic BOM (GBOM) concept has been developed (Hegge and Wortmann 1991; van Veen 1992). The GBOM provides a means of describing, with a limited amount of data, a large number of variants within a product family, while leaving the product structure unimpaired. Underlying the GBOM is a generic variety structure for characterizing variety, as schematically illustrated in Figure 8. This structure has three aspects:

1. *Product structure:* All product variants of a family share a common structure, which can be described as a hierarchy containing constituent items (I_i) at different levels of abstraction, where $\{I_i\}$ can be either abstract or physical entities. Such a breakdown structure (AND tree) of $\{I_i\}$ reveals the topology for end-product configuration (Suh 1997). Different sets of I_i and their interrelationships (in the form of a decomposition hierarchy) distinguish different common product structures and thus different product families.
2. *Variety parameters:* Usually there is a set of attributes, A, associated with each I_i. Among them, some variables are relevant to variety and thus are defined as variety parameters, $\{P_j\} \subset A$. Like attribute variables, parameters can be inherited by child node(s) from a parent node. Different instances of a particular P_j, e.g., $\{V_k\}$, embody the diversity resembled by, and perceived from, product variants.

   Two types of class-member relationships can be observed between $\{P_j\}$ and $\{V_k\}$. A leaf P_j (e.g., *P32*) indicates a binary-type instantiation, meaning whether *I32* is included in *I3* (*V32* = 1), or not (*V32* = 0). On the other hand, a node P_j (e.g., *P2*) indicates a selective type instantiation, that is, *I2* has several variants in terms of values of *P2*, i.e., *V2* ~ {*V2_1*, *V2_2*}.
3. *Configuration constraints:* Two types of constraint can be identified. Within a particular view of product families, such as the functional, behavioral, or physical view, restrictions on the

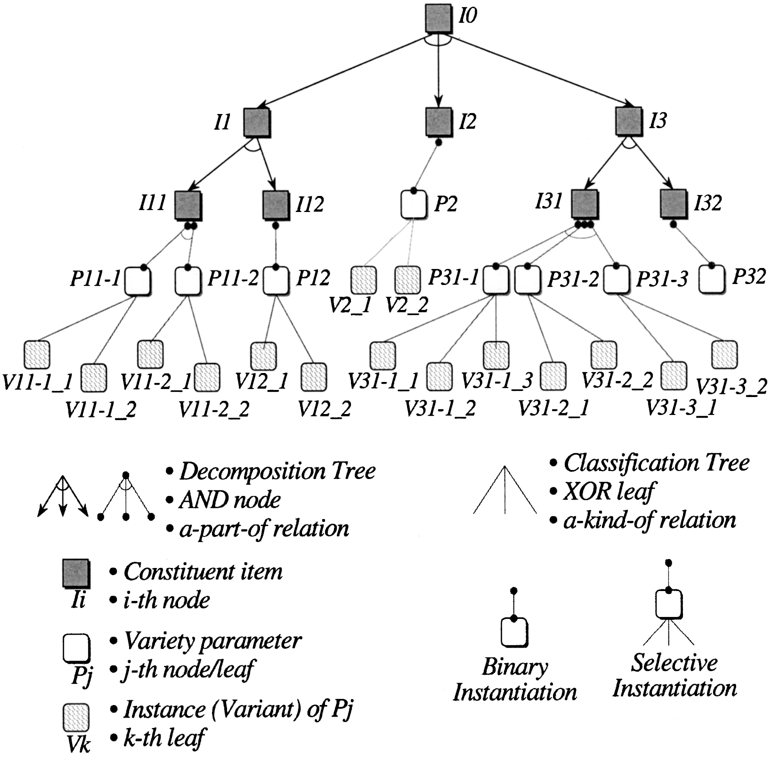

I0
I1
I2
I3
I11
I12
P2
I31
I32
P11-1
P11-2
P12
V2_1
V2_2
P31-1
P31-2
P31-3
P32
V11-1_1
V11-1_2
V11-2_1
V11-2_2
V12_1
V12_2
V31-1_1
V31-1_2
V31-1_3
V31-2_1
V31-2_2
V31-3_1
V31-3_2
• Decomposition Tree
• AND node
• a-part-of relation
• Classification Tree
• XOR leaf
• a-kind-of relation
Ii
• Constituent item
• i-th node
Pj
• Variety parameter
• j-th node/leaf
Vk
• Instance (Variant) of Pj
• k-th leaf
Binary Instantiation
Selective Instantiation

Figure 8 A Generic Structure for Characterizing Variety.

combination of parameter values, $\{V_k\}$, are categorized as Type I constraints. For example, *V11-1_1* and *V31-3_2* are incompatible, that is, only one of them can be selected for a product variant, indicating an exclusive all (XOR) relationship. The mapping relationships of items and their variety parameters across the functional, behavioral, and structural views are referred to as Type II constraints. This type of constraint deals mostly with configuration design knowledge. Usually they are described as rules instead of being graphically depicted in a generic structure. While the functional and behavioral views of product families are usually associated with product family design, the major concern of managing variety in production is Type I constraints which mostly involves the structural view.

Table 2 illustrates the above generic variety structure using a souvenir clock example. As far as variant handling is concerned, the rationale of the generic variety structure lies in the recognition of the origin and subsequent propagation of variety. Three levels of variation have been indicated, that is, at the structure, variety parameter, and instance levels. Different variation levels have different variety implications. To understand the "generic" concept underlying such a variety representation, two fundamental issues need to be highlighted:

1. *Generic item:* A generic item represents a set of similar items (called variants) of the same type (a family). The item may be an end product, a subassembly, an intermediate part, or a component part (van Veen, 1992). It may also be a goes-into-relationship or an operation. For example, a red front plate (I_1^*), a blue front plate (I_2^*) and a transparent front plate (I_3^*) are three individual variants, whereas a generic item, *I*, represents such a set of variants (a family of front plates), that is $\boldsymbol{I} \sim \{I_1^*, I_2^*, I_3^*\}$. However, these variants are similar in that they share a common structure (e.g., BOM structure) in configuring desk clocks.
2. *Indirect identification:* Instead of using part numbers (referred to as direct identification), the identification of individual variants from a generic item (within a family) is based on variety parameters and their instances (a list if parameter values). Such identification is called indirect identification (Hegge and Wortmann. 1991). In the above example, a variety parameter, color, and its value list, , can be used for an indirect identification of a particular variant, i.e., $I_1^* \sim$

TABLE 2 The Generic Variety Structure for Souvenir Clocks

Structural Items $\{I_i\}$	Variety Parameter $\{P_j\}$	Variety Instance $\{V_k\}$
..3/Hands	Setting Type	Two-hand Setting, Three-hand Setting
	Color	White, Grey, etc.
	Size	Large, Medium, Small
..3/Dial	Pattern	Logo, Mosic, Scenery, Customized Photo, etc.
	Size	Large, Medium, Small
...4/Transmission	Alarm	Yes, No
...4/Core	Alarm	Yes, No
..3/Base	Shape	Round, Rectangular, Hexagonal
	Material	Acrylic, Aluminum, etc.
	Color	Transparent, Red, etc.
..3/Front Plate	Shape	Rectangular, Round, Rhombus
	Material	Acrylic, Aluminum, etc.
	Color	Transparent, Red, etc.
..1/Label Sticker	Pattern	HKUST, Signature, etc.
..1/Paper Box	Type	Ordinary, Deluxe, etc.
Constraint #	Constraint Fields	Constraint Type
1	Hands.Size Dial.Size	Size Compatible
2	Transmission.Alarm Core.Alarm	Type Compatible
3	Base.Material FrontPlate.Material	Material Compatible
4	Base.Color FrontPlate.Color	Color Compatible

{$\boldsymbol{I}$|color = "red"} and $I_3^* \sim$ {$\boldsymbol{I}$|color = "transparent"}. On the other hand, the identification of product family "front plate" is $\boldsymbol{I}$.

3.2. Coordination in Manufacturing Resource Allocation

Mass customization manufacturing is characterized by shortened product life cycle with high-mixed and low-volume products in a rapidly changing environment. In customer-oriented plants, orders consisting of a few parts, or even one part, will be transferred directly from vendors to producers, who must respond quickly to meet short due dates. In contrast to mass production, where the manufacturer tells consumers what they can buy, mass customization is driven by customers telling the manufacturer what to manufacture. As a result, it is difficult to use traditional finite capacity scheduling tools to support the new style of manufacturing. Challenges of manufacturing resource allocation for mass customization include:

1. The number of product variety flowing through the manufacturing system is approaching an astronomical scale.
2. Production forecasting for each line item and its patterns is not often available.
3. Systems must be capable of rapid response to market fluctuation.
4. The system should be easy for reconfiguration—ideally, one set of codes employed across different agents.
5. The addition and removal of resources or jobs can be done with little change of scheduling systems.

Extensive research on coordination of resource allocation has been published in connection with scenarios of multiple resource providers and consumers. In the research, existence of demand patterns is the prerequisite for deciding which algorithm is applicable. The selection of a certain algorithm is often left to empirical judgment, which does not alleviate difficulties in balancing utilization and level of services (e.g., meeting due dates).

As early as 1985, Hatvany (1985) pointed out that the rigidity of traditional hierarchical structures limited the dynamic performance of systems. He suggested a heterachical system, which is described as the fragmentation of a system into small, completely autonomous units. Each unit pursues its own goals according to common sets of laws, and thus the system possesses high modularity, modifiability, and extendibility. Following this idea, agent-based manufacturing (Sikora and Shaw 1997) and holonic manufacturing (Gou et al. 1994), in which all components are represented as different agents and holons, respectively, are proposed to improve the dynamics of operational organizations.

From traditional manufacturing perspectives, mass customization seems chaotic due to its large variety, small batch sizes, random arrival orders, and wide span of due dates. It is manageable, however, owing to some favorable traits of modern manufacturing systems, such as inherent flexibility in resources (e.g., increasing use of machining centers and flexible assembly workstations) and similarities among tools, production plans, and product designs. The challenge is thus how to encode these characteristics into self-coordinating agents so that the invisible hand, in the sense of Adam Smith's market mechanism (Clearwater 1996), will function effectively.

Market-like mechanisms have been considered as an appealing approach for dealing with the coordination of resource allocation among multiple providers and consumers of resources in a distributed system (Baker 1991; Malone et al. 1988; Markus and Monostori 1996). Research on such a distributed manufacturing resource-allocation problem can be classified into four categories: the bidding/auction approach (Shaw 1988; Upton and Barash 1991), negotiation approach (Lin and Solberg 1992), cooperative approach (Burke and Prosser 1991) and pricing approach (Markus and Monostori 1996).

Major considerations of scheduling for resource allocation include:

1. Decompose large, complex scheduling problems into smaller, disjointed allocation problems.
2. Decentralize resource access, allocation, and control mechanisms.
3. Design a reliable, fault-tolerant, and robust allocation mechanism.
4. Design scalable architectures for resource access in a complex system and provide a plug-and-play resource environment such that resource providers and consumers can enter or depart from the market freely.
5. Provide guarantees to customers and applications on performance criteria.

In this regard, the agent-based, market-like mechanism suggests itself as a means of decentralized, scalable, and robust coordination for resource allocation in a dynamic environment (Tseng et al. 1997). In such a collaborative scheduling system, each workstation is considered as an autonomous agent seeking the best return. The individual work order is considered as a job agent that vies for

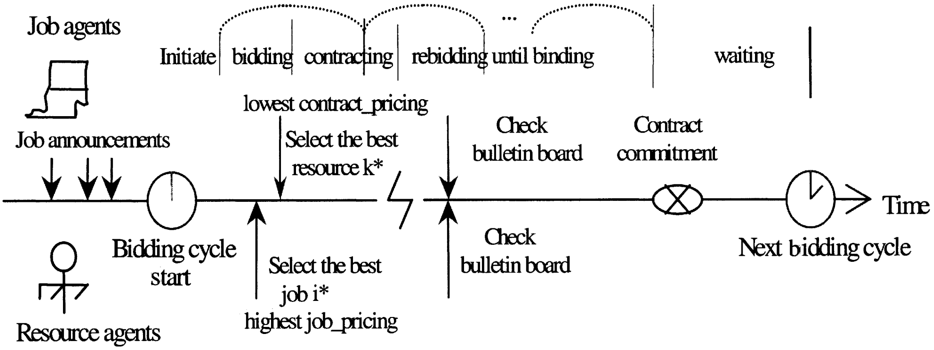

Job agents
Initiate
bidding
contracting
rebidding until binding
...
waiting
lowest contract_pricing
Select the best resource k*
Job announcements
Check bulletin board
Contract commitment
Time
Bidding cycle start
Select the best job i*
highest job_pricing
Check bulletin board
Next bidding cycle
Resource agents

Figure 9 A Market Structure for Collaborative Scheduling.

the lowest cost for resource consumption. System scheduling and control are integrated as an auction-based bidding process with a price mechanism that rewards product similarity and response to customer needs. A typical market model consists of agents, a bulletin board, a market system clock, the market operating protocol, the bidding mechanism, pricing policy, and the commitment mechanism. Figure 9 illustrates how the market operating protocol defines the rules for synchronization among agents.

The satisfaction of multiple criteria, such as costs and responsiveness, cannot be achieved using solely a set of dispatching rules. A price mechanism should be constructed to serve as an invisible hand to guide the coordination in balancing diverse requirements and maximizing performance in a dynamic environment. It is based on market-oriented programming for distributed computation (Adelsberger and Conen 1995). The economic perspective on decentralized decision making has several advantages. It overcomes the narrow view of dispatching rules, responds to current market needs, uses maximal net present value as the objective, and coordinates agents' activities with minimal communication. In collaborative scheduling, objectives of the job agent are transformed into a set of evaluation functions. The weights of the functions can be adjusted dynamically on basis of system states and external conditions. Resource agents adjust the charging prices based on their capability and utilization and the state of current system. Mutual selection and mutual agreement are made through two-way communication. Figure 10 depicts the market price mechanism. In the market model, the job agents change routings (i.e., select different resource agents), and adjust Job_Price as a pricing tool to balance the cost of resources and schedule exposure. Resource agents adjust their prices according to market demands on their capability and optimal utilization and returns. For example, a powerful machine may attract many job agents, and thus the queue will build up and

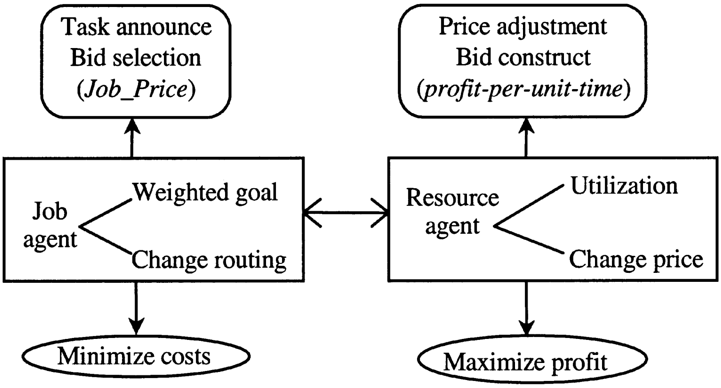

Task announce
Bid selection
(Job_Price)
Price adjustment
Bid construct
(profit-per-unit-time)
Job agent
Weighted goal
Change routing
Resource agent
Utilization
Change price
Minimize costs
Maximize profit

Figure 10 The Price Mechanism of a Market Model.

waiting time will increase. When a resource agent increases its price, the objective of bidding job will decrease, thus driving jobs to other resources and diminishing the demand for this resource.

Specifically, the price mechanism can be expressed as follows:

1. The job agent calculates the price to be paid for the i operation:

$$\text{Job\_price} = \text{Basic\_price} + \text{Penalty\_price} \tag{1}$$

where the Penalty_price reflects the due date factor:

$$\text{Penalty\_price} = \text{penalty} \times e^{-(d_i - c_i)} \tag{2}$$

where d_i represents the due date of the operation i and c_i represents the completion time of the operation i.

2. The resource agent gets the Job_price and then starts to rank the job according to profit per unit time:

$$\text{Profit/time} = (\text{Job\_price} - \text{PFAindex} \times s - \text{Opportunity\_cost})/tp \tag{3}$$

where s represents the setup cost, tp denotes the processing time of the operation, and PFAindex represents the product family consideration in setting up the manufacturing system. For instance, we can let PFAindex = 0, if two consecutive jobs, i and j, are in the same product family, and hence the setup charge can be eliminated in the following job. Otherwise, PFAindex = 1. The Opportunity_cost in Eq. (3) represents the cost of losing particular slack for other job agents due to the assignment of resource for one job agent's operation i, as expressed below:

$$\text{Opportunity\_cost} = \sum_j e^{-c(\min(0,t_{sj}))}\,(e^{c\delta t_{sj}} - 1) \tag{4}$$

where $t_{sj} = t_{dj} - t - w_j$, t_{dj} represents the due date of the corresponding job, t represents the current time (absolute time), w_j represents the remaining work content of the job, and c is a system opportunity cost parameter, setting to 0.01. In Eq. (4), δt_{sj} represents the critical loss of slack of operation j due to the scheduling of the operation i before j, as depicted below:

$$\delta t_{sj} = \begin{cases} tp_i, & t_{sj} \le 0 \\ 0 & tp_i \ge t_{sj} \\ tp_i - t_{sj}, & t_{sj} \ge tp_i \end{cases} \tag{5}$$

Meanwhile, the resource agent changes its resource price p_i:

$$p_i = q \sum_k \text{Job\_price}_k \tag{6}$$

where q is a normalized constant.

3. The job agent, in a period of time, collects all bids and responds to its task announcement and selects the resource agent with minimal Actual_cost to confirm the contract:

$$\text{Actual\_cost} = p_i \times tp_i + \max((c_i - d_i),0) \times \text{penalty} \tag{7}$$

Based on the above formulations, the collaborative control can be modeled as a message-based simulation, as shown in Figure 11. The control process is driven by an event and/or an abnormal event, such as machine breakdown or a new due date. All events can be represented as messages.

3.3. High-Variety Shop-Floor Control

Mass customization manufacturing motivates a new generation of shop-floor control systems that can dynamically respond to customer orders and unanticipated changes in the production environment. The requirements of the new control systems include reconfigurability, decomposability, and scalability to achieve make-to-order with very short response time. A systematic approach has been developed to design control system by leveraging recent progresses in computing and communication hardware and software, including new software engineering methods and control technologies, such as smart sensors and actuators, open architectures, and fast and reliable networks (Schreyer and Tseng

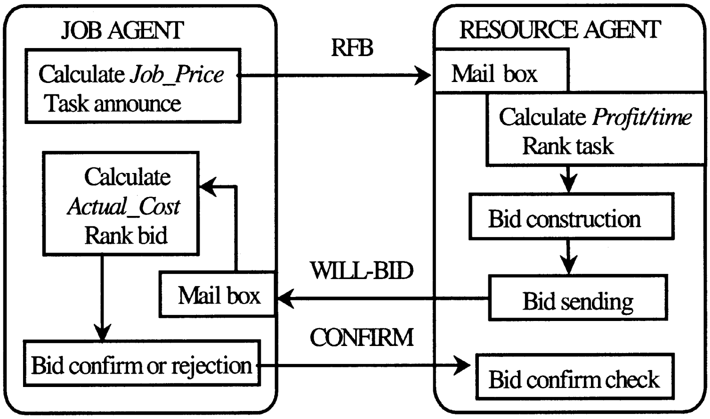

JOB AGENT
Calculate *Job_Price*
Task announce
Calculate
Actual_Cost
Rank bid
Mail box
Bid confirm or rejection
RFB
WILL-BID
CONFIRM
RESOURCE AGENT
Mail box
Calculate *Profit/time*
Rank task
Bid construction
Bid sending
Bid confirm check

Figure 11 Message-Based Bidding and Dynamic Control.

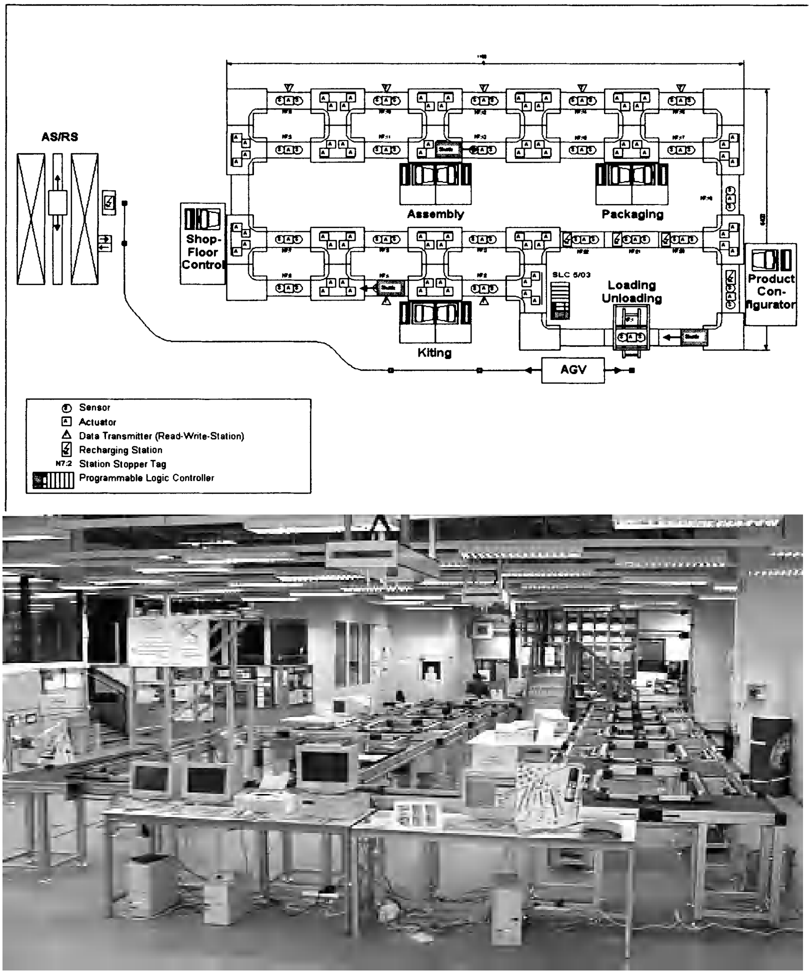

AS/RS
Shop-Floor Control
Assembly
Packaging
SLC 5/03
Loading Unloading
Product Con-figurator
Kitting
AGV
Sensor
Actuator
Data Transmitter (Read-Write-Station)
Recharging Station
N7:2 Station Stopper Tag
Programmable Logic Controller

Figure 12 Example of Installation of Mass Customization Manufacturing System at the Hong Kong University of Science and Technology.

1998). Figure 12 illustrates an actual example of installation of a mass customization manufacturing system in the Hong Kong University of Science and Technology.

4. SALES AND MARKETING FOR MASS CUSTOMIZATION

In the majority of industries and sectors, customer satisfaction is a powerful factor in the success of products and services. In the era of mass production, customers were willing to constrain their choices to whatever was available, as long as the price was right. However, customers today are liberated and better informed. This leads them to be choosy about their purchases and less willing to compromise with what is on the shelf. Meeting customer requirements requires full understanding of customers' values and preferences. In addition, it is important that customers know what the company can offer as well as their possible options and the consequences of their choices, such as cost and schedule implications.

4.1. Design by Customers

The setup time and its resulting economy of scale have been widely accepted as the foundation for the mass production economy, where batch size and lead time are important instruments. Consequently, the popular business model of today's firms is design *for* customers. Companies design and then produce goods for customers through conforming a set of specifications that anticipates customer's requirements. Often the forecasting of end users' requirements is developed by the marketing department. It is usually carried out through aggregating the potential needs of customers with the consideration of market direction and technology trends. Given the complexities of modern products, the dynamic changes in customers' needs, and the competitive environment in which most businesses have to operate, anticipating potential customers' needs can be very difficult. Chances are that forecasting will deviate from the reality by a high margin. Three major economic deficiencies are often encountered.

Type A is the possibility of producing something that no customers want. The waste is presented in the form of inventory, obsolescence, or scrap. Although a significant amount of research and development has been conducted on improving the forecast accuracy, inventory policy, increased product offerings, shortened time-to-market, and supply chain management, avoiding the possibility of producing products that no customer want is still a remote, if not impossible, goal.

Type B waste comes from not being able to provide what customers need when they are ready to purchase. It often presents itself in the form of shortage or missing opportunity. The costs of retail stores, marketing promotion, and other sales expenditures on top of the goods themselves will disappoint the customers who are ready to purchase. The cost of missing opportunities can be as significant as the first type.

Type C deficiency results from customers making compromises between their real requirements and existing SKUs (stock keeping units), that is, what is available on the shelf or in the catalogue. Although these compromises are usually not explicit and are difficult to capture, they lead to customer dissatisfaction, reduce willingness to make future purchases, and erode the competitiveness of a company.

To minimize the effect of these deficiencies, one approach is to revise the overall systems design of a manufacturing enterprise. Particularly with the growing flexibility in production equipment, manufacturing information systems and workforce, the constraints of setup cost and lead time in manufacturing have been drastically reduced. The interface between customers and product realization can be reexamined to ensure that total manufacturing systems produce what the customers want and customers are able to get what the systems can produce within budget and schedule. Furthermore, with the growing trends of cultural diversity and self-expression, more and more customers are willing to pay more for products that enhance their individual sizes, tastes, styles, needs, comfort, or expression (Pine 1993).

With the rapid growth of Internet usage and e-commerce comes an unprecedented opportunity for manufacturing enterprise to connect directly customers scattered around the world. In addition, through the Internet and business-to-business e-commerce, manufacturing enterprise can now acquire access to the most economical production capabilities on a global basis. Such connectivity provides the necessary condition for customers to become connected to the company. However, by itself it will not enhance effectiveness.

In the last decade, concurrent engineering brought together design and manufacturing, which has dramatically reduced the product development life cycle and hence improved quality and increased productivity and competitiveness. Therefore, design *by* customers has emerged as a new paradigm to further extend concurrent engineering by extending connectivity with customers and suppliers (Tseng and Du 1998). The company will be able to take a proactive role in helping customers define needs and negotiate their explicit and implicit requirements. Essentially, it brings the voice of customers into design and manufacturing, linking customer requirements with the company's capabilities and extending the philosophy of concurrent engineering to sales and marketing as part of an integrated product life cycle. Table 3 summarizes the comparison of these two principles for customer-focused product realization.

The rationale of design by customers can be demonstrated by the commonly accepted value chain concept (Porter 1986). The best match of customer needs and company capability requires several technical challenges:

1. Customers must readily understand the capability of a company without being a design engineer or a member in the product-development team.
2. Company must interpret the needs of customers accurately and suggest alternatives that are closest to the needs.
3. Customers must make informed choices with sufficient information about alternatives.
4. Company must have the ability to fulfill needs and get feedback.

To tackle these challenges, it is necessary that customers and the company share a context-coherent framework. Product configuration has been commonly used as a viable approach, primarily because it enables both sides to share the same design domain. Based on the product configuration approach, the value chain, which includes the customer interface, can be divided into four stages:

1. *Formation:* Presenting the capability that a company can offer in the form of product families and product family structure.
2. *Selection:* Finding customers' needs and then matching the set of needs by configuring the components and subassemblies within the constraints set by customers.
3. *Fulfillment:* Includes logistics, manufacturing and distribution so that customer' needs can be satisfied within the cost and time frame specified.
4. *Improvement:* Customers' preferences, choices, and unmet expressed interests are important inputs for mapping out the future improvement plan.

Formation and selection are new dimensions of design for customer. They are explained further below.

4.2. Helping Customers Making Informed Choices: Conjoint Analysis

Design *by* customers assumes customers are able to spell out what they want with clarity. Unfortunately, this is often not the case. To begin with, customers may not be able to know what is possible. Then the system needs to pull the explicit and implicit needs from customers. Conjoint analysis is a set of methods in marketing research originally designed to measure consumer preferences by assessing the buyers' multiattribute utility functions (Green and Krieger, 1989; IntelliQuest 1990). It assumes that a product could be described as vectors of M attributes, $Z_1, Z_2, \ldots, Z_M$. Each attribute can include several discrete levels. Attribute Z_m can be at any one of the L_m levels, $Z_{m1}, Z_{m2}, \ldots, Z_{m,Lm}$, $m \in [1, M]$. A utility functions is defined as (McCullagh and Nelder 1989):

$$U_r = \sum_{m=1}^{M} W_m \left(\sum_{l=1}^{L_m} d_{ml} X_{rml} \right) = \sum_{m=1}^{M} \sum_{l=1}^{L_m} U_{ml} X_{rml} \tag{8}$$

$$U_{ml} = W_m * d_{ml} \tag{9}$$

where U_r = customer's utility for profile r, $r \in [1, R]$
W_m = importance of attribute Z_m for the customer
d_{rml} = desirability for l^{th} level of attribute m, $l \in [1, L_m]$, $m \in [1, M]$
U_{ml} = utility of attribute m's l^{th} level

TABLE 3 A Comparison of Design for Customers and Design by Customers

Principle	Design for Customers	Design by Customers
Manufacturing Practice	Mass production	Mass customization
Design	Anticipate what customers would buy	Create platforms and capabilities
Manufacturing	Build to forecast	Build to order
Sales	Promote what is available	Assist customers to discover what can be done and balance their needs

In the above formulation, X_{rml} is a dummy variable indicating whether or not the particular level of an attribute is selected, as expressed by:

$$X_{rml} = \begin{cases} 1 & \text{if attribute } m \text{ is on } l^{\text{th}} \text{ level;} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Usually, a large number of attributes, discrete levels, and their preference indices is required to define the preferred products through customer interaction, and thus the process may become overwhelming and impractical. There are several approaches to overcoming this problem. Green et al. (1991) and IntelliQuest (1990) have proposed adaptive conjoint analysis to explore customers' utility with iterations. Customers are asked to rate the relative importance of attributes and refine the tradeoffs among attributes in an interactive setting through comparing a group of testing profiles. Other approaches, such as Kano diagram and analytic hierarchy process (AHP), can also be applied to refine the utility value (Urban and Hauser 1993).

With the utility function, U_{ml}, customers can find the relative contribution of each attribute to their wants and thus make necessary tradeoffs. Customers can finalize their design specifications by maximizing their own personal value for the unit price they are spending.

4.3. Customer Decision-Making Process

Design by customers allows customers to directly express their own requirements and carry out the mapping to the physical domain. It by no means gives customers free hand to design whatever they want in a vacuum. Instead, it guides customers in navigating through the capabilities of a firm and defining the best alternative that can meet the cost, schedule, and functional requirements of the customers. Figure 13 illustrates the process of design by customers based on a PFA platform. In the figure, arrows represent data flows, ovals represent processes, and variables in uppercase without subscript represent a set of relevant variables. This process consists of two phases: the front-end customer interaction for analyzing and matching customer needs, and the back-end supporting process for improving the compatibility of customer needs and corporate capabilities. There are two actors in the scenario: the customers and the system supported by the PFA.

4.3.1. Phase I: Customer Needs Acquisition

1. *Capability presentation:* In order to make informed decisions, customers are first informed of the capabilities of the firm, which is in the form of the spectrum of product offerings, product attributes, and their possible levels. By organizing these capabilities, the PFA provides a systematic protocol for customers to explore design options.
2. *Self-explication:* Customers are then asked to prioritize desired attributes for their requirements according to their concern about the difference. Customers must assess the value they attach to each attribute and then specify their degree of relative preference between the most desirable and the least desirable levels. The results of this assessment are a set of W_m reflecting the relative importance of each attribute.
3. Utility exploration: Based on W_m, the next task is to find a set of $d_{rml}^{(0)}$ that reflect the desirability of attribute levels. Response surface can be applied here to create a set of testing profiles to search for the value of desirability of each selected level. The AHP can be used to estimate $d_{rml}^{(0)}$. Substituting W_m and $d_{rml}^{(0)}$ in Eq. (9), the utility of each attribute level can be derived.

4.3.2. Phase II: Product Design

1. *Preliminary design:* With $d_{rml}^{(0)}$ and W_m, $U_{ml}^{(0)}$ can be calculated with Eq. (8). A base product (BP) can be determined in accordance with a utility value close to $U_{ml}^{(0)}$. Base product selection can be further fine-tuned through iterative refinement of U_{ml}.
2. *Customization:* Customers can modify the attributes from Z to $Z + \Delta Z$ through the customization process of adding building blocks. Z will be adjusted, and the utility will be recalculated, until the customer gets a satisfactory solution.
3. *Documentation:* After it is confirmed by the customers, the design can be delivered. The results include refined Z and ΔZ and customized BP and ΔBP. These will be documented for the continuous evolution of PFA. Over time, the PFA can be updated so that prospective customers can be better served. This includes changes not only in the offerings of product families but also in production capabilities so that the capabilities of a firm can be better focused to the needs of its customers.

In practice, customers may found this systematic selection process too cumbersome and time consuming. Research in the area of customer decision-making process is still undergoing.

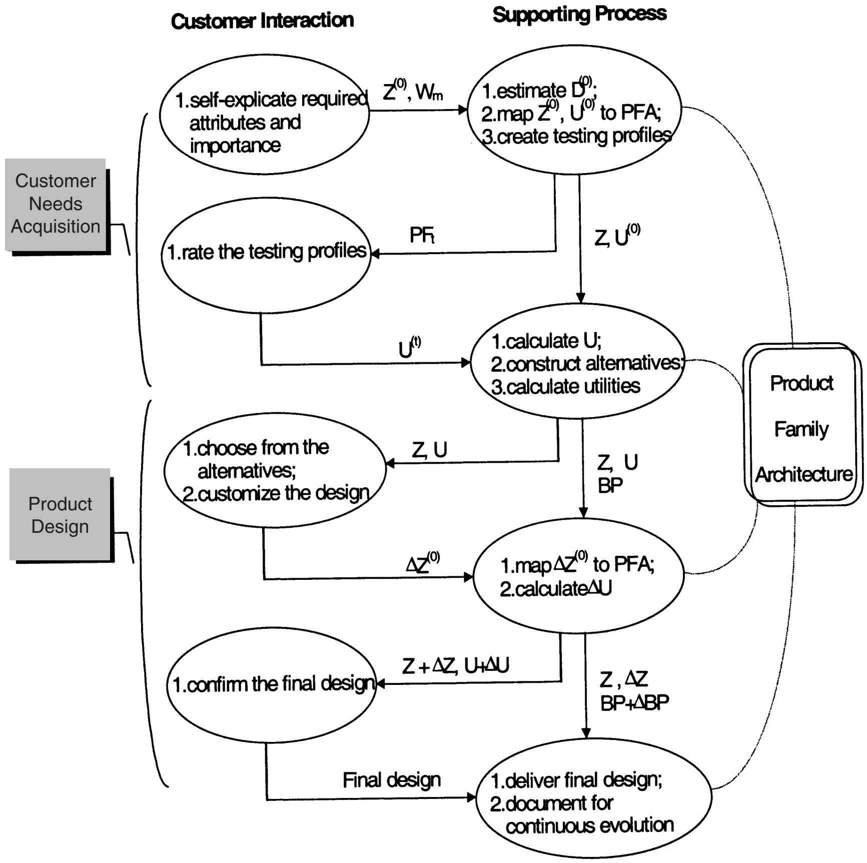

Customer Interaction
Supporting Process
Customer Needs Acquisition
Product Design
1.self-explicate required attributes and importance
Z(0), Wm
1.estimate D(0);
2.map Z(0), U(0) to PFA;
3.create testing profiles
PFt
1.rate the testing profiles
Z, U(0)
U(t)
1.calculate U;
2.construct alternatives;
3.calculate utilities
Product Family Architecture
1.choose from the alternatives;
2.customize the design
Z, U
Z, U BP
ΔZ(0)
1.map ΔZ(0) to PFA;
2.calculate ΔU
1.confirm the final design
Z+ΔZ, U+ΔU
Z, ΔZ BP+ΔBP
Final design
1.deliver final design;
2.document for continuous evolution

Figure 13 The Process of Customer Decision Making.

4.4. One-to-One Marketing

With the rapid transmission of mass media, globalization, and diverse customer bases, the market is no longer homogeneous and stable. Instead, many segmented markets now coexist simultaneously and experience constant changes. Customers cast their votes through purchasing to express their preferences on products. With the fierce competition, customers can easily switch from one company to another. In addition, they become less loyal to a particular brand. Because a certain portion of customers may bring more value-added to the business than the average customers, it is imperative to keep customer loyalty by putting customers at the center point of business. Such a concept has been studied by many researchers (e.g., Greyser 1997) Peppers and Rogers (1997, p. 22) state: "The 1:1 enterprise practices 1:1 marketing by tracking customers individually, interacting with them, and integrating the feedback from each customer into its behavior towards that customer."

With the growing popularity of e-commerce, customers can directly interact with product and service providers on a real-time basis. With the help of system support, each individual customer can specify his or her needs and make informed choices. In the meantime, the providers can directly conduct market research with more precise grasp of customer profiles. This will replace old marketing models (Greyser 1997). At the beginning, the concern is the capability of manufacturers to make products—that is, it is production oriented. Then the focus shifts towards the capability to sell products that have already been made—that is, it is sales oriented. Later on, the theme is customers' preferences and how to accommodate these preferences with respect to company capabilities—that is, it is marketing oriented. With the paradigm shift towards mass customization, manufacturers aim at providing best values to meet customers' individual needs within a short period. The closed-loop interaction between clients and providers on a one-to-one basis will increase the efficiency of matching buying and selling.

Furthermore, the advent of one-to-one marketing transcends geographical and national boundaries. The profile of customers' individual data can be accessed from anywhere, and in turn the company can serve the customers from anywhere at any time. Nonetheless, it also creates a borderless global market that leads to global competition. To be able to sustain, companies are putting more attention on one-to-one marketing.

5. MASS CUSTOMIZATION AND E-COMMERCE

The Internet is becoming a pervasive communication infrastructure connecting a growing number of users in corporations and institutions worldwide and hence providing immense business opportunities for manufacturing enterprises. The Internet has shown its capability to connect customers, suppliers, producers, logistics providers, and almost every stage in the manufacturing value chain. Leading companies have already started to reengineer their key processes, such as new product development and fulfillment, to best utilize the high speed and low cost of the Internet. Impressive results have been reported with significant reduction in lead time, customer value enhancements, and customer satisfaction improvement. Some even predict that a new industrial revolution has already quietly started, geared towards e-commerce-enabled mass customization (*Economist* 2000; Helander and Jiao 2000).

In essence, mass customization attempts to bring customers and company capabilities closer together. With the Internet, customers and providers in different stages of production can be connected at multiple levels of the Web. How this new capability will be utilized is still at a very early stage. For examples, customers can be better informed about important features and the related costs and limitations. Customers can then make educated choices in a better way. In the meantime, through these interactions the company will then be able to acquire information about customers' needs and preferences and can consequently build up its capabilities in response to these needs. Therefore, e-commerce will be a major driving force and an important enabler for shaping the future of mass customization.

Rapid communication over the Internet will revolutionize not only trade but also all the business functions. A paradigm of electronic design and commerce (eDC) can be envisioned as shown in Figure 14. Further expanded to the entire enterprise, it is often referred to as electronic enterprise (eEnterprise). Three pillars support eDC or eEnterprise: the integrated product life cycle, mass customization, and the supply chain.

The integrated product life cycle incorporates elements that are essential to companies, including marketing/sales, design, manufacturing, assembly, and logistics. Using the Internet, some of these activities may be handed over to the supply chain. There may also be other companies similar to the regular supply chain that supplies services. These constitute business-to-service functions.

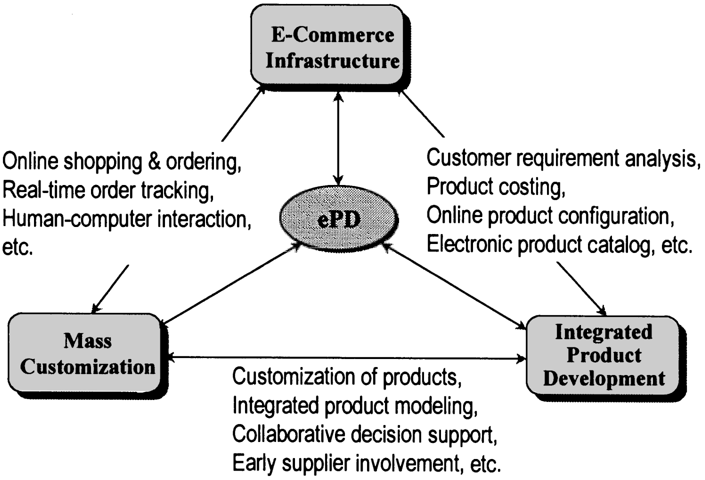

E-Commerce Infrastructure
Online shopping & ordering,
Real-time order tracking,
Human-computer interaction,
etc.
ePD
Customer requirement analysis,
Product costing,
Online product configuration,
Electronic product catalog, etc.
Mass Customization
Integrated Product Development
Customization of products,
Integrated product modeling,
Collaborative decision support,
Early supplier involvement, etc.

Figure 14 A Systems Model of eDC/eEnterprise.

With the communication and interactivity brought by the Internet, the physical locations of companies may no longer be important. The common business model with a core company that engages in design, manufacturing, and logistics will become less common. Manufacturing, as well as design and logistics, may, for example, be conducted by outside service companies. As a result, in view of the supply chain and service companies, the core of business-to-business e-commerce is flourishing.

Figure 14 also illustrates a systems approach to manufacturing. It is a dynamic system with feedback. For each new product or customized design, one must go around the loop. The purpose is to obtain information from marketing and other sources to estimate customer needs. The product life cycle is therefore illustrated by one full circle around the system.

The company can sell products to distributors and/or directly to customers through business-to-customer e-commerce. In some cases, products may be designed by the customer himself or herself. This is related to mass customization. Customer needs are then captured directly through the customers' preferences—the customers understand what they want and can submit their preferred design electronically. A well-known example is Dell Computers, where customers can select the elements that constitute a computer according to their own preferences.

Usually information about customer needs may be delivered by sales and marketing. Typically, they rely on analyses of customer feedback and predictions for the future. These remain important sources of information for new product development. From this kind of information, the company may redesign existing products or decide to develop a new one. The design effort has to take place concurrently, with many experts involved representing various areas of expertise and parties that collaborate through the system. The supply chain companies may also participate if necessary.

For manufacturing the product, parts and/or other services may be bought from the supply chain and delivered just-in-time to manufacturing facilities. These constitute typical business-to-business e-commerce.

Some important technical issues associated with eDC include human–computer interaction and usability (Helander and Khalid 2001), the customer decision-making process over the Internet (So et al. 1999), product line planning and electronic catalog, and Web-based collaborative design modeling and design support.

6. SUMMARY

Mass customization aims at better serving customers with products and services that are closer to their needs and building products upon economy of scale leading to mass production efficiency. To this end, an orchestrated effort in the entire product life cycle, from design to recycle, is necessary. The challenge lies in how to leverage product families and how to achieve synergy among different functional capabilities in the value chain. This may lead to significant impact on the organizational structure of company in terms of new methods, education, division of labor in marketing, sales, design, and manufacturing. The technological roadmap of mass customization can also lead to redefinition of job, methodology, and investment strategies as witnessed in current practice. For instance, the sales department will be able to position itself to sell its capabilities instead of a group of point products.

As a new frontier of business competition and production paradigm, mass customization has emerged as a critical issue. Mass customization can best be realized by grounding up, instead of by directly synthesizing, existing thrusts of advanced manufacturing technologies, such as JIT, flexible, lean and agile manufacturing, and many others. Obviously, much needs to be done. This chapter provides materials for stimulating an open discussion on further exploration of mass customization techniques.

REFERENCES

Adelsberger, H. H., and Conen, W. (1995), "Scheduling Utilizing Market Models," in *Proceedings of the 3rd International Conference on CIM Technology* (Singapore), pp. 695–702.

Baker, A. D. (1991), "Manufacturing Control with a Market-Driven Contract Net," Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, NY.

Baldwin, R. A., and Chung, M. J. (1995), "Managing Engineering Data for Complex Products," *Research in Engineering Design,* Vol. 7, pp. 215–231.

Boynton, A. C., and Bart, V. (1991), "Beyond Flexibility: Building and Managing the Dynamically Stable Organization," *California Management Review,* Vol. 34, No. 1, pp. 53–66.

Burke, P., and Prosser, P. (1991), "A Distributed Asynchronous System for Predictive and Reactive Scheduling," *Artificial Intelligence in Engineering,* Vol. 6, No. 3, pp. 106–124.

Choi, S. C., and Desarbo, W. S. (1994), "A Conjoint-Based Product Designing Procedure Incorporating Price Competition," *Journal of Product Innovation Management,* Vol. 11, pp. 451–459.

Choi, S. C., Desarbo, W. S., and Harker, P. T. (1990), "Product Positioning under Price Competition," *Management Science,* Vol. 36, pp. 175–199.

Clearwater, S. H., Ed. (1996), *Market-Based Control: A Paradigm for Distributed Resource Management,* World Scientific, Singapore.

Davis, S. M. (1987), *Future Perfect,* Addison-Wesley, Reading, MA.

Economist, The (2000), "All yours: The Dream of Mass Customization," Vol. 355, No. 8164, pp. 57–58.

Erens, F. J., and Hegge, H. M. H. (1994), "Manufacturing and Sales Co-ordination for Product Variety," *International Journal of Production Economics,* Vol. 37, No. 1, pp. 83–99.

Erens, F., McKay, A., and Bloor, S. (1994), "Product Modeling Using Multiple Levels of Abstraction Instances as Types," *Computers in Industry,* Vol. 24, No. 1, pp. 17–28.

Erixon, G. (1996), "Design for Modularity," in *Design for X: Concurrent Engineering Imperatives,* G. Q. Huang, Ed., Chapman & Hall, New York, pp. 356–379.

Feitzinger, E., and Lee, H. L. (1997), "Mass Customization at Hewlett-Packard: The Power of Postponement," *Harvard Business Review,* Vol. 75, pp. 116–121.

Green, P. E., and Krieger, A. M. (1989), "Recent Contributions to Optimal Product Positioning and Buyer Segmentation," *European Journal of Operational Research,* Vol. 41, No. 2, pp. 127–141.

Green, P. E., Krieger, A. M., and Agarwal, M. K. (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research,* Vol. 28, No. 2, pp. 215–222.

Greyser, S. A. (1997), "Janus and Marketing: The Past, Present, and Prospective Future of Marketing," in *Reflections on the Futures of Marketing,* D. R. Lehmann and K. E. Jocz, Eds., Marketing Science Institute, Cambridge, MA.

Gou, L., Hasegawa, T., Luh, P. B., Tamura, S., and Oblak, J. M. (1994), "Holonic Planning and Scheduling for a Robotic Assembly Testbed," in *Proceedings of the 4th International Conference on CIM and Automation Technology* (Troy, NY, October), IEEE, New York, pp. 142–149.

Hatvany, J. (1985), "Intelligence and Cooperation in Heterarchic Manufacturing Systems," *Robotics and Computer Integrated Manufacturing,* Vol. 2, No. 2, pp. 101–104.

Hegge, H. M. H., and Wortmann, J. C. (1991) "Generic Bills-of-Material: A New Product Model," *International Journal of Production Economics,* Vol. 23, Nos. 1–3, pp. 117–128.

Helander, M., and Jiao, J. (2000), "E-Product Development (EPD) for Mass Customization," in *Proceedings of IEEE International Conference on Management of Innovation and Technology,* Singapore, pp. 848–854.

Helander, M. G., and Khalid, H. M. (2001), "Modeling the Customer in Electronic Commerce," Applied Ergonomics, (In Press).

IntelliQuest (1990), *Conjoint Analysis: A Guide for Designing and Integrating Conjoint Studies,* Marketing Research Technique Series Studies, American Marketing Association, Market Research Division, TX.

Ishii, K., Juengel, C., and Eubanks, C. F. (1995a), "Design for Product Variety: Key to Product Line Structuring," in *Proceedings of Design Engineering Technical Conferences,* ASME, DE-Vol. 83, pp. 499–506.

Ishii, K., Lee, B. H., and Eubanks, C. F. (1995b), "Design for Product Retirement and Modularity Based on Technology Life-Cycle," in *Manufacturing Science and Engineering,* MED-Vol. 2-2/ MH-Vol. 3-2, ASME, pp. 921–933.

Krause, F. L., Kimura, F., Kjellberg, T., and Lu, S. C.-Y. (1993), "Product Modeling," *Annals of the CIRP,* Vol. 42, No. 1, pp. 695–706.

Lin, G. Y., and Solberg, J. J. (1992), "Integrated Shop Floor Control Using Autonomous Agents," *IIE Transactions,* Vol. 24, No. 3, pp. 57–71.

Malone, T. W., Fikes, R. E., Grant, K. R., and Howard, M. T. (1988), "Enterprise: A Market-Like Task Scheduler for Distributed Computing Environments," in *The Ecology of Computation,* B. A. Huberman, Ed., North-Holland, Amsterdam, pp. 177–205.

Markus, A., and Monostori, L. (1996), "A Market Approach to Holonic Manufacturing," *Annals of the CIRP,* Vol. 45, No. 1, pp. 433–436.

Martin, M. V., and Ishii, K. (1996), "Design for Variety: A Methodology for Understanding the Costs of Product Proliferation," in *Proceedings of ASME Design Engineering Technical Conferences,* DTM-1610, Irvine, CA.

Martin, M. V., and Ishii, K. (1997), "Design for Variety: Development of Complexity Indices and Design Charts," in *Proceedings of ASME Design Engineering Technical Conferences,* DFM-4359, Sacramento, CA.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models,* 2nd Ed., Chapman & Hall, London.

Meyer, M., and Lehnerd, A. P. (1997), *The Power of Product Platforms: Building Value and Cost Leadership,* Free Press, New York.

Meyer, M. H., and Utterback, J. M. (1993), "The Product Family and the Dynamics of Core Capability," *Sloan Management Review,* Vol. 34, No. 3, pp. 29–47.

Newcomb, P. J., Bras, B., and Rosen, D. W. (1997), "Implications of Modularity on Product Design for the Life Cycle," in *Proceedings of ASME Design Engineering Technical Conferences,* DETC96/DTM-1516, Irvine, CA.

Page, A. L., and Rosenbaum, H. F. (1987), "Redesigning Product Lines with Conjoint Analysis: How Sunbeam Does It," *Journal of Product Innovation Management,* Vol. 4, pp. 120–137.

Peppers, D., and Rogers, M. (1997), *Enterprise One to One,* Currency Doubleday, New York.

Pine, B. J. (1993), *Mass Customization: The New Frontier in Business Competition,* Harvard Business School Press, Boston.

Porter, M. (1986), *Competition in Global Industries,* Harvard Business School Press, Boston.

Rezendes, C. (1997), "More Value-Added Software Bundled with Your Scanners This Year," *Automatic I.D. News,* Vol. 13, No. 1, pp. 29–30.

Roberts, E. B., and Meyer, M. H. (1991), "Product Strategy and Corporate Success," *IEEE Engineering Management Review,* Vol. 19, No. 1, pp. 4–18.

Ryan, N. (1996), "Technology Strategy and Corporate Planning in Australian High-Value-Added Manufacturing Firms," *Technovation,* Vol. 16, No. 4, pp. 195–201.

Sanderson, S., and Uzumeri, M. (1995), "Managing Product Families: The Case of the Sony Walkman," *Research Policy,* Vol. 24, pp. 761–782.

Schreyer, M., and Tseng, M. M. (1998), "Modeling and Design of Control Systems for Mass Customization Manufacturing," in *Proceedings of the 3rd Annual Conference on Industrial Engineering Theories, Applications, and Practice* (Hong Kong).

Shaw, M. J. (1988), "Dynamic Scheduling in Cellular Manufacturing Systems: A Framework for Networked Decision Making," *Journal of Manufacturing Systems,* Vol. 7, No. 2, pp. 83–94.

Sikora, R., and Shaw, M. J. P. (1997), "Coordination Mechanisms for Multi-agent Manufacturing Systems: Application to Integrated Manufacturing Scheduling," *IEEE Transactions on Engineering Management,* Vol. 44, No. 2, pp. 175–187.

So, R. H. Y., Ling, S. H., and Tseng, M. M. (1999), *Customer Behavior in Web-Based Mass-Customization Systems: An Experimental Study on the Buying Decision-Making Process over the Internet,* Hong Kong University of Science and Technology.

Subrahmanian, E., Westerberg, A., and Podnar, G. (1991), "Towards a Shared Computational Environment for Engineering Design," in *Computer-Aided Cooperative Product Development,* D. Sriram, R. Logcher, and S. Fukuda, Eds., Springer, Berlin.

Suh, N. P. (1990), *The Principles of Design,* Oxford University Press, New York.

Suh, N. P. (1997), "Design of Systems," *Annals of the CIRP,* Vol. 46, No. 1, pp. 75–80.

Suzue, T., and Kohdate, A. (1990), *Variety Reduction Program: A Production Strategy for Product Diversification,* Productivity Press, Cambridge, MA.

Teresko, J. (1994), "Mass Customization or Mass Confusion," *Industrial Week,* Vol. 243, No. 12, pp. 45–48.

Toffler, A. (1971), *Future Shock,* Bantam Books, New York.

Tseng, M. M., and Jiao, J. (1996), "Design for Mass Customization," *Annals of the CIRP,* Vol. 45, No. 1, pp. 153–156.

Tseng, M. M., and Du, X. (1998), "Design by Customers for Mass Customization Products," *Annals of the CIRP,* Vol. 47, No. 1, pp. 103–106.

Tseng, M. M., Lei, M., and Su, C. J. (1997), "A Collaborative Control System for Mass Customization Manufacturing," *Annals of the CIRP,* Vol. 46, No. 1, pp. 373–377.

Ulrich, K. (1995), "The Role of Product Architecture in the Manufacturing Firm," *Research Policy,* Vol. 24, pp. 419–440.

Ulrich, K. T., and Eppinger, S. D. (1995), *Product Design and Development,* McGraw-Hill, New York.

Ulrich, K. T., and Seering, W. P. (1990), "Function Sharing in Mechanical Design," *Design Studies,* Vol. 11, pp. 223–234.

Ulrich, K., and Tung, K. (1991), "Fundaments of Product Modularity," in *Issues in Mechanical Design International 1991,* A. Sharon, Ed., ASME DE-39, New York, pp. 73–79.

Upton, D. M., and Barash, M. M. (1991), ‘‘Architectures and Auctions in Manufacturing,’’ *International Journal of Computer Integrated Manufacturing,* Vol. 4, No. 1, pp. 23–33.

Urban, G. L., and Hauser, J. R. (1993), *Design and Marketing of New Products,* 2nd Ed., Prentice Hall, Englewood Cliffs, NJ.

van Veen, E. A. (1992), *Modeling Product Structures by Generic Bills-of-Material,* Elsevier, New York.

CHAPTER 26

Client/Server Technology

On Hashida
University of Tsukuba

Hiroyuki Sakata
NTT DATA Corporation

1. INTRODUCTION 711
- 1.1. Concept of C/S Systems 711
- 1.2. Roles of Client and Server 711
- 1.3. Computer Technology Trends 712
  - 1.3.1. Web Technologies 712
  - 1.3.2. Open System Technologies 714

2. FEATURES OF C/S SYSTEMS 714
- 2.1. Advantages 714
- 2.2. Disadvantages 714

3. C/S SYSTEM ARCHITECTURES 715
- 3.1. Functional Elements of C/S Systems 715
- 3.2. Two-Tier Architecture 715
- 3.3. Three-Tier Architecture 716

4. FUNDAMENTAL TECHNOLOGIES FOR C/S SYSTEMS 718
- 4.1. Communication Methods 718
  - 4.1.1. Socket 718
  - 4.1.2. Remote Procedure Call 719
  - 4.1.3. CORBA 719
  - 4.1.4. Other Communication Methods 721
- 4.2. Distributed Transaction Management 721
- 4.3. Distributed Data Management 723

5. CAPACITY PLANNING AND PERFORMANCE MANAGEMENT 723
- 5.1. Objectives of Capacity Planning 723
- 5.2. Steps for Capacity Planning and Design 725
- 5.3. Performance Objectives and System Environment 726
- 5.4. Performance Criteria 726
- 5.5. Workload Modeling 727
  - 5.5.1. Workload Characterization 727
  - 5.5.2. Workload Modeling Methodology 727
- 5.6. Performance Evaluation 728
  - 5.6.1. Analytical Models 728
  - 5.6.2. Simulation 728
  - 5.6.3. Benchmarking 729
  - 5.6.4. Comparing Analysis and Simulation 729

6. MAINTENANCE AND ADMINISTRATION OF C/S SYSTEMS 729
- 6.1. Architecture of System Management 729
  - 6.1.1. OSI Management Framework 729
  - 6.1.2. System Management Architecture 730
- 6.2. Network Management Protocol 730
- 6.3. Security Management 732
  - 6.3.1. Threats 732
  - 6.3.2. Security Services 732
  - 6.3.3. Security Technologies 733

7. A PRACTICAL EXAMPLE: INTERNET BANKING SYSTEM 735

ADDITIONAL READING 736

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

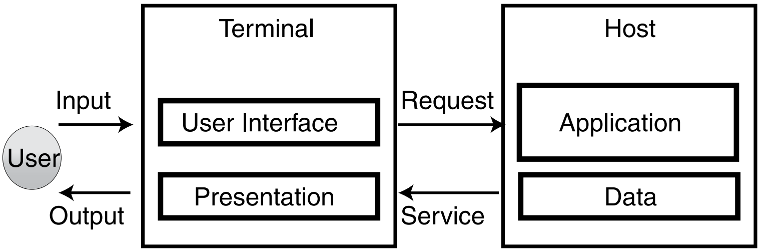


Figure 1 Host-Centered Processing System.

1. INTRODUCTION

1.1. Concept of C/S Systems

In a traditional, centralized mainframe system, a user interacts with an application via a dumb terminal that has a keyboard for entering commands and a display to show the results. All applications and databases are placed on the mainframe (host), and the terminal does not have any ability to process applications. That is, the functions of the user interface and the presentation are placed on the terminal and other applications are placed on the host side, as shown in Figure 1.

In a client/server (C/S) system, by contrast, a user advances processing by using a program called a server, from a workstation or personal computer connected to the server computer by network. The program that receives services from the server is called a client. As in the host-centric system, the functions of user interface and presentation are placed on the user (client) side. Databases that are used for processing requests are placed on the server side. Each application is divided into several processes that run on the client and server sides and cooperate for processing the application. The difference between the two systems is the way of distribution of applications. As shown in Figure 2, some applications can be placed on the user side in the C/S system. Figure 3 shows the physical and logical structures for a host-centric system and a C/S system.

In the physical structure of the host-centric system, applications and databases are placed on a mainframe. That is, the physical structure is almost the same as the logical structure described above. On the other hand, in the C/S system, all the processes that comprise the service are not necessarily placed in the same computer. In many cases, they are distributed in separate computers connected to each other via network.

The features of C/S systems are:

- They contain the basic functional components: presentation, application service, and database.
- Generally, each application service is divided into several processes.
- In many cases, the service is offered by cooperative work of several processes placed on separate computers.

1.2. Roles of Client and Server

There is a many-to-one relationship between clients and servers, but the relationship is relative. In some cases, a client may pass a reference to a callback object when it invokes a service. This lets

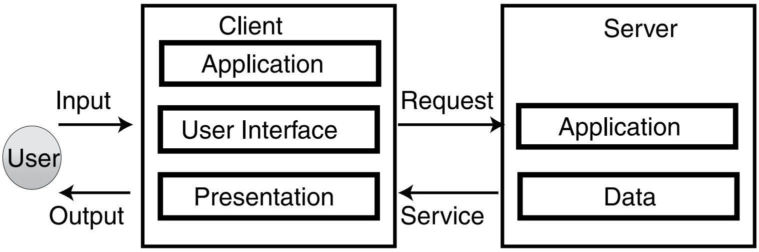


Figure 2 Relationship of Client and Server.

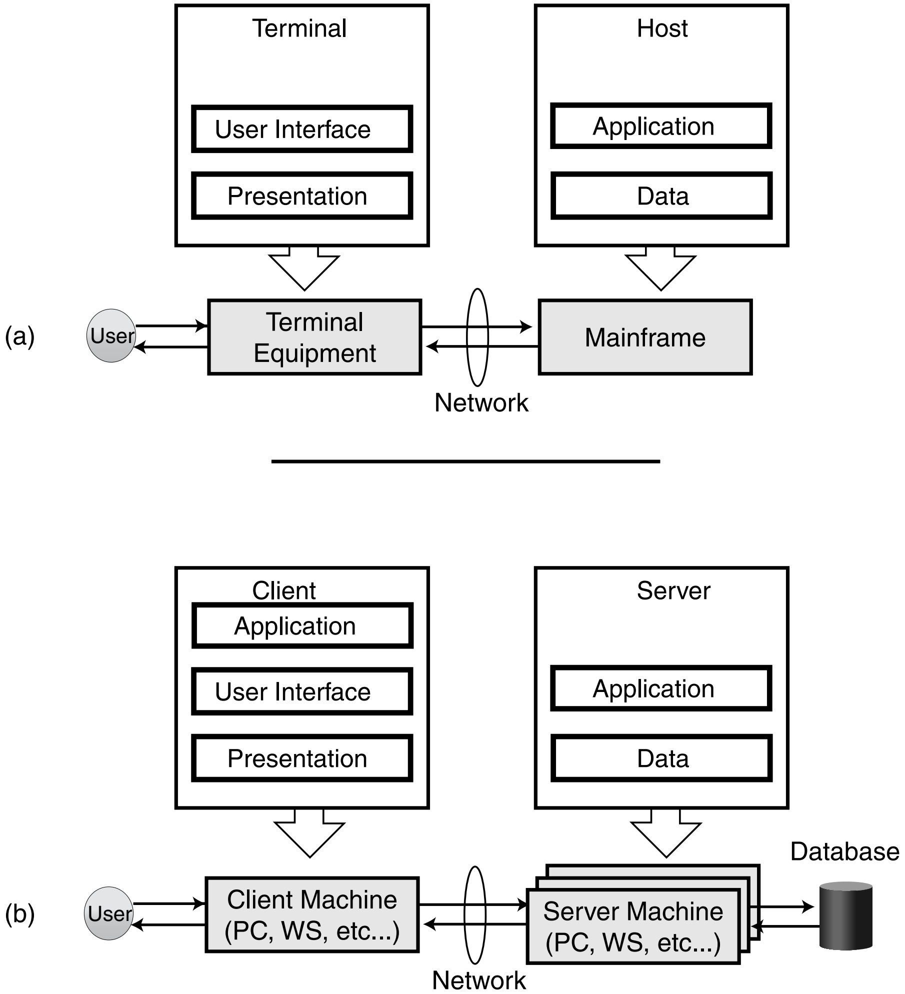

Terminal
User Interface
Presentation
Host
Application
Data
(a)
User
Terminal Equipment
Mainframe
Network
Client
Application
User Interface
Presentation
Server
Application
Data
Database
(b)
User
Client Machine (PC, WS, etc...)
Server Machine (PC, WS, etc...)
Network

Figure 3 Logical Structure vs. Physical Structure. (*a*) Host processing model. (*b*) C/S processing model.

the server call back the client, so the client becomes a server. For example, in an Internet banking system, a user program (WWW browser) is a client and a program located at the bank that offers services to the client is a server. However, the role of the program located at the bank is not always fixed to the server. To offer a fund transfer service to the user, the program located at bank A asks for a fund acceptance process that is located at bank B. In this case, the program at bank A becomes a client and the program at bank B becomes a server. Thus, the roles of the client and server change dynamically depending on the circumstances. In this chapter, ‘‘client’’ means a program that is located on the user side and provides the user interface and presentation functions, and ‘‘server’’ means a program that receives requests from the client and provides services to the client.

1.3. Computer Technology Trends

1.3.1. Web Technologies

A C/S system is realized by a network over which clients and servers work together to accomplish a task. To develop the network applications, it is important for a developer to select an appropriate communication protocol from the viewpoint of cost and performance. The communication protocol is a set of rules by which reliable end-to-end communication between two processes over a network is achieved. In recent years, the TCP/IP protocol stack has been adopted for constructing many network applications and has become the standard protocol for the Internet. World Wide Web (or Web) technologies using HTTP (Hypertext Transfer Protocol), an application layer protocol based on TCP/IP, have been applied to C/S systems. At present, Web technologies are the most widespread for C/S applications. The client side of the Web application is called the browser. Microsoft's Internet Explorer and Netscape's Navigator are the most widely used browsers.

Figure 4 shows the progression of the Web structure of C/S applications:

- At the initial stage, the main service of the Web system was distribution of static information such as a text or an image stored in the storage device of the server machine (the Web server). (Figure 4[*a*]).
- In the middle stage, in addition to distribution of information, more interactive services were provided to the user by applications added to the Web server. For example, in an online shopping service, monthly expenditure is calculated using inputs of the purchase amount of daily shopping (Figure 4[*b*]).
- In the recent stage, external systems such as legacy systems are placed on the back end of the Web server and more advanced services can be provided. An Internet banking service is realized by this structure (Figure 4[*c*]).

The advantage of a C/S system using the Web technologies is that users can unify various kinds of operations and the maintenance of the client software becomes easier. In the future, the proliferation of the Internet means that Web technologies will gain more popularity for realizing C/S systems.

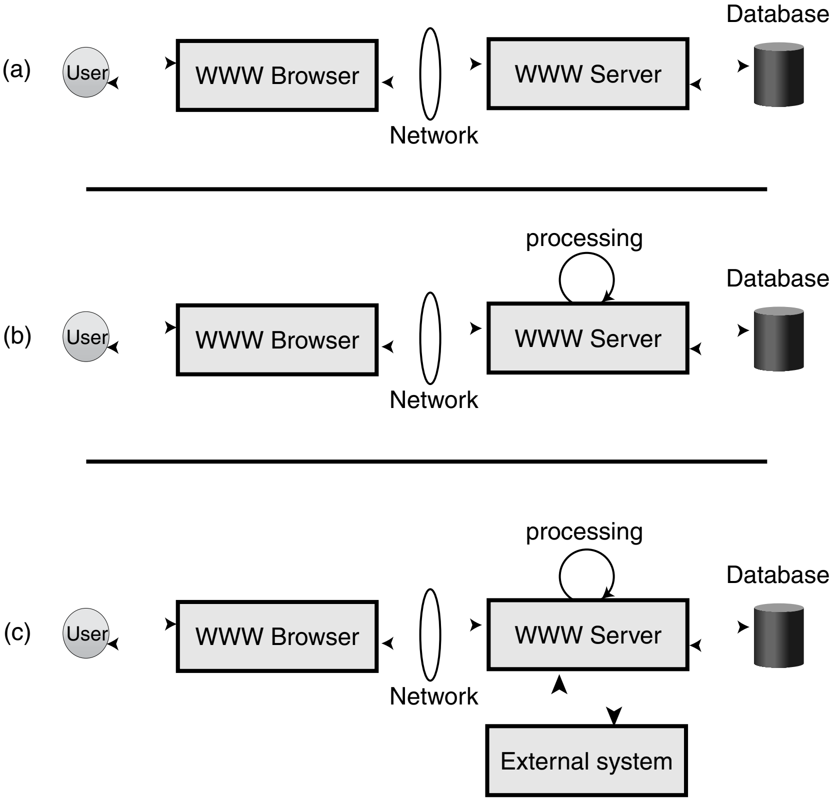


Figure 4 Progressive Stages of Web-Style C/S System. (*a*) WWW information-providing system. (*b*) WWW information-processing system. (*c*) WWW information-processing system with external systems processing.

1.3.2. Open System Technologies

A C/S system is one of distributed models of computing in the environment of heterogeneous computers. This is because open system technologies such as communication protocols and developing languages are required. For example, the adoption of the Java language as the developing language has increased, as has the adoption of CORBA (common object request broker architecture) as interconnection technology. Java is a new object-oriented programming language from Sun Microsystems and a portable operating system environment that enables the writing of portable components. CORBA is a standard architecture created by the OMG (Object Management Group) for a message broker called object request broker, a software component that enables communication among objects in a distributed environment. The common feature of Java and CORBA is platform independence. The most commonly used operating system on which C/S systems are developed is UNIX, which mainly works on workstations or Windows NT, which mainly runs on server-class personal computers. Java and CORBA are designed for running on both operating systems and cooperate mutually. In the future, technology like Java or CORBA that works in heterogeneous computing environments will become more and more important.

Due to the recent progress in hardware technologies, the processing capacity of a personal computer has become almost equal to that of a workstation. Therefore, there are too many choices among various technologies to allow appropriate system components to be selected. The system configuration has become more complex because of a combination of distributed components as well as heterogeneity.

2. FEATURES OF C/S SYSTEMS

2.1. Advantages

The advantages of C/S systems derive from their distributed processing structure, which C/S systems adopt, where processing is handled by two or more cooperating geographically distinct processors.

1. *Performance tuning:* In a distributed processing environment, the service requests from a client are not concentrated on a single server but can be distributed among several servers. Therefore, it becomes possible to distribute required workload among processing resources and improve the performance, such as response time to the client. Furthermore, adopting an efficient algorithm of request distribution such as an adaptive algorithm taking account of current loads of all servers makes it possible to improve the throughput (the amount of work processed per unit time) of the system.
2. *Availability improvement:* Clients are sensitive to performance of the system, but they do not perceive the physical structure of the distribution for requests. In other words, the group of networked servers seems to the client like one server. In a distributed processing environment, even if one server breaks, other servers can substitute for the broken server and continue to process its tasks. This can guarantee the high availability of services to the client compared to a host-centric processing environment.
3. *Scalability and cost-efficiency:* Scalability is the capacity of the system to perform more total work in the same elapsed time when its processing power is increased. With a distributed system, the processing capacity of the system can be scaled incrementally by adding new computers or network elements as the need arises, and excessive investment in a system can be avoided at the introduction stage of the C/S system. On the other hand, with a host-centric system, if the load level is going to saturate the system, the current system will be replaced by a more powerful computer. If further growth is planned, redundant computer capacity will need to be built into the current system to cope with future growth in requirements.

2.2. Disadvantages

1. *Strong dependence on the network infrastructure:* Because a distributed environment is largely based on a networking infrastructure, the performance and availability of the distributed system are strongly influenced by the state of the network. For example, if a network failure, such as of the router or transmission line, occurs, performance and availability for services may seriously deteriorate. The system designer and manager should design the fault-tolerant system and plan the disaster recovery taking account of the networking infrastructure.
2. *Security:* From the viewpoint of security, the possibility exists that confidential information stored on the database may leak out through the network. The system manager should take security measures such as authentication, access control, and cryptographic control according to the security level of information.
3. *Complexity of system configuration:* C/S systems are composed of several components from multiple vendors. Distributed applications often have more complex functionality than cen-

tralized applications, and they are built from diverse components. Multitier architectures, which will be explained in the next section, provide a nearly limitless set of choices for where to locate processing functions and data. In addition to choices about locations, there are also more hardware and software choices, and more data sharing occurs. These cause the difficulty in developing a C/S system and the complexity of managing it.

3. C/S SYSTEM ARCHITECTURES

3.1. Functional Elements of C/S Systems

Since vendors began releasing RDBMS (relational database management system) products in the 1970s, the processing model in which various business data are divided into distributed databases and are accessed via network has been widely adopted. Client/server systems are composed of the various functional elements associated with data processing.

The logical functions of C/S systems are roughly divided into three layers:

1. *Presentation logic:* This function contains all the logic needed to manage screen formats, the content of windows, and interaction with the user, that is, the user interface. In recent years, the use of graphical user interface (GUI) features such as buttons has become general.
2. *Application logic:* This is a generic name for application functions that do not belong to other layers. It includes business logic and the flow of control among application components.
3. *Data logic:* This contains all the logic relating to the storage and retrieval of data, and enforcing business rules about data consistency. Generally, databases are treated and are maintained by a DBMS (database management system).

In addition to these application-functional layers, some functions are needed to support interactions between clients and servers via networks. All the distributed software that supports interaction between application components and network software is called middleware. Middleware is a generic term for all the software components that allow us to connect separate layers or components and put them into a complete distributed system. It provides an application programming interface (API) that isolates application codes from the underlying network communication formats and protocols. It also supplies intermediate system services such as security, naming, directory, messaging, and transaction management services:

1. *Naming service:* In C/S systems, names of clients and servers must be unique within the range in which they are used. A naming service uniquely names a physical entity such as a client or server and associates logical structure such as a tree with physical entities.
2. *Directory service:* This is a directory service that provides a way for clients to locate servers and their services on the network and controls the address of messages and processing requests through a tree structure.
3. *Security service:* This service provides authentication, authorization, access control, and user account management.
4. *Messaging service:* This service provides the ability to send and receive messages between and among applications and users.
5. *Transaction service:* This service manages distributed transaction processing, such as consistency control, that maintains a database in a consistent state and enables recovery control from failures.

3.2. Two-Tier Architecture

Client/server architectures can be characterized by how the applications are distributed between the client and the server. In a two-tier architecture, the application logic is placed on either the client or the server undivided or split into two parts, which are placed on the client and the server respectively. In most cases, the presentation logic and the application logic are placed on the client side and the data logic is placed on the server side. Examples of two-tier C/S systems are file servers and database servers with stored procedures. Figure 5 shows a two-tier architecture.

The advantage of the two-tier architecture is ease of development. Because the two-tier architecture is simple, a developer can create applications quickly using a 4GL (fourth-generation language) development environment such as Microsoft's Visual Basic or Inprise's Delphi. However, as C/S systems grew up to run mission-critical or enterprise-wide applications, shortcomings of the two-tier architecture emerged:

1. *Performance:* Performance may deteriorate markedly as the number of clients, the size of the database, or the volume of transferred data increases. This deterioration is caused by a lack of capacity of resources such as the processing unit, the memory area, and the network bandwidth

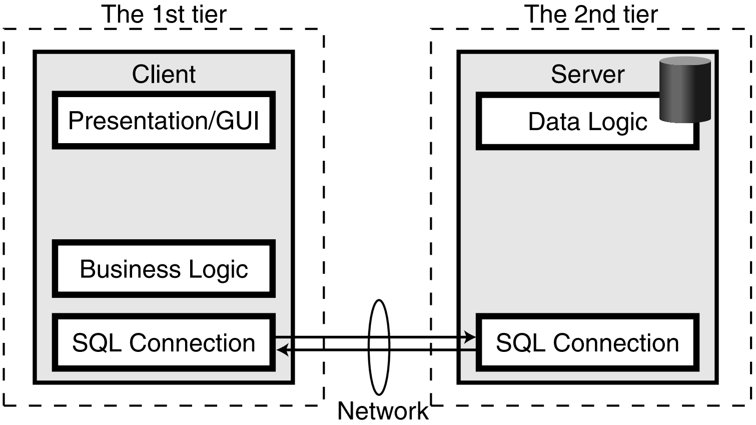


Figure 5 Two-Tier Architecture.

when the processing load is concentrated on the database server. Placing a part of application logic on the database server allows network traffic between the client and the database server to be reduced and the performance even improved. For example, a set of database access commands is compiled and saved in the database of the server. This scheme is called stored procedure. Some DBMS packages support this scheme. But there remains the problem that the application maintenance becomes complicated.

2. *Maintenance:* Because the business logic is placed on the client side in most two-tier C/S systems, version control of applications becomes complex. For example, in the case of an online shopping system, the business logic (or rule), such as tax rate, is implemented on the client side. Although this structure is effective in reducing network traffic between the client and the server, a system manager has to replace the application of all the clients every time the business rule or version of application is changed. This becomes a big problem in the total cost of ownership of two-tier C/S systems: costs of maintaining, operating, and managing large-scale systems accommodating many users and services.

3.3. Three-Tier Architecture

To solve some problems of the two-tier architecture and improve the reliability and performance of the system, a three-tier architecture has been adopted. In a three-tier architecture, the business logic is separated from the presentation and data layers and becomes the middle layer between them. The presentation logic resides in the first tier (the client side), and the data logic in the third tier (the server side), and the business logic in the second tier (the middle) between both tiers. This solves the problems occurring with the two-tier architecture. Figure 6 shows a three-tier architecture where the business logic and some connection functions are placed in the second tier.

In a three-tier architecture, the second tier plays the most important role. Many functions such as data access and connection with other systems are located in the second tier and can be used by any client. That is, the second tier becomes the gateway to other systems. The second tier is often called the application server.

The advantages of three-tier architecture derived from the application server are:

1. *Reduction of network traffic:* Concentrating the function of accessing database and other systems on the application server makes it possible to reduce network traffic between the client and the server. That is, instead of interacting with the database directly, the client calls the business logic on the application server, and then the business logic accesses the database on the database server on behalf of the client. Therefore, only service requests and responses are sent between the client and the server. From the viewpoint of network traffic, comparisons of two-tier architecture and three-tier architecture are shown in Figure 7.
2. *Scalability:* Adding or removing application servers or database servers allows the system to be scaled incrementally according to the number of clients and volume of requests.
3. *Performance:* Because workloads can be distributed across application servers and database servers, this architecture can prevent an application server from becoming a bottleneck and can keep the performance of the system in a permissible range.

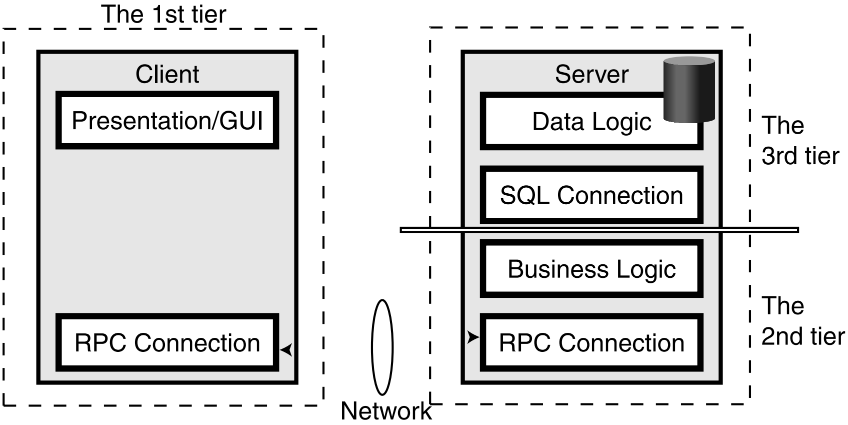

The 1st tier
Client
Presentation/GUI
RPC Connection
Network
Server
Data Logic
SQL Connection
Business Logic
RPC Connection
The 3rd tier
The 2nd tier

Figure 6 Three-Tier Architecture.

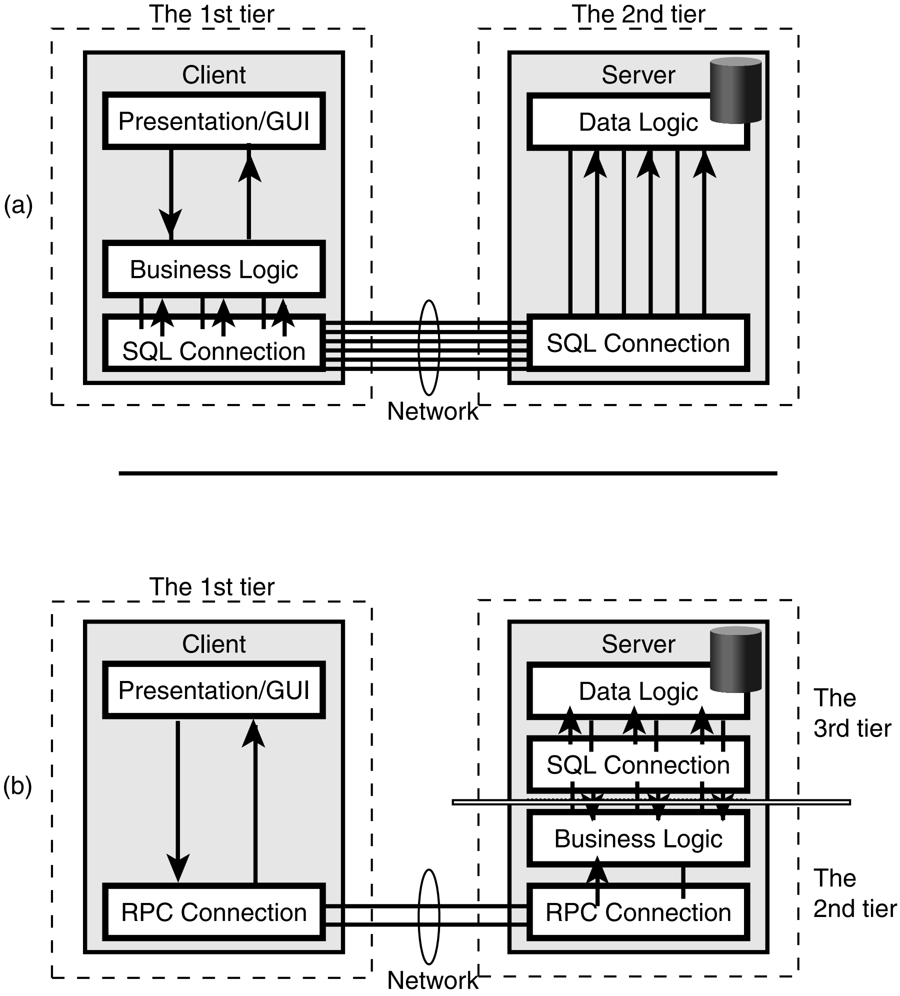

The 1st tier
The 2nd tier
(a)
Client
Presentation/GUI
Business Logic
SQL Connection
Network
Server
Data Logic
SQL Connection
The 1st tier
(b)
Client
Presentation/GUI
RPC Connection
Network
Server
Data Logic
SQL Connection
Business Logic
RPC Connection
The 3rd tier
The 2nd tier

Figure 7 Network Traffic Comparison. (*a*) Two-tier architecture model. (*b*) Three-tier architecture model.

4. *Availability:* Even if an application server fails during processing the business logic, the client can restart the business logic on other application servers.

4. FUNDAMENTAL TECHNOLOGIES FOR C/S SYSTEMS

4.1. Communication Methods

The communication method facilitating interactions between a client and a server is the most important part of a C/S system. For easy development of the C/S systems, it is important for processes located on physically distributed computers to be seen as if they were placed in the same machine. Socket, remote procedure call, and CORBA are the main interfaces that realize such transparent communications between distributed processes.

4.1.1. Socket

The socket interface, developed as the communication port of Berkeley UNIX, is an API that enables communications between processes through networks like input/output access to a local file. Communications between two processes by using the socket interface are realized by the following steps (see Figure 8):

1. For two processes in different computers to communicate with each other, a communication port on each computer must be created beforehand by using ‘‘socket’’ system call.
2. Each socket is given a unique name to recognize the communication partner by using ‘‘bind’’ system call. The named socket is registered to the system.
3. At the server process, the socket is prepared for communication and it is shown that the server process is possible to accept communication by using ‘‘listen’’ system call.
4. The client process connects to the server process by using ‘‘connect’’ system call.

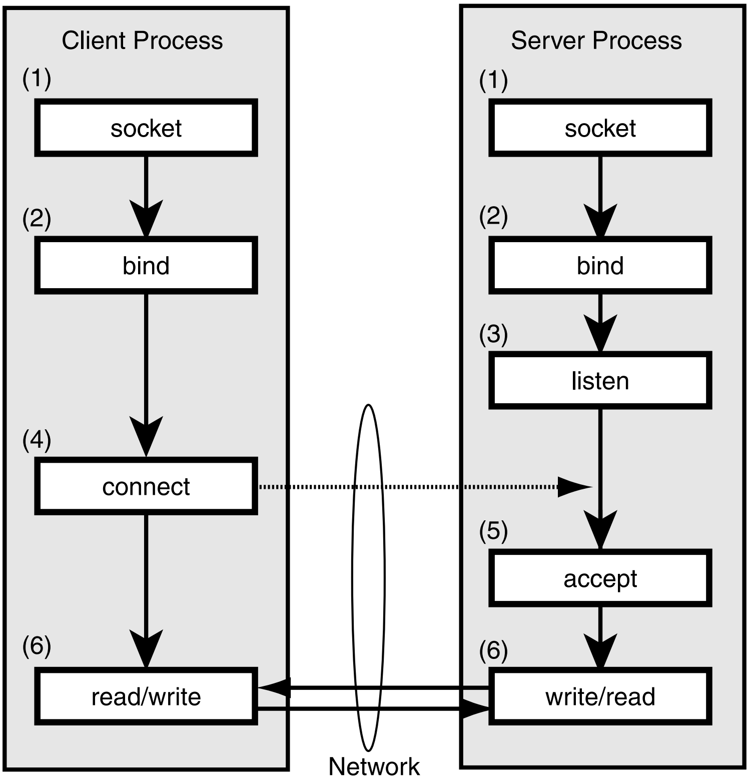


Figure 8 Interprocess Communication via Socket Interface.

5. The communication request from the client process is accepted by using "accept" system call in the server process. Then the connection between the client process and the server process is established.
6. Finally, each process begins communicating mutually by using "read" or "write" system calls.

4.1.2. Remote Procedure Call

Because the application programming interface (API) for socket programming uses system calls, the overhead associated with an application that communicates through the socket interface is rather small. However, because API is very primitive, socket programming depends on operating systems and the development of a system with a socket interface becomes complicated. For example, where programmers develop a system in which communications between processes on different platforms (such as Windows and UNIX) are required, they should master several versions of socket API supported by each platform.

One of the earliest approaches to facilitating easier use of sockets is realized by a remote procedure call (RPC). An RPC is a mechanism that lets a program call a procedure located on a remote server in the same fashion as a local one within the same program. It provides a function-oriented interface, and necessary preparation for communication is offered beforehand. An RPC is realized by a mechanism called a stub. The functions of stubs at client and server sides are to mimic the missing code, convert the procedure's parameters into messages suitable for transmission across the network (a process called marshaling) and unmarshal the parameters to a procedure, and dispatch incoming calls to the appropriate procedure.

Communications between two processes by using an RPC is realized by the following steps (see Figure 9):

1. The client program invokes a remote procedure function called the client stub.
2. The client stub packages (marshals) the procedure's parameters in several RPC messages and uses the runtime library to send them to the server.
3. At the server, the server stub unpacks (unmarshals) the parameters and invokes the requested procedure.
4. The procedure processes the request.
5. The results are sent back to the client through the stubs on both sides that perform the reverse processing.

The sequence from 1 to 5 is concealed from application programmers. One of the advantages of an RPC is that it hides the intricacies of the network and these procedures behave much the same as ordinary procedures to application programmers.

4.1.3. CORBA

An object is an entity that encapsulates data and provides one or more operations (methods) acting on those data; for example, "the bank object" has data from his client's account and operations by

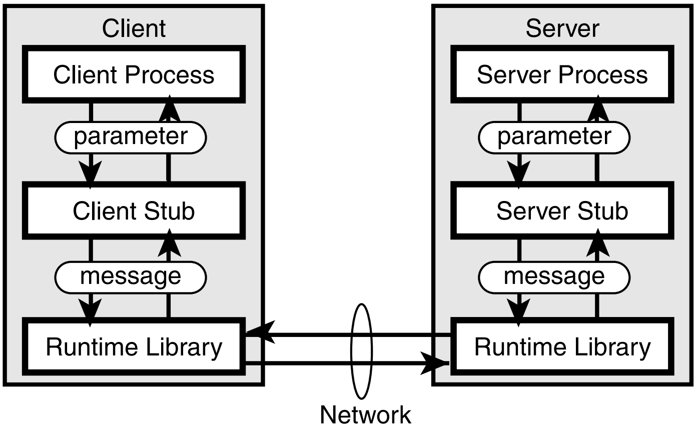


Figure 9 Interprocess Communication via an RPC.

which to manipulate them. Object-oriented computing is enabling faster software development by promoting software reusability, interoperability, and portability. In addition to enhancing the productivity of application developers, object frameworks are being used for data management, enterprise modeling, and system and network management. When objects are distributed, it is necessary to enable communications among distributed objects.

The Object Management Group, a consortium of object technology vendors founded in 1989, created a technology specification named CORBA (Common Object Request Broker Architecture). CORBA employs an abstraction similar to that of RPC, with a slight modification that simplifies programming and maintenance and increases extensibility of products.

The basic service provided by CORBA is delivery of requests from the client to the server and delivery of responses to the client. This service is realized by using a message broker for objects, called object request broker (ORB). An ORB is the central component of CORBA and handles distribution of messages between objects. Using an ORB, client objects can transparently make requests to (and receive responses from) server objects, which may be on the same computer or across a network.

An ORB consists of several logically distinct components, as shown in Figure 10.

The Interface Definition Language (IDL) is used to specify the interfaces of the ORB as well as services that objects make available to clients. The job of the IDL stub and skeleton is to hide the details of the underlying ORB from application programmers, making remote invocation look similar to local invocation. The dynamic invocation interface (DII) provides clients with an alternative to using IDL stubs when invoking an object. Because in general the stub routine is specific to a particular operation on a particular object, the client must know about the server object in detail. On the other hand, the DII allows the client to dynamically invoke a operation on a remote object. The object adapter provides an abstraction mechanism for removing the details of object implementation from the messaging substrate: generation and destruction of objects, activation and deactivation of objects, and invocation of objects through the IDL skeleton.

When a client invokes an operation on an object, the client must identify the target object. The ORB is responsible for locating the object, preparing it to receive the request, and passing the data needed for the request to the object. Object references are used for the ORB of the client side to identify the target object. Once the object has executed the operation identified by the request, if there is a reply needed, the ORB is responsible for returning the reply to the client.

The communication process between a client object and a server object via the ORB is shown in Figure 11.

1. A service request invoked by the client object is handled locally by the client stub. At this time, it looks to the client as if the stub were the actual target server object.
2. The stub and the ORB then cooperate to transmit the request to the remote server object.
3. At the server side, an instance of the skeleton instantiated and activated by the object adapter is waiting for the client's request. On receipt of the request from the ORB, the skeleton passes the request to the server object.
4. Then the server object executes the requested operations and creates a reply if necessary.
5. Finally, the reply is sent back to the client through the skeleton and the ORB that perform the reverse processing.

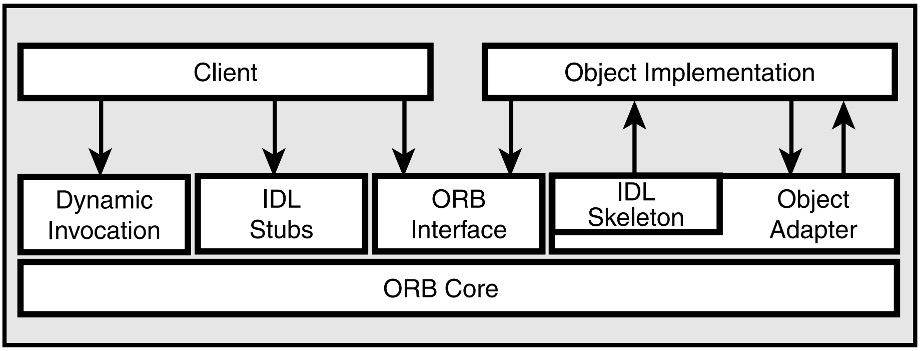


Figure 10 The CORBA Architecture.

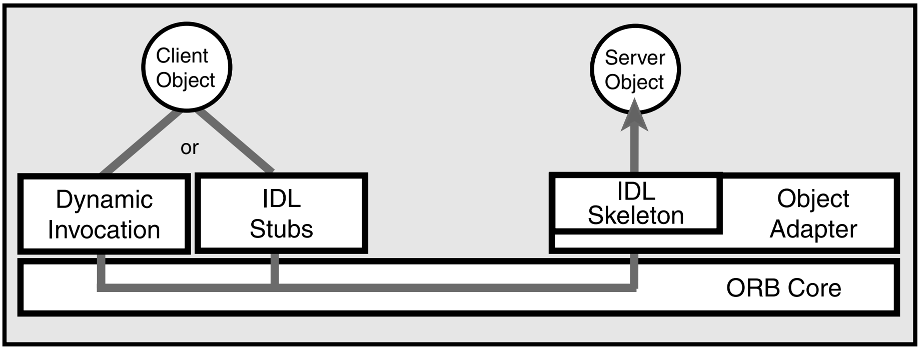


Figure 11 Communication between Objects in the CORBA Environment.

The main features of the CORBA are as follows:

1. *Platform independence:* Before the ORB transmits the message (request or result) from the client object or the server object into the network, the ORB translates the operation and its parameters into the common message format suitable for sending to the server. This is called marshaling. The inverse process of translating the data is called unmarshaling. This mechanism realizes communications between objects on different platforms. For example, a client object implemented on the Windows operating system can communicate with a server object implemented on the UNIX operating system.
2. *Language independence:* Figure 12 shows the development process of a CORBA C/S application. The first step of developing the CORBA object is to define the interface of the object (type of parameters and return values) by using the intermediate language IDL. Then the definition file described in IDL is translated into the file containing rough program codes in various development languages such as C, C++, Java, and COBOL.

IDL is not a programming language but a specification language. It provides language independence for programmers. The mapping of IDL to various development languages is defined by OMG. Therefore, a developer can choose the most appropriate one from various development languages to implement objects. Also, this language independence feature enables the system to interconnect with legacy systems developed by various languages in the past.

4.1.4. Other Communication Methods

In addition to Socket, RPC, and CORBA, there are several methods for realizing communication between processes.

- Java RMI (remote method invocation) is offered as a part of the JAVA language specification and will be a language-dependent method for the distributed object environment. Although there is the constraint that all the environments should be unified into Java, the developing process with it is a little easier than with CORBA or RPC because several functions for distributed computing are offered as utility objects.
- DCOM (distributed component object model) is the ORB that Microsoft promotes. A DCOM is implemented on the Windows operating system, and the development of applications can be practiced by using a Microsoft development environment such as Visual Basic.

4.2. Distributed Transaction Management

In a computer system, a transaction consists of an arbitrary sequence of operations. From a business point of view, a transaction is an action that involves change in the state of some recorded information related to the business. Transaction services are offered on both file systems and database systems.

The transaction must have the four properties referred to by the acronym ACID: atomicity, consistency, isolation, and durability:

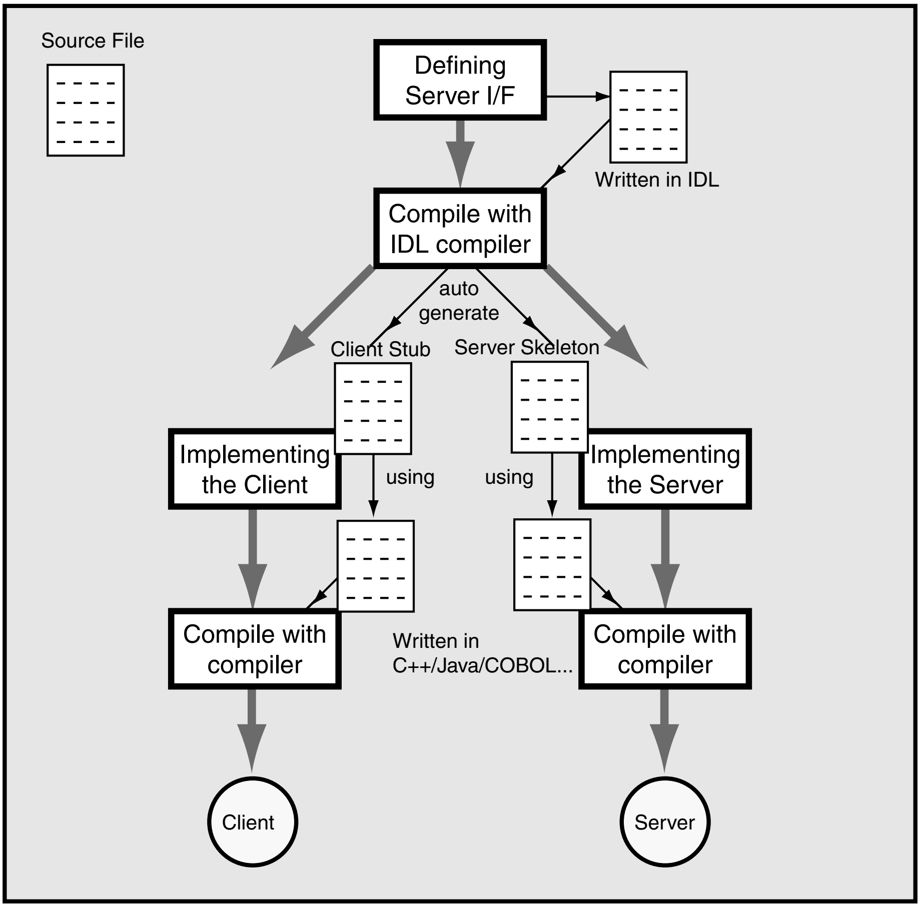

Source File
Defining Server I/F
Written in IDL
Compile with IDL compiler
auto generate
Client Stub
Server Skeleton
Implementing the Client
Implementing the Server
using
using
Compile with compiler
Written in C++/Java/COBOL...
Compile with compiler
Client
Server

Figure 12 Development Process of a CORBA C/S Application.

- *Atomicity* means that all the operations of the transaction succeed or they all fail. If the client or the server fails during a transaction, the transaction must appear to have either completed successfully or failed completely.
- *Consistency* means that after a transaction is executed, it must leave the system in a correct state or it must abort, leaving the state as it was before the execution began.
- *Isolation* means that operations of a transaction are not affected by other transactions that are executed concurrently, as if the separate transactions had been executed one at a time. The transaction must serialize all accesses to shared resources and guarantee that concurrent programs will not affect each other's operations.
- *Durability* means that the effect of a transaction's execution is permanent after it completes commitments. Its changes should survive system failures.

The transaction processing should ensure that either all the operations of the transaction complete successfully (commit) or none of them commit. For example, consider a case in which a customer transfers money from an account A to another account B in a banking system. If the account A and the account B are registered in two separate databases at different sites, both the withdrawal from account A and the deposit in account B must commit together. If the database crashes while the updates are processing, then the system must be able to recover. In the recovery procedure, both the updates must be stopped or aborted and the state of both the databases must be restored to the state before the transaction began. This procedure is called rollback.

In a distributed transaction system, a distributed transaction is composed of several operations involving distributed resources. The management of distributed transactions is provided by a transaction processing monitor (TP monitor), an application that coordinates resources that are provided by other resources.

A TP monitor provides the following management functions:

- *Resource management:* it starts transactions, regulates their accesses to shared resources, monitors their execution, and balances their workloads.
- *Transaction management:* It guarantees the ACID properties to all operations that run under its protection.
- *Client/server communications management:* It provides communications between clients and servers and between servers in various ways, including conversations, request-response, RPC, queueing, and batch.

Available TP monitor products include IBM's CICS and IMS/TP and BEA's Tuxedo.

4.3. Distributed Data Management

A database contains data that may be shared between many users or user applications. Sometimes there may be demands for concurrent sharing of the data. For example, consider two simultaneous accesses to an inventory data in a multiple-transaction system by two users. No problem arises if each demand is to read a record, but difficulties occur if both users attempt to modify (write) the record at the same time. Figure 13 shows the inconsistency that arises when two updates on the same data are processed at nearly the same time.

The database management must be responsible for this concurrent usage and offer concurrency control to keep the data consistent. Concurrency control is achieved by serializing multiple transactions through use of some mechanism such as locking or timestamp.

Also, in case the client or the server fails during a transaction due to a database crash or a network failure, the transaction must be able to recover. Either the transaction must be reversed or else some previous version of the data must be available. That is, the transaction must be rolled back. ‘‘All conditions are treated as transient and can be rolled back anytime’’ is the fundamental policy of control in the data management.

In some C/S systems, distributed databases are adopted as database systems. A distributed database is a collection of data that belong logically to the same system but are spread over several sites of a network. The main advantage of distributed databases is that they allow access to remote data transparently while keeping most of the data local to the applications that actually use it.

The primary concern of transaction processing is to maintain the consistency of the distributed database. To ensure the consistency of data at remote sites, a two-phase commit protocol is sometimes used in a distributed database environment. The first phase of the protocol is the preparation phase, in which the coordinator site sends a message to each participant site to prepare for commitment. The second phase is the implementation phase, in which the coordinator site sends either an abort or a commit message to all the participant sites, depending on the responses from all the participant sites. Finally, all the participant sites respond by carrying out the action and sending an acknowledgement message. Because the commitment may fail in a certain site even if commitments are completed in other sites, in the first phase of the two phase commit protocol, temporary ‘‘secure’’ commitment that can be rolled back anytime is done at each site. After it is confirmed that all the sites succeeded, the ‘‘official’’ commitment is performed.

The two-phase commit protocol has some limitations. One is the performance overhead that is introduced by all the message exchanges. If the remote sites are distributed over a wide area network, the response time could suffer further. The two-phase commit is also very sensitive to the availability of all sites at the time of update, and even a single point of failure could jeopardize the entire transaction. Therefore, decisions should be based on the business needs and the trade-off between the cost of maintaining the data on a single site and the cost of the two-phase commit when data are distributed at remote sites.

5. CAPACITY PLANNING AND PERFORMANCE MANAGEMENT

5.1. Objectives of Capacity Planning

Client/server systems are made up of many hardware and software resources, including client workstations, server systems, and network elements. User requests share the use of these common resources. The shared use of these resources gives rise to contention that degrades the system behavior and worsens users' perception of performance.

In the example of the Internet banking service described in the Section 7, the response time, the time from when a user clicks the URL of a bank until the home page of the bank is displayed on

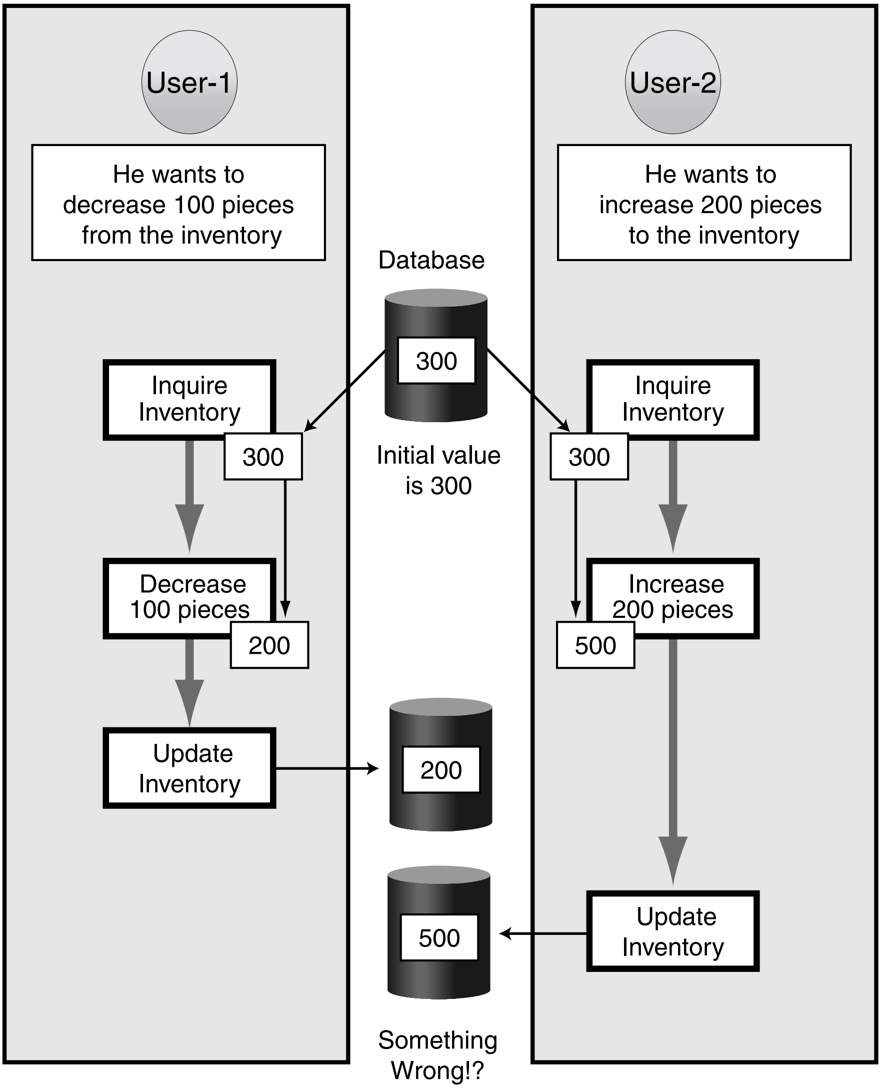

User-1
He wants to decrease 100 pieces from the inventory
User-2
He wants to increase 200 pieces to the inventory
Database
300
Initial value is 300
Inquire Inventory
300
Inquire Inventory
300
Decrease 100 pieces
200
Increase 200 pieces
500
Update Inventory
200
500
Update Inventory
Something Wrong!?

Figure 13 Inconsistency Arising from Simultaneous Updates.

the user's Web page, affects user's perception of performance and quality of service. In designing a system that treats requirements from multiple users, some service levels may be set; for example, average response time requirements for requests must not exceed 2 sec, or 95% of requests must exhibit a response time of less than 3 sec. For the given service requirements, service providers must design and maintain C/S systems that meet the desired service levels. Therefore, the performance of the C/S system should be evaluated as exactly as possible and kinds and sizes of hardware and software resources included in client systems, server systems, and networks should be determined. This is the goal of capacity planning.

The following are resources to be designed by capacity planning to ensure that the C/S system performance will meet the service levels:

- Types and numbers of processors and disks, the type of operating system, etc.
- Type of database server, access language, database management system, etc.
- Type of transaction-processing monitor

- Kind of LAN technology and bandwidth or transmission speed for clients and servers networking
- Kind of WAN and bandwidth when WAN is used between clients and servers

In the design of C/S systems, processing devices, storage devices, and various types of software are the building blocks for the whole system. Because capacity can be added to clients, servers, or network elements, addition of capacity can be local or remote. The ideal approach to capacity planning is to evaluate performance prior to installation. However, gathering the needed information prior to specifying the elements of the system can be a complicated matter.

5.2. Steps for Capacity Planning and Design

Capacity planning consists of the steps shown in Figure 14, which are essentially the same as those for capacity planning of general information systems.

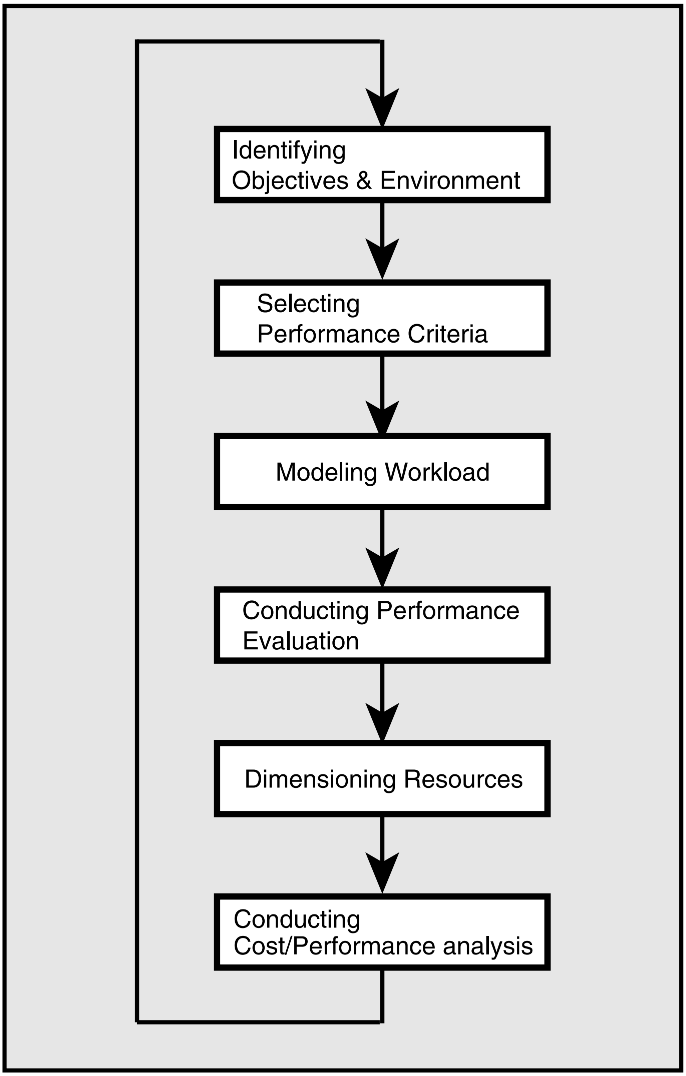

Identifying
Objectives & Environment
Selecting
Performance Criteria
Modeling Workload
Conducting Performance
Evaluation
Dimensioning Resources
Conducting
Cost/Performance analysis

Figure 14 The Steps in Capacity Planning and Design.

Because new technology, altered business climate, increased workload, change in users, or demand for new applications affects the system performance, these steps should be repeated at regular intervals and whenever problems arise.

5.3. Performance Objectives and System Environment

Specific objectives should be quantified and recorded as service requirements. Performance objectives are essential to enable all aspects of performance management. To manage performance, we must set quantifiable, measurable performance objectives, then design with those objectives in mind, project to see whether we can meet them, monitor to see whether we are meeting them, and adjust system parameters to optimize performance. To set performance objectives, we must make a list of system services and expected effects.

We must learn what kind of hardware (clients and servers), software (OS, middleware, applications), network elements, and network protocols are presented in the environment. Environment also involves the identification of peak usage periods, management structures, and service-level agreements. To gather these information about the environment, we use various information-gathering techniques, including user group meetings, audits, questionnaires, help desk records, planning documents, and interviews.

5.4. Performance Criteria

1. *Response time* is the time required to process a single unit of work. In interactive applications, the response time is the interval between when a request is made and when the computer responds to that request. Figure 15 shows a response time example for a client/server database access application. As the application executes, a sequence of interactions is exchanged among the components of the system, each of which contributes in some way to the delay that the user experiences between initiating the request for service and viewing the DBMS's response to the query. In this example, the response time consists of several time components, such as (a) interaction with the client application, (b) conversion of the request into a data stream by an API, (c) transfer of the data stream from the client to the server by communication stacks, (d) translation of the request and invocation of a stored procedure by DBMS, (e) execution of SQL (a relational data-manipulation language) calls by the stored procedure, (f) conversion of the results into a data stream by DBMS, (g) transfer of the data stream from the server to the client by the API, (h) passing of the data stream to the application, and (i) display of the first result.
2. *Throughput* is a measure of the amount of work a component or a system performs as a whole or of the rate at which a particular workload is being processed.
3. *Resource utilization* normally means the level of use of a particular system component. It is defined by the ratio of what is used to what is available. Although unused capacity is a waste of resources, a high utilization value may indicate that bottlenecks in processing will occur in the near future.

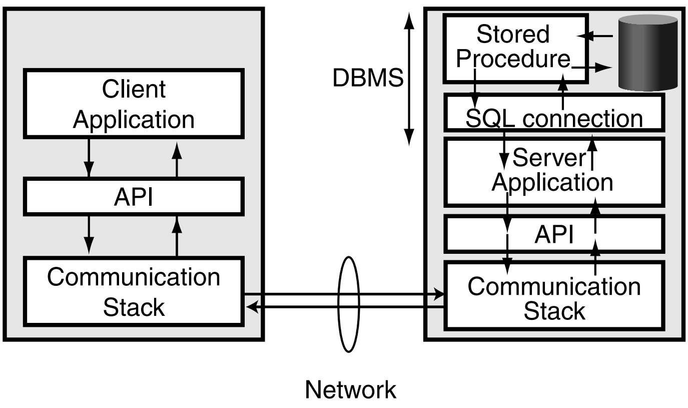


Figure 15 Response Time Example for a Database Access.

4. *Availability* is defined as the percentage of scheduled system time in which the computer is actually available to perform useful work. The stability of the system has an effect on performance. Unstable systems often have specific effects on the performance in addition to the other consequences of system failures.
5. *Cost–Performance ratio:* Once a system is designed and capacity is planned, a cost model can be developed. From the performance evaluation and the cost model, we can make an analysis regarding cost–performance trade-offs.

In addition to these criteria, resource queue length and resource waiting time are also used in designing some resources.

Response time and availability are both measures of the effectiveness of the system. An effective system is one that satisfies the expectations of users. Users are concerned with service effectiveness, which is measured by response time for transaction processing, elapsed time for batch processing, and query response time for query processing. On the other hand, a system manager is concerned with optimizing the efficiency of the system for all users. Both throughput and resource utilization are measures for ensuring the efficiency of system operations. If the quality of performance is measured in terms of throughput, it depends on the utilization levels of shared resources such as servers, network elements, and application software.

5.5. Workload Modeling

Workload modeling is the most crucial step in capacity planning. Misleading performance evaluation is possible if the workload is not properly modeled.

5.5.1. Workload Characterization

Workload refers to the resource demands and arrival intensity characteristics of the load brought to the system by the different types of transactions and requests. A workload consists of several components, such as C/S transactions, web access, and mail processing. Each workload component is further decomposed into basic components such as personnel transactions, sales transactions, and corporate training.

A real workload is one observed on a system being used for normal operations. It cannot be repeated and therefore is generally not suitable for use as a test workload in the design phase. Instead, a workload model whose characteristics are similar to those of real workload and can be applied repeatedly in a controlled manner, is developed and used for performance evaluations.

The measured quantities, service requests, or resource demands that are used to characterize the workload, are called workload parameters. Examples of workload parameters are transaction types, instruction types, packet sizes, source destinations of a packet, and page-reference patterns. The workload parameters can be divided into workload intensity and service demands. Workload intensity is the load placed on the system, indicated by the number of units of work contending for system resources. Examples include arrival rate or interarrival times of component (e.g., transaction or request), number of clients and think times, and number of processors or threads in execution simultaneously (e.g., file reference behavior, which describes the percentage of accesses made to each file in the disk system) The service demand is the total amount of service time required by each basic component at each resource. Examples include CPU time of transaction at the database server, total transmission time of replies from the database server in LAN, and total I/O time at the Web server for requests of images and video clips used in a Web-based learning system.

5.5.2. Workload Modeling Methodology

If there is a system or service similar to a newly planned system or service, workloads are modeled based on historical data of request statistics measured by data-collecting tools such as monitors.

A monitor is used to observe the performance of systems. Monitors collect performance statistics, analyze the data, and display results. Monitors are widely used for the following objectives:

1. To find the frequently used segments of the software and optimize their performance.
2. To measure the resource utilization and find the performance bottleneck.
3. To tune the system; the system parameters can be adjusted to improve the performance.
4. To characterize the workload; the results may be used for the capacity planning and for creating test workloads.
5. To find model parameters, validate models, and develop inputs for models.

Monitors are classified as software monitors, hardware monitors, firmware monitors, and hybrid monitors. Hybrid monitors combine software, hardware, or firmware.

If there is no system or service similar to a newly planned system or service, workload can be modeled by estimating the arrival process of requests and the distribution of service times or processing times for resources, which may be forecast from analysis of users' usage patterns and service requirements.

Common steps in a workload modeling process include:

1. Specification of a viewpoint from which the workload is analyzed (identification of the basic components of the workload of a system)
2. Selecting the set of workload parameters that captures the most relevant characteristics of the workload
3. Observing the system to obtain the raw performance data
4. Analyzing and reducing of the performance data
5. Constructing of a workload model.

The basic components that compose the workload must be identified. Transactions and requests are the most common.

5.6. Performance Evaluation

Performance models are used to evaluate the performance of a C/S system as a function of the system description and workload parameters. A performance model consists of system parameters, resource parameters, and workload parameters. Once workload models and system configurations have been obtained and a performance model has been built, the model must be examined to see how it can be used to answer the questions of interest about the system it is supposed to represent. This is the performance-evaluation process. Methods used in this process are explained below:

5.6.1. Analytical Models

Because generation of users' requests and service times vary stochastically, analytical modeling is normally done using queueing theory. Complex systems can be represented as networks of queues in which requests receive services from one or more groups of servers and each group of servers has its own queue. The various queues that represent a distributed C/S system are interconnected, giving rise to a network of queues, called a queueing network. Thus, the performance of C/S systems can be evaluated using queueing network models. Figure 16 shows an example of queueing network model for the Web server, where the client and the server are connected through a client side LAN, a WAN, and a server-side LAN.

5.6.2. Simulation

Simulation models are computer programs that mimic the behavior of a system as transactions flow through the various simulated resources. Simulation involves actually building a software model of

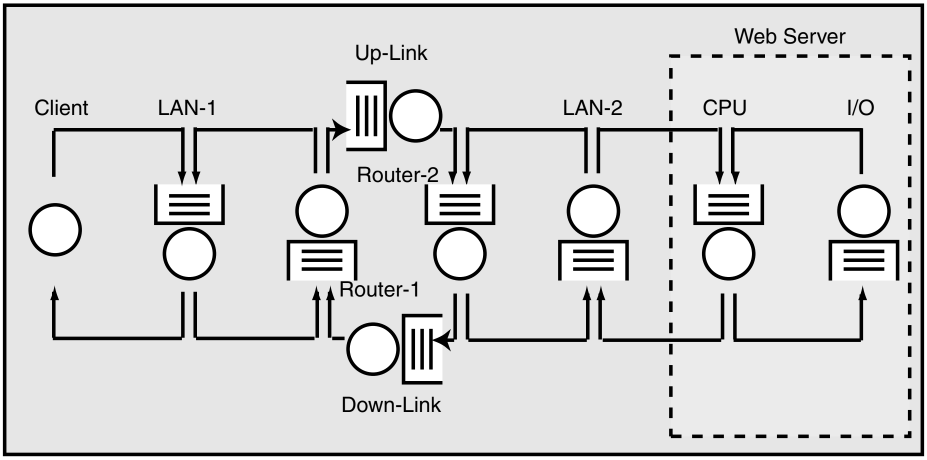


Figure 16 An Example of a Queueing Network Model for a Web Server.

each device, a model of the queue for that device, model processes that use the devices, and a model of the clock in the real world. Simulation can be accomplished by using either a general programming languages such as FORTRAN or C or a special-purpose language such as GPSS, SIMSCRIPT, or SLAM. For network analysis, there are special-purpose simulation packages such as COMNET III and BONeS are available.

5.6.3. Benchmarking

A benchmark is a controlled test to determine exactly how a specific application performs in a given system environment. While monitoring supplies a profile of performance over time for an application already deployed, either in a testing or a production environment, benchmarking produces a few controlled measurements designed to compare the performance of two or more implementation choices.

Some of the most popular benchmark programs and most widely published benchmark results come from groups of computer hardware and software vendors acting in consort. RISC workstation manufacturers sponsor the Systems Performance Evaluation Cooperative (SPEC), and DBMS vendors operate the transaction Processing Council (TPC). The TPC developed four system-level benchmarks that measure the entire system: TPC-A, B, C, and D. TPC-A and TPC-B are a standardization of the debit/credit benchmark. TPC-C is a standard for moderately complex online transaction-processing systems. TPC-D is used to evaluate price/performance of a given system executing decision support applications. The SPEC developed the standardized benchmark SPECweb, which measures a system's ability to act as a web server for static pages.

5.6.4. Comparing Analysis and Simulation

Analytic performance modeling, using queueing theory, is very flexible and complements traditional approaches to performance. It can be used early in the application development life cycle. The actual system, or a version of it, does not have to be built as it would be for a benchmark or prototype. This saves tremendously on the resources needed to build and to evaluate a design. On the other hand, a simulation shows the real world in slow motion. Simulation modeling tools allow us to observe the actual behavior of a complex system; if the system is unstable, we can see the transient phenomena of queues building up and going down repeatedly.

Many C/S systems are highly complex, so that valid mathematical models of them are themselves complex, precluding any possibility of an analytical solution. In this case, the model must be studied by means of simulation, numerically exercising the model for the inputs in question to see how they affect the output measures of performance.

6. MAINTENANCE AND ADMINISTRATION OF C/S SYSTEMS

Although system maintenance and administration is a complicated task even for a single centralized system, the complexity increases significantly due to the scalability, heterogeneity, security, distribution, naming, and so on in a C/S system. Efficient network and system-management tools are critical for reliable operation of distributed computing environments for C/S systems. In recent years, open and multivendor technologies have been adopted in construction of C/S systems. To administrate and maintain those systems totally, a standardized system management is needed for equipment from different vendors.

6.1. Architecture of System Management

6.1.1. OSI Management Framework

The OSI has defined five functional areas of management activities that are involved in distributed system management:

1. *Configuration management:* This involves collecting information on the system configuration and managing changes to the system configuration. Inventory control, installation, and version control of hardware and software are also included in this area.
2. *Fault management:* This involves identifying system faults as they occur, isolating the cause of the faults, and correcting them by contingency fallback, disaster recovery and so on.
3. *Security management:* This involves identifying locations of sensitive data and securing the system access points as appropriate to limit the potential for unauthorized intrusions. Encryption, password requirements, physical devices security, and security policy are also included in this area.
4. *Performance management:* This involves gathering data on the usage of system resources, analyzing these data, and acting on the performance prediction to maintain optimal system

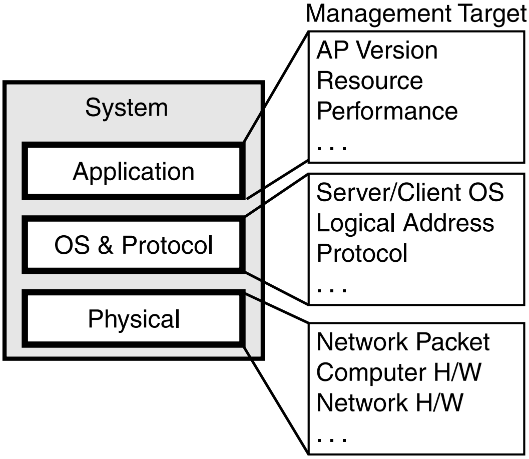


Figure 17 System Management Items.

performance. Real-time and historical statistical information about traffic volume, resource usage, and congestion are also included in this area.

5. *Accounting management:* This involves gathering data on resource utilization, setting usage shares, and generating charging and usage reports.

6.1.2. System Management Architecture

The items that are managed by system management can be classified into three layers: physical layer, operating system and network protocol layer, and application layer, as shown in Figure 17.

1. The physical layer includes client devices, server devices, and network elements, including LANs, WANs, computing platforms, and systems and applications software.
2. The operating system and network protocol layer includes the items for managing communication protocols to ensure interoperability between some units. In recent years, the adoption of TCP/IP standard protocol for realizing the Internet has been increasing. In a system based on the TCP/IP protocol stack, SNMP can be used for system management.
3. The application layer includes the items for managing system resources, such as CPU capacity and memory.

The system management architecture consists of the following components. These components may be either physical or logical, depending on the context in which they are used:

- A *network management station* (NMS) is a centralized workstation or computer that collects data from agents over a network, analyzes the data, and displays information in graphical form.
- A *managed object* is a logical representation of an element of hardware or software that the management system accesses for the purpose of monitor and control.
- An *agent* is a piece of software within or associated with a managed object that collects and stores information, responds to network management station requests, and generates incidental messages.
- A *manager* is software contained within a computer or workstation that controls the managed objects. It interacts with agents according to rules specified within the management protocol.
- A *management information base* (MIB) is a database containing information of use to network management, including information that reflects the configuration and behavior of nodes, and parameters that can be used to control its operation.

6.2. Network Management Protocol

An essential function in achieving the goal of network management is acquiring information about the network. A standardized set of network management protocols has been developed to help extract the necessary information from all network elements. There are two typical standardized protocols for network management: simple network management protocol (SNMP), developed under Internet sponsorship, and common management information protocol (CMIP), from ISO (International Organization for Standardization) and ITU-T (International Telecommunication Union-Telecommunication Standardization Sector).

SNMP is designed to work with the TCP/IP protocol stack and establishes standards for collecting and for performing security, performance, fault, accounting, and configuration functions associated with network management. The communication protocol for the SNMP is UDP, which is a very simple, unacknowledged, connectionless protocol. CMIP is designed to support a richer set of network-management functions and work with all systems conforming to OSI standards. Both SNMP and CMIP use an object-oriented technique to describe information to be managed, where the software describing actions is encapsulated with the rest of agent code within the managed object. CMIP requires considerably more overhead to implement than SNMP.

Because SNMP is the most widely implemented protocol for network management today, an SNMP management system is described below. An SNMP management system consists of the following components, as shown in Figure 18:

1. An SNMP agent is a software entity that resides on a managed system or a target node, maintains the node information, and reports on its status to managers.
2. An SNMP manager is a software entity that performs management tasks by issuing management requests to agents.
3. An MIB is a database containing the node information. It is maintained by an agent.

SNMP is an asynchronous request/response protocol that supports the following operations (Version 2):

- Get: a request issued by a manager to read the value of a managed object
- GetNext: a request made by a manager to traverse an MIB tree

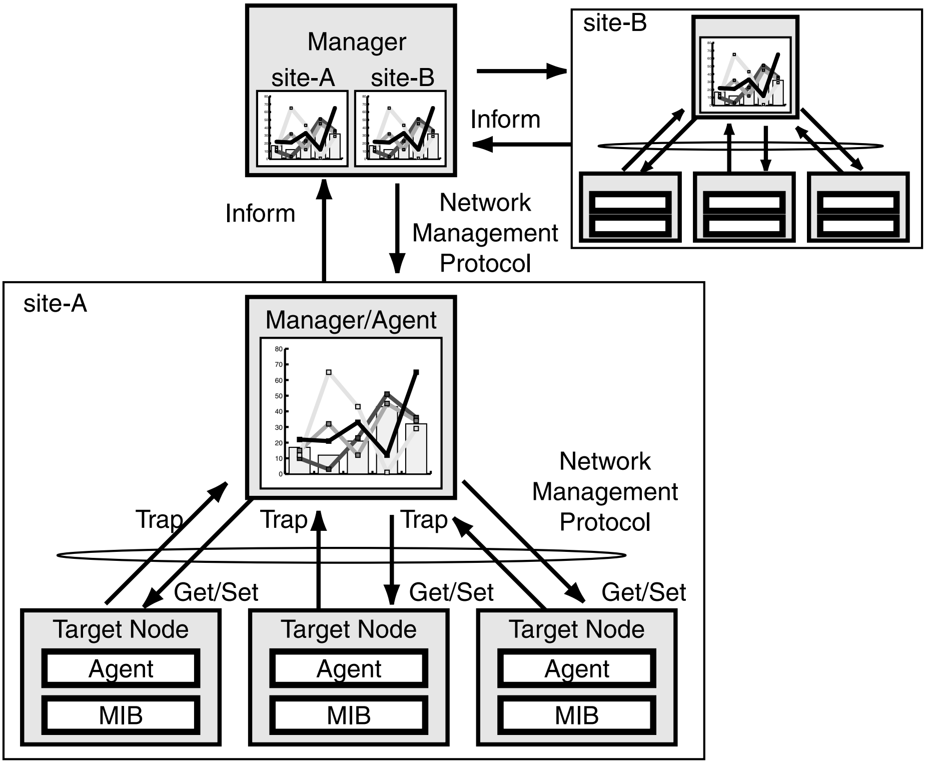

Manager
site-A
site-B
Inform
site-B
Inform
Network
Management
Protocol
site-A
Manager/Agent
Network
Management
Protocol
Trap
Trap
Trap
Get/Set
Get/Set
Get/Set
Target Node
Agent
MIB
Target Node
Agent
MIB
Target Node
Agent
MIB

Figure 18 SNMP Management System.

- GetBulk: a command issued by a manager, by which an agent can return as many successor variables in the MIB tree as will fit in a message
- Set: a request issued by a manager to modify the value of a managed object
- Trap: a notification issued from an agent in the managed system to a manager that some unusual event has occurred
- Inform: a command sent by a manager to other managers, by which managers can exchange management information

In this case, the managed system is a node such as a workstation, personal computer, or router. HP's OpenView and Sun Microsystem's SunNet Manager are well-known commercial SNMP managers.

System management functions are easily decomposed into many separate functions or objects that can be distributed over the network. It is a natural idea to connect those objects using CORBA ORB for interprocess communications. CORBA provides a modern and natural protocol for representing managed objects, defining their services, and invoking their methods via an ORB. Tivoli Management Environment (TME) is a CORBA-based system-management framework that is rapidly being adopted across the distributed UNIX market.

6.3. Security Management

C/S systems introduce new security threats beyond those in traditional host-centric systems. In a C /S system, it is more difficult to define the perimeter, the boundary between what you are protecting and the outside world. From the viewpoint of distributed systems, the problems are compounded by the need to protect information during communication and by the need for the individual components to work together. The network between clients and servers is vulnerable to eavesdropping crackers, who can sniff the network to obtain user IDs and passwords, read confidential data, or modify information. In addition, getting all of the individual components (including human beings) of the system to work as a single unit requires some degree of trust.

To manage security in a C/S system, it is necessary to understand what threats or attacks the system is subject to. A threat is any circumstance or event with the potential to cause harm to a system. A system's security policy identifies the threats that are deemed to be important and dictates the measures to be taken to protect the system.

6.3.1. Threats

Threats can be categorized into four different types:

1. *Disclosure or information leakage:* Information is disclosed or revealed to an unauthorized person or process. This involves direct attacks such as eavesdropping or wiretapping or more subtle attacks such as traffic analysis.
2. *Integrity violation:* The consistency of data is compromised through any unauthorized change to information stored on a computer system or in transit between computer systems.
3. *Denial of service:* Legitimate access to information or computer resources is intentionally blocked as a result of malicious action taken by another user.
4. *Illegal use:* A resource is used by an unauthorized person or process or in an unauthorized way.

6.3.2. Security Services

In the computer communications context, the main security measures are known as security services. There are some generic security services that would apply to a C/S system:

- *Authentication:* This involves determining that a request originates with a particular person or process and that it is an authentic, nonmodified request.
- *Access control:* This is the ability to limit and control the access to information and network resources by or for the target system.
- *Confidentiality:* This ensures that the information in a computer system and transmitted information are accessible for reading only by authorized persons or processes
- *Data integrity:* This ensures that only authorized persons or processes are able to modify data in a computer system and transmitted information.
- *Nonrepudiation:* This ensures that neither the sender nor the receiver of a message is able to deny that the data exchange occurred.

6.3.3. Security Technologies

There are some security technologies fundamental to the implementation of those security services.

6.3.3.1. Cryptography Cryptographic systems or cryptosystems can be classified into two distinct types: symmetric (or secret-key) and public-key (or asymmetric) cryptosystems. In a symmetric cryptosystem, a single key and the same algorithm are used for both encryption and decryption. The most widely used symmetric cryptosystem is the Data Encryption Standard (DES), which is the U.S. standard for commercial use. In a public-key cryptosystem, instead of one key in a symmetric cryptosystem, two keys are employed to control the encryption and the decryption respectively. One of these keys can be made public and the other is kept secret. The best-known the public-key cryptosystem is RSA, developed by Rivest, Shamir, and Adleman at MIT (1978).

The major problem in using cryptography is that it is necessary to disseminate the encryption/decryption keys to all parties that need them and ensure that the key distribution mechanism is not easily compromised. In a public-key cryptosystem, the public key does not need to be protected, alleviating the problem of key distribution. However, a public key also needs to be distributed with authentication for protecting it from frauds. Public-key cryptosystems have some advantages in key distribution, but implementation results in very slow processing rates. For example, encryption by RSA is about 1000 times slower than by DES in hardware and about 100 times slower than DES in software. For these reasons, public-key cryptosystems are usually limited to use in key distribution and the digital signature, and symmetric cryptosystems are used to protect the actual data or plaintexts.

Data integrity and data origin authentication for a message can be provided by hash or message digest functions. Cryptographic hash functions involve, instead of using keys, mapping a potentially large message into a small fixed-length number. Hash functions are used in sealing or digital signature processes, so they must be truly one-way, that is, it must be computationally infeasible to construct an input message hashed to a given digest or to construct two messages hashed to the same digest. The most widely used hash function is message digest version 5 (MD5).

6.3.3.2. Authentication Protocol In the context of a C/S system, authentication is the most important of the security services because other security services depend on it in some way. When a client wishes to establish a secure channel between the client and a server, the client and the server will wish to identify each other by authentication. There are three common protocols for implementing authentication: three-way handshake authentication, trusted-third-party authentication, and public-key authentication. One of the trusted third-party protocols is Kerberos, a TCP/IP-based network authentication protocol developed as a part of the project Athena at MIT. Kerberos permits a client and a server to authenticate each other without any message going over the network in the clear. It also arranges for the secure exchange of session encryption keys between the client and the server. The trusted third-party is sometimes called an authentication server.

A simplified version of the third-party authentication in Kerberos is shown in Figure 19. Kerberos protocol assumes that the client and the server each share a secret key, respectively K_c and K_s, with the authentication server. In Figure 19, $[M]K$ denotes the encryption of message M with key K.

1. The client first sends a message to the authentication server that identifies both itself and the server.
2. The authentication server then generates a timestamp T, a lifetime L, and a new session key K and replies to the client with a two-part message. The first part, $[T, L, K, ID_s]K_c$, encrypts the three values T, L, and K, along with the server's identifier ID_s, using the key K_c. The second part, $[T, L, K, ID_c]K_s$, encrypts the three values T, L, and K, along with the client's identifier ID_c using the key K_s.
3. The client receives this message and decrypts only the first part. The client then transfers the second part to the server along with the encryption $[ID_c, T]K$ of ID_c and T using the session key K, which is decrypted from the first part.
4. On receipt of this message, the server decrypts the first part, $[T, L, K, ID_c]K_s$, originally encrypted by the authentication server using K_s, and in so doing recovers T, K, and ID_c. Then the server confirms that ID_c and T are consistent in the two halves of the message. If they are consistent, the server replies with a message $[T + 1]K$ that encrypts $T + 1$ using the session key K.
5. Now the client and the server can communicate with each other using the shared session key K.

6.3.3.3. Message Integrity Protocols There are two typical ways to ensure the integrity of a message. One uses a public-key cryptosystem such as RSA to produce a digital signature, and the other uses both a message digest such as MD5 and a public-key cryptosystem to produce a digital

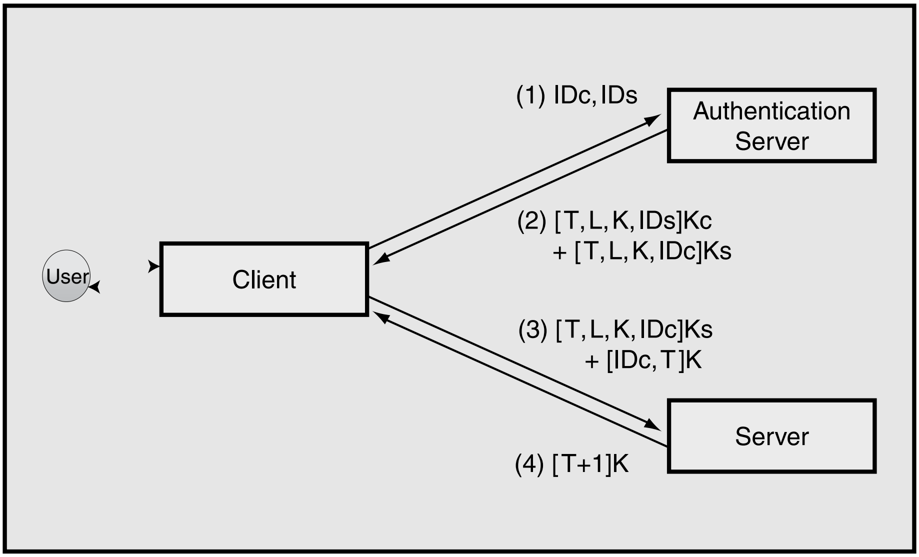

(1) IDc, IDs
Authentication Server
(2) [T, L, K, IDs]Kc + [T, L, K, IDc]Ks
User
Client
(3) [T, L, K, IDc]Ks + [IDc, T]K
Server
(4) [T+1]K

Figure 19 Third-Party Authentication in Kerberos.

signature. In the latter type, a hash function is used to generate a message digest from the message content requiring protection. The sender encrypts the message digest using the public-key cryptosystem in the authentication mode; the encryption key is the private key of the sender. The encrypted message digest is sent an appendix along with the plaintext message. The receiver decrypts the appendix using the corresponding decryption key (the public key of the sender) and compares it with the message digest that is computed from the received message by the same hash function. If the two are the same, then the receiver is assured that the sender knew the encryption key and that the message contents were not changed en route.

6.3.3.4. Access Control Access control contributes to achieving the security goals of confidentiality, integrity, and legitimate use. The general model for access control assumes a set of active entities, called subjects, that attempt to access members of a set of resources, called objects. The access-control model is based on the access control matrix, in which rows correspond to subjects (users) and columns correspond to objects (targets). Each matrix entry states the access actions (e.g., read, write, and execute) that the subject may perform on the object. The access control matrix is implemented by either:

- *Capability list:* a row-wise implementation, effectively a ticket that authorizes the holder (subject) to access specified objects with specified actions
- *Access control list* (ACL): a column-wise implementation, also an attribute of an object stating which subjects can invoke which actions on it

6.3.3.5. Web Security Protocols: SSL and S-HTTP As the Web became popular and commercial enterprises began to use the Internet, it became obvious that some security services such as integrity and authentication are necessary for transactions on the Web. There are two widely used protocols to solve this problem: secure socket layer (SSL) and secure HTTP (S-HTTP). SSL is a general-purpose protocol that sits between the application layer and the transport layer. The security services offered by the SSL are authentication of the server and the client and message confidentiality and integrity. The biggest advantage of the SSL is that it operates independently of application-layer protocols. HTTP can also operate on top of SSL, and it is then often denoted HTTPS. Transport Layer Security (TLS) is an Internet standard version of SSL and is now in the midst of the IETF standardization process. Secure HTTP is an application-layer protocol entirely compatible with HTTP and contains security extensions that provide client authentication, message confidentiality and integrity, and nonrepudiation of origin.

6.3.3.6. Firewall Because the Internet is so open, security is a critical factor in the establishment and acceptance of commercial applications on the Web. For example, customers using an Internet banking service want to be assured that their communications with the bank are confidential and not tampered with, and both they and the bank must be able to verify each other's identity and to keep authentic records of their transactions. Especially, corporate networks connected to the Internet are liable to receive attacks from crackers of the external network. The prime technique used commercially to protect the corporate network from external attacks is the use of firewalls.

A firewall is a collection of filters and gateways that shields the internal trusted network within a locally managed security perimeter from external, untrustworthy networks (i.e., the Internet). A firewall is placed at the edge of an internal network and permits a restricted set of packets or types of communications through. Typically, there are two types of firewalls: packet filters and proxy gateways (also called application proxies).

- A packet filter functions by examining the header of each packet as it arrives for forwarding to another network. It then applies a series of rules against the header information to determine whether the packet should be blocked or forwarded in its intended direction.
- A proxy gateway is a process that is placed between a client process and a server process. All incoming packet from the client is funneled to the appropriate proxy gateway for mail, FTP, HTTP, and so on. The proxy then passes the incoming packets to the internal network if the access right of the client is verified.

7. A PRACTICAL EXAMPLE: INTERNET BANKING SYSTEM

Here we will explain an Internet banking system as a practical example of a C/S system. An Internet banking service is an online financial service that is offered on the Internet. The system offers various financial services, such as inquiry for bank balance, inquiry for payment details, and fund transfer to customers via the Internet.

We will explain the flow of transactions in C/S processing through an example of fund transfer service between a customer's account in bank A and another account in bank B. Figure 20 shows the flow of processing for the example.

Step 1: A customer connects to the Internet from a desktop personal computer in his home or office and accesses the site of bank A, which offers the Internet banking service, by using a Web browser.

Step 2: At the top menu of the Web page, the user is asked to enter his or her user ID and password already registered. According to the indication on the Web page, the user inputs user ID and password into the input field on the Web page.

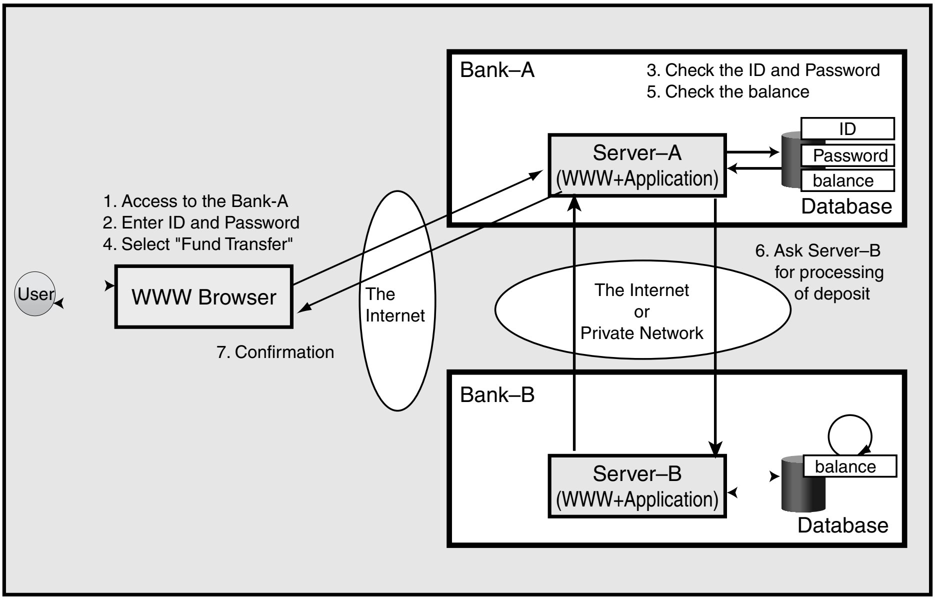

Bank–A
3. Check the ID and Password
5. Check the balance
ID
Password
balance
Server–A
(WWW+Application)
Database
1. Access to the Bank-A
2. Enter ID and Password
4. Select "Fund Transfer"
User
WWW Browser
The Internet
7. Confirmation
6. Ask Server–B for processing of deposit
The Internet or Private Network
Bank–B
Server–B
(WWW+Application)
balance
Database

Figure 20 Internet Banking System.

Step 3: The entered user ID and password are sent back to the server via the Internet by the secured communication protocol HTTPS. The server receives that information and checks information on the database that accumulates information about the client and other various data. If the user ID and password correspond to those already registered, the customer is permitted to use the service.

Step 4: The customer selects ‘‘fund transfer’’ from the menu displayed on the Web page.

Step 5: The server receives the above request, searches the database, and retrieves the customer’s account information.

Step 6: After confirming that there is enough money for payment in the customer’s account, the server asks another server in bank B for processing of deposit. In this case, the server of bank A becomes ‘‘client’’ and the server of bank B becomes ‘‘server.’’

Step 7: After the server of bank A confirms that the processing was completed normally in the server of bank B, it updates (withdraws) the account information of the customer in the database. The acknowledgement that the request for fund transfer is successfully completed is also displayed on the customer’s Web page.

ADDITIONAL READING

Crowcroft, J., *Open Distributed Systems,* UCL Press, London, 1996.

Davis, T., Ed., *Securing Client/Server Computer Networks*, McGraw-Hill, New York, 1996.

Edwards, J., *3-Tier Client/Server at Work,* Rev. Ed., John Wiley & Sons, New York, 1999.

Menascé, D. A., Almeida, V. A. F., and Dowdy, L. W., *Capacity Planning and Performance Modeling: From Mainframes to Client–Server Systems,* Prentice Hall, Englewood Cliffs, NJ, 1994.

Orfali, R., Harkey, D., and Edwards, J., *Client/Server Survival Guide,* 3d Ed., John Wiley & Sons, New York, 1999.

Renaud, P. E., *Introduction to Client/Server Systems,* John Wiley & Sons, New York, 1993.

Vaughn, L. T., *Client/Server System Design and Implementation,* McGraw-Hill, New York, 1994.

CHAPTER 27

Industrial Engineering Applications in Health Care Systems

Swatantra K. Kachhal
University of Michigan-Dearborn

1. INTRODUCTION TO HEALTH CARE DELIVERY SYSTEMS 737
  1.1. History of Health Care Delivery Systems 737
  1.2. Government Regulations 738
  1.3. Health Care Financing 738
  1.4. Emerging Trends 738
2. INTRODUCTION TO INDUSTRIAL ENGINEERING IN HEALTH CARE DELIVERY SYSTEMS 739
3. APPLICATIONS OF METHODS IMPROVEMENT AND WORK SIMPLIFICATION 740
  3.1. Introduction 740
  3.2. Application of a Flowchart to Document Patient Flow in an Outpatient Clinic 740
4. APPLICATION OF STAFFING METHODOLOGIES 740
  4.1. Introduction 740
  4.2. A Case Study in the Ambulatory Surgery Center 742
5. APPLICATION OF SCHEDULING METHODOLOGIES 742
  5.1. Introduction 742
  5.2. Work Scheduling 742
    5.2.1. Appointment Scheduling in a Primary Care Clinic 743
  5.3. Personnel Scheduling 744
6. APPLICATION OF QUEUING AND SIMULATION METHODOLOGIES 744
7. APPLICATION OF STATISTICAL METHODS 745
  7.1. Use of Regression Models 745
  7.2. Determination of Sample Size 745
  7.3. Use of Control Charts 746
8. APPLICATION OF OPTIMIZATION MODELS 746
9. APPLICATION OF QUALITY-IMPROVEMENT TOOLS 747
10. APPLICATION OF INFORMATION SYSTEMS/DECISION SUPPORT TOOLS 747
11. APPLICATION OF OTHER INDUSTRIAL ENGINEERING TECHNIQUES 748
12. FUTURE TRENDS 748
REFERENCES 748

1. INTRODUCTION TO HEALTH CARE DELIVERY SYSTEMS

This section gives a brief history of health care delivery in the United States for readers who may be unfamiliar with the health care industry. It also addresses government regulations and health care financing issues that impact the delivery of health care.

1.1. History of Health Care Delivery Systems

Historically, most health care has been delivered in hospitals. Initially, hospitals were founded, many of them by charitable organizations, to house poor people during epidemics. As medical technology

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

advanced, hospitals actually became centers for treating patients. They also became resources for community physicians, enabling them to share high-cost equipment and facilities that they could not afford by themselves.

Some hospitals are government-owned, including federal, state, county, and city hospitals. Federal hospitals include Army, Navy, Air Force, and Veterans Administration hospitals. Nongovernment hospitals are either not-for-profit or for-profit. The not-for-profit hospitals are owned by church-related groups or communities and constitute the bulk of the hospitals in the country. Investor-owned for-profit hospitals are mostly owned by chains such as Hospital Corporation of America (HCA) and Humana.

1.2. Government Regulations

A number of federal, state, county, and city agencies regulate hospitals. These regulations cover various elements of a hospital, including physical facilities, medical staff, personnel, medical records, and safety of patients. The Hill-Burton law of 1948 provided matching funds to towns without community hospitals. This resulted in the development of a large number of small hospitals in the 1950s and 1960s. The enactment of Medicare legislation for the elderly and the Medicaid program for the poor in 1965 resulted in rapid growth in hospitals. Because hospitals were paid on the basis of costs, much of the cost of adding beds and new expensive equipment could be recovered. Many states developed ‘‘certificate-of-need’’ programs that required hospitals planning to expand or build a new facility to obtain approval from a state agency (Mahon 1978). To control expansion and associated costs, the National Health Planning and Resources Development Act was enacted in 1974 to require hospitals that participated in Medicare or Medicaid to obtain planning approval for capital expenditures over $100,000. The result was a huge bureaucracy at local, state, and national levels to implement the law. Most of the states have revised or drastically reduced the scope of the law. Many states have changed the review limit to $1 million. In other states, only the construction of new hospitals or new beds is reviewed (Steinwelds and Sloan 1981).

1.3. Health Care Financing

Based on the method of payment for services, patients can be classified into three categories. The first category includes patients that pay from their own pocket for services. The second category includes patients covered by private insurance companies. The payment for services for patients in the last category is made by one of the government programs such as Medicare. A typical hospital receives approximately 35% of its revenue from Medicare patients, with some inner-city hospitals receiving as much as 50%. A federal agency called the Health Care Financing Administration (HCFA) manages this program. Medicare formerly paid hospitals on the basis of cost for the services rendered to Medicare patients. Under this system of payment, hospitals had no incentive to control costs and deliver the care efficiently. To control the rising costs, a variety of methods were used, including cost-increase ceilings, but without much success. In 1983, a new means of payment was introduced, known as the diagnostic related group (DRG) payment system. Under this system, hospital cases are divided into 487 different classes based on diagnosis, age, complications, and the like. When a patient receives care, one of these DRGs is assigned to the patient and the hospital receives a predetermined amount for that DRG irrespective of the actual cost incurred in delivering the services. This system of payment encouraged the hospitals to deliver care efficiently at lower cost. Many of the commercial insurance companies such as Blue Cross and Blue Shield have also adopted this method of payment for the hospitals.

1.4. Emerging Trends

During the past decade, a few key trends have emerged. A shift to managed health care from traditional indemnity insurance is expected to continue. This has resulted in the growth of health maintenance organizations (HMOs) and preferred provider organizations (PPOs), which now account for as much as 60–70% of the private insurance in some markets. Employers have also been changing health care benefits for their employees, increasingly going to so-called cafeteria plans, where employers allocate a fixed amount of benefit dollars to employees and allow them to allocate these dollars among premiums for various services. This has resulted in increased out of pocket costs and copayments for health care for the employees.

Medicare has reduced payments for DRGs and medical education to the hospitals. It has put many health care systems in serious financial trouble. The increase in payments is expected to stay well below the cost increases in the health care industry. These trends are expected to continue in the near future (Sahney et al. 1986.)

Since the advent of the DRG system of payment, the length of stay has continually dropped in hospitals from an average of over 10 days in the 1970s to below 5 days. Hospital admission rates have also dropped because many of the procedures done on an inpatient basis are now being performed on an ambulatory care basis. Inpatient admissions and length of stay are expected to continue to decline, and ambulatory care is expected to continue to grow over the next decade.

Medically indigent people who have neither medical insurance nor coverage by federal and/or state programs continue to place a serious financial burden on health care institutions. In the past, hospitals were able to transfer the cost of indigent care to other payers under a cost-based payment system. But under the DRG-based payment systems, hospitals are unable to do so.

Hospitals have been experiencing shortages in RN staffing, especially in inpatient units and emergency rooms. The nursing shortages can be attributed to low starting salaries for nurses and potential nursing students opting for other carriers with regular daytime work hours. Nurses also have other opportunities within health care aside from the inpatient setting that do not require night shift rotation or weekend coverage. The nursing shortage is projected to continue for positions in inpatient units.

Another trend in the health care industry has been an increase in ambulatory care clinics and ambulatory surgical centers. Hospitals have developed freestanding ambulatory care centers as a means of penetrating new markets by providing easy access to primary care services to the growing number of HMO patients.

The other major trend is consolidation within the industry. Hospitals have been facing a bleak financial future. Hospitals have been consolidating to form larger and leaner health care delivery systems to take advantage of economies of scale. Usually there is a reduction in the number of excess inpatient beds and an elimination of duplicate services in a community as a result of these mergers. Industry consolidation is expected to continue in the coming years. It is expected that in most large cities only a few large health systems will account for most of the inpatient health care services.

2. INTRODUCTION TO INDUSTRIAL ENGINEERING IN HEALTH CARE DELIVERY SYSTEMS

Frank Gilbreth is considered to be the first to use industrial engineering tools for methods improvement in a hospital situation. He applied his motion study techniques to surgical procedures in 1913 (Nock 1913); (Gilbreth 1916). In the 1940s, Lillian Gilbreth published articles explaining the benefits of using methods-improvement techniques in hospitals and in nursing (Gilbreth 1945, 1950). In 1952, a two-week workshop was conducted at the University of Connecticut on the application of methods improvement in hospitals (Smalley 1982). Also in 1952, the American Hospital Association (AHA) created a Committee on Methods Improvement. This committee prepared several papers on methods-improvement activities in hospitals and published them in its interim report in 1954.

In the late 1950s, the American Hospital Association started offering workshops on methods improvement around the country. Universities added courses on the application of methods improvement in hospital administration and industrial engineering programs. Gradually, other industrial engineering techniques were studied and applied to various hospital problems. The growing use of industrial engineering techniques in hospitals resulted in the foundation of the Hospital Management Systems Society (HMSS) in Atlanta in 1961, which subsequently moved to Chicago in 1964. The Institute of Industrial Engineers also recognized the expanding role of industrial engineering techniques and formed a hospital section in 1964. This section changed its name to the Health Services Division in 1977, reflecting the broader scope of the field. In 1988, the Institute of Industrial Engineers approved the formation of the Society for Health Systems (SHS) to replace the Health Services Division. In 1987, HMSS changed its name to Healthcare Information and Management Systems Society (HIMSS) to recognize the growing role of information systems in health care. Both HIMSS and SHS offer a number of seminars and workshops throughout the year related to the application of information systems and industrial engineering in health care. They also cosponsor annual HIMSS conference in February and the annual SHS conference in September/October of each year. Many of the industrial engineers working in the health care industry are members of both organizations.

Educational institutions started offering industrial engineering programs with specialization in health care systems. Many industrial engineers today work for hospitals and health systems, while others work as consultants in health care. The industrial engineers working in health care systems are usually referred as management engineers or operations analysts. Smalley (1982) gives a detailed history of the development of the use of industrial engineering in hospitals.

Industrial engineers gradually realized that many industrial engineering techniques initially applied to manufacturing/production systems were equally applicable in service systems such as health care systems. Almost all of the industrial engineering tools and techniques have been applied to health care systems. In this chapter, the application of only some of these techniques to health care systems will be discussed. From the examples presented here, readers should be able to appreciate the application of other techniques to health care systems. The application of the following techniques in health care systems is discussed here:

1. Methods improvement and work simplification
2. Staffing analysis
3. Scheduling
4. Queuing and simulation

5. Statistical analysis
6. Optimization
7. Quality improvement
8. Information systems/decision support systems

The emphasis will be on the application in health care, not on the techniques themselves, because the details of industrial engineering techniques are given in the other chapters of this Handbook. Applications will be discussed using various hospital departments as examples. For individuals not familiar with health care systems, Goldberg and Denoble (1986) describe various hospital departments.

3. APPLICATIONS OF METHODS IMPROVEMENT AND WORK SIMPLIFICATION

3.1. Introduction

The terms *methods improvement, methods engineering, operations analysis,* and *work simplification* have been used synonymously in industrial engineering literature. These techniques use a systematic procedure to study and improve methods for carrying out any set of activities to accomplish a task. In health care systems, these could be the work methods used in the actual delivery of health care to the patients or the work methods utilized in support activities. One common tool used to document an existing process or work method is a flowchart, which can also be used to document a new process as it is developed prior to implementation. A flowchart allows a critical examination of the various steps of the process and assists in the identification of unnecessary steps and inefficiencies in the process. Flowcharts show each step of the process as well as decision points and various courses of action based upon the decision.

Other tools available for methods improvement that have been used in health care systems include a flow diagram and a paperwork simplification chart. A flow diagram can be used to visualize the flow of material and patients in a facility. It can help identify areas of congestion and assist in the planning of the physical layout of facilities. Paperwork-simplification charts are used to analyze and improve paper documents to provide the needed control and communication.

3.2. Application of a Flowchart to Document Patient Flow in an Outpatient Clinic

Figure 1 shows the flowchart for patient flow in an outpatient clinic starting from the time when a patient arrives at the reception counter to the time the patient leaves the clinic after the visit. This chart shows all the steps a patient goes through to complete a visit with the physician. Management engineers can critically examine each step to study the possibility of eliminating, simplifying, or modifying it. For example, this flowchart indicates that patients have to stand in line till a receptionist is available to serve them. If this wait is excessive, one alternative could be to have a sign-up sheet so the patients could sign it and then take a seat. When available, the receptionists could then call the patients to the counter to serve them.

4. APPLICATION OF STAFFING METHODOLOGIES

4.1. Introduction

Staffing refers to the number of employees of a given skill level needed in a department to meet a given demand for services. The determination of the staffing is basically a problem of work measurement. It requires listing all the different tasks done by a certain skill level in the department. The departmental tasks are classified as constant or variable. Constant tasks are those tasks that are not directly related to the departmental output, that is, they are independent of the level of demand for services. For example, cleaning the work area and equipment every morning in a blood-testing laboratory is a constant task that does not depend upon the number of specimens tested. Variable tasks are those that are directly related to the output of the department. For example, a variable task may be the actual testing of a specimen. The total time spent during a day in testing specimens would be equal to the time taken to test one specimen multiplied by the total number of specimens tested in a day.

After the identification of constant and variable tasks, the work content associated with each task is measured. The work content could be measured using any of the traditional work-measurement techniques such as stopwatch time study, predetermined motion time systems, and work sampling. Work content could also be established using the Delphi approach in which a group of individuals familiar with a well-defined task try to reach a consensus about the time required to accomplish the task. The frequency at which each variable task occurs per day is also determined. The total number of hours of work to be performed for each skill level is then determined by adding the time taken

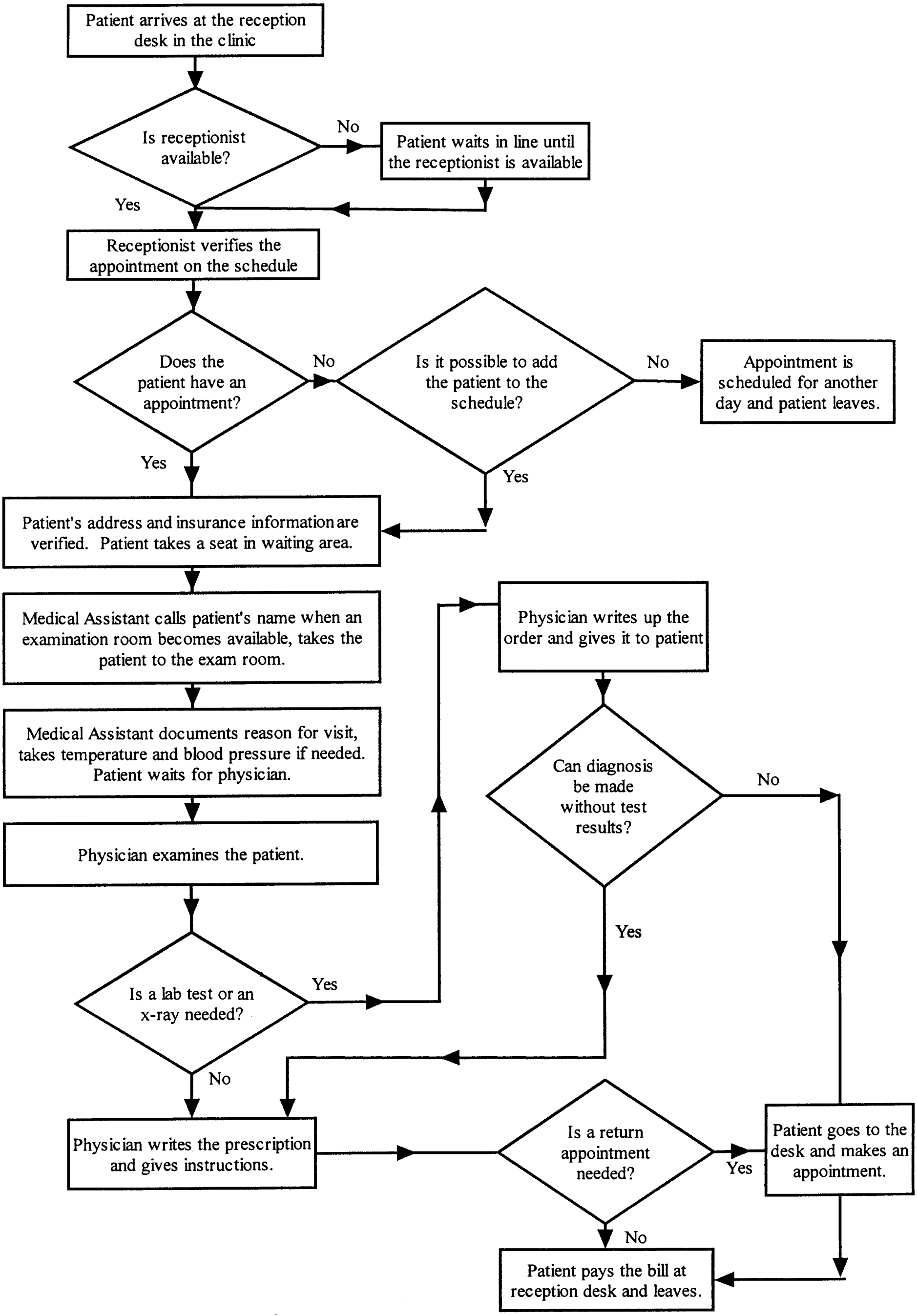

Patient arrives at the reception desk in the clinic
Is receptionist available?
No
Patient waits in line until the receptionist is available
Yes
Receptionist verifies the appointment on the schedule
Does the patient have an appointment?
No
Is it possible to add the patient to the schedule?
No
Appointment is scheduled for another day and patient leaves.
Yes
Yes
Patient's address and insurance information are verified. Patient takes a seat in waiting area.
Medical Assistant calls patient's name when an examination room becomes available, takes the patient to the exam room.
Medical Assistant documents reason for visit, takes temperature and blood pressure if needed. Patient waits for physician.
Physician examines the patient.
Physician writes up the order and gives it to patient
Can diagnosis be made without test results?
No
Yes
Is a lab test or an x-ray needed?
Yes
No
Physician writes the prescription and gives instructions.
Is a return appointment needed?
Yes
Patient goes to the desk and makes an appointment.
No
Patient pays the bill at reception desk and leaves.

Figure 1 Patient Flowchart in an Outpatient Clinic.

to perform the constant tasks and the variable tasks at a given frequency level associated with a given demand for services.

One disadvantage of using the common work-measurement techniques is that the process is very cumbersome and time-consuming. Moreover, any time the method changes, the work content of various tasks has to be measured again. Another alternative may be to use one of the standard-staffing methodologies available in the health care industry. These methodologies describe the typical tasks that are performed in a particular hospital department and assign a standard time to each task. Resource Monitoring System (RMS), developed by the Hospital Association of New York State, is an example of one such system.

The time available per full time equivalent (FTE) is also calculated. If the department is open 8 am to 5 pm with a one-hour lunch break and if two 15-minute breaks are allowed during the shift, the time available per FTE per day will be 7.5 hours. The actual required FTE depends on demand variability, coverage issues, scheduling constraints, skill level requirements, and similar factors. One approach is to determine the efficiency level (a number less than one) that could be expected because of these factors. Efficiency is defined as the ratio of the actual hours of work done to the actual number of hours used to do the work, including idle time, if any. The FTE needed is then determined using the following equation:

$$\text{FTE needed} = \frac{\text{total number of hours of work to be performed per day}}{\text{time available per FTE per day} \times \text{efficiency level}}$$

4.2. A Case Study in the Ambulatory Surgery Center

An ambulatory surgery center can be defined as a specialized facility organized to perform surgical cases in an ambulatory setting. Ambulatory surgery centers are normally staffed by registered nurses (RNs), who provide care in preoperative, operating, and recovery rooms. A staff of nurse anesthetists (CRNAs) works in the operating rooms and also performs patient assessments in the preoperative area before the patient enters surgery. Nurses' aides clean and package surgical instruments and assist with room turnover between cases. Room turnover involves cleaning and restocking the operating room between cases. The patient registration and all the initial laboratory work are done prior to the day of the surgery. On the day of surgery, the patient is asked to arrive one hour prior to the scheduled surgery time. After checking in with the receptionist, the patient is sent to preoperative area for assessment by CRNAs. After the surgery, the patient is in the recovery room until the patient is stable enough to be sent home.

The objective of the study was to determine the staffing requirement for each skill level including RNs, CRNAs, and nursing aides. As stated above, all the constant and variable tasks were identified for each skill level. The time required to perform each task was determined using a combination of approaches including direct observation and Delphi consensus-building methodology. The calculations for determining staffing requirements for nursing aides are shown in Table 1.

The actual efficiency should be computed and monitored each week. If there are definite upward or downward trends in efficiency, adjustments in staffing levels should be made. The work level at which a department is staffed is a critical parameter. One approach could be to perform a sensitivity analysis and compute the FTE requirement at various levels of case activity and then decide on the appropriate staffing level.

5. APPLICATION OF SCHEDULING METHODOLOGIES

5.1. Introduction

In health care systems, the term *scheduling* is used in a number of different ways. It can refer to the way in which appointments are scheduled for patient visits, testing procedures, surgical procedures, and the like. This form of scheduling is called work scheduling because patients represent work. In certain departments, it may be possible to move the workload from peak periods to periods of low demand to balance the workload. An example is an outpatient clinic where only a given number of appointment slots are made available. The number of appointment slots in a session determines the maximum number of patients that can be seen during that session, thereby restricting the workload to a defined level. In other departments, such as an emergency room, the patients or workload cannot be scheduled.

Another use of the this term is in the scheduling of personnel in a department based on the varying demand for services during the various hours of the day and days of the week while providing equitable scheduling for each individual over weekends and holidays. It must be noted that in the inpatient setting in health care, certain staff, such as nursing staff, need to be scheduled for work even over holidays. In the scheduling of personnel, each individual is given a schedule indicating what days of the week and what hours of the day he or she would be working over a period.

TABLE 1 Staffing Requirements for Nursing Aides

(A) Variable Activities		
Room Turnover		10.0 min/case
Cleaning of Instruments		12.0 min/case
Sorting and Wrapping of Instrument Sets		6.0 min/case
Total for Variable Activities		28.0 min/case
(B) Constant Activities		Minutes per Day
Testing Equipment	5 min/day	5 min/day
Morning Set-up	10 min/day	10 min/day
Cleaning of Sterilizers	30 min × 9 sterilizers per week	54 min/day
Running Spordi Test	90 min/day	90 min/day
Stocking Supply Room	15 min/day	15 min/day
Checking for Outdates	2 hrs/week	24 min/day
Operation of Sterilizers	3 hrs/day	180 min/day
Stocking Scrub Area	20 min/day	20 min/day
Prepare Suction Containers	30 min/day	30 min/day
Taking Specimens to Lab	1 hr/day	60 min/day
Taking Supplies to Station	1 hr/day	60 min/day
Prepare Wrapper/Stock	1 hr/week	12 min/day
Other Misc. Tasks	1 hr/week	12 min/day
	Total Constant Activity Time	572 min/day
(C) Total Work Per Day at 100 Cases per Day		
Constant Activity Time		572 min/day
Variable Activity Time	23 min/case × 100 cases/day	2300 min/day
	Total Work per Day	2872 min/day
(D) Staffing Requirement at 90% Efficiency		
Time Available per FTE per Day	7.5 hr/day × 60 min/hr	450 min/day
Required FTE	= (Work per Day in Minutes)/(Available min/day/FTE × Efficiency)	
	= (2872 min/day)/(450min/day/FTE × 0.9)	
	= 7.1 FTE	

5.2. Work Scheduling

Some of the areas where work scheduling is possible are elective admission scheduling, case scheduling in operating rooms, appointment scheduling in outpatient clinics, and appointments for radiological procedures and other testing such as amniocentesis. These are the areas where there is not an immediate and urgent need for services. To help determine the demand for services and understand the variability in demand, the workload data are collected by day of the week and hour of the day. In certain areas, it may be possible to smooth out the peaks and valleys in demand. For example, workload at a call center, where patients call by phone to schedule an appointment, can be shifted from busy hours to slow hours by playing a message on the phone as the patients wait to be served advising them to call during specific periods for reduced wait. After the workload has been smoothed, a decision is made to meet the demand a certain percentage of the time. After the staffing requirements are determined for this demand level using the methodologies discussed in the last section, staff is scheduled. Workload will then basically be restricted by the staff schedule.

Appointment slots are created based upon the type of work to be done and the number of providers available by hour of the day and day of the week. Appointments are scheduled as patients call for appointments. Once the appointment slots are filled, no additional work is scheduled. Appointment scheduling can be done using a manual system where an appointment book is maintained and the appointments are made by filling in the available slots. A number of computerized systems are on the market that basically allow the computerization of manual appointment books. These systems have the capability to search for available appointment slots meeting the time restrictions of the patient.

5.2.1. Appointment Scheduling in a Primary Care Clinic

Most people have a specific physician whom they see when they have a health-related problem. Individuals belonging to an HMO (health maintenance organization) select a primary care physician who takes care of all of their health care needs. For infants and young children, this would be a pediatrician, and for adults and teenagers, an internist or a family practitioner. Appointment scheduling or work scheduling in a primary care clinic (internal medicine) is discussed here.

An internal medicine clinic had been struggling with the appointment scheduling issue for a period of time. In particular, they wanted to determine the number of slots to be provided for physical exams, same-day appointments, and return appointments. The first step was to estimate the total demand for physical exams and for other appointments in the clinic. Data on average number of visits per person per year to the internal medicine clinic were available from past history. Multiplying this number by the total number of patients seen in the clinic yielded an estimate of total number of visits to be handled in the clinic.

The number of visits to the clinic varied by month. For example, clinic visits were higher during the flu season than in the summer months. Past data helped in the estimation of the number of visits to be handled each month. It was determined that the number of visits varied by day of the week, too. Mondays were the busiest and Fridays the least busy. The number of visits to be handled each day of the month was estimated based on this information. Staff scheduling was done to meet this demand. The total number of appointment slots was divided among physical exams, same-day visits, and return visits based on historical percentages. Other issues such as no-show rates and overbooking were also taken into account in appointment scheduling.

5.3. Personnel Scheduling

Once the work has been scheduled in areas where work scheduling is possible and the staffing requirements have been determined based upon the scheduled work, the next step is to determine which individuals will work which hours and on what days. This step is called personnel scheduling. In other areas of the hospital, such as the emergency room, where work cannot be scheduled, staffing requirements are determined using historical demand patterns by day of the week and hour of the day. The next step again is to determine the schedule for each staff member.

The personnel schedule assigns each staff member to a specific pattern of workdays and off days. There are two types of scheduling patterns: cyclical and noncyclical. In cyclical patterns, the work pattern for each staff is repeated after a given number of weeks. The advantage of this type of schedule is that staff members know their schedules, which are expected to continue until there is a need for change due to significant change in the workload. The disadvantage is the inflexibility of the schedule in accommodating demand fluctuations and the individual needs of workers. In noncyclical schedules, a new schedule is generated for each scheduling period (usually two to four weeks) and is based on expected demand and available staff. This approach provides flexibility to adjust the staffing levels to expected demand. It can also better accommodate requests from workers for days off during the upcoming scheduling period. But developing a new schedule every two to four weeks is very time consuming.

A department may allow working five 8-hour days, four 10-hour days, or three 12-hour shifts plus one 4-hour day. The personnel schedule has to be able to accommodate these various patterns. The schedule must also provide equitable scheduling for each individual when coverage is needed over weekends and holidays.

Personnel schedules are developed by using either a heuristic, trial-and-error approach or some optimization technique. Several computerized systems are available using both types of approaches.

Warner (1974) developed a computer-aided system for nurse scheduling that maximizes an expression representing the quality of schedule subject to a constraint that minimum coverage be met. Sitompul and Randhawa (1990) give detailed review and bibliography of the various nurse-scheduling models that have been developed.

6. APPLICATION OF QUEUING AND SIMULATION METHODOLOGIES

There are a number of examples of queues or waiting lines in health care systems where the customers must wait for service from one or more servers. The customers are usually the patients but could be hospital employees too. When patients arrive at the reception desk in a clinic for an appointment, they may have to wait in a queue for service. A patient calling a clinic on the phone for an appointment may be put on hold before a clinic service representative is able to answer the call. An operating room nurse may have to wait for service at the supply desk before a supply clerk is able to handle the request for some urgently needed supply. These are all examples of queuing systems where the objective may be to determine the number of servers to obtain a proper balance between the average customer waiting time and server idle time based on cost consideration. The objective may also be to ensure that only a certain percentage of customers wait more than a predetermined amount of time (for example, that no more than 5% of the patients wait in the queue for more than five minutes.)

In simple situations, queueing models can be applied to obtain preliminary results quickly. Application of queueing models to a situation requires that the assumptions of the model be met. These assumptions relate to the interarrival time and service time probability distributions. The simplest model assumes Poisson arrivals and exponential service times. These assumptions may be approximately valid for certain queues in health care systems. Incoming phone calls to a call center for scheduling appointments have been shown to follow Poisson distribution. Usually the service times are also approximately exponentially distributed. Queueing models with Poisson arrivals and exponential service times have also been applied to emergency rooms for determining proper staffing of nurses and physicians (Hale 1988). From the data collected over a period of time, he was able to verify the validity of assumptions made in the model and determine the average waiting time and average number of patients waiting for service.

In situations where a simple queueing model is not applicable or a more accurate result is desired, simulation models are used. Simulation models are also used when there are complex probabilistic systems with a number of entities interacting with each other. These models are also useful during the design stage to investigate the effect of changes in various parameters in a system.

Simulation models have been extensively used in the analysis of emergency rooms (Weissberg 1977; Saunders et al. 1989). Ortiz and Etter (1990) used general-purpose simulation system (GPSS) to simulate the emergency room. They found simulation modeling to be a better approach even for simple, less complex systems because of its flexibility and capacity to investigate various alternatives. Dawson et al. (1994) used simulation models to determine optimal staffing levels in the emergency room. McGuire (1997) used a simulation model to investigate various process changes and choose a solution that significantly reduced the length of stay for patients in the emergency room.

Simulation models have also been extensively used in other areas of health care systems. Roberts and English (1981) published a bibliography of over 400 articles in literature related to simulation in health care. The *Journal of the Society for Health Systems* (1992b, 1997) published two issues dedicated to health care applications of simulation. Bressan et al. (1988) and Mahachek (1992) discussed generic structure for simulation in health care. Dumas and Hauser (1974) developed a simulation model to study various hospital admission policies. Kachhal et al. (1981) used GPSS to simulate outpatient internal medicine clinics in order to investigate various clinic consolidation options. They used another model to determine optimal staffing for audiologists administering hearing tests ordered by ear, nose, and throat (ENT) physicians. Dumas (1984) used simulation in hospital bed planning. Hunter et al. (1987) used a model to simulate movement of surgical patients in a facility. Levy et al. (1989) developed a simulation model using SIMAN to assist in the design of a new outpatient service center. Wilt and Goddin (1989) simulated staffing needs and the flow of work in an outpatient diagnostic center. Lowery and Martin (1993) used a simulation model in a critical care environment. Cirillo and Wise (1996) used simulation models to test the impact of changes in a radiology department. Schroyer (1997) used a simulation model to plan an ambulatory surgery facility. Benneyan (1997) conducted a simulation study of a pediatric clinic to reduce patient wait during a visit to the clinic.

7. APPLICATION OF STATISTICAL METHODS

Several statistical methods have been used in the analysis and improvement of health care systems. Variability is the fact of life in any kind of data collected in health care, and statistical methods are the tools needed to understand this data. In this section, some common examples of the use of statistical analysis and modeling in health care will be presented.

As stated above, use of queuing and simulation models requires collection of data regarding the probability distribution of interarrival times for customers and service times. Use of a specific queuing model requires that the assumptions regarding the probability distributions in the model are valid. Industrial engineers must be able to verify that the collected data fit the assumed probability distribution. Similarly, simulation models need the data regarding various service times and other probabilistic elements of the model. Statistical methods have also been used to study and manage variability in demand for services (Sahney 1982).

7.1. Use of Regression Models

Regression models can be used to predict important response variables as a function of variables that could be easily measured or tracked. Kachhal and Schramm (1995) used regression modeling to predict the number of primary care visits per year for a patient based upon patient's age and sex and a measure of health status. They used ambulatory diagnostic groups (ADGs) encountered by a patient during a year as a measure of health status of the patient. All the independent variables were treated as 0–1 variables. The sex variable was 0 for males and 1 for females. The age of the patient was included in the model by creating ten 0–1 variables based on specific age groups. The model was able to explain 75% of the variability in the response variable. Similar models have been used to predict health care expenses from patient characteristics.

7.2. Determination of Sample Size

Industrial engineers frequently encounter situations in health care systems where they need to determine the appropriate amount of data needed to estimate the parameters of interest to a desired level of accuracy. In health care systems, work-sampling studies are frequently done to answer questions such as “What percentage of a nurse’s time is spent on direct patient care?” or “What fraction of a clinic service representative’s time is spent on various activities such as answering phones, checking-in the patients, looking for medical records, and the like?” In work-sampling studies, the person being studied is observed a predetermined number of times at random time intervals. For each observation, the task being performed is recorded. The estimate of the fraction of time spent on an activity is obtained by dividing the number of occurrences of that activity by the total number of observations.

The number of observations needed for a work-sampling study can be calculated from the following formula:

$$N = \frac{Z^2 \times p(1 - p)}{I^2}$$

where N = number of observations
Z = normal probability distribution factor, based on confidence level
p = unknown fraction to be estimated
I = desired margin of error in estimation

It is common to use a 95% confidence level for which the value of Z in the equation is 1.96. The value of I needs to be selected based upon the situation. For example, if the fraction being estimated is in 0.4–0.6 range, a value of I as ± 0.02 may be acceptable, but ± 0.05 may be too high. The value of p to be used in the equation poses a problem because this is the unknown fraction one is attempting to estimate from the data. If, from past experience or from literature, an approximate value of p is known, it can be used in the calculation of N. Another approach is to use $p = 0.5$ because it results in the largest value of N for fixed values of Z and I. If the computed value of N is too large to observe due to time and cost constraints, the only choice is to tolerate a larger possible error or a lower level of confidence.

A similar formula is also available for determining the sample size in situations that require determination of actual time to perform a task. For example, for a staffing study in a call center, one may need to estimate average time needed to service a phone call.

7.3. Use of Control Charts

Control charts have been used in health care as visual tools to observe the performance of a process over a period of time and alert the managers when an investigation may be necessary to identify an external cause that may have affected the performance. Individual (X) and moving range (MR) charts have been used for measurement data such as patient wait time, turnaround time for a test, number of times an error is made during a day, and length of stay for a particular diagnostic related group (DRG). P-charts have been used for fractions or percentages. Examples include patient satisfaction data, percentage utilization of a resource, bed occupancy expressed as % of total beds, and fraction of patient transport requests handled within 15 minutes of request. Cumulative sum charts have been used to detect small but significant changes in the mean. Kachhal and Schramm (1995) describe a bed model where auto regressive integrated moving average was used on a control chart with economic control limits to manage bed capacity at a hospital.

8. APPLICATION OF OPTIMIZATION MODELS

The best-known application of optimization models has been in the context of nurse scheduling. Warner (1976) formulated the problem of determining the optimal nurse schedule as a two-phase multiple-choice programming problem and solved it using a modified version of Balintfy’s algorithm. In his formulation, the objective is the determination of the schedule that maximizes an expression representing the quality of the schedules, quantifying the degree to which the nurses like the schedules they would work. The solution is determined subject to the constraints that the minimum coverage requirement for each shift and for each skill level be met. The popular ANSOS nurse-scheduling software uses this model. Warner et al. (1990) discuss automated nurse scheduling software.

Trivedi (1976) studied the problem of the reallocation of nursing resources at the beginning of each shift. He developed a mathematical model that minimizes the total severity index for the hospital, which is a measure of the need for an additional staff member by skill level. The model was solved using a branch-and-bound algorithm to determine the optimal assignment of nurses by skill level to various inpatient units.

Fries and Marathe (1981) used mathematical models to determine the optimal variable-sized multiple-block appointment system. Calichman (1990) used a linear programming model to balance the bed demand among various surgical services.

9. APPLICATION OF QUALITY-IMPROVEMENT TOOLS

The concepts of quality control, quality improvement, and quality management in the United States largely evolved from the efforts of Dr. W. Edward Deming, Dr. Joseph Juran, and Philip Crosby. Manufacturing industries were first to apply these concepts to achieve significant improvements in the quality of manufactured goods while reducing costs. During the late 1980s, the health care industry started using some of these methods to improve quality of health care delivery and reduce costs. Many hospitals started continuous quality improvement (CQI) and total quality management (TQM) programs. Sahney et al. (1989) summarize 10 common points from the quality philosophies of Deming, Juran, and Crosby and conclude that they can be applied to any industry, including the health care industry, to improve quality.

Continuous quality improvement focuses on improvement of various processes in a health care organization. It starts with identifying the customers and their needs for each of the processes. The objective is to improve the processes to meet the needs of the customers. All the changes are data driven. Key quality characteristics are identified and measured to track improvements. The physicians and hospital employees are trained in quality concepts and tools for improvement. Some of these tools are the Pareto chart, the cause–effect (or fishbone) diagram, flowcharts, run charts, and control charts. These tools assist in the documentation, analysis, and improvement of processes. Sahney and Warden (1991), and Gaucher and Coffey (1993) give a good overview of the application of CQI approach in health care. A popular CQI methodology is FOCUS-PDCA. Griffith et al. (1995) give a description of this methodology.

Another term in the quality improvement effort that has gained popularity during the late 1990s is *process reengineering.* While the CQI approach looks at making incremental improvements in the processes, reengineering efforts look at radical redesign of systems and processes to achieve major improvements in the cost and quality of services (Hammer and Champy 1995). Kralovec (1993) provides a good introduction to applying reengineering techniques to health care. The proceedings of the Quest for Quality and Productivity in Health Services Conferences that have been held each year since 1989 provide numerous examples of the application of total quality management, CQI, and reengineering methods in health care systems.

10. APPLICATION OF INFORMATION SYSTEMS/DECISION SUPPORT TOOLS

Many projects conducted by industrial engineers in health care deal with the use of information systems to improve quality and performance. There is a need to take a team approach in information system-related projects. A project team is needed that has individuals from management engineering and information services as well as the user department. The role of an industrial engineer on this team is to first interview the client department to determine the user requirements. The requirements are then used in the request for proposal (RFP) process to obtain proposals from vendors of information systems. Subsequently, the evaluation of proposals is also done on the basis of meeting these requirements as well as other factors. Industrial engineers are also used in the implementation phase to redesign the manual processes in the client department because these processes usually require major changes after the implementation of an information system. Kachhal and Koch (1989) report on the development of user requirements for a management information system for an operating room and the evaluation of the systems available in the market in meeting these requirements.

Decision support systems (DSS) have become a standard component of hospital information systems. They allow managers to look at financial data on a product line basis. A product line is a collection of one or more diagnostic related groups (DRGs). DSS are able to integrate various sources of data and give managers the capability for ad hoc reporting, building models to project revenues and costs, analyzing various reimbursement scenarios, and the like. The *Journal of the Society for Health Systems* (1991) published an issue with focus on decision support systems that presents a good overview of DSS in health care.

The other type of DSS used in health care are the clinical decision support systems (CDSS), which use a set of clinical findings such as signs, symptoms, laboratory data, and past history to assist the care provider by producing a ranked list of possible diagnoses. Some CDSS act as alerting systems by triggering an alert when an abnormal condition is recognized. Another set of these systems act as critiquing systems by responding to an order for a medical intervention by a care provider. Several systems that detect drug allergies and drug interactions fall into this category. Another use of CDSS is in consulting systems that react to a request for assistance from a physician or a nurse to provide suggestions about diagnoses or concerning what steps to take next. The application of CDSS in implementing clinical practice guidelines has been a popular use of these systems. Doller (1999) provides a current status of the use of these systems as medical expert systems in health care industry. The author predicts that the major use of CDSS in future will be in implementing clinical guidelines and allowing nonphysicians to handle routine care while directing the attention of true medical experts to cases that are not routine.

While some industrial engineers may be involved in the development of home-grown DSS or CDSS, the role of other industrial engineers is still in the evaluation and cost–benefit analysis of the standard systems available in the market.

11. APPLICATION OF OTHER INDUSTRIAL ENGINEERING TECHNIQUES

The application of other industrial engineering techniques such as forecasting, inventory control, materials management, facilities planning, economic analysis, and cost control is also common in health care systems. Examples of many of these applications are available in the proceedings of the annual Quest for Productivity and Productivity Conferences of the Society for Health Systems and the Healthcare Information and Management Systems Society Conferences. See also *Journal of the Society for Health Systems* (1992a), and Sahney (1993) on the use of industrial engineering tools in health care.

12. FUTURE TRENDS

Most healthcare systems around the country are facing serious financial problems due to the reductions in Medicare and Medicaid reimbursements. Health maintenance organizations (HMOs) have also reduced payments to health systems for services due to reductions in health care premiums forced by competition from other HMOs and by employers looking for cost reductions in their own organizations. Health care institutions are looking for ways to reduce costs while maintaining or improving quality. Industrial engineers are being asked to conduct benchmarking studies to identify areas where costs are higher than the norm and need reduction.

The financial problems are also leading to mergers and takeovers of financially troubled hospitals by health care systems in better financial health. Industrial engineers are being asked to work with individuals from finance on projects related to mergers and consolidation of services to evaluate the financial impact of alternative courses of action.

Many health care systems have undertaken TQM programs that include training of managers and supervisors in quality improvement tools such as flowcharts, cause–effect diagrams, run charts, and control charts. Quality improvement teams (QITs) consisting of employees from the departments in which the process resides are now conducting many of the studies previously conducted by industrial engineers. Industrial engineers are being asked to serve on these teams as facilitators or in staff capacity to conduct quantitative analysis of the alternatives.

Although simulation tools have been available to health care systems for decades, it is only recently that simulation models have been increasingly used as tools to model real systems to predict the impact of various policies and resource levels. Administrators and department managers are aware of the capabilities of these models and are requesting the use of simulation models to make sound decisions. The availability of reasonably priced, user-friendly simulation packages has increased the use of simulation modeling. The same is true for advance statistical tools. With the availability of user-friendly statistical software, more sophisticated statistical models are being built to guide in decision making.

Another area gaining attention is supply chain management. As health care systems are looking for ways to reduce costs without impacting the quality of patient care, procurement of supplies has stood out as an area where substantial reductions can be achieved through supply chain management. Some hospitals are having some of their supplies shipped directly from manufacturers to the user departments, bypassing all the intermediate suppliers and distributors. Other hospitals have eliminated their warehouses and distribution centers, giving the responsibility for daily deliveries and restocking of departments to external vendors. Industrial engineers are evaluating the alternative courses of action in supply chain management.

Industrial engineers are also working for consulting companies contracted by health care systems to assist them out of financial problems. The use of sophisticated industrial engineering tools in health care still lags behind the use in manufacturing industry. Over the next decade, it is expected that the use of these tools will become more prevalent in health care.

REFERENCES

Benneyan, J. (1997), “An Introduction to Using Simulation in Health Care: Patient Wait Case Study,” *Journal for the Society for Health Systems,* Vol. 5, No. 3, pp. 1–16.

Bressan, C., Facchin, P., and Romanin Jacur, G. (1988), “A Generalized Model to Simulate Urgent Hospital Departments,” in *Proceedings of the IMACS Symposium on System Modeling and Simulation,* pp. 421–425.

Calichman, M. (1990), "The Use of Linear Programming to Balance Bed Demand," in *Proceedings of the 1990 Annual Healthcare Information and Management Systems Conference*, AHA, pp. 385–390.

Cirillo, J., and Wise, L. (1996), "Testing the Impact of Change Using Simulation," in *Proceedings of the 1996 Annual Healthcare Information and Management Systems Conference,* AHA, pp. 51–64.

Dawson, K., O'Conner, K., Sanchez, P., and Ulgen, O. (1994), "How to Conduct a Successful Emergency Center Staffing Simulation Study," in *Proceedings of the 1994 Annual Healthcare Information and Management Systems Conference,* AHA, pp. 273–287.

Doller, H. J. (1999), "Medical Expert Systems: How Do We Get There from Here?" in *Proceedings of the 1999 Annual HIMSS Conference*, Healthcare Information and Management Society, pp. 173–185.

Dumas, M. B. (1984), "Simulation Modeling for Hospital Bed Planning," *Simulation,* Vol. 43, No. 2, pp. 69–78.

Dumas, M., and Hauser, N. (1974), "Hospital Admissions System Simulation for Policy Investigation and Improvement," in *Proceedings of the 6th Pittsburgh Conference on Modeling and Simulation* (Pittsburgh), pp. 909–920.

Fries, B., and Marathe, V. (1981), "Determination of Optimal Variable Sized Multiple-Block Appointment Systems," *Operations Research*, Vol. 29, pp. 324–345.

Gaucher, E., and Coffey, R. (1993), *Total Quality in Health Care,* Jossey-Bass, San Francisco.

Gilbreth, F. B. (1916), "Motion Study in Surgery," *Canadian Journal of Medical Surgery*, Vol. 40, No. 1, pp. 22–31.

Gilbreth, L. M. (1945), "Time and Motion Study," *Modern Hospital*, Vol. 65, No. 3, pp. 53–54.

Gilbreth, L. M. (1950), "Management Engineering and Nursing," *American Journal of Nursing*, Vol. 50, No. 12, pp. 780–781.

Goldberg, A. J., and Denoble, R. A., Eds. (1986), *Hospital Departmental Profiles*, American Hospital Publishing, Chicago, IL.

Griffith, J., Sahney, V., and Mohr, R. (1995), *Reengineering Health Care: Building on CQI,* Health Administration Press, Ann Arbor, MI.

Hale, J. (1988), "Queuing Theory in the Emergency Department," in *Proceedings of the 1988 Annual Healthcare Systems Conference*, vol. 1, AHA, pp. 1–7.

Hammer, M., and Champy, J. (1993), *Reengineering the Corporation: A Manifesto for Business Revolution,* HarperCollins, Inc., New York.

Hunter, B., Asian, S., and Wiget, K. (1987), "Computer Simulation of Surgical Patient Movement in a Medical Care Facility," in *Proceedings of the 11th Annual Symposium on Computer Applications in Medical Care,* pp. 692–697.

Journal of the Society for Health Systems (1991), Special Issue: Decision Support Systems, Vol. 3, No. 1.

Journal of the Society for Health Systems (1992a), Special Issue: IE Contributions, Vol. 3, No. 4.

Journal of the Society for Health Systems (1992b), Special Issue: Simulation, Vol. 3, No. 3.

Journal of the Society for Health Systems (1997), Special Issue: Simulation, Vol. 5, No. 3.

Kachhal, S., and Koch, F. (1989), "Evaluation of the Operating Room Management Information System Packages Available in the Market," in *Proceedings of the 1989 Annual Healthcare Systems Conference*, AHA, pp. 431–441.

Kachhal, S. K., and Schramm, W. S. (1995), "Using Statistical Methods to Improve Healthcare Systems," in *Proceedings of the Quest for Quality and Productivity in Health Services Conference,* Institute of Industrial Engineers, pp. 219–228.

Kachhal S., Klutke, G., and Daniels, E. (1981), "Two Simulation Applications to Outpatient Clinics," in *Winter Simulation Conference Proceedings*, IEEE, pp. 657–665.

Kralovec, O. (1993), "Applying Industrial Reengineering Techniques to Health Care," *Healthcare Information Management,* Vol. 7, No. 2, pp. 3–10.

Levy, J., Watford, B., and Owen, V. (1989), "Simulation Analysis of an Outpatient Service Facility," *Journal of the Society for Health Systems*, Vol. 17, No. 2, pp. 35–46.

Lowery, J., and Martin, J. (1992), "Design and Validation of a Critical Care Simulation Model," *Journal of the Society for Health Systems,* Vol. 3, No. 3, pp. 15–36.

Mahachek, A. R. (1992), "An Introduction to Patient Flow Simulation for Health Care Managers," *Journal of the Society for Health Systems,* Vol. 3, No. 3, pp. 73–81.

Mahon, J. J., Ed. (1978), *Hospitals,* Mahon's Industry Guides for Accountants and Auditors, Guide 7, Warren, Gorham & Lamont, Boston.

McGuire, F. (1997), ''Using Simulation to Reduce the Length of Stay in Emergency Departments,'' *Journal of the Society for Health Systems,* Vol. 5, No. 3, pp. 81–90.

Nock, A. J. (1913), '' Frank Gilbreth's Great Plan to Introduce Time-Study into Surgery,'' *American Magazine*, Vol. 75, No. 3, pp. 48–50.

Ortiz, A., and Etter, G. (1990), ''Simulation Modeling vs. Queuing Theory,'' in *Proceedings of the 1990 Annual Healthcare Information and Management Systems Conference*, AHA, pp. 349–357.

Roberts, S. D., and English, W. L., Eds. (1981), *Survey of the Application of Simulation to Healthcare,* Simulation Series, Vol. 10, No. 1, Society for Computer Simulation, La Jolla, CA.

Sahney, V. (1982), ''Managing Variability in Demand: A Strategy for Productivity Improvement in Healthcare,'' *Health Care Management Review*, Vol. 7, No. 2, pp. 37–41.

Sahney, V. (1993), ''Evolution of Hospital Industrial Engineering: From Scientific Management to Total Quality Management,'' *Journal of the Society for Health Systems*, Vol. 4, No. 1, pp. 3–17.

Sahney, V., and Warden, G. (1989), ''The Role of Management in Productivity and Performance Management,'' in *Productivity and Performance Management in Health Care Institutions*, AHA, pp. 29–44.

Sahney, V., and Warden, G. (1991), ''The Quest for Quality and Productivity in Health Services,'' *Frontiers of Health Services Management,* Vol. 7, No. 4, pp. 2–40.

Sahney, V. K., Peters, D. S., and Nelson, S. R. (1986), ''Health Care Delivery System: Current Trends and Prospects for the Future,'' *Henry Ford Hospital Medical Journal,* Vol. 34, No. 4, pp. 227–232.

Sahney, V., Dutkewych, J., and Schramm, W. (1989), ''Quality Improvement Process: The Foundation for Excellence in Health Care,'' *Journal of the Society for Health Systems*, Vol. 1, No. 1, pp. 17–29.

Saunders, C. E., Makens, P. K., and Leblanc, L. J. (1989), ''Modeling Emergency Department Operations Using Advanced Computer Simulation Systems,'' *Annals of Emergency Medicine,* Vol. 18, pp. 134–140.

Schroyer, D. (1997), ''Simulation Supports Ambulatory Surgery Facility Design Decisions,'' in *Proceedings of the 1997 Annual Healthcare Information and Management Systems Conference,* AHA, pp. 95–108.

Sitompul, D., and Randhawa, S. (1990), ''Nurse Scheduling Models: A State-of-the-Art Review,'' *Journal of the Society for Health Systems,* Vol. 2, No. 1, pp. 62–72.

Smalley, H. E. (1982), *Hospital Management Engineering*, Prentice Hall, Englewood Cliffs, NJ.

Steinwelds, B., and Sloan, F. (1981) ''Regulatory Approvals to Hospital Cost Containment: A Synthesis of the Empirical Evidence,'' in *A New Approach to the Economics of Health Care*, M. Olsin, Ed., American Enterprise Institute, Washington, DC., pp. 274–308.

Trivedi, V. (1976), ''Daily Allocation of Nursing Resources,'' in *Cost Control in Hospitals,* J. R. Griffith, W. M. Hancock, and F. C. Munson, Eds., Health Administration Press, Ann Arbor, MI, pp. 202–226.

Warner, D. M. (1976), ''Computer-Aided System for Nurse Scheduling,'' in *Cost Control in Hospitals,* J. R. Griffith, W. M. Hancock, and F. C. Munson, Eds., Health Administration Press, Ann Arbor, MI, pp. 186–201.

Warner, M., Keller, B., and Martel, S. (1990), ''Automated Nurse Scheduling,'' *Journal of the Society for Health Systems,* Vol. 2, No. 2, pp. 66–80.

Weissberg, R. W. (1977), ''Using Interactive Graphics in Simulating the Hospital Emergency Department,'' in *Emergency Medical Systems Analysis,* T. Willemain and R. Larson, Eds., Lexington Books, Lexington, MA, pp. 119–140.

Wilt, A., and Goddin, D. (1989), ''Healthcare Case Study: Simulating Staffing Needs and Work Flow in an Outpatient Diagnostic Center,'' *Industrial Engineering*, May, pp. 22–26.

CHAPTER 28

Industrial Engineering Applications in Financial Asset Management

R. McFall Lamm, Jr.
Deutsche Bank

1. INTRODUCTION 751
2. THE ASSET-MANAGEMENT PROBLEM 752
- 2.1. Origins 752
- 2.2. Problem Structure and Overview 752
- 2.3. Implementation 753

3. AN ILLUSTRATION 753
- 3.1. The Efficient Frontier 753
- 3.2. Investor Risk Preferences 753

4. THE OPTIMIZATION PROBLEM 755
5. CAVEATS 756
- 5.1. Shortcomings of Mean-Variance Analysis 756
- 5.2. Dangers of Extrapolating from History 756
- 5.3. Asset Selection 757

6. NEW ASSET CLASSES 758
- 6.1. Desirable Asset Characteristics 758
- 6.2. Hedge Funds 759
- 6.3. Private Equity and Venture Capital 759
- 6.4. Inflation-Protected Securities 761
- 6.5. Other Assets 761

7. THE FORECASTING PROBLEM 761
8. ENGINEERING CLIENT-TAILORED SOLUTIONS: APPLYING PORTFOLIO RESTRICTIONS 763
9. TAXATION 764
- 9.1. Tax-Efficient Optimization 764
- 9.2. Alternative Tax Structures 764
- 9.3. Results of Tax-Efficient MV Optimizations 764

10. TIME HORIZON 766
11. EXTENSIONS AND NEW FRONTIERS 767
12. COMBINING MEAN-VARIANCE ANALYSIS WITH OTHER TECHNIQUES—CONSTRUCTING OPTIMAL HEDGE FUND PORTFOLIOS 768
13. CONCLUSION 769
REFERENCES 770

1. INTRODUCTION

Over the last decade, the challenges faced by asset managers have become significantly more complex. The array of investment choices has expanded tremendously in response to globalization and financial engineering. New products are now available such as collateralized debt, insurance-linked securities, and exotic structured products. At the same time, the need for investment planning has shifted more to individuals as companies and public entities relinquish responsibility for retirement programs.

Because of this incredible change in the investment world, institutions such as pension plan administrators, endowments, foundations, and asset managers are under greater pressure than ever to produce higher portfolio returns. This is the genesis for embracing new assets, which hopefully will

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

generate superior investment performance than plain vanilla portfolios of stocks and bonds. Better returns reduce the need for future cash injections to fund obligations.

Unfortunately, market risks have increased as the array of available investment instruments has broadened. For example, the Mexican peso crisis in 1994, the Asian currency debacle and recession beginning in 1997, the Russian debt default and the unprecedented hedge fund bail-out coordinated by the Federal Reserve Bank in 1998, and a 30% drop in the price of technology shares early in 2000 all had major repercussions for financial markets. Where is an investor to find solace in such an unfriendly and disturbing environment?

The obvious answer to heightened complexity and uncertainty lies in utilizing financial engineering techniques to manage asset portfolios. This chapter reviews the current state of the art from a practitioner's perspective. The prime focus is on mean-variance optimization techniques, which remain the principal application tool. The key message is that while the methods employed by today's specialists are not especially onerous mathematically or computationally, there are major issues in problem formulation and structure. It is in this arena that imagination and inventiveness take center stage.

2. THE ASSET-MANAGEMENT PROBLEM

The job of investment managers is to create portfolios of assets that maximize investment returns consistent with risk tolerance. In the past, this required simply selecting a blend of stocks, bonds, and cash that matched the client's needs. Asset managers typically recommended portfolios heavily weighted in stocks for aggressive investors desiring to build wealth. They proffered portfolios overweight in bonds and cash for conservative investors bent on wealth preservation. Aggressive stock-heavy portfolios would be expected to yield higher returns over time but with considerable fluctuations in value. Concentrated cash and bond portfolios would be less volatile, thus preserving capital, but produce a lower return.

2.1. Origins

Until recently, a casual rule-of-thumb approach sufficed and was deemed adequate to produce reasonable performance for investors. For example, a portfolio consisting of 65% equities and 35% bonds generated returns and risk similar to a 75% equities and 25% bonds portfolio. However, following the inflation trough in the 1990s, bond returns declined and investors with low equity exposure suffered. In addition, those investors ignoring alternative assets such as hedge funds and venture capital found themselves with lagging performance and greater portfolio volatility.

Studies have consistently shown that selection of the asset mix is the most important determinant of investment performance. Early influential research by Brinson et al. (1986, 1991) and a more recent update by Ibbotson and Kaplan (1999) indicate that asset allocation explains up to 90% of portfolio returns. Security selection and other factors explain the remainder. Consequently, the asset blend is the key intellectual challenge for investment managers and should receive the most attention. Traditional rules of thumb no longer work in a dynamic world with many choices and unexpected risks.

2.2. Problem Structure and Overview

Markowitz was the first to propose an explicit quantification of the asset-allocation problem (Markowitz 1959). Three categorical inputs are required: the expected return for each asset in the portfolio, the risk or variance of each asset's return, and the correlation between asset returns. The objective is to select the optimal weights for each asset that maximizes total portfolio return for a given level of portfolio risk. The set of optimum portfolios over the risk spectrum traces out what is called the efficient frontier.

This approach, usually referred to as mean-variance (MV) analysis, lies at the heart of modern portfolio theory. The technique is now mainstream, regularly taught in investment strategy courses. A massive literature exists exploring the methodology, its intricacies, and variations. "Black box" MV optimization programs now reside on the desks of thousands of brokers, financial advisors, and research staff employed by major financial institutions.

The principal advantage of MV analysis is that it establishes portfolio construction as a process that explicitly incorporates risk in a probabilistic framework. In this way, the approach acknowledges that asset returns are uncertain and requires that precise estimates of uncertainty be incorporated in the problem specification.

On the surface, MV analysis is not especially difficult to implement. For example, it is very easy to guess at future stock and bond returns and use historical variances and correlations to produce an optimum portfolio. It is not so simple to create a multidimensional portfolio consisting of multiple equity and fixed income instruments combined with alternative assets such as private equity, venture capital, hedge funds, and other wonders. Sophisticated applications require a lot of groundwork, creativity, and rigor.

2.3. Implementation

Because the vast majority of MV users are neophytes who are not fully aware of its subtleties, nuances, and limitations, they are sometimes dismayed by the results obtained from MV optimization programs. The reason is that achieving sensible outcomes is highly dependent on quality input. Very often the return, risk, and correlations injected into MV models are empirically and theoretically inconsistent. This generates fallacious and highly distorted portfolios, leading many novices to reject the approach as capricious and unrealistic.

Mean-variance analysis must be integrated with a specific investment process if the results are to be useful. The steps required include:

1. Specifying the investment alternatives to be considered
2. Accurately forecasting expected asset returns, variances, and correlations
3. Executing the optimization
4. Choosing the appropriate implementation vehicles that deliver the performance characteristics embedded in the analysis

Optimization results must be carefully reviewed to ensure that assumptions satisfy consistency requirements and examined for solution sensitivity to changes in inputs.

3. AN ILLUSTRATION

3.1. The Efficient Frontier

To illustrate the essence of the asset-allocation problem, I begin with a simple example that includes the following assets: U.S. stocks, international stocks, U.S. bonds, international bonds, and cash. This constitutes a broad asset mix typical of that used by many professional managers. Expected returns are 11.0%, 12.6%, 6.0%, 6.9%, and 5.0%, respectively. The expected standard deviations (risk) of these returns are 12.5%, 14.7%, 4.3%, 7.8%, and 0.0%. Note that the standard deviation of cash returns is zero because this asset is presumed to consist of riskless U.S. Treasury securities. The correlation matrix for asset returns is:

$$\mathbf{R} = \begin{vmatrix} 1.0 & 0.55 & 0.30 & 0.00 & 0.00 \\ & 1.0 & 0.10 & 0.30 & 0.00 \\ & & 1.0 & 0.30 & 0.00 \\ & & & 1.0 & 0.00 \\ & & & & 1.0 \end{vmatrix}$$

Optimum portfolios are obtained by selecting the asset weights that maximize portfolio return for a specific portfolio risk. The problem constraints are that (1) the portfolio return is a linear combination of the separate asset returns, (2) portfolio variance is a quadratic combination of weights, asset risks, and asset correlations, and (3) all weights are positive and sum to one.

The results of producing MV portfolios are presented in Table 1. The corresponding efficient frontier is shown in Figure 1. The allocations obtained are typical for analyses of this type. That is, to obtain higher portfolio returns, the investor must take on more risk. This is accomplished by weighting equities more heavily. The most conservative portfolios have both low returns and low risk. They consist primarily of cash and bonds. For moderate levels of risk, investors should blend stocks and bonds together in more equal proportions. In addition, note that the most risky portfolios skew to higher weights for international equities, which are more risky than U.S. stocks.

It should be obvious that the production of MV portfolios is not extraordinarily input intensive. Efficient frontiers for five asset portfolios require only five predicted returns, five standard deviations, and a five-by-five symmetrical matrix of correlation coefficients. Yet the process yields indispensable information that allows investors to select suitable portfolios.

3.2. Investor Risk Preferences

Once the efficient frontier is established, the issue of investor risk preferences must be addressed. Individuals exhibit markedly different attitudes towards risk. Some are extremely risk averse, tolerating nothing but near certainty in life. Others relish risk taking. Most are somewhere between these two extremes.

Risk attitudes depend on personality, life experience, wealth, and other socioeconomic factors. By default, extremely risk-averse investors are not interested in building wealth via asset accumulation, because they are unwilling to tolerate the risk necessary to obtain high returns. The opposite is true for adventurous souls ready to gamble for large payoffs. In this regard, “old-money” investors

TABLE 1 Selected Optimal Mean-Variance Portfolios

Return	Risk	Sharpe Ratio[a]	U.S. Stocks	International Stocks	U.S. Bonds	International Bonds	Cash
			Portfolio Allocation				
12.6%	15.0%	0.84	—	100%	—	—	—
12.5%	14.0%	0.89	9%	91%	—	—	—
12.2%	13.0%	0.94	25%	75%	—	—	—
11.8%	12.0%	0.99	43%	55%	—	2%	—
11.3%	11.0%	1.03	42%	48%	—	10%	—
10.8%	10.0%	1.08	40%	40%	—	20%	—
10.3%	9.0%	1.14	37%	34%	6%	23%	—
9.7%	8.0%	1.21	32%	29%	17%	22%	—
9.1%	7.0%	1.30	27%	25%	19%	19%	9%
8.5%	6.0%	1.42	23%	22%	17%	17%	22%
7.9%	5.0%	1.59	19%	18%	14%	14%	35%
7.4%	4.0%	1.84	16%	15%	11%	11%	48%
6.8%	3.0%	2.26	12%	11%	8%	8%	61%
6.2%	2.0%	3.09	8%	7%	6%	6%	74%
5.6%	1.0%	5.59	4%	4%	3%	3%	87%

Source: Results of optimization analysis.

[a] The Sharpe ratio is the portfolio return less the risk-free Treasury rate divided by portfolio risk.

are usually more risk averse than “new-money” investors. They require portfolios heavily weighted in conservative assets. New-money investors are comfortable with inherently risky equities (Figure 2). Age often plays an important role. Older investors are commonly more risk averse because they are either retired or close to retirement and dissaving (Figure 3).

How does one ascertain risk tolerance to guarantee a relevant portfolio is matched to the investor’s needs? There are a number of ways. One is to estimate the risk-aversion parameter based on the investor’s response to a battery of questions designed to trace out their return preferences with different payoff probabilities.

Some financial managers simply use personal knowledge of the investor to deduce risk tolerance. For example, if an investor with a $10 million portfolio is distraught if the portfolio’s value plunges 20% in a year, he or she is not a candidate for an all-equity portfolio. Yet another approach used by some managers is to work backwards and determine the risk necessary for the investor to reach specific wealth-accumulation goals.

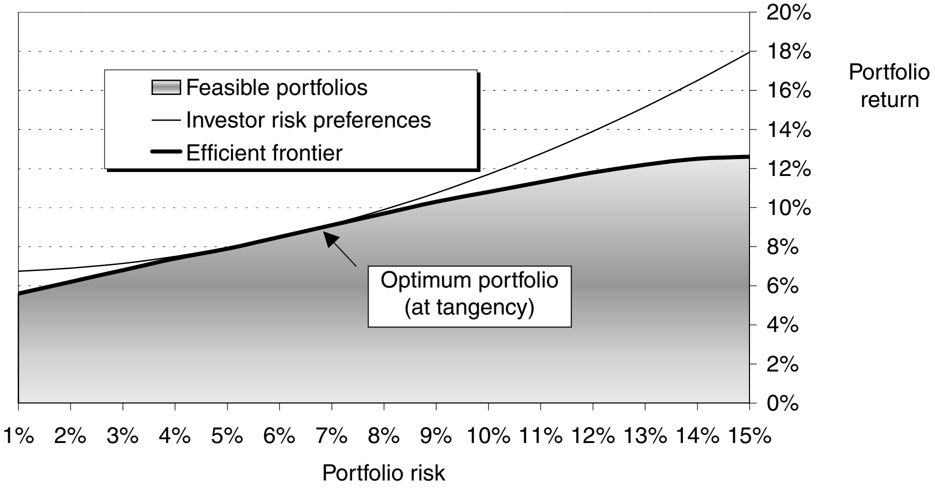


Figure 1 The Efficient Frontier and Investor Risk Preferences.

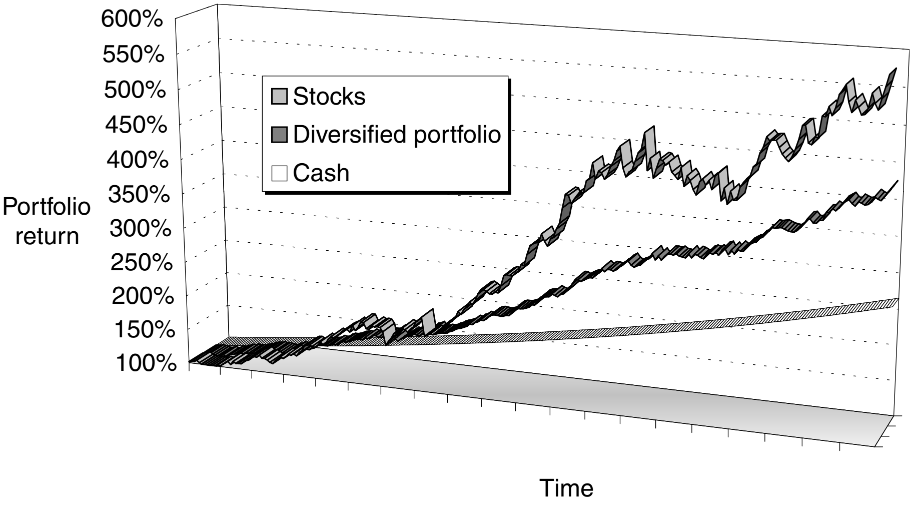


Figure 2 Wealth Creation vs. Wealth Preservation.

4. THE OPTIMIZATION PROBLEM

The specific mathematical formulation for the standard MV problem is straightforward:

$$\max \mathrm{U}(\mathbf{w}) = (r - \lambda\sigma^2) \quad \text{subject to} \tag{1}$$

$$r = \mathbf{w}^{\mathrm{T}}\mathbf{r} \tag{2}$$

$$\sigma = \mathbf{w}^{\mathbf{T}}\mathbf{S}^{\mathbf{T}}\mathbf{R}\mathbf{S}\mathbf{w} \tag{3}$$

$$\text{colsum } \mathbf{w} = 1 \tag{4}$$

$$w_j \geq 0 \tag{5}$$

where $\mathbf{w} = [w_1\ w_2\ .\ .\ .\ w_J]^{\mathrm{T}}$ = the portfolio weights for each asset
$\mathbf{r} = [r_1\ r_2\ .\ .\ .\ r_J]^{\mathrm{T}}$ = the return vector
$\mathbf{S}$ = the risk matrix with diagonal standard deviations $s_1\ s_2\ .\ .\ .\ s_J$
$\mathbf{R}$ = the $J \times J$ asset correlation matrix
λ = the investor's risk-aversion parameter

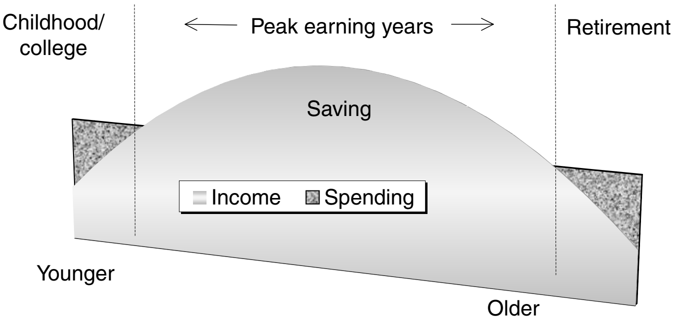

Childhood/ college
⟵ Peak earning years ⟶
Retirement
Saving
Income
Spending
Younger
Older

Figure 3 The Investor's Life Cycle.

Equation (1) incorporates the investor's attitude to risk via the objective function U. Equation (2) represents the portfolio return. Equation (3) is portfolio risk. Constraint (4) requires that the portfolio weights sum to 100%, while constraint (5) requires that weights be positive. This latter restriction can be dropped if short selling is allowed, but this is not usually the case for most investors.

This formulation is a standard quadratic programming problem for which an analytical solution exists from the corresponding Kuhn–Tucker conditions. Different versions of the objective function are sometimes used, but the quadratic version is appealing theoretically because it allows investor preferences to be convex.

The set of efficient portfolios can be produced more simply by solving

$$\max\ r = \mathbf{w}^{\mathrm{T}}\mathbf{r} \text{ subject to } \sigma = \mathbf{w}^{\mathbf{T}}\mathbf{S}^{\mathbf{T}}\mathbf{RSw},\ \text{colsum } \mathbf{w} = 1,\ w_j \geq 0 \text{ for target } \sigma \tag{6}$$

or alternatively,

$$\max\ \sigma = \mathbf{w}^{\mathbf{T}}\mathbf{S}^{\mathbf{T}}\mathbf{RSw} \text{ subject to } r = \mathbf{w}^{\mathrm{T}}\mathbf{r},\ \text{colsum } \mathbf{w} = 1,\ w_j \geq 0 \text{ for target } r \tag{7}$$

A vast array of alternative specifications is also possible. Practitioners often employ additional inequality constraints on portfolio weights, limiting them to maximums, minimums, or linear combinations. For example, total equity exposure may be restricted to a percentage of the portfolio or cash to a minimum required level. Another common variation is to add inequality constraints to force solutions close to benchmarks. This minimizes the risk of underperforming.

With respect to computation, for limited numbers of assets (small J), solutions are easily obtained (although not necessarily efficiently) using standard spreadsheet optimizers. This works for the vast majority of allocation problems because most applications typically include no more than a dozen assets. More specialized optimizers are sometimes necessary when there are many assets. For example, if MV is applied to select a stock portfolio, there may be hundreds of securities used as admissible "assets."

5. CAVEATS

5.1. Shortcomings of Mean-Variance Analysis

One must be cognizant that the MV approach has several important shortcomings that limit its effectiveness. First, model solutions are sometimes very sensitive to changes in the inputs. Second, the number of assets that can be included is generally bounded. Otherwise, collinearity problems can result that produce unstable allocations and extreme asset switching in the optimal portfolios. Third, the asset allocation is only as good as forecasts of prospective returns, risk, and correlation. Inaccurate forecasts produce very poorly performing portfolios. The first two limitations can be addressed through skillful model specification. The third requires that one have superlative forecasting ability.

A common mistake committed by naive users of MV analysis is to use recent returns, risk, and correlation as predictors of the future. Portfolios produced using such linear extrapolation methods normally exhibit poor performance. Table 2 shows historical returns and risk for various asset classes over the last decade. A variety of extraneous market developments, shocks, and one-time events produced these results. Rest assured that future time paths will not closely resemble those of the 1990s. For this reason, while one cannot ignore history, extending past performance into the future is a poor technique.

5.2. Dangers of Extrapolating from History

As an example of the extrapolation fallacy, consider portfolio performance over the last two decades. If one constructed an efficient portfolio in 1990 based on the 1980s history, large allocations would have been made to international equities. This is primarily due to the fact that Japanese stocks produced the best returns in the world up to 1989. Yet in the 1990s, Japanese equities fell by more than 50% from their 1989 peak, and the best asset allocation would have been to U.S. equities. Using the 1980s history to construct MV portfolios would have produced dismal portfolio returns (Table 3).

Empirical work by Chopra and Ziembra (1993) demonstrated that the most critical aspect of constructing optimal portfolios is the return forecast. For this reason, a shortcut employed by some practitioners is to concentrate on the return forecast and use historical risk and correlation to construct optimum portfolios. This may prove satisfactory because correlations and risk are more stable than returns and are therefore more easily predicted. However, this line of attack may be ineffective if return forecasts substantially deviate from history and are combined with historical risk and correlations. In this case, the optimal allocations can skew overwhelmingly to the high return assets

Whatever method is used to obtain forecasts, once the optimum portfolio is determined, the manager can momentarily relax and wait. Of course, actual outcomes will seldom match expectations.

TABLE 2 Annual Returns and Risk for Various Assets, 1990–99

Group	Asset	Return	Volatility	Sharpe Ratio
Equity	Large caps	17.7%	13.4%	0.94
	Mid caps	17.2%	16.0%	0.75
	Small caps	14.1%	17.3%	0.52
	Nasdaq	24.2%	20.6%	0.92
	International	8.6%	17.1%	0.20
	Europe	14.7%	14.6%	0.65
	Japan	2.6%	26.0%	−0.10
	Emerging markets	13.5%	23.8%	0.35
Fixed income	Treasury bonds	7.2%	4.0%	0.51
	Treasury notes	6.4%	1.7%	0.74
	Municipals	6.7%	4.2%	0.36
	International bonds hedged	8.1%	3.5%	0.83
	International bonds unhedged	8.4%	8.7%	0.37
	U.S. high yield debt	10.5%	5.5%	0.96
	Emerging market debt	16.3%	17.0%	0.65
Alternative	Passive commodities	−0.9%	7.5%	−0.81
	Active commodities (CTAs)	5.2%	11.7%	0.00
	Hedge funds	14.8%	4.3%	2.22
	Venture capital	28.1%	19.0%	1.21
	Private equity	15.7%	8.0%	1.32
	REITs	8.5%	12.1%	0.28
	Cash (3-month Treasuries)	5.2%	0.4%	0.00

Source: Deutsche Bank.

No investment manager possesses perfect foresight. Errors in forecasting returns, risk, and correlation will produce errors in portfolio composition. Obviously, managers with the best forecasting ability will reap superior results.

5.3. Asset Selection

If a manager does not forecast extremely well, it is possible to produce superior investment performance via asset selection. That is, choosing an exceptional array of candidate assets, can enhance portfolio returns due to diversification benefits that other managers miss.

For example, consider a simple portfolio of U.S. equities and bonds. Normally managers with the best forecasts will achieve better performance than other managers investing in the same assets. But another manager, who may not forecast U.S. equity and bond returns extremely well, can outperform by allocating funds to assets such as international equities and bonds. These assets possess different returns, risks, and correlations with each other and U.S. assets. Their inclusion shifts the efficient frontier upward beyond that resulting when only U.S. stocks and bonds are considered.

TABLE 3 Optimum Portfolios: The 1980s vs. the 1990s

		Asset				Portfolio	
Period	Measure	S&P	International Stocks	Bonds	Cash	Return	Risk
1980s	Weight	15%	47%	34%	4%		
	Returns	18%	22%	12%	9%	17.4%	
	Risk	16%	18%	12%	0%		10.5%
1990s	Weight	75%	0%	25%	0%		
	Returns	18%	7%	8%	5%	15.2%	
	Risk	14%	18%	4%	0%		10.5%
1980s portfolio performance in 1990s:						8.9%	10.8%

Source: Deutsche Bank.

The primary distinction between asset selection and asset allocation is that the thought processes differ. In asset selection, a manager focuses on defining the candidate universe broadly. In asset allocation, assets are typically viewed as given and the effort is on forecast accuracy.

A deep knowledge of markets and investment possibilities is necessary to identify the broadest possible asset universe. A manager who incorporates new assets with enticing features has a larger information set than a manager laboring in a narrowly defined world. This is why astute investment managers are constantly searching for new assets—their goal is to gain an edge over competitors by shifting the efficient frontier outward (Figure 4). Neglecting the opportunity to employ an ubiquitous asset domain is a common failure of beginners who rely on black box MV solutions they do not fully understand.

6. NEW ASSET CLASSES

A virtually endless stream of new investment products is paraded forth by financial service companies each year. Although promoted as innovative with superlative benefits, in reality most such creations are often narrowly focused with high risk or are reconfigured and redundant transformations of assets already available. Sorting through the muck and discovering innovative products that shift the efficient frontier outward is not an easy task

6.1. Desirable Asset Characteristics

In general, new assets that effectively shift the efficient frontier outward must possess differential return and risk profiles, and have low correlation with the existing assets in the portfolio. Incorporating an asset that has similar returns and risk and is collinear with an asset already in the portfolio is duplicative. One needs only one of the two.

Differential returns, risk, and correlation are not enough, however. A new asset that has a lower return and significantly higher volatility will not likely enter optimal portfolios even if it has a low correlation with existing assets. In contrast, a new asset with a higher return, lower risk, and modest correlation with existing assets will virtually always enter at least some of the optimal portfolios on the efficient frontier.

In decades past, adding real estate, commodities, and gold to stock and bond portfolios produced the desired advantages. Low returns for these assets in the 1980s and subsequent years led to their eventual purging from most portfolios by the mid-1990s. They were replaced with assets such as international stocks and bonds, emerging market securities, and high-yield debt.

Today, most managers employ multiple equity classes in their portfolios to achieve asset diversification. They usually include U.S. as well as international and emerging market stocks. Some managers split the U.S. equities category into large, midsized, and small capitalization securities, or alternatively into growth and value stocks. Some managers use microcap stocks. The fixed-income assets commonly used include U.S. bonds, international bonds, high-yield debt, and emerging market debt. Some managers further break U.S. bonds into short, medium, and long-duration "buckets." A

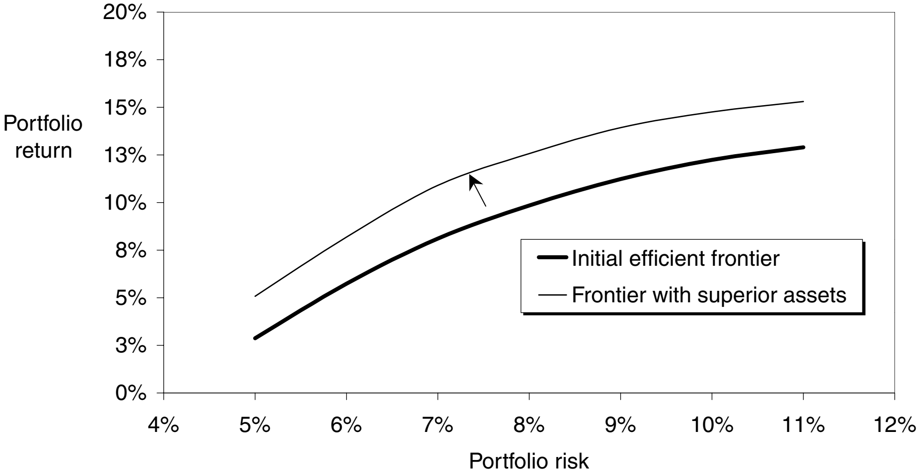


Figure 4 Improving Investment Performance via Asset Selection.

few investment managers retain real estate as an investable asset, while a smaller minority includes commodities as an inflation hedge.

One problem is that the correlation between different equity and fixed income subcategories is fairly high. Thus, the advantages gained from dividing equities and fixed income into components are not that great. In fact, cross-market correlation has increased recently due to globalization. For example, the performance of large, small, and midcap equities, as well as that of international equities, now moves much more in parallel than in the past.

Where are the new opportunities to improve portfolio performance by creatively expanding the investable asset domain? There are several assets that are now increasingly accepted and utilized in portfolios. The most important are hedge funds, private equity, and venture capital. Other new assets have shown the ability to improve portfolio performance but for various reasons have not yet made the leap to widespread recognition. These include inflation-protected securities and insurance-linked products.

6.2. Hedge Funds

Over the last several years, hedge funds have garnered tremendous attention as an attractive addition to classic stock and bond portfolios. A virtual avalanche of published MV studies indicates that hedge funds extend the efficient frontier. Examples include research by Agarwal and Naik (2000), Lamm and Ghaleb-Harter (2000b), Lamm (1999a), Purcell and Crowley (1999), Edwards and Lieu (1999), and Schneeweis and Spurgin (1999). Other work supporting the inclusion of hedge funds in portfolios was done by Tremont Partners and TASS Investment Research (1999), Goldman Sachs and Financial Risk Management (1999), Yago et al. (1999), Schneeweis and Spurgin (1998), and Fung and Hseih (1997). The general conclusions are that substantial hedge fund allocations are appropriate, even for conservative investors.

All of these studies presume that hedge fund exposure is accomplished via broad-based portfolios. That is, like equities, hedge fund investments need to be spread across a variety of managers using different investment strategies. This means at least a dozen or more. Investing in only one hedge fund is an extremely risky proposition, much like investing in only one stock.

The primary advantage of hedge fund portfolios is that they have provided double-digit returns going back to the 1980s with very low risk. Indeed, hedge fund portfolio volatility is close to that of bonds. With much higher returns and low correlation compared with traditional asset classes, they exhibit the necessary characteristics required to enhance overall portfolio performance.

Hedge funds are private partnerships and are less regulated than stocks. For this reason, a strict due diligence process is essential to safeguard against the rare occasion when unscrupulous managers surface. Because most investors cannot or do not want to perform comprehensive evaluations of each hedge fund, funds of hedge funds are increasingly used as implementation vehicles. Large investment banks or financial institutions usually sponsor these funds. They have professional staffs capable of performing a laborious investigative and review process.

Table 4 reproduces the results of a MV optimization by Lamm and Ghaleb-Harter to assess the impact of adding hedge funds to portfolios of traditional assets. Results are given for a conservative range of prospective hedge fund returns varying from 6.5–10.5%. The volatility of the hedge fund portfolio is close to 4%, not much different from history. The assumed correlation with stocks and bonds is 0.3, also close to historical experience. The other assumptions are similar to those presented in the Section 3 example.

The key findings are that (1) hedge funds primarily substitute for bonds and (2) substantial allocations are appropriate for practically all investors, even conservative ones. This is true even if relatively low hedge fund returns are assumed. The conclusion that hedge funds are suitable for conservative investors may be counterintuitive to many readers, who likely perceive hedge funds as very risky investments. However, this result is now strongly supported by MV analysis and occurs primarily because hedge fund managers practice diverse investment strategies that are independent of directional moves in stocks and bonds. These strategies include merger and acquisition arbitrage, yield curve arbitrage, convertible securities arbitrage, and market-neutral strategies in which long positions in equities are offset by short positions.

In response to recent MV studies, there has been a strong inflow of capital into hedge funds. Even some normally risk-averse institutions such as pension funds are now making substantial allocations to hedge funds. Swensen (2000) maintains that a significant allocation to hedge funds as well as other illiquid assets is one of the reasons for substantial outperformance by leading university endowment funds over the last decade.

6.3. Private Equity and Venture Capital

As in the case of hedge funds, managers are increasingly investing in private equity to improve portfolio efficiency. Although there is much less conclusive research on private equity and investing, the recent performance record of such investments is strong.

TABLE 4 Optimum Portfolios with Hedge Funds and Conventional Assets

			Allocation					
Hedge Fund Return	Total Portfolio		Equities		Bonds		Hedge Funds	Cash
	Risk	Return	U.S.	International	U.S.	International		
10.5%	12.5%	12.2%	—	83%	—	—	17%	—
	10.5%	11.9%	—	66%	—	—	34%	—
	8.5%	11.5%	—	50%	—	—	50%	—
	6.5%	11.2%	—	31%	—	—	69%	—
	4.5%	10.6%	—	4%	—	—	96%	—
9.5%	12.5%	12.1%	15%	76%	—	—	9%	—
	10.5%	11.6%	11%	62%	—	—	27%	—
	8.5%	11.0%	7%	46%	—	—	46%	—
	6.5%	10.5%	3%	29%	—	—	67%	—
	4.5%	9.6%	—	7%	—	—	90%	—
8.5%	12.5%	12.0%	33%	66%	—	—	1%	—
	10.5%	11.3%	26%	53%	—	—	21%	—
	8.5%	10.6%	19%	40%	—	—	41%	—
	6.5%	9.8%	11%	25%	—	—	64%	—
	4.5%	8.8%	5%	8%	—	10%	77%	—
7.5%	12.5%	12.4%	35%	65%	—	—	—	—
	10.5%	11.7%	34%	48%	—	—	18%	—
	8.5%	10.7%	26%	36%	—	3%	35%	—
	6.5%	9.7%	19%	23%	—	11%	48%	—
	4.5%	8.4%	10%	8%	—	19%	62%	—
6.5%	12.5%	12.4%	35%	65%	—	—	—	—
	10.5%	11.6%	41%	44%	—	15%	—	—
	8.5%	10.4%	34%	31%	6%	22%	6%	—
	6.5%	9.2%	23%	22%	15%	20%	20%	—
	4.5%	7.9%	16%	14%	12%	14%	17%	27%

Source: Lamm and Ghaleb-Harter 2000b.

There are varying definitions of private equity, but in general it consists of investments made by partnerships in venture capital (VC) leveraged buyout (LBO) and mezzanine finance funds. The investor usually agrees to commit a minimum of at least $1 million and often significantly more. The managing partner then draws against commitments as investment opportunities arise.

In LBO deals, the managing partner aggregates investor funds, borrows additional money, and purchases the stock of publicly traded companies, taking them private. The targeted companies are then restructured by selling off noncore holdings, combined with similar operating units from other companies, and costs reduced. Often existing management is replaced with better-qualified and experienced experts. After several years, the new company is presumably operating more profitably and is sold back to the public. Investors receive the returns on the portfolio of deals completed over the partnership's life.

The major issue with private equity is that the lock-up period is often as long as a decade. There is no liquidity if investors want to exit. Also, because the minimum investment is very high, only high-net-worth individuals and institutions can participate. In addition, investments are often concentrated in only a few deals, so that returns from different LBO funds can vary immensely. Efforts have been made to neutralize the overconcentration issue with institutions offering funds of private equity funds operated by different managing partners.

VC funds are virtually identical in structure to LBO funds except that the partnership invests in start-up companies. These partnerships profit when the start-up firm goes public via an initial public offering (IPO). The partnership often sells its shares during the IPO but may hold longer if the company's prospects are especially bright. VC funds received extraordinary attention in 1999 during the internet stock frenzy. During this period, many VC funds realized incredible returns practically overnight as intense public demand developed for dot-com IPOs and prices were bid up speculatively.

Because LBO and VC funds are illiquid, they are not typically included in MV portfolio optimizations. The reason is primarily time consistency of asset characteristics. That is, the risk associated with LBOs and VC funds is multiyear in nature. It cannot be equated directly with the risk of stocks, bonds, and hedge funds, for which there is ample liquidity. A less egregious comparison requires multiyear return, risk, and correlation calculations.

6.4. Inflation-Protected Securities

Treasury inflation-protected securities (TIPs) are one of the most innovative financial products to appear in recent years. They pay a real coupon, plus the return on the Consumer Price Index (CPI). Thus, they automatically protect investors 100% against rising inflation, a property no other security possesses. Further, in a rising inflation environment, returns on TIPs are negatively correlated with those of bonds and other assets whose prices tend to decline when inflation rises.

In an MV study of combining TIPs with conventional assets, Lamm (1998a, b) finds that an overweighting in these securities vs. traditional bonds is appropriate only when inflation is expected to rise. He suggests a 50/50 weighting vs. traditional bonds when inflation direction is uncertain and underweighting when inflation is expected to fall.

In years past, some investment managers touted gold, commodities, or real estate as inflation hedges. The availability of TIPs neutralizes arguments made in favor of these "real" assets because their correlation with inflation is less than perfect. TIPs provide a direct one-to-one hedge against rising inflation.

Despite the attractive characteristics of TIPs, they have met with limited market acceptance. This is a consequence partly of a lack of awareness but also because TIPs significantly underperformed conventional Treasury securities as inflation declined immediately after their introduction in 1997. However, for the first time in years, inflation is now rising and TIPs have substantially outperformed Treasuries over the last year. This will likely change investor attitudes and lead to greater acceptance of the asset class.

6.5. Other Assets

Another new asset proposed for inclusion in portfolios is insurance-linked products such as catastrophe bonds. These securities were evaluated in MV studies done by Lamm (1998b,1999b). They are issued by insurance companies as a way of protecting against heavy losses that arise as a consequence of hurricanes or earthquakes. The investor receives a large payoff if the specified catastrophe does not occur or a low return if they do. Because payoffs are linked to acts of nature, the correlation of insurance-linked securities with other assets is close to zero. Unfortunately, the market for insurance-linked securities has failed to develop sufficient liquidity to make them broadly accessible to investors.

Another asset occasionally employed by investment managers is currency. While currencies meet liquidity criteria, their expected returns are fairly low. In addition, exposure to foreign stocks and bonds contains implicit currency risk. For this reason, currencies are not widely viewed as a distinct standalone asset. Even managers that do allocate to currencies tend to view positions more as short-term and tactical in nature.

Finally, there are a number of structured derivative products often marketed as new asset classes. Such products are usually perturbations of existing assets. For example, enhanced index products are typically a weighted combination of exposure to a specific equities or bond class, augmented with exposure to a particular hedge fund strategy. Similarly, yield-enhanced cash substitutes with principal guarantee features are composed of zero coupon Treasuries with the residual cash invested in options or other derivatives. Such products are part hedge fund and part primary asset from an allocation perspective and are better viewed as implementation vehicles rather than incorporated explicitly in the asset allocation.

7. THE FORECASTING PROBLEM

Beyond asset selection, the key to successful investment performance via MV asset allocation depends largely on the accuracy of return, risk, and correlation forecasts. These forecasts may be subjective or quantitative. A subjective or purely judgmental approach allows one the luxury of considering any number of factors that can influence returns. The disadvantage of pure judgment is that it is sometimes nebulous, not always easily explained, and may sometimes be theoretically inconsistent with macroeconomic constraints.

Model forecasts are rigorous and explicit. They derive straightforwardly from a particular variable set. Most quantitative approaches to forecasting are based on time series methods, explored in great depth in the finance literature over the years. Beckers (1996) provides a good review of such methods used to forecast returns, while Alexander (1996) surveys risk and correlation forecasting. In addition, Lummer et al. (1994) describe an extrapolative method as a basis for long-term asset allocation. An alternative to pure time series is causal models. For example, Connor (1996) explored macroeconomic factor models to explain equity returns. Lamm (2000) proposes a modified bivariate Garch model as one way of improving MV forecast accuracy.

In the case of a two-asset stock and bond portfolio, the modified Garch model proposed by Lamm is simply:

$$r_{1t} = \varphi_{11} + \varphi_{12} r_{1,t-1} + \Sigma \gamma_k x_{k,t-1} + \varepsilon_{1t} \tag{8}$$

$$r_{2t} = \varphi_{21} + \varphi_{22} r_{2,t-1} + \Sigma \gamma_k x_{k,t-1} + \varepsilon_{2t} \tag{9}$$

$$\sigma^2_{1t} = \alpha_{10} + \alpha_{11} \varepsilon^2_{1,t-1} + \beta_{11} \sigma^2_{1,t-1} \tag{10}$$

$$\sigma^2_{2t} = \alpha_{20} + \alpha_{21} \varepsilon^2_{2,t-1} + \beta_{21} \sigma^2_{2,t-1} \tag{11}$$

$$\sigma_{12t} = \alpha_{30} + \alpha_{31} \varepsilon_{1,t-1} \varepsilon_{2,t-1} + \beta_{31} \sigma_{12,t-1} \tag{12}$$

where r_{it} = the returns on equities and bonds, respectively
σ^2_{it} = the corresponding variances
ε_{it} = residuals
σ_{12t} = the covariance between stocks and bonds
x_k = exogenous factors (e.g., inflation, industrial production, corporate earnings, and interest rates) that determine market returns

The other symbols are estimated parameters. The first two equations predict returns, while the second two forecast associated risk. The correlation between stocks and bonds in any period is simply $\rho = \sigma_{12t} / \sigma_{1t}\sigma_{2t}$, which is derived from the last three equations. This model postulates that when extraneous changes occur that are not reflected in economic variables, the resulting prediction errors push up risk and reduce correlation—exactly the pattern observed in response to market shocks through time.

Lamm reports that augmenting Garch with exogenous variables significantly improves forecast accuracy. These findings have important implications for portfolio management. Critically, augmented Garch provides a more logical and systematic basis for reallocation decisions through the economic cycle (Figure 5) and changing inflation scenarios, which shift efficient frontiers (Figure 6). The augmented Garch process also allows one to distinguish economic influences from purely unexpected shocks, which are often event driven. A pure time series approach provides no such delineation.

If one desires to focus only on return forecasting, a useful approach is vector autoregression (VAR). Although largely atheoretic, except regarding system specification, such models have been shown to have superior predictive capability. In particular, VAR forecasts are more accurate the longer the periodicity of the data. For example, monthly VAR models typically explain 5–10% of the variation in returns, while annual VAR models often explain 90% or more. Quarterly forecasting models produce results in between. VAR models thus provide a reasonably good method for annual asset allocation while providing only a slight edge for monthly allocation.

VAR models are fairly simple and are specified as:

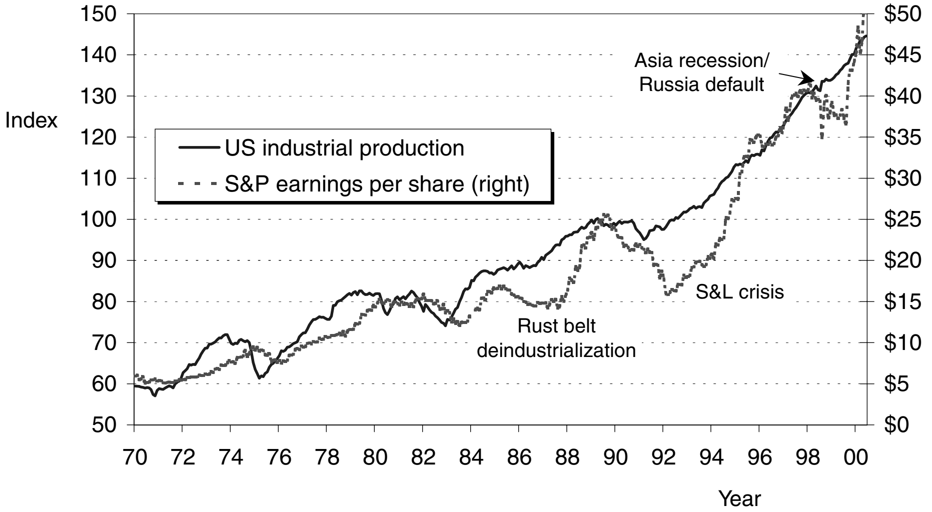


Figure 5 Corporate Earnings and Stock Prices Follow Economic Growth Cycles.

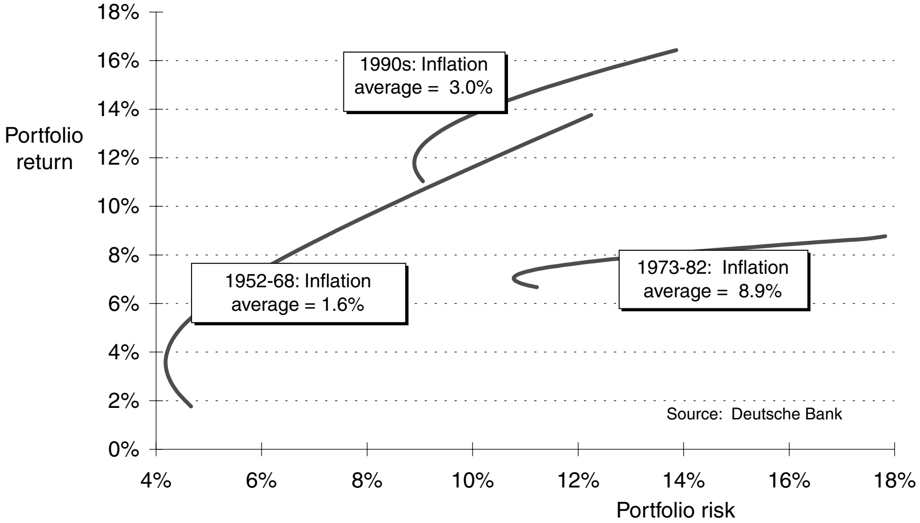


Figure 6 Efficient Frontiers in Different Macroeconomic Environments.

$$\mathbf{y}_t = \mathbf{A}L(\mathbf{y}_{t-1}) + \mathbf{e}_t \tag{13}$$

where $\mathbf{y}_t$ = the vector of system variables (including returns) that are to be predicted
$L(\ldots)$ = the lag operator
$\mathbf{e}_t$ = a vector of errors
$\mathbf{A}$ = the parameter matrix.

If some of the system variables are policy variables (such as Federal Reserve interest rate targets), they can be endogenized for making forecasts and evaluating scenarios. The forecasting system then becomes:

$$\mathbf{z}_{t+1} = \mathbf{B}L(\mathbf{z}_t) + \mathbf{C}L(\mathbf{x}_t) \tag{14}$$

where $\mathbf{z}$ contains the system variables to be forecast from $\mathbf{y} = [\mathbf{z}|\mathbf{x}]^{\mathrm{T}}$
$\mathbf{x}$ contains the expected policy values
$\mathbf{B}$ and $\mathbf{C}$ are the relevant coefficient submatrices contained in $\mathbf{A}$

8. ENGINEERING CLIENT-TAILORED SOLUTIONS: APPLYING PORTFOLIO RESTRICTIONS

The discussion so far has focused on asset allocation as generally applicable to a broad cross-section of investors. In reality, the vast majority of individual investors face special restrictions that limit their flexibility to implement optimal portfolios. For example, many investors hold real estate, concentrated stock holdings, VC or LBO partnerships, restricted stock, or incentive stock options that for various reasons cannot be sold. For these investors, standard MV optimization is still appropriate and they are well advised to target the prescribed optimum portfolio. They should then utilize derivative products such as swaps to achieve a synthetic replication.

Synthetic rebalancing cannot always be done, however. This is likely to be the case for illiquid assets and those with legal covenants limiting transfer. For example, an investor may own a partnership or hold a concentrated stock position in a trust whose position cannot be swapped away. In these situations, MV optimization must be amended to include these assets with their weights restricted to the prescribed levels. The returns, risk, and correlation forecasts for the restricted assets must then be incorporated explicitly in the analysis to take account of their interaction with other assets. The resulting constrained optimum portfolios will comprise a second-best efficient frontier but may not be too far off the unconstrained version.

An interesting problem arises when investors own contingent assets. For example, incentive stock options have no liquid market value if they are not yet exercisable, but they are nonetheless worth something to the investor. Because banks will not lend against such options and owners cannot realize value until exercise, it can be argued that they should be excluded from asset allocation, at least until the asset can be sold and income received. Proceeds should then be invested consistent with the investor's MV portfolio. An alternative is to probabilistically discount the potential value of the option to the present and include the delta-adjusted stock position in the analysis. To be complete, the analysis must also discount the value of other contingent assets such as income flow after taxes and living expenses.

9. TAXATION

Taxes are largely immaterial for pension funds, foundations, endowments, and offshore investors because of their exempt status. Mean-variance asset allocation can be applied directly for these investors as already described. However, for domestic investors, taxes are a critical issue that must be considered in modeling investment choices.

9.1. Tax-Efficient Optimization

Interest, dividends, and short-term capital gains are taxed at ''ordinary'' income rates in the United States. The combined federal, state, and local marginal tax rate on such income approaches 50% in some jurisdictions. In contrast, long-term capital gains are taxed at a maximum of 20%, which can be postponed until realization upon sale. For this reason, equities are tax advantaged compared with bonds and investments that deliver ordinary income.

In an unadulterated world, MV analysis would simply be applied to after-tax return, risk, and correlations to produce tax-efficient portfolios. Regrettably, tax law is complex and there are a variety of alternative tax structures for holding financial assets. In addition, there are tax-free versions of some instruments such as municipal bonds. Further complicating the analysis is the fact that one must know the split between ordinary income and long-term gains to forecast MV model inputs. This muddles what would otherwise be a fairly straightforward portfolio-allocation problem and requires that tax structure subtleties be built into the MV framework to perform tax-efficient portfolio optimization.

9.2. Alternative Tax Structures

As one might imagine, modeling the features of the tax system is not a trivial exercise. Each structure available to investors possesses its own unique advantages and disadvantages. For example, one can obtain equity exposure in a number of ways. Stocks can be purchased outright and held. Such a ''buy-and-hold'' strategy has the advantage that gains may be realized at the investor's discretion and the lower capital gains rate applied.

Alternatively, investors may obtain equity exposure by purchasing mutual funds. Mutual fund managers tend to trade their portfolios aggressively, realizing substantially more short-term and long-term gains than may be tax efficient. These gains are distributed annually, forcing investors to pay taxes at higher rates and sacrificing some of the benefits of tax deferral. If portfolio trading produces higher returns than a simple buy-and-hold strategy, such turnover may be justified. But this is not necessarily the case.

A third option for equity exposure is to purchase stocks via retirement fund structures using pretax contributions. Tax vehicles such as 401k, IRA, Keogh, and deferred compensation programs fall in this category. The major benefit of this approach is that it allows a larger initial investment to compound. However, upon withdrawal, returns are taxed at ordinary income tax rates.

A fourth alternative for achieving equity exposure is through annuity contracts offered by insurance companies. These are often referred to as wrappers. This structure allows investors to purchase equity mutual funds with after-tax dollars and pay no taxes until distribution. Thus, gains are sheltered. Even so, all income above basis is taxed at ordinary rates upon distribution. Investors receive deferral benefits but lose the advantage of paying lower rates on returns arising from long-term capital gains.

Although I have described four alternative tax vehicles for equity exposure, the same general conclusions apply for other assets. The one difference is that consideration of ordinary vs. capital gains differentials must explicitly be considered for each type of asset. Table 5 summarizes the tax treatment of investment income through currently available tax structures.

9.3. Results of Tax-Efficient MV Optimizations

Another perplexing feature of the tax system is that after-tax returns are affected by the length of the holding period. For example, it is well known that the advantage of owning equities directly increases the longer is the holding period. This is because deferring capital gains produces benefits via tax-free compounding (Figure 7). Similarly, assets in insurance wrappers must be held over very long periods before the positive effects of tax-free compounding offset the potential negatives of converting long-term gains to ordinary tax rates.

TABLE 5 Alternative Tax Structures for Asset Holding

	Annual Flow Taxation			
Structure	Dividends and Interest	Capital Gains	Annual Pre-Tax Return	Annualized After-Tax Return
Direct ownership	Ordinary rates	Deferred	$b_t = d_t + (r_T)^{1/\mathrm{T}} - 1$	$\tau d_t + [\kappa(r_T)^{1/\mathrm{T}} - 1]$
Mutual fund	Ordinary rates	Paid as gains realized	$f_t = d_t + r_t$	$\tau d_t + [\tau\alpha + \kappa(1 - \alpha)]r_t$
Retirement fund structure	Deferred	Deferred	$s_t = (2 - \tau)f_t$ or $s_t = (2 - \tau)b_t$	$(2 - \tau)\tau(d_t + r_t)$
Wrapper	Deferred	Deferred	$w_t = d_t + r_t$	$\tau(d_t + r_t)$

Source: Lamm and Ghaleb-Harter 2000c.

d_t = dividend return, r_T = gain from period 0 to T equal to $(p_T - p_0)/p_0$; r_t = gain from prior period, τ = one minus ordinary income tax rate, κ = one minus capital gains tax rate, α = percentage of gains that are short term, and T = investment horizon.

When it comes to applying the MV approach to taxable investing, the methodology is the same as that used for tax-free portfolios, except that four different after-tax return streams, risk measures, and four-square correlations are required for each asset. Table 6 compares the pre-tax and after-tax returns for assets typically considered in taxable portfolio allocation studies.

A complete review of after-tax-efficient frontiers would be beyond the scope of this chapter. However, there are four general conclusions. First, a basic conclusion of most after-tax allocation studies is that investors are well advised to hold the maximum possible portion of their investable wealth in tax-deferred vehicles such as 401k programs and deferred compensation plans. The 401k structure is particularly attractive because employers often match employee contributions.

Second, after maximizing contributions to tax-deferred vehicles, investors should utilize buy-and-hold strategies for equities and hold municipal bonds as the appropriate fixed income investment. Third, hedge funds enter most optimal allocations directly even though as much as 75% of the income is taxed at ordinary rates. Optimal allocations to hedge funds are lower than for tax-exempt investors, however. Fourth, for long horizons of at least a decade and more, wrapped assets often enter optimum portfolios, particularly for hedge funds.

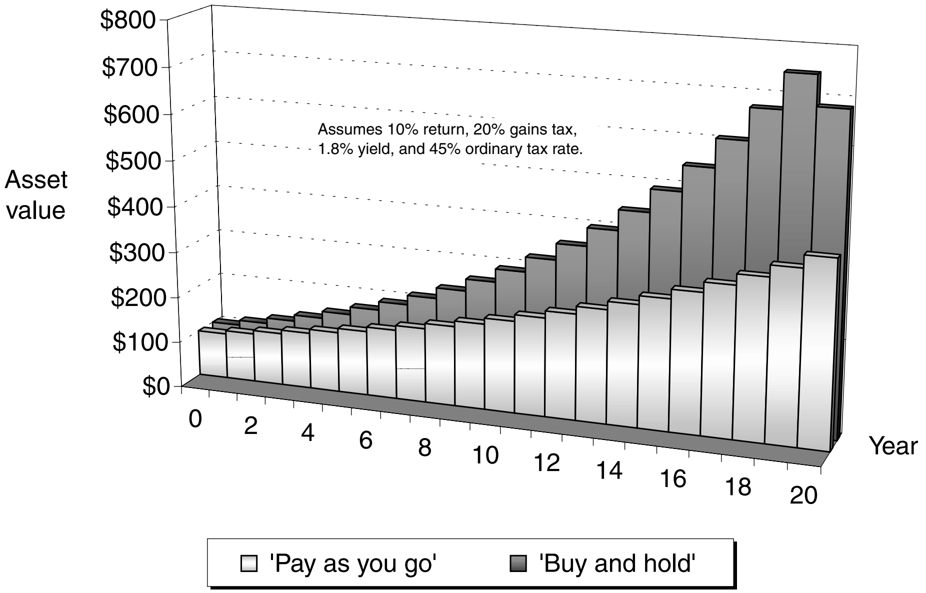


Figure 7 The Benefit of Deferring Capital Gains Taxes.

TABLE 6 After-Tax Returns

		5-Year Horizon			10-Year Horizon			20-Year Horizon		
Asset	Pre-Tax Return	Buy/ Hold	Fund	Wrap	Buy/ Hold	Fund	Wrap	Buy/ Hold	Fund	Wrap
U.S. stocks	11.5%	8.6%	7.8%	5.9%	9.0%	7.8%	6.6%	9.6%	7.8%	7.6%
International stocks	13.1%	9.8%	8.8%	6.9%	10.3%	8.8%	7.8%	11.0%	8.8%	9.1%
U.S. bonds	6.5%	3.5%	3.5%	2.7%	3.5%	3.5%	3.0%	3.5%	3.5%	3.4%
International bonds	7.4%	4.0%	4.0%	3.3%	4.0%	4.0%	3.6%	4.0%	4.0%	4.1%
Muni bonds	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%
REITs	8.25%	4.8%	4.8%	4.1%	4.8%	4.8%	4.5%	4.8%	4.8%	5.3%
Hedge funds	10.5%	6.2%	6.2%	5.2%	6.2%	6.2%	5.8%	6.2%	6.2%	6.8%
Cash	5.5%	3.0%	3.0%	2.1%	3.0%	3.0%	2.3%	3.0%	3.0%	2.6%
TE cash	3.2%	3.2%	3.2%	3.2%	3.2%	3.2%	3.2%	3.2%	3.2%	3.2%

Source: Lamm and Ghaleb-Harter 2000c.

Ordinary tax rates are assumed to be 46% (state and federal). Incremental costs of 1% annually are deducted for the wrap structure.

These generalizations presume investors are in the top marginal tax bracket (federal, state, and local) both at the beginning of the horizon and at the end. The extent to which investors move into lower brackets via diminished post-retirement income or by migrating to lower tax jurisdictions make the benefits of deferral even greater.

10. TIME HORIZON

Beyond the taxation issue, a key component of MV problem specification is properly defining the horizon over which the optimization is applied. In particular, risk and correlation can differ significantly, depending on the time frame considered. For example, equities are much less risky over five-year horizons than over one year (Figure 8). This has led some analysts to argue for larger equity allocations for long-term investors because stock risk declines relative to bonds over lengthy horizons.

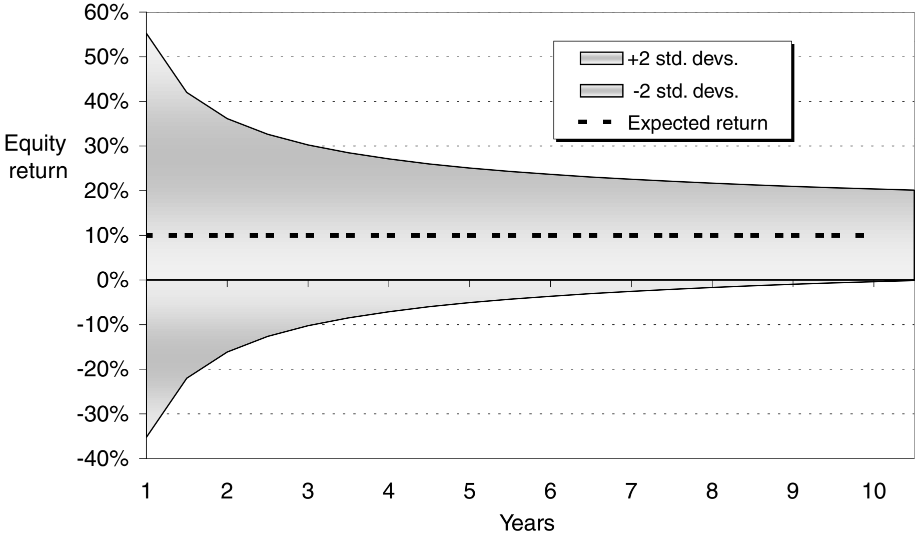


Figure 8 Equity Risk Is Lower over Long Time Horizons.

This issue is especially pertinent if illiquid assets such as private equity are to be included in the portfolio. Using annual volatility for private equity makes no sense—10-year risk must be considered.

In general, the longer the time horizon, the lower the risk for assets that appear to be risky when evaluated based on short-term data. As a consequence, long-horizon investors will allocate more to assets such as stocks and private equity. The reason for this is that annual risk is related to risk over other intervals by the identity $\sigma_a = \sigma_T T^{1/2}$, where T is the periodicity of the data. Note that $T = 4$, 12, or 52 for quarterly, monthly, and weekly data and $T = 1/2$, $1/5$, or $1/10$ for 2-, 5-, and 10-year intervals. Thus, assets with high annual risk have exponentially declining risk over lengthy time intervals. This is why long-horizon portfolios often have relatively small allocations to fixed-income instruments and higher allocations to what are perceived as riskier assets (Figure 9). Some managers go as far as to suggest that a long-horizon investor hold 100% equities if they desire liquidity and the ability to exit portfolios at a time of their choosing.

11. EXTENSIONS AND NEW FRONTIERS

There are numerous extensions of the basic MV model beyond those already described. For example, by dropping the constraint that portfolio weights be positive, the model can be used to ascertain short and long positions that should be held for various assets. Similarly, incorporating borrowing rates allows MV models to be used to design optimal leverageable portfolios. Furthermore, the MV approach can be applied to specific asset classes to build portfolios of individual securities such as equities or bonds. Beyond this, MV analysis has even broader strategic uses. For example, it can be applied by corporations to design portfolios of businesses.

One additional application of MV analysis is to use the technique to reengineer implied market returns. This requires the problem be reformulated to select a set of returns given asset weights, risk, and correlations. The weights are derived from current market capitalizations for equities, bonds, and other assets. The presumption is that today's market values reflect the collective portfolio optimizations of all investors. Thus, the set of returns that minimizes risk is the market's forecast of future expected returns. This information can be compared with the user's own views as a reliability check. If the user's views differ significantly, they may need to be modified. Otherwise the user can establish the portfolio positions reflecting his or her outlook under the premise his or her forecasts are superior.

Other variants of MV optimization have been proposed to address some of its shortcomings. For example, MV analysis presumes normal distributions for asset returns. In actuality, financial asset return distributions sometimes possess "fat tails." Alternative distributions can be used. In addition, the MV definition of risk as the standard deviation of returns is arbitrary. Why not define risk as first- or third-order deviation instead of second? Why not use absolute or downside deviation?

Recently, Duarte (1999) has suggested that value-at-risk approaches be used to derive efficient frontiers. This approach differs from MV analysis in that it uses Monte Carlo simulation to determine the dollar value of the portfolio that is at risk with a particular degree of statistical confidence. That

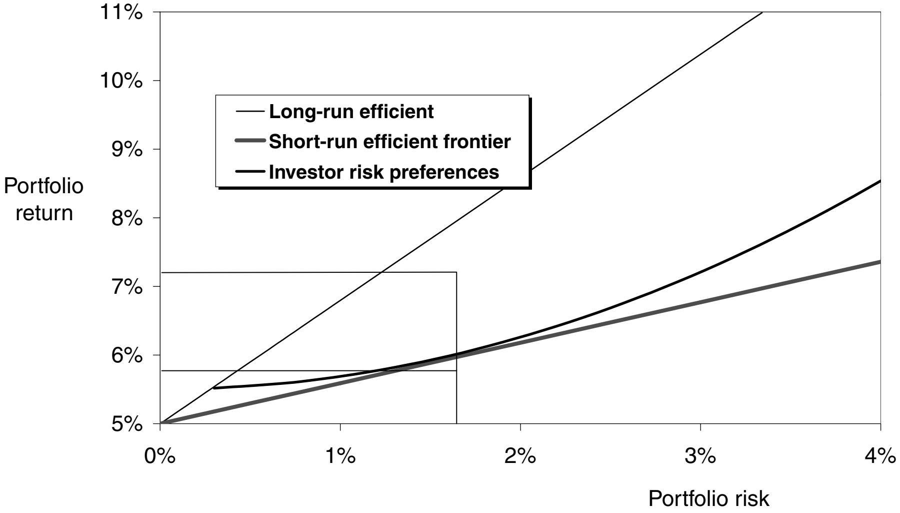


Figure 9 Long-Run vs. Short-Run Time Horizons.

is, a \$10 million portfolio might be found to have a value at risk on a specific day of \$1.1 million with 95% confidence. This means the value of the portfolio would be expected to fall more than this only 5% of the time. Changing the composition of the portfolio allows a value-at-risk efficient frontier to be traced out.

Duarte also proposes a generalized approach to asset allocation that includes mean semi-variance, mean absolute deviation, MV, and value-at-risk as special cases. While cleverly broad, Duarte's method relies on simulation techniques and is computationally burdensome. In addition, because simulation approaches do not have explicit analytical solutions, the technique loses some of the precision of MV analysis. For example, one can examine solution sensitivities from the second-order conditions of MV problems, but this is not so easy with the simulation. It remains to be seen whether simulation approaches receive widespread acceptance for solving portfolio problems. Nonetheless, simulation techniques offer tremendous potential for future applications.

12. COMBINING MEAN-VARIANCE ANALYSIS WITH OTHER TECHNIQUES—CONSTRUCTING OPTIMAL HEDGE FUND PORTFOLIOS

Because of the difficulty of identifying new assets that shift the efficient frontier upward, some investment managers have taken to creating their own. These assets are often "bottoms-up" constructions that in aggregate possess the desired return, risk, and correlation characteristics

In the case of funds of hedge funds, the current industry practice is simply to assemble judgmentally a montage consisting of well-known managers with established track records. Although some effort is expended towards obtaining diversification, a quantitative approach is seldom employed. As a consequence, most portfolios of hedge funds exhibit positive correlation with other asset classes such as stocks. This became obvious in 1998 when many hedge fund portfolios produced negative performance at the same time equity prices fell following the Russian debt default.

Lamm and Ghaleb-Harter (2000a) have proposed a methodology for constructing portfolios of hedge funds that have zero correlation with other asset classes. Their approach brings more rigor to the design of hedge fund portfolios by using an MV optimization format under the constraint that aggregate correlation with traditional assets is zero.

The steps in this process are as follows. First, hedge fund returns are decomposed into two components using a reformulation of Sharpe's (1992) style analysis. Specifically:

$$h_j = \alpha_j + \Sigma\beta_{jk}r_k \qquad j = 1, \ldots, J;\ k = 1, \ldots, K \tag{15}$$

where h_j = the return for the jth hedge fund in a portfolio of J funds
α_j = the skill performance of the fund
β_{jk} = the exposure of the fund to the kth "traditional" asset class
r_k = the traditional asset return

Time subscripts, which apply to h_j and r_k, are dropped.

Equation (15) can be written more simply as:

$$\mathbf{h} = \mathbf{a} + \mathbf{Br} \tag{16}$$

where $\mathbf{h} = [h_1\ h_2\ \ldots\ h_J]^{\mathrm{T}}$
$\mathbf{a} = [\alpha_1\ \alpha_2\ \ldots\ \alpha_J]^{\mathrm{T}}$
$\mathbf{r} = [r_1\ r_2\ \ldots\ r_k]^{\mathrm{T}}$

and the asset exposure matrix is:

$$\mathbf{B} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1K} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{J1} & \beta_{J2} & \cdots & \beta_{JK} \end{bmatrix} \tag{17}$$

The return for the hedge fund portfolio is:

$$h = \mathbf{w}^{\mathrm{T}}\mathbf{h} = \mathbf{w}^{\mathrm{T}}(\mathbf{a} + \mathbf{Br}) \tag{18}$$

where $\mathbf{w} = [w_1\ w_2\ \ldots\ w_J]^{\mathrm{T}}$ are the portfolio weights assigned to each hedge fund.

The variance of the hedge fund portfolio is:

$$\sigma = \mathbf{w}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{R}\mathbf{S}\mathbf{w} \tag{19}$$

where $\mathbf{S}$ = the risk matrix with diagonal standard deviations $s_1\ s_2\ \ldots\ s_J$
$\mathbf{R}$ = the $J \times J$ hedge fund correlation matrix.

To select the optimum MV hedge fund portfolio that produces the highest return with a target zero beta exposure to other assets, the problem is to:

$$\max \mathrm{U}(\mathbf{w}) = (h - \lambda\sigma^2) \quad \text{subject to} \tag{20}$$

$$h = \mathbf{w}^{\mathrm{T}}(\mathbf{a} + \mathbf{B}\mathbf{r}) \tag{21}$$

$$\sigma = \mathbf{w}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{R}\mathbf{S}\mathbf{w} \tag{22}$$

$$\text{colsum } \mathbf{w}^{\mathrm{T}}\mathbf{B} = \mathbf{0} \tag{23}$$

$$\text{colsum } \mathbf{w} = 1 \tag{24}$$

$$w_j \geq 0 \tag{25}$$

$$w_j \leq 1/n \tag{26}$$

where $1/n$ = the maximum weight allowed for any hedge fund
λ = the risk aversion parameter.

This is the same as the standard MV model except the portfolio return takes a more complex form (21) and an additional constraint (23) is added that forces net exposure to traditional assets to zero. Also, constraint (26) forces sufficient diversification in the number of hedge funds.

Obtaining a feasible solution for this problem requires a diverse range of beta exposures across the set of hedge funds considered for inclusion. For example, if all equity betas for admissible hedge funds are greater than zero, a feasible solution is impossible to attain. Fortunately, this problem is alleviated by the availability of hundreds of hedge funds in which to invest.

A significant amount of additional information is required to solve this problem compared to that needed for most MV optimizations. Not only are individual hedge fund returns, risk, and correlations necessary, but returns must also be decomposed into skill and style components. This requires a series of regressions that link each hedge fund's return to those of traditional assets. If 200 hedge funds are candidates, then the same number of regressions is necessary to estimate the α and β parameters.

Table 7 presents optimum hedge fund portfolios that maximize returns for given risk levels while constraining beta exposure to traditional assets to zero. The optimization results are shown for (1) no constraints; (2) restricting net portfolio beta exposure to traditional assets to zero; (3) restricting the weight on any hedge fund to a maximum of 10% to assure adequate diversification; and (4) imposing the maximum weight constraint and the requirement that net beta exposure equals zero. Alternatives are derived for hedge fund portfolios with 3% and 4% risk.

Adding more constraints shifts the hedge fund efficient frontier backward as one progressively tightens restrictions. For example, with total portfolio risk set at 3%, hedge fund portfolio returns fall from 27.4% to 24.4% when zero beta exposure to traditional assets is required. Returns decline to 23.0% when hedge fund weights are restricted to no more than 10%. And returns decrease to 20.4% when net beta exposure is forced to zero and hedge fund weights can be no more than 10%.

13. CONCLUSION

This chapter has presented a brief introduction to asset management, focusing on primary applications. The basic analytical tool for portfolio analysis has been and remains MV analysis and variants of the technique. Mean-variance analysis is intellectually deep, has an intuitive theoretical foundation, and is mathematically efficient. Virtually all asset-management problems are solved using the approach or modified versions of it.

The MV methodology is not without shortcomings. It must be used cautiously with great care paid to problem formulation. The optimum portfolios it produces are only as good as the forecasts used in their derivation and the asset universe from which the solutions are derived. Mean-variance analysis without a concerted effort directed to forecasting and asset selection is not likely to add significant value to the quality of portfolio decisions and may even produce worse results than would otherwise be the case.

While MV analysis still represents the current paradigm, other approaches to portfolio optimization exist and may eventually displace it. Value-at-risk simulation methodologies may ultimately prove more than tangential. Even so, for many practitioners there is still a long way to go before forecasting techniques, asset identification, and time horizon considerations are satisfactorily inte-

TABLE 7 Optimized Portfolios—Hedge Fund Allocations under Various Constraints

		$\sigma = 3\%$				$\sigma = 4\%$			
Fund Strategy	Manager ID	None	$\beta_k = 0$	$w_i = 0.1$	$w_i < 0.1$ $\beta_k = 0$	None	$\beta_k = 0$	$w_i = 0.1$	$w_i < 0.1$ $\beta_k = 0$
Relative value	L/S #1	6%	—	10%	10%	—	—	10%	10%
	CA #1	—	4%	—	—	—	—	—	—
	CA #2	—	—	2%	—	—	—	2%	—
	BH #1	—	5%	—	—	—	3%	—	—
	RO #1	58%	53%	10%	10%	75%	60%	10%	10%
	AG #1	—	—	10%	10%	—	—	10%	10%
	AG #2	—	—	10%	7%	—	—	10%	7%
Event driven	DA #1	3%	4%	8%	9%	3%	2%	10%	10%
	DS #1	15%	—	10%	10%	2%	—	2%	—
	ME #1	—	—	2%	—	—	—	3%	—
Equity hedge	DM #1	3%	3%	7%	8%	—	—	5%	8%
	DM #2	—	3%	4%	10%	—	2%	5%	10%
	DM #3	2%	—	5%	—	3%	—	8%	—
	DM #4	—	—	2%	—	—	—	4%	—
	OP #1	—	—	—	—	—	—	—	3%
	GI #1	—	3%	—	—	—	—	—	2%
	GI #2	—	—	—	—	—	3%	—	—
Global asset allocation	Sy #1	5%	7%	3%	5%	6%	11%	5%	10%
	Sy #2	—	—	2%	—	—	2%	4%	7%
	Sy #3	—	2%	—	4%	—	2%	—	10%
	Sy #4	—	—	3%	—	—	—	5%	—
Short-sellers	SS #1	2%	4%	—	—	—	—	—	—
	SS #2	—	3%	6%	7%	2%	6%	6%	6%
	SS #3	3%		—	—	3%	—	—	—
	Others	3%	9%	7%	7%	6%	8%	6%	6%
Total		100%	100%	100%	100%	100%	100%	100%	100%
	Funds	16	19	24	19	14	15	19	17
Portfolio	Return	27.4	24.4	23.0	20.4	30.6	26.6	23.5	21.2

Source: Lamm and Ghaleb-Harter 2000b.

L/S = long/short; CA = convertible arb; BH = bond hedge; RO = rotational; AG = aggressive; DA = deal arb; DS = distressed; ME = multi-event; DM = domestic; GI = global/international; Sy = systematic; SS = short sellers.

grated with core MV applications. Greater rewards are likely to come from efforts in this direction rather than from devising new methodologies that will also require accurate forecasting and asset selection in order to be successful.

REFERENCES

Agarwal, V., and Naik, N. (2000), "On Taking the Alternative Route: The Risks, Rewards, and Performance Persistence of Hedge Funds," *Journal Of Alternative Investments,* Vol. 3, No. 2, pp. 6–23.

Alexander, C. (1996), "Volatility and Correlation Forecasting," in *The Handbook of Risk Management and Analysis,* C. Alexander, Ed., John Wiley & Sons, New York, pp. 233–260.

Beckers, S. (1996), "A Survey of Risk Measurement Theory and Practice," in *The Handbook of Risk Management and Analysis,* Carol Alexander, Ed., John Wiley & Sons, New York, pp. 171–192

Brinson, G. P., Randolph Hood, L., and Beebower, G. L. (1986), "Determinants of Portfolio Performance," *Financial Analysts Journal,* Vol. 42, No. 4, pp. 39–44.

Brinson, G. P., Singer, B. D., and Beebower, G. L. (1991), "Determinants of Portfolio Performance II: An Update," *Financial Analysts Journal,* Vol. 47, No. 3, pp. 40–48.

Chopra, V., and Ziembra, W. (1993), "The Effect of Errors in Means, Variances, and Covariances on Portfolio Choices," *Journal of Portfolio Management,* Vol. 20, No. 3, pp. 51–58.

Connor, G. (1996), "The Three Types of Factor Models: A Comparison of Their Explanatory Power," *Financial Analysts Journal,* Vol. 52, No. 3, pp. 42–46.

Duarte, A. M. (1999), "Fast Computation of Efficient Portfolios," *Journal of Risk,* Vol. 1, No. 1, pp. 71–94.

Edwards, F. R., and Lieu, J. (1999), "Hedge Funds versus Managed Futures as Asset Classes," *Journal of Derivatives,* Vol. 6, No. 4, pp. 45–64.

Fung, W., and Hseih, D. A. (1997), "Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds," *Review of Financial Studies,* Vol. 10, No. 2, pp. 275–302.

Goldman Sachs & Co. and Financial Risk Management, Ltd. (1999), "The Hedge Fund Industry and Absolute Return Funds," *Journal of Alternative Investments,* Vol. 1, No. 4, pp. 11–27.

Ibbotson, R., and Kaplan, P. (1999), "Does Asset Allocation Explain 40%, 90%, or 100% of Performance?" unpublished paper, Ibbotson Associates, April.

Lamm, R. M. (1998a), "Asset Allocation Implications of Inflation-Protection Securities: Adding Real Class to Portfolios," *Journal of Portfolio Management,* Vol. 24, No. 4, pp. 93–100.

Lamm, R. M. (1998b), "Inflation-Protected and Insurance Linked Securities," *Treasurer,* February, pp. 16–19.

Lamm, R. M. (1999a), "Portfolios of Alternative Assets: Why Not 100% Hedge Funds?" *Journal of Investing,* Vol. 8, No. 4, pp. 87–97.

Lamm, R. M. (1999b), "The Exotica Portfolio: New Financial Instruments Make Bonds Obsolete," in *Insurance and Weather Derivatives: From Exotic Options to Exotic Underlyings,* Helyette Geman, Ed., Financial Engineering, London, pp. 85–99.

Lamm, R. M. (2000), "Economic Foundations and Risk Analysis in Investment Management," *Business Economics,* Vol. 35, No. 1, pp. 26–32.

Lamm, R. M., and Ghaleb-Harter, T. E. (2000a), *Optimal Hedge Fund Portfolios,* Deutsche Asset Management research monograph, February 8.

Lamm, R. M., and Ghaleb-Harter, T. E. (2000b), *Hedge Funds as an Asset Class: An Update on Performance and Attributes,* Deutsche Asset Management research monograph, March 6.

Lamm, R. M., and Ghaleb-Harter, T. E. (2000c), *Do Hedge Funds Belong in Taxable Portfolios?* Deutsche Asset Management research monograph, August 30.

Lummer, S. L., Riepe, M. W., and Siegel, L. B. (1994), "Taming Your Optimizer: A Guide through the Pitfalls of Mean-Variance Optimization," in *Global Asset Allocation: Techniques for Optimizing Portfolio Management,* J. Lederman and R. Klein, Eds., John Wiley & Sons, New York, pp. 7–25.

Markowitz, H. M. (2000), *Mean-Variance Analysis in Portfolio Choice and Capital Markets,* Frank J. Fabozzi Associates, New Hope, PA.

Purcell, D., and Crowley, P. (1999), "The Reality of Hedge Funds," *Journal of Investing,* Vol. 8, No. 3, Fall, 26–44.

Schneeweis, T., and Spurgin, R. (1998), "Multi-Factor Analysis of Hedge Funds, Managed Futures, and Mutual Funds," *Journal of Alternative Investments,* Vol. 3, No. 4, Winter, pp. 1–24.

Schneeweis, T., and Spurgin, R. (1999), "Alternative Investments in the Institutional Portfolio," in *The Handbook of Alternative Investment Strategies,* T. Schneeweis and J. Pescatore, Eds., Institutional Investor, New York, pp. 205–214.

Sharpe, W. (1992), "Asset Allocation: Management Style and Performance Measurement," *Journal of Portfolio Management,* Vol. 18, No. 2, pp. 7–19.

Swensen, D. F. (2000), *Pioneering Portfolio Management: An Unconventional Approach to Institutional Investment,* Simon & Schuster, New York.

Tremont Partners, Inc. and TASS Investment Research (1999), "The Case for Hedge Funds," *Journal of Alternative Investments,* Vol. 2, No, 3, pp. 63–82.

Yago, G., Ramesh, L., and Hochman, N. (1999), "Hedge Funds: Structure and Performance," *Journal of Alternative Investments,* Vol. 2, No. 2, pp. 43–56.

CHAPTER 29

Industrial Engineering Applications in Retailing

Richard A. Feinberg
Purdue University

Tim Christiansen
Montana State University

1. **HOW BOB GOT HIS BOOK** **773**
2. **INTRODUCTION TO RETAIL LOGISTICS** **773**
   - 2.1. The Retail Supply Chain 773
   - 2.2. Strategic Advantages through Retail Supply Chain Management 774
   - 2.3. What Is Supply Chain Management? 775
   - 2.4. A Short History of Supply Chain Management 776
   - 2.5. The Current State of Supply Chain Management 776
3. **RETAIL SUPPLY CHAIN COMPONENTS** **776**
   - 3.1. Product Selection and Sourcing 777
   - 3.2. Inbound Transportation 777
   - 3.3. Processing of Retail Goods 777
   - 3.4. Warehouse Management Technologies 777
   - 3.5. Distribution 777
   - 3.6. Outbound Transportation 777
   - 3.7. Outsourcing 778
   - 3.8. Storage 778
     - 3.9.1. Inventory Management 779
     - 3.9.2. Store Operations 779
     - 3.9.3. Customer Support Logistics 779
4. **STRATEGIC OBJECTIVES FOR RETAIL SUPPLY CHAIN MANAGEMENT** **779**
   - 4.1. Improved Forecasting Ability 779
   - 4.2. Faster and More Accurate Replenishment 780
   - 4.3. Flexible Channel Capability 781
5. **THE EMERGING PARADIGM FOR RETAIL SUPPLY CHAINS** **781**
   - 5.1. Relationships/Alliances/Partnerships 781
   - 5.2. Integrated Forecasting, Planning, and Execution 781
   - 5.3. Statistical Techniques 782
   - 5.4. The Role of Senior Management 782
   - 5.5. Information Technology 782
6. **RETAIL OPPORTUNITIES ON THE HORIZON** **782**
   - 6.1. The Global Marketplace 782
   - 6.2. E-commerce: The Virtual Retailer 782
     - 6.2.1. Scope of E-commerce 783
     - 6.2.2. The Internet Mindset 784
   - 6.3. One to One Marketing 784
   - 6.4. The Product Comes Back: Reverse Logistics 784
7. **THE FUTURE FOR SUPPLY CHAIN MANAGEMENT** **784**
   - 7.1. Conclusions 785

REFERENCES **785**

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. HOW BOB GOT HIS BOOK

Bob decides that he needs a copy of Gordon MacKenzie's *Orbiting the Giant Hairball,* published in 1996 by Penguin Books. He gets in his car and drives to the local book superstore. After unsuccessfully searching in the superstore for the book, then someone to ask about the book, Bob gets a Starbucks and goes to the smaller chain bookstore in the nearby mall. As it turns out, this store does not have the book either, but the employee says he thinks they can get it for him—but cannot tell him for sure when. Frustrated with both stores, yet very satisfied with the Starbucks, he drives home.

At home, Bob decides that he will check Amazon.com. Within two clicks, he finds that *Orbiting* can be sent within 24 hours and at 30% off the price at the superstore. Bob also sees that it has a five-star rating and is one of the most popular purchases at Nike and Ernst & Young (not bad company, he thinks). Finally, he reads a review from a reader that says the following, "Here is the passage that was on his funeral leaflet (it seems that Mr. McKenzie has recently died): 'You have a masterpiece inside you, too, you know. One unlike any that has ever been created, or ever will be. And remember: If you go to your grave without painting your masterpiece, it will not get painted. No one else can paint it. Only you.' " Taken by this statement, Bob decides the book may be better than he heard and orders five copies (one to keep and four to give away). He clicks and within 30 minutes receives an e-mail notification that the order was received. The next day he receives an e-mail notification that the order was shipped. In three days, he receives the package from Amazon. That night he reads the book, finds it to be a great book on leadership, and orders 20 more from Amazon to give out as Christmas gifts. Thinking about the experience, Bob concludes that there may be no reason to ever go to a book store again (except for the Starbucks).

Superstore bookstores stock many single copies of books that sell infrequently and multiple copies of books that sell quickly. They have a sophisticated sales/inventory system so books can be replaced rather quickly from regional distribution centers. Each of the superstores has 130,000 different books, about 10 checkout lines, and a staff of about 20 working in the stores at any one time. When Bob went to purchase *Orbiting* and it was not there, there was no system or person who was able to tell Bob if they could get it to him in a reasonable amount of time. It seems that this superstore doesn't care that Bob did not get his book. These stores are profitable carrying a large inventory with many popular books. The lifetime value of Bob's business does not seem to be important to them.

When Bob goes to Amazon.com, he is welcomed as a regular client and asked if he wants to see some recommendations. (Most of the time he clicks "yes" and purchases a book or two he had not heard of, maybe a new book from an author he likes, and maybe a CD that he did not know about.) This time he simply typed in the name of the book he wanted and clicked on a button at the bottom of the page that tells Amazon.com how Bob wants to pay for the book and where to ship it based upon his previous orders. Amazon then assigned the order to one of its seven U.S. distribution centers (five have opened in 1999—Amazon has 3 million square feet of floor space) that had the book. A red light on a shelf where the book was stored automatically goes off. Workers move from bulb to bulb, retrieving the items ordered then pressing a button to turn off the light. Computers determine which rows the workers go to. Bob's order is placed in a moving crate that contains any number of orders. The crate moves along at 2.9 ft per sec through 10 miles of conveyor. Machines and workers (all of whom get Amazon stock options by the way) scan the bar codes on the items at least 15 times. At the end of the trail, the bar-coded books get matched with the orders. Bob's book goes down a three-foot chute to a box with a bar code that identifies the order. The box is packed, weighed, and labeled before leaving the warehouse in a truck. One of Amazon's newest facilities, in McDonough, Georgia, can ship as many as 200,000 boxes a day. One to seven days later, Bob is another one of over 13 million customers satisfied with Amazon.

The moral of the story—Bob really didn't care if the companies cared or what they had to go through to get the book to him. HE JUST WANTED THE BOOK.

2. INTRODUCTION TO RETAIL LOGISTICS

If your business has all the time in the world, no sales or profit pressures, and a strong competitive advantage in the marketplace, then you do not have to be concerned about the efficiency and effectiveness of your supply chain management. This is not the world of retailing, where margins are thin and competition plentiful (and aggressive).

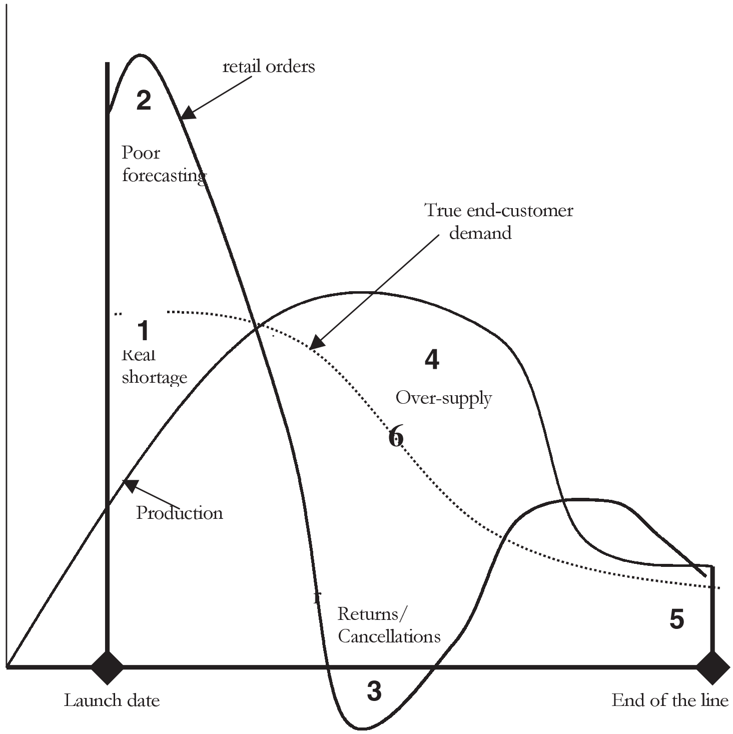
2.1. The Retail Supply Chain

By and large, consumers don't know (or really care) how the product made it to the retail shelf (or to their door). They do care that retailers have the right products, in the right sizes, in the right colors, at the right prices, where and when they want it. The business challenge for retailers is to manage the thousands of paths that exist between raw material to putting the product on the shelf, from the small one-store operation to the 2600+ store chain Wal-Mart.

The retailer's role is to manage these complex relationships and make it possible for the consumer to make a choice easily among the many thousands of vendors and manufacturers creating products that consumers would want. By reducing the complexity of finding and acquiring the right merchandise, retailers help deliver something of value to the consumer. The goal of the retail supply chain should be to minimize expense and time between the manufacturer and the consumer so that consumers find what they want at a competitive price. Customer relationships are only developed by designing product paths that deliver what the customer wants, when they want it, at the price they want it. This is what shareholders pay the executives to do. Executives who do not marshal the supply chain well will have a business that will underperform (see Figure 1).

2.2. Strategic Advantages through Retail Supply Chain Management

Today an executive whose goal is not lowering costs is not shepherding the resources of the company well. Procurement, distribution, and storing accounts for about 55% of all costs. By contrast, labor



1. Production cannot meet initial projected demand, resulting in real shortages. Retailers frustrated because they cannot get the merchandise they want. Consumers dissatisfied because they cannot find what they want.
2. Retailers overorder in an attempt to meet demand and stock their shelves. Retailers are angry at manufacturer and lose confidence.
3. As supply catches up with demand, orders are cancelled or returned. Retailers lose money and lose customers.
4. Financial and production planning are not aligned with real demand; therefore production continues. Overproduction means manufacturers lose money.
5. As demand declines, all parties attempt to drain inventory to prevent writedown
6. Supply and demand match closely. Everyone maximizes profit.

Figure 1 Supply–Demand Misalignment. (From J. Gattorna, Ed., *Strategic Supply Chain Alignment: Best Practice in Supply Chain Management.* Reprinted by permission of the publisher, Gower Publishing, UK.)

accounts for 6%. Thus, leveraging cost savings on supply chain issues creates a greater return than almost any other single category. A 5% savings in supply chain expense adds almost 3% to net profits (a 5% labor savings adds less than 0.3%). To achieve the bottom-line impact of a 5% supply chain saving, you would have to cut personnel costs by 46%. According to a study by Computer Sciences Corporation and *Consumer Goods Manufacturer Magazine,* almost 75% of executives responding to their survey said that reducing supply chain costs was one of the most significant market forces driving their companies and that integrating systems with their customers (e.g., retailers) was a top information system concern (57%) (see Figure 2).

How could such a large area be almost invisible? Quite simply, lack of interest, knowledge, and the clear priorities of sales and marketing drive it into invisibility. The issues in the supply chain seem almost coincidental to the business. It is clear that many senior executives don't quite understand the nature, scope, and bottom-line impact of their supply chain.

But the issue is not simply cost savings. The real benefits of supply chain management come when the retailer develops relationships with suppliers that add value for customers. Wal-Mart is the number one retailer in the world—$160 billion in sales in 1999. Its advantage in the marketplace is based on its ability to give customers what they want at the lowest price every day. The reason for its success is not really its outstanding customer service (although customer service doesn't hurt). The reason is that it manages it supply chain better than any other retailer does. Wal-Mart dictates to vendors the manner and costs by which goods will be delivered to its distribution centers and stores. Wal-Mart's ability to track and know what consumers buy, and its control over the process, result in over 50% of all Wal-Mart products on the shelves and out the door before it has to pay for them. Its efforts to increase this to 100% of all products sold will lead Wal-Mart's dominance into the 21st century. Wal-Mart's ability to manage its supply chain allows it to have the lowest cost structure in the industry, allowing lower prices and maximizing profitability.

2.3. What Is Supply Chain Management?

Supply chain management goes beyond mere transportation, logistics, warehousing, and distribution. It is a relationship among manufacturers, suppliers, vendors, retailers, and customers working together to provide a product that the customer wants at a price that the customer is willing to pay. The "extended enterprise" makes optimal use of shared resources to achieve greater operating efficiency than any could achieve alone. The resulting product is the highest quality, lowest cost, fastest delivery, and maximal customer satisfaction.

The best retailers are focusing on the complete supply chain, developing and maximizing whatever partnerships can exist to get what customers want, before they know they want it (Kuglin 1998). The new terms going into the 2000s are *quick response, category management, continuous replenishment,*

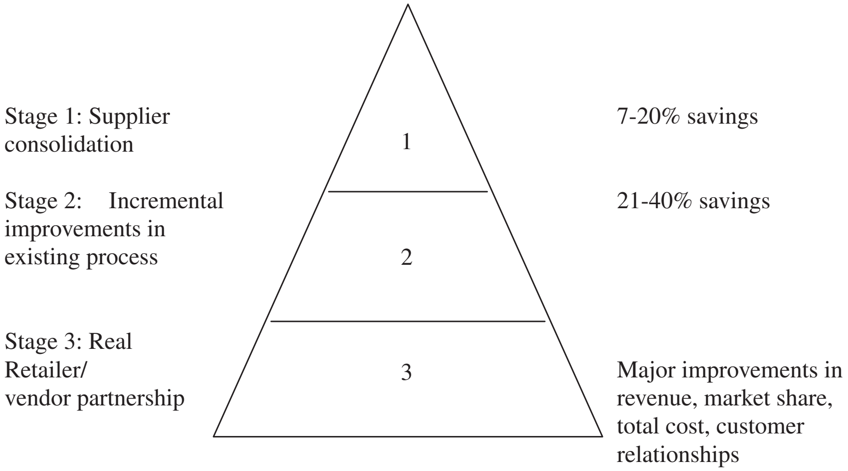

Stage 1: Supplier consolidation
1
7-20% savings
Stage 2: Incremental improvements in existing process
2
21-40% savings
Stage 3: Real Retailer/ vendor partnership
3
Major improvements in revenue, market share, total cost, customer relationships

Figure 2 Supply Management Benefits. (Adapted from *The Executive's Guide to Supply Chain Management Strategies,* Copyright © 1998 David Riggs and Sharon Robbins. Used with permission of the publisher, AMACOM, a division of American Management Association International, New York, NY. All rights reserved. http://www.amacombooks.com)

supply chain partnerships, efficient consumer response, flow-through merchandising (*cross-docking*), and *enterprise resource planning.*

2.4. A Short History of Supply Chain Management

In the beginning, it was sufficient for the owner/proprietor simply to order the products he or she thought the customer would want, and these would be delivered to the store. As business got more complex (and larger), the need arose for a division of labor in which a separate expert department/function was born for the major areas of the business. One part of the organization and set of people was empowered to order, track, plan, negotiate, and develop contacts. The greater the size of the retailer, the greater the number of buying functions—accessories, socks, underwear, coats sportswear, petites, men's wear, furnishings. Each department worthy of its name had an "expert" buyer and assistant buyer. Each buyer reported to a divisional merchandise manager (DMM). Each DMM reported to a general merchandise manager (GMM), and the GMM reported to a senior vice president.

Small retailers have direct store deliveries from the manufacturer or middleman. Larger retailers have centralization of storing, transporting, and distributing from regional distribution facilities (distribution). Technology development over the past 15 years has focused on more efficient and faster ways to get "stuff" into the regional distribution facilities and out to the stores or end user. Success was measured in the number of days or hours it took merchandise to hit the distribution center from the manufacturer or supplier and go out the distribution door. The end result of these developments is the situation today (Fernie 1990; Riggs and Robbins 1998):

- Many small transactions and high transaction costs
- Multiple suppliers delivering similar goods with the same specification but little price competition
- Very low levels of leveraging purchasing or achieving economies of scale
- Too many experts and executives focusing time and effort on very narrow areas that frequently have nothing to do with the core competency of the business, essentially wasting time that could be otherwise be devoted to improving the business's competitive advantage
- Little time for strategy because the experts are involved in daily transactions

2.5. The Current State of Supply Chain Management

We have seen what was. Now here is what some retailers have achieved. Expected demand for a product is forecasted (forecasts can be updated as changes in demand reflect change is needed). Stock levels are constantly monitored. The frequency of production and shipment is based on the difference between demand and on-hand goods. When stock falls below some number, the product is ordered by electronic data interchange (EDI). There is immediate electronic acknowledgment with a promised delivery date and mode of shipping. The mode will be automatically selected by comparing the current level of inventory and demand. When the product is shipped, it is bar coded and a packing skip is EDIed to the company. When the products arrive, bar codes are matched to the order. Any discrepancies are handled electronically. The product is cross-shipped, inventoried, picked, and entered into the purchaser's inventory electronically and is immediately available. Store shipment goes out according to store inventory levels and cost considerations.

There are a number of state-of-the-art processes and technologies at work here. Forecasting and monitoring is used to minimize inventory levels while maintaining very high levels of in-stock positions for the consumer. Money is saved by reducing the inventory at each distribution center. Inventory can be assessed at any of the distribution centers. Almost all handling and clerical labor is eliminated and greater accuracy is achieved. Lower handling costs per unit translate into savings and efficiency. Finally, if you don't have excessive inventory in the stores, it may be possible to raise the percentage of goods sold prior to the need to pay for them. This reduction of cash to cash cycle time is a major benefit of supply chain management.

3. RETAIL SUPPLY CHAIN COMPONENTS

Retail supply chains are different than other industry models. Many of the components of the supply chain are the same: product sourcing, inbound transportation, processing, location and storage of inventory, outbound transportation, company operations, and information. However, retailers are at the end of the chain, just before the products touch the consumer. As a result, the retailer is at the end of the cumulative efficiencies and deficiencies of all the chain partners. It may be that retail supply chains are just a bit more complex. Imagine the thousands of vendors, each with their own ideas and operations, all moving with a thousand different retailers' set of unique requirements and multiply this by the 90,000+ different stock keeping units (SKUs) in the typical large discount store.

3.1. Product Selection and Sourcing

Retailers must establish competence in selecting products that consumers want. The retailer must discern the needs and wants of its consumers and translate that into category assortments. Sourcing describes the manner in which the retailer forms relationships with manufacturers or vendors in order to deliver the products at the right time. Guess wrong and the retailer suffers a season of lower sales and profitability (e.g., short skirts when long is in). Guess right and the retailer becomes a hero.

Retailers have two broad choices as they strategize sourcing. With branded merchandise, the retailer selects manufacturers who have established some image equity with the consumer. The supply chain is simplified for the retailer, for all it has to do is inform the manufacturer how much and when it wants to get the ‘‘stuff’’ to the store. With private-label merchandise, the retailer must manage the whole process, from design to material to manufacturing to shipping, all the way to selling, and advertising the brand equity. Retailers choose to develop private-label merchandise because of the market advantage that this strategy allows. First, there is added profit margin. Retailers can make greater profit margins, even though private label goods are generally priced lower than manufacturer brands. Second, with private labels, retailers have the ability to develop products that are unique in the marketplace because the consumer can get the private label only at that retailer.

3.2. Inbound Transportation

The issues here revolve around getting the merchandise to the retailer’s warehouse and distribution centers in the quickest time, minimizing the handling. The faster goods can be at the distribution center, the faster the speed to market.

3.3. Processing of Retail Goods

All products require some type of value-added service to make it ready for the shelf. Ten years ago, retailers received the merchandise in centralized warehouses and the warehouses tagged and priced the products before the products left for the stores. Today, retailers are forcing manufacturers to deliver the items already tagged and priced. Indeed, the greater the ability of the manufacturer to create shelf-ready merchandise, the greater the retailer’s ability to develop just-in-time strategies for merchandise replenishment. Manufacturers with the ability to meet country-specific requirements for labeling and presentation will have a better chance of biting off a piece of the global marketplace.

3.4. Warehouse Management Technologies

Warehousing is a critical part of retail success and can add value as a crucial link between supply and demand. The key issues in warehousing are efficient handling, inventory management, product flow, transportation, and delivery.

Let us follow merchandise as it comes into the warehouse. Systems are needed to plan the breakdown of merchandise into manageable orders. Weight, size, and shipping dates need to be scheduled. Dock area management facilitates the loading and unloading of product. There is a need to make certain that products coming in match what was ordered. Discrepancies with merchandise not caught here will cause problems as the merchandise flows through the retail organization. The second significant function of the warehouse system is ‘‘put away’’ and replenishment. Inventory that will not be cross-docked has to be stored and available for retrieval. Merchandise that has been stored will need to be found and shipped out to stores that need it. In modern warehouses, this function is highly computerized, with automated picking tools and belts. In less sophisticated facilities, a lot is still done slowly and by hand. Finally, the warehouse system must pack the items and make certain accurate paperwork and order consolidation occur.

Given the complexity of the warehouse, it is easy to see how systems and technologies can be of considerable benefit. An effective warehouse-management system saves labor costs, improves inventory management by reducing inaccuracies, speeds delivery of merchandise to store or consumer, and enhances cross-docking management.

3.5. Distribution

Distribution is the set of activities involved in storing and transporting goods and services. The goal has been to achieve the fastest throughput of merchandise at a distribution center with minimal cost. These efforts are measured by internal standards with very little involvement by manufacturers/vendors and consumers. The service standards are typically set by pursuing efficiencies, not necessarily making sure that consumers have what they want, when and how they want it.

3.6. Outbound Transportation

Moving the correct quantity and type of merchandise to stores or direct to the consumer can create great economies, flexibility, and control. Inefficient transportation can be disastrous. Systems that do not get the correct merchandise to the proper stores during a selling season or when an advertising campaign is scheduled (e.g., winter goods three weeks after the consumer starts shopping for winter clothing) face very poor sales. The effect is more than simply the store not selling a particular piece of merchandise. Each time the consumer does not find what he or she wants, the customer's chance of shopping at a competitor increases.

The diversity of merchandise and the number of stores create several important challenges for retailers. How do you transport merchandise to reach all the stores efficiently at the lowest cost and time? How do you fill the trucks to maximize space in the trucks? Differences in demand and geographic concerns make empty trucks very expensive.

Every day, merchandise moves into the city from a variety of vendors to a variety of stores. Yet for most stores, full truckloads have not been purchased. Third-party distributors that combine loads from different vendors to different merchants allow economies of scale to develop but nothing like the advantage of a large store that receives and buys full truckloads many times a week. The delivery expense to cost is a significant one for retailers. It cuts into the profit and makes prices for the smaller retailer less competitive.

Getting the goods from point A to point B is neither simple nor pedestrian. Tomorrow's business leaders will see this as part of the seamless experience leading to short lead times and reliable service at the lowest price. It has recently been estimated that the cost of transporting goods is 50% of the total supply chain costs. Significant improvements in the movement of goods have been advanced by improvements in software that aid in planning transportation, vehicle routing and scheduling, delivering tracking and execution, and managing the enterprise.

Cross-docking of merchandise occurs when the delivery of product to the distribution center is put directly into the trucks heading for the stores. For cross-docking to work well, suppliers and retailers must have fully integrated information systems. For example, Federal Express manages the complex task of getting all the component parts to Dell at the same time so that a particular Dell Computer order can be manufactured without Dell having to have inventories on hand.

3.7. Outsourcing

Many retailers make use of third-party logistics in the form of warehouse management, shipment consolidation, information systems, and fleet management. Outsourcing allows a retail management to focus on core competencies of purchasing, merchandising, and selling.

3.8. Storage

Consumers buy merchandise that is available—on the shelf—not merchandise in some backroom storage area. Storage costs money (which must be added to the price that consumers pay). Not every retailer is Wal-Mart, which claims to be able to restock products in any of its 2600+ stores within 24 hours. As a result, Wal-Mart stores can have minimal storage areas and maximal selling areas with shelves that always have the merchandise. Just-in-time, quick response, and vendor-managed inventory hold the promise that less merchandise will have to be stored at the local level because systems will dictate that manufacturers or central storage facilities will fill shelves with product as the product starts to decline. Most retailers must still carry inventory so that they may achieve a responsible level of service for the consumer. As a result, compared to Wal-Mart, they have fewer products to sell, more out-of-stock positions, greater consumer frustration, and greater cost structure because they have more nonproductive space used for storing inventory.

The Internet is increasing the pressure on retailers to have better in-stock positions. The only real sustainable competitive advantage that retailers have right now over the Internet is their ability to get product in consumers' hands immediately. If the consumer cannot find what he or she wants in the stores, and its takes just as long for the store to order the merchandise and get it to the customer, the customer might just take advantage of the Internet. Because information and pricing are more transparent on the Internet, if stores do not have the product, the competitive advantage goes to the Internet. This is assuming that the Internet has the same product available at a lower price. The consumer can use a shopping agent like My Simon (www.mysimon.com) to compare all the prices for a particular product on the Internet and choose the lowest price.

As this chapter was being written (December 1999), Toys "R" Us reported very significant problems in getting Internet orders delivered to consumers. Not only will this impact their bottom line in cancelled orders (all customers who were promised delivery that now could not be met were offered cancellation), but all affected customers received a certificate worth $100 on their next Toys "R" Us purchase. This will cost Toys "R" Us millions of dollars which would have gone directly toward their bottom line. More importantly, we cannot estimate the number of customers who will never go to a Toys "R" Us website again.

Internet retailers understand this, and independent or third-party businesses are offering almost immediate delivery in major metropolitan areas. If this is successful, Internet retailers will be removing one of the only barriers and will be competing head-on with store retailers. This suggests large multibrand/manufacturer warehousing in central urban locations with technology and systems to support it.

Storage is also a major issue for Internet retailers. While Internet retailers have a competitive financial advantage because they do not have to support stores on their balance sheets, they still have to maintain extensive storage and distribution facilities. Their efficiency in building and maintaining and efficiently running these massive facilities will actually determine their success to a greater extent than any other single factor.

3.8.1. Inventory Management

The goal of managing inventory is to have what customers want in the right styles, colors, and prices, while holding costs down. The cost of maintaining inventory is from 25% to 40% of the value of the inventory. Too much inventory and you are not profitable. Not enough inventory and you miss sales and increase the likelihood that consumers will shop in other places. Money tied up in inventory that is not selling takes up money that could be profitably invested in inventory that would be selling. Because of inefficiencies in the supply chain, retailers often have more products in inventory than required by demand.

Vendor-managed inventory (VMI) means that the retailer shifts to the manufacturer the role of counting and replenishing dwindling stocks. Many vendors believe that this is simply an effort by retailers to reduce retailer costs by making the vendor pick up the tab for work the retailer used to do. VMI is most common in large mass merchandisers and in grocery stores. VMI helps reduce stock outs, a benefit to the vendor because when the item is out, consumers may try and switch to a competing brand.

3.8.2. Store Operations

Merchandise is not cost neutral when it arrives in the stores. The expense associated with unloading, unpacking, and storing the merchandise is considerable. Retailers have aggressively tried to minimize the amount of effort required once the merchandise enters the back room. Five years ago, it was common for central distribution facilities to receive the merchandise, unpack it, tag or bar code it, repack it, and ship to stores. Larger retailers have tried to shift many of these functions back up the supply chain to the manufacturer and demand that merchandise be priced, tagged, and packed for individual stores. However, not all suppliers are advanced enough to accomplish these simple tasks. For example, 20% of Saks' 12,000 suppliers are unable to accomplish these back room functions. In about 25% of these cases, the ability of Saks' to process products is hindered.

Retailers also have costs and problems of matching supply to demand across the chain. Adherence to a customer-satisfaction goal may require retailers to ship merchandise from store to store to meet different demands. This reshipping is quite expensive and inefficient. Systems and technologies that better predict what is needed, when, in what size and color, and in what quantity is essential for progress on this issue.

3.8.3 Customer Support Logistics

This refers to the support of customer service operations. Services that require logistical foundations include service parts management (parts shipment and inventories), service management (equipment, technology, scheduling involved in servicing and repairing products), and customer contact management (problem resolution, contact management, training). For example, a large Japanese consumer electronics company believed its products to be of superior quality. As a result, they saw no need to inventory parts in the United States. When product repairs were needed, they had to ship the required parts to the United States. Not only was this expensive, but consumer satisfaction was low with the repair process and the electronics manufacturer. The low satisfaction had an impact on repurchase of the product in that product category and other product categories carrying that brand name.

4. STRATEGIC OBJECTIVES FOR RETAIL SUPPLY CHAIN MANAGEMENT

4.1. Improved Forecasting Ability

Collaborative forecasting and replenishment is the term used to represent the joint partnership between retailers and vendors/manufacturers to exchange information and synchronize the demand and supply. With improved forecasting, the amount of material and inventory that vendors, manufacturers, and retailers have to have in hand is reduced. Most forecasting, if it is done (much more likely in larger retails), is done separately. It is only within the past five years that retailers have discovered the value or implemented any joint forecasting partnerships (e.g., Procter & Gamble and Wal-Mart. P&G maintains an office in Bentonville, Arkansas, where Wal-Mart world headquarters is located, to serve this important partnership).

Supply chain management advances rest firmly on the flow of information. Overwhelming paper flow, separate computer systems and databases, and nonnetworked computer systems are unacceptable in the supply chain era. Linking computers on a single platform of information with fast access to all information needed to make decisions fosters enterprise-wide solutions that make up supply chain management.

The success of the Procter & Gamble/Wal-Mart relationship has been looked to as the model for what will be happening in retailing over the next 10 years. It has also served as the model that P&G will be using with its business partners to grow and prosper. Since supply chain management procedures were implemented, market share for P&G went from 24.5% in 1992 to 28% in 1997, while net margin increased from 6.4% to 9.5% (Drayer 1999).

Typically it is the retailer that develops a forecast and hands it to the vendor/manufacturer. Trusting mutually beneficial relationships are not the operational reality between retailers and vendors/manufacturers (see Figure 3). Even greater efficiency will be achieved when multiple retailers in a geographic region combine their information seamlessly with multiple vendors/manufacturers to allow the most efficient delivery and manufacturing of product. Why should Procter & Gamble plan and ship with Wal-Mart, then do it again with Target, with K-Mart, and with the hundreds of smaller retailers in the area, when if they did it together, the costs would be lower for all?

4.2. Faster and More Accurate Replenishment

Direct product replenishment offers to increase the efficiency and effectiveness of retail operations. According to a Coopers & Lybrand study (1996), direct replenishment improves retail operations by allowing:

- Integration of suppliers with mission and function of store
- More reliable operations
- Synchronized production with need
- Cross-docking and associated savings and efficiencies
- Continuous replenishment and fewer out-of-stock positions
- Automated store ordering and more attention and money available for more important areas of the business

Retailers would need fewer people to be experts in a product area. Fewer mistakes would be made. Markdowns (need to sell inventory that did not or would not sell) would be minimized making margin enhancement possible. Goods would reach the consumer faster. Less inventory storage would be needed. There would be lower distribution expenses since much of the inventory would be stored at the manufacturer until needed.

Since inventory is a retailer's number one asset, making that asset more productive feeds the bottom line directly. Continuous and time-minimal replenishment schemes are easier with basic everyday items that move through the store quickly. Grocery stores have widely implemented electronic data system techniques to maximize the efficient replenishment of fast-moving merchandise.

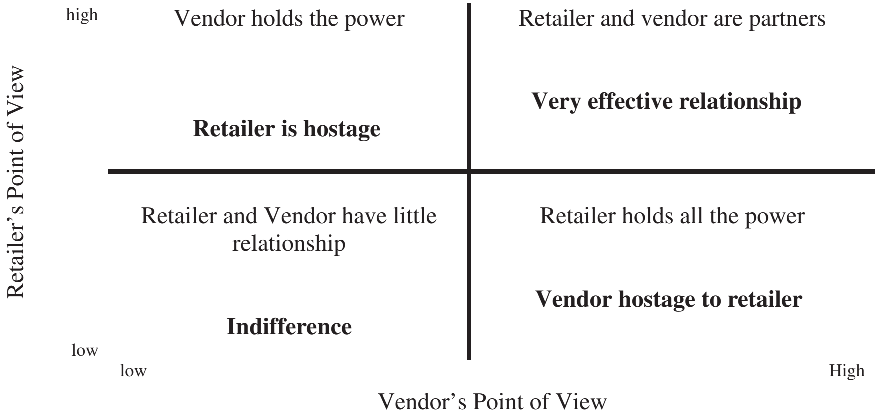

high
Vendor holds the power
Retailer is hostage
Retailer and vendor are partners
Very effective relationship
Retailer's Point of View
Retailer and Vendor have little relationship
Indifference
Retailer holds all the power
Vendor hostage to retailer
low
low
High
Vendor's Point of View

Figure 3 Retailer/Vendor Affairs in the Retail Supply Chain.

Other retailers are just beginning to learn how to use point-of-sale information to minimize quantities of merchandise stored (needed to buffer poor forecasting and slow replenishment). Ernst & Young estimates that savings in inventory management would give the economy a $35 billion boost.

4.3. Flexible Channel Capability

The emerging e-commerce pressures are forcing retailers to build supply chain flexibility. Getting merchandise to the consumer outside of the store chain while maintaining the stores (frequently called a clicks-and-mortar strategy) forces retailers to understand how to deliver items directly to the consumer as well as to the consumer from the stores. Federated Department Stores (parent company of such retailers as Macy's and Bloomingdales) realized the different nature of e-commerce and purchased one of the better mail-order fulfillment houses (Fingerhut) to build the expertise for their emerging e-commerce initiatives. Retailers who have the flexibility to provide multiple access points (stores, catalog, Internet) will be better positioned for success than solely store-based or Internet-based retailers. Petsmart will deliver Internet-ordered goods from their local stores. They have an Internet distribution system built into their already vast network of stores.

5. THE EMERGING PARADIGM FOR RETAIL SUPPLY CHAINS

Up to this point, retail supply chains have largely generated greater profit margins through driving out inefficiencies. For example, retailers have pushed the function of maintaining an inventory back onto the vendor or middleman. Vendors, in turn, have required lower prices and faster deliveries from their suppliers. This type of approach has been effective, but for many supply chains there is precious little fat still left to cut out of the system. To reach the next level of efficiency, productivity, and profitability, a new perspective on what retail competition means will have to be adopted.

5.1. Relationships/Alliances/Partnerships

Perhaps the factor that will have the greatest impact on supply chain management issues is not a technology or a system but a way of thinking. The gap between winning and losing retail organizations in the last decade was defined by their ability to manage the supply chain better than the competition: squeaking out margins and profits from cost savings and efficiencies in the supply chain. The winners in the 21st century will be those retailers who structure, organize, coordinate, and manage partnerships between manufacturers and stores to meet better, faster, and more closely the needs of the consumer. A survey of U.S. companies by Forrester Research shows that 50% of companies share inventory data with supply chain partners. Only 30% share demand histories and forecasts. It is this latter category that holds the greatest promise for all chain partners . . . and U.S. business has a long way to go.

The commoditization of products and services dictates that the real competition will come from organizations that will compete as configurations of partnerships. These partnerships allow greater probabilities of stores having the right item, at the right price, in the right size, in the right color, JUST before the customer wants it. Moreover, supply chain partnerships mean that retailers can spend less of their time and effort in accomplishing this portion of their operation, freeing time, energy, people, and financial resources for developing and maintaining customer relationships. According to Lewis (1995), these partnerships allow:

- Ongoing cost reductions
- Quality improvements
- Faster design cycle times
- Increased operating flexibility
- More value to the customer's customer
- More powerful competitive strategies

A partnership starts with a discussion based on “You are important to us and we are important to you. Now how can we do this better so we can both make more money.” To be successful in this new paradigm will require skills that have not been well developed as yet.

5.2. Integrated Forecasting, Planning, and Execution

Multiple forecasts for the same line of business within the organization are common (if any planning is even done). The gap between what is planned and what actually happens represents lost profits and lost opportunities. The new paradigm for retail supply chain management begins with an accurate view of customer demand. That demand drives planning for inventory, production, and distribution within some understood error parameters. Consumers will never be completely predictable. At the same time, prediction is bounded by limitations in our statistical and modeling sciences. We can predict only as well as our tools allow us. Computer advances will allow easier and affordable management of millions of data points so that demand can be measured. Effective demand management represents an untapped and significant opportunity for most if not all retailers. Effective demand modeling allows greater forecast accuracy, increases supply chain effectiveness, reduces costs, and improves service levels, and all reflect greater profit. The result—lower inventories, higher service levels, better product availability, reduced costs, satisfied customers, and more time available to respond to other areas of the organization.

5.3. Statistical Techniques

Understandable, usable, and affordable software to accommodate millions and millions of individual consumer purchases has only recently become available. But that is only part of the equation. It is still unusual to find company executives who understand the capability and use of these sophisticated software packages. Of equal importance is the need to have a uniform database representing the forecasting target. Typically, different data needed to accomplish accurate forecasting resides in isolated silos of databases that cannot talk to each other.

In addition to decisions regarding the supply chain, these tools would:

- Provide the CEO with an important strategic view of the business
- Provide opportunities for cost savings
- Aid in developing long-range strategic plans
- Assist in resource allocation
- Identify inadequate areas in existing management and operations

5.4. The Role of Senior Management

Logistic, warehouse, and other supply chain management efficiency can result from a piecemeal decision-making process. But true partnerships must flow from a strategic alignment with the mission and values of the organization. Wal-Mart's industry-leading supply chain operation did not result from chance but from Sam Walton's understanding of the business and a strategic view that saw that investments in state-of-the-art technology and systems will pay off handsomely in the future. For most companies, the appointment of a supply chain czar is needed to push supply chain management through the organization.

5.5. Information Technology

No advances in supply chain management can occur without a layer of middleware. Information and its collection, management, availability, and use serve as the foundation for all advances in supply chain management. Sustained success of retailers comes from an understanding of how information technology (IT) creates competitive advantage. Prepackaged enterprise-wide IT solutions are available, but they make it difficult to be different. Unique and strategic IT enterprise-wide solutions are difficult to integrate and implement.

The IT system should allow new and better customer–vendor relations, provide new insights into a company's consumers and markets, maximally exploit efficiencies in the supply chain, transform product development and delivery, increase the capability to make real-time decisions, increase planning ability, manage increasingly complex national and global markets, and eliminate traditional and costly processes.

6. RETAILER OPPORTUNITIES ON THE HORIZON

6.1. The Global Marketplace

The globalization of retailing presents a host of problems and opportunities for retailing. The ability of companies to meet cultural needs, transport goods and services across boundaries (sometimes oceans), while controlling costs and maintaining efficiencies is not clear . . . yet. Some companies may have already met this challenge. Dell Computers' direct-to-customer model is a global reality. Since a Dell computer is manufactured only when it is ordered, it does not matter to Dell if the computer is selected by a consumer sitting in Peoria, Illinois, or Kuala Lumpur, Malaysia. The computers are built and the appropriate shipping is arranged from the closest facility. Trained technicians answer technical and support questions over the phone and can be anywhere in the world (questions are also answered from a central website).

6.2. E-commerce: The Virtual Retailer

Fifty years ago, catalog shopping was deemed the beginning of the end of store retailing, in much the same way that e-commerce is proclaimed to spell the beginning of the end for store retailing today. Offering products and services directly to the consumer who has money but little time is nothing new. The e-commerce alternative is simply another way for consumers to get what they want, when they want it. The e-commerce alternative is not a threat to the existence of retailing but is a challenge to supply chain management of the store-based retailer.

The promise of direct to consumer retailing is clear—fast, individualized, convenient delivery of what the customer wants. The reality is less clear. Few if any direct to consumer e-commerce initiatives have proven successful . . . yet. The reason is simple. The Internet is an easy way to sell but a lousy way to distribute. Everything purchased on the Internet has a price to which a shipping charge must be added as well as the time to deliver it. The Internet's ability to provide immediate purchase gratification is limited.

It is interesting to note when we review e-commerce discussions that it is as if home delivery were a new concept discovered by these Internet pioneers. In fact, home delivery has been a burgeoning trend in many industries—furniture, appliances, video, Chinese food, pizza, and now the grocery business. A study by Anderson Consulting predicted that delivery can capture 12% of the U.S. grocery business, about $85 billion. But even in this area of e-commerce there is plenty of need for, and money for, the store-based retailer. And it is still not clear whether grocery e-retailers will be able to make a profit in the near future. Peapod is the most successful example of this business model in groceries. The company started in 1989 in a suburb of Chicago, went online in 1990, and now has over 40,000 products available for delivery in eight metropolitan areas. The company delivers what is ordered at a prearranged time. Revenues for 1999 were estimated to be $70 million. However, Peapod has yet to show a profit.

6.2.1. Scope of E-commerce

Toyota is considering a direct electronic channel where customers submit their specifications, as well as financing information, via the Internet (or 800 number). Toyota would respond within the first 24 hours after researching their supply chain and retail distribution with the closest matches and estimated delivery time.

Cisco, Autodesk, Intel, Dell, and Gateway all allow their customers to access inventory and order information from the Internet. They can place and track orders without any human intervention (hence lower costs).

Anderson Windows has supplied point-of-sale kiosks to its retailers. Consumer and retailers design the windows, get order quotes, and place orders at the same time. The system shows customers what they have designed and links the retailers to the supply chain and distributors.

Levi's stores have a minimal inventory, but with a computer model, they measure the consumer at the key design points of their personal fit jeans. The measurements are transmitted to the production facility and a personal pair of jeans is mailed to the consumer within nine days. Consumers are willing to pay slightly more for the perfect pair of jeans. The customer simply cannot buy jeans that fit as well at any other retailer. The customer can now call Levi's from any place in the world and a custom-fit pair of jeans will be shipped.

Grainger is not the business you might think of when you think of an e-business. Since 1927, Grainger has sold industrial supplies (motors, cleaners, lightbulbs) through a 7-lb thick read catalog and 500 stores. It has made a spectacular transition to the Web. Its advertising says it all: "The Red Book [its catalog] has .com of age." They further say, "Our award winning web site (www.grainger.com) carries over 560,000 products online, 24 hours a day, 7 days a week. Our powerful search engine finds exactly what you want. Fast. And you can take advantage of your company's specific pricing." In 1998, its Internet sales were $13.5 million. In 1999, they were $100 million. Only Cisco, Dell, Amazon, and IBM have a larger sales volume. Grainger builds online customized catalogs for the many businesses that negotiate selling prices. When the customer write his or her own order, there are fewer errors made than with placing the order by calling the 800 number. The site tells customers whether the product is in stock at a local store that will deliver it the same day (or overnight) without the overnight shipping fee. The cost of selling a product is a fraction of what the costs are in other areas of Grainger's business, allowing greater discounts and greater profits. To encourage sales, account managers encourage their accounts to order on the Internet, and the salespeople will still get commission on their accounts, even when products are ordered from the website.

One of the biggest problems that retailers face in Internet retailing is the ability of manufacturers to go directly to the consumer. Why pay for a GE microwave from an appliance store on the Web when you can buy it from GEs website (probably for less money). What is particularly interesting about Grainger is that by selling a wide variety of manufactured products, it gives the customer a side-by-side comparison of all available brands when the customer is searching. The customer can buy all the brands he or she wants conveniently and easily instead of having to go to each manufacturer's website and buy separately.

6.2.2. The Internet Mindset

Yes, the Internet will spur many direct-to-consumer businesses. Yes, the Internet will allow existing retailers to expand their markets and provide a new channel for their existing companies. And yes, the Internet will allow retailers to develop efficiencies of time and money with their vendors and manufacturers. But more importantly, the Internet will force retail leaders to develop an Internet frame of mind that will allow challenges to the way business is being done at all levels in all ways. The advances that will come from this change of mindset cannot be predicted but can be expected.

6.3. One to One Marketing

A company's success at individualized product offerings means greater market penetration, greater market share, greater share of the consumer's wallet, and improved satisfaction as customers get exactly what they want instead of settling for what the retailer has. Mass customization/one to one marketing is the real final frontier for retailers. Responsive manufacturing, IT, and efficient replenishment systems allow Levi's to offer the consumer a pair of jeans made to his or her exact specifications.

Here we have a vision of the ultimate supply chain management. Customers have exactly what they want. A long-term relationship is established so that the customer has no reason to desert to a competitor. Because a customer is getting a unique product made just for him or her, a competitor cannot steal that customer simply by offering price inducement. The jeans are not manufactured until they are needed, and a third party delivers the product. The very issues that supply chain management has been concerned with are reduced and eliminated by one to one marketing.

6.4. The Product Comes Back: Reverse Logistics

If the supply chain is relatively invisible, the reverse supply chain (getting back to the manufacturer products that need repair or replacement) is practically ethereal. However, the costs of getting products back through these reverse channels makes process and management issues extremely important to retailers. According to the National Retail Federation, the average return rate from products bought in specialty stores is 10%, and in department stores, 12%. Catalogs have a return rate three times higher. Early indications are that Internet purchases are twice that. Reverse logistics is maximizing the value from returned products and materials.

Frequently, the products/material put into the reverse supply chain can be refurbished or remanufactured. Xerox reports that recovering and remanufacturing saves over $200 million annually, which can then be passed along to customers in lower prices or shareholders in higher earnings and profits. A product returned for repair can either be replaced (and the product broken down and its parts used in other manufacturing) or repaired. Repaired or returned products that cannot be sold as new can be resold, frequently as refurbished, using parts from other returned products. Unfortunately, retailers and manufacturers have made poor use of the strategic information available in returned products to change and improve design. Mostly the repairs are made or the broken product sits in a warehouse. The information about your product and the company sits useless.

A growing area of concern is how to conscientiously dispose of products or their components. Black & Decker avoided significant landfill costs and made money by selling recyclable commodities. When there are no alternatives except disposal, the product or material needs to be appropriately scrapped. In an era of environmental concern (which will continue to grow and drive marketplace decisions), companies should take proactive action to develop environmentally sound landfill or incineration programs. Not only will this allow favorable consumer reactions, but it may forestall government regulation and control.

7. THE FUTURE FOR SUPPLY CHAIN MANAGEMENT

The logic of supply chain management is compelling. Its benefits are well understood. The trend toward supply chain management is clear. Companies are excelling in parts of the equation.

As retailing moves into the 21st century, more retailers will be adopting the following lessons from supply chain management as we see it today:

- Advances in and use of IT to drive supply chain decisions. Large integrated stock-replenishment systems will control the storage and movement of large numbers of different goods into the stores.
- A restructuring of existing distribution facilities and strategic placement and development of new distribution facilities to reduce inventory and move the inventory more efficiently.
- Greater and more effective adoption of quick response. More frequent delivery of less. Greater use of cross-docking so that merchandise hits the docks of the distribution centers and is immediately loaded into destination trucks. EDI and POS electronics will track what is selling and transmit directly to distribution center to plan shipment.

- Greater diffusion of efficient consumer response, defining a process of close collaboration of retailer and manufacturer/supplier to coordinate activities to achieve maximal efficiency and better service levels for the ultimate customer.

7.1. Conclusions

It is easy to see how retailing will need to improve the front-end selling and merchandising function through stores, catalogs, television, and the Internet. What is not so clear to an outsider is that the real difference to retail success and bottom-line earnings and profits will not be in the store or on the Internet but in the middle. The invisible chain between manufacturer and consumer will be where the most significant competitive advantages will take place.

According to Gattorna (1998), the major improvements in the supply chain, over the next 10 years will include:

- Whole industries will rethink their sales and marketing channels, and these new channels to customers will require newly reengineered processes and technologies, in turn demanding significant changes at all levels.
- Leading companies will recognize the close relationship between customer relationship management and the supply chain; taken together, these two will open up new avenues for shaping trading terms from both supplier and reseller/customer perspectives.
- A much more aggressive search for additional organization design options for supply chains in particular industries will eliminate polarized thinking in terms of insourcing or outsourcing, opening the way to new solutions and combinations along this continuum.
- The best supply chains will have fully integrated enterprise systems and indeed will go beyond transactional information and decision-support systems to knowledge management. In turn, this will affect organization design and lead to improved coordination (rather than ''control'').
- Companies looking to embrace supply chain regimes at the global level will require a much better understanding of country cultures as a necessary ingredient for success.
- Strategic sourcing approaches at the supply end, and mass customization approaches at the consumption end, are likely to be fertile areas for relatively quick large-scale benefits.
- Reverse logistics will loom large on the agenda, with issues such as extending product usage life cycles, product-recovery processes, and bidirectional logistics channels coming to the fore in the search for new competitive dimensions.
- Organizations other than product companies will begin to recognize the huge untapped potential that the application of logistics and supply chain principles to their businesses will release, such as telecommunications, utilities, health care, education, entertainment, and financial services.

On the one hand, supply chain management is not hard. Any business can get anything to stores. Any store or e-retailer can get merchandise to consumers. What is hard is achieving an efficiency and effectiveness that maximizes a retailer's competitive advantage in the marketplace. Until we develop the *Star Trek* process of transportation, where we can dematerialize and materialize objects at command (now that's a one to one deliver system of the future), supply chain issues will be a leading-edge practice for the company and the driver for corporate strategy. Collaborative planning forecasting replenishing (CPFR) will be the strategy of the next decade in retailing. The bottom line is, of course, profits and shareholder value (Anderson and Lee 1999; Quinn 1999). Incremental improvements in the supply chain have a greater impact on profits than incremental improvements in other areas. Technology and the Web will be driving the most innovative changes and applications in retail supply chain management in the next 10 years, and this will drive retail performance and excellence.

REFERENCES

Anderson, D., and Lee, H. (1999), ''Synchronized Supply Chains: The New Frontier,'' Achieving Supply Chain Excellence Through Technology Project, www.ascet.com/ascet, Understanding the New Frontier.

Coopers & Lybrand (1996), European Value Chain Analysis Study, Final Report, ECR Europe, Utrecht.

Drayer, R. (1999), ''Procter & Gamble: A Case Study,'' Achieving Supply Chain Excellence through Technology Project, www.ascet.com/ascet, Creating Shareholder Value.

Fernie, J. (1990), *Retail Distribution Management,* Kogan Page, London.

Fernie, J., and Sparks, L. (1999). *Logistics and Retail Management: Insights into Current Practice and Trends from Leading Experts,* CRC Press, Boca Raton.

Gattorna, J., Ed. (1998), *Strategic Supply Chain Management,* Gower, Hampshire.

Kuglin, F. (1998), Customer-Centered Supply Chain Management: A Link-by-Link Guide, AMACOM, New York.

Lewis, J. (1995), *The Connected Corporation,* Free Press, New York.

Quinn, F. (1999), ''The Payoff Potential in Supply Chain Management,'' Achieving Supply Chain Excellence through Technology Project, www.ascet.com/ascet, Driving Successful Change.

Riggs, D., and Robbins, S. (1998), *The Executive's Guide to Supply Chain Strategies: Building Supplies Chain Thinking into All Business Processes,* AMACOM, New York.

CHAPTER 30

Industrial Engineering Applications in Transportation

Chryssi Malandraki
David Zaret
Juan R. Perez
Chuck Holland
United Parcel Service

1. OVERVIEW **788**
2. INTRODUCTION **788**
3. TRANSPORTATION AND INDUSTRIAL ENGINEERING **788**
  3.1. Transportation as a System 788
4. THE PARAMETERS AND FUNCTIONS ASSOCIATED WITH TRANSPORTATION **789**
5. THE ROLE OF THE INDUSTRIAL ENGINEER IN TRANSPORTATION PLANNING AND TRANSPORTATION OPERATIONS **790**
6. TRANSPORTATION AND THE SUPPLY CHAIN **790**
7. TRANSPORTING GOODS **791**
  7.1. Cost, Time, and Quality Optimization in Transportation 791
  7.2. Integrating Customer Needs in Transportation Planning 792
  7.3. Forecasting in Transportation Planning 792
8. PICKUP AND DELIVERY **793**
  8.1. Pickup-and-Delivery Operations 793
  8.2. Modeling 794
  8.3. Heuristic Construction Algorithms 795
  8.4. A Sequential Route-Construction Heuristic 795
  8.5. A Parallel Route-Construction Heuristic 795
  8.6. A Numerical Example 797
  8.7. Infeasible Problems 798
  8.8. Work Breaks 799
  8.9. Route-Improvement Heuristics 800
  8.10. Preassigned Routes and Preassigned Territories 801
  8.11. Implementation Issues 803
9. LARGE-SCALE TRANSPORTATION NETWORK PLANNING **803**
  9.1. Line-Haul Movement of Packages 803
  9.2. A Network-Optimization Problem 804
  9.3. Modeling 804
    9.3.1. Notation 805
  9.4. Network Design Formulation 806
  9.5. Package-Routing Problem 807
  9.6. Lagrangian Relaxation of the Package-Routing Problem 808
  9.7. Package-Routing Heuristic Algorithm 809
  9.8. Trailer-Assignment Problem 810
  9.9. Trailer-Assignment Formulation 810
  9.10. Lagrangian Relaxation Algorithm for the Trailer-Assignment Problem 811
  9.11. Subgradient Optimization Algorithm 811
  9.12. Extensions of the Network-Design Problem 812

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

10. DRIVER SCHEDULING 812
10.1. Tractor-Trailer-Driver Schedules 812
10.2. Driver-Scheduling Problem 813
10.2.1. Notation 813
10.3. Set-Partitioning Formulation with Side Constraints 813
10.4. Set-Covering Formulation with Soft Constraints 814
10.5. Column-Generation Methodology 814
10.6. Iterative Process for Solving the Driver-Scheduling Problem with Column Generation 815
10.7. Integrated Iterative Process for Solving the Driver-Scheduling Problem 815
10.8. Generation of Schedules 816
10.9. Beyond Algorithms 816
11. QUALITY IN TRANSPORTATION 817
12. TECHNOLOGY 819
12.1. Vehicle Routing 819
12.2. Information Gathering and Shipment Tracking 819
12.3. New Trends: Intelligent Transportation Systems (ITS) 819
REFERENCES 822

1. OVERVIEW

Transportation and distribution play a critical role in the successful planning and implementation of today's supply chains. Although many view the transportation of goods as a non-value-added activity, effective transportation planning and execution will not only enhance a company's productivity but will also increase customer satisfaction and quality.

In this chapter, we will explore the factors that impact the transportation of goods and the different tools and techniques that the industrial engineer can apply in the development of effective transportation networks and systems to reduce or minimize costs, improve cycle time, and reduce service failures. A similar but inherently different aspect of transportation is that of transporting people. Although this chapter is concerned with the transportation of goods, the industrial engineer also plays an important role in designing these types of systems.

Today's logistics activities are concerned with the movement of goods, funds, and information. Information technology has now become an integral component of any transportation system. Technology is being used for scheduling and creating complex delivery and pickup routes and also for providing customers with up-to-the-minute information on the status of their shipments. The industrial engineer will not only aid in the development of efficient delivery routes and schedules, but will also help in the design of state-of-the-art transportation information systems.

2. INTRODUCTION

Transport is the process of transferring or conveying something from one place to another. Transportation is the process of transporting. The transportation of people, goods, funds, and information plays a key role in any economy, and the industrial engineer can play a key role in properly balancing the parameters and constraints that affect the effectiveness and efficiency of transportation systems.

This chapter will cover certain applications of industrial engineering in transportation. The emphasis is placed on the movement of goods. However, the methodologies described in the chapter can be applied to a variety of transportation problems.

Transportation plays a critical role in today's increasingly small world. Economies, businesses, and personal travel are, in one word, global. Information on the status of the items or persons being moved is as crucial as the movement itself. Confirmation of delivery in a timely, electronic form is often as important as on-time, damage-free, value-priced arrival.

The industrial engineer can apply a variety of mathematical and engineering tools and techniques in the planning and management of effective transportation networks and systems in order to reduce or minimize costs, improve cycle time, reduce service failures, and so on. The industrial engineer plays a critical role in the development of efficient delivery routes, schedules, and plans and also helps in the design and implementation of transportation information systems.

3. TRANSPORTATION AND INDUSTRIAL ENGINEERING

3.1 Transportation as a System

Designing a transportation system means, in most cases, the design of multiple integrated systems—systems to move goods or people, systems to move information about goods or people, and systems to move funds associated with goods or people. Within each of these three types of systems there may be multiple subsystems.

Today, in the package delivery business, transportation companies offer several types of services in which the primary distinction is the time it takes to move the package from its point of origin to its final destination. Services range from same-day or next-day delivery to multi-week delivery, utilizing different transportation modes such as airplanes, trucks, railcars, boats, and even bicycles. The packages may all originate at the same location (e.g., shipper) going to the same destination, or each package in a shipment may have a different destination. In either case, the transportation systems must be capable of supporting a variety of service offerings, depending on the customer's needs. The success of such systems is measured by the systems' effectiveness in meeting the promised service guarantees.

In air travel, the service offerings vary from first class to coach travel. Dividing the aircraft into separate travel compartments allows multiple systems to utilize the same asset, route, flight crew, and schedule to offer variations in service. Therefore, the air travel industry must also utilize multiple integrated systems when designing the processes to facilitate the movement of people.

Providing information about the goods or people being moved also involves the implementation and use of multiple integrated systems. Many customers today provide information about shipments to the carrier at the time of shipment or earlier. This information is introduced into multiple systems for a variety of uses. The shipper has information about the shipment: its contents, the carrier, the mode of transportation, the expected date of arrival, the value of the shipment, shipping charges, and so on. The carrier and shipper can, through integrated systems, track the status of the shipment as it moves to its final destination. Upon delivery, the carrier can provide proof of delivery proactively or upon request. The receiver can be prealerted of the upcoming shipment, its contents, and the expected date and time of arrival. For international shipments, information can be sent to customs before the item is moved to the customs site. The movement and timely availability of all this information increase the efficiency of the modern supply chain.

As with the movement of goods, people, or information, the transfer of funds makes use of integrated systems. The electronic billing of the carrier's charges to the shipper is integrated with the system providing information about the shipment itself. The payment and transfer of funds to the carrier can be initiated by the pickup or the delivery or through periodic billing cycles. Payments upon delivery for the contents of the shipment can be handled electronically and triggered by the electronic transmission of the signature at delivery.

In designing a transportation system, it is crucial to facilitate the integration of the multiple systems needed for the movement of the goods, people, information, and funds. Industrial engineers are often the catalysts in facilitating such integration.

4. THE PARAMETERS AND FUNCTIONS ASSOCIATED WITH TRANSPORTATION

There are two basic parameters that affect the design of freight transportation processes: the territory to be covered and the frequency with which the transportation events occur. These parameters are not static. Demand (e.g., number of shipments per week or day) fluctuations make these parameters dynamic in nature. Demand in transportation is often seasonal. The Thanksgiving holiday in the United States is one of the highest demand days for air travel. The Saturday prior to Mother's Day usually presents a substantial increase in package delivery volume compared to other Saturdays. Demand can vary by time of day or day of the week. Traffic volume in a major city varies greatly between Monday morning at 7:30 am and Thursday morning at 3:00 am. The transportation system must be designed to handle these variations in demand effectively.

It is important to note that what drives demand in the private freight transportation sector is cost and quality of service. The design (planning), execution, and measurement of the transportation system have an impact on the cost and quality of the transportation process. The better the costs and quality of service, the greater the demand and therefore, the greater the need to alter the service territory and the frequency of service.

Transportation planning, execution, and measurement are the fundamental functions associated with transportation and are integral to the success of the transportation system.

Planning is needed across all areas of the transportation system. Planning the overall distribution network, regional (e.g., certain geography) planning, and site (e.g., local) planning are crucial. Asset planning, including buildings and facilities, vehicles, aircraft, trailers, and materials-handling equipment, is also required. Demand variations drive decisions for asset quantity and capacity. The use of owned vehicles supplemented by leased vehicles during peak (high-demand) times is an example of the decisions that need to be made during the planning process. The impact of demand on scheduling, vehicle routing, and dispatching drives labor and asset needs. Labor can often be the largest component of cost in the transportation industry. The transportation planning activity is charged with developing plans that, while minimizing costs, meet all service requirements.

The execution of the plan is as important as the planning of the transportation process itself. An excellent plan properly executed results in lower costs and higher quality of service. This in turn drives demand and therefore the need for new plans.

Finally, the proper measurement of the effectiveness of the transportation system, in real time, offers a mechanism by which transportation managers, supply chain specialists, and industrial engineers can constantly reduce costs and improve service. With new technologies such as wireless communication, global positioning systems (GPS), and performance-monitoring technology, measurement systems allow the users to improve transportation processes as needed.

5. THE ROLE OF THE INDUSTRIAL ENGINEER IN TRANSPORTATION PLANNING AND TRANSPORTATION OPERATIONS

The role of the industrial engineer in the transportation industry is primarily to aid the organization in providing a high level of service at a competitive price. The industrial engineer has the skills necessary to assist in many areas that impact the effectiveness of the transportation system:

- *Work measurement and methods analysis:* Labor is a large component of the total cost of service in the transportation industry. The pilot, driver, captain, or engineer literally controls the ‘‘container’’ of goods or people in transit. The design of effective work methods and the development of the appropriate work measurement offer tools that aid management in the control and execution of transportation processes and provide a mechanism by which the performance of the system can be measured. Methods design and work measurement are often used to develop comprehensive work scheduling and vehicle routing systems aimed at reducing costs and meeting all service commitments.
- *Facility design and location:* The determination of the number of facilities required to move materials and finished goods from one point to another, their capacity, and their location is often considered one of the traditional roles of the industrial engineer. The use of single (e.g., decentralized) vs. regional distribution center locations and the determination of the territory served by a local terminal are part of the transportation system design process. In addition, industrial engineers will often aid in the decision process to determine whether a facility should be automated or be built using manual sorting systems.
- *System design:* The integration of the components of the transportation process into a highly efficient system also involves the industrial engineer.
- *Equipment:* Requirements, design and selection of trucks, trailers, containers, aircraft, scanners, communication devices, materials-handling systems, etc. are tasks undertaken by the industrial engineer.
- *Facility layout:* The industrial engineer is often responsible for the design of facility layouts that will offer the most effective arrangement of the physical components and materials-handling equipment inside distribution centers and delivery terminals.
- *Asset utilization and control:* The design of systems and procedures to balance and manage the number of trucks, trailers, containers, aircraft, scanners, communication devices, and materials-handling systems required to facilitate the transportation processes is one more area requiring the attention of the industrial engineer.
- *Measurement systems:* Industrial engineers also participate in the development of effective transportation performance measures, including customer satisfaction, cost, and plan accuracy.

As indicated by this list, the industrial engineer adds great value to the transportation industry. Industrial engineering principles are highly applicable in this complex industry. Industrial engineers continue to serve as key members of the teams responsible for the design and integration of systems to facilitate the movement of goods, people, funds, and information in this increasingly competitive industry.

6. TRANSPORTATION AND THE SUPPLY CHAIN

Supply chain management is a comprehensive concept capturing objectives of functional integration and strategic deployment as a single managerial process. Figure 1 depicts the supply chain structure.

This structure has been in place for decades. Even when the manufacturing activity takes place in minutes, the final delivery of a product may take days, weeks, or months, depending on the efficiency of the supply chain. The operating objectives of a supply chain are to maximize response, minimize variance, minimize inventory, maximize consolidation, maintain high levels of quality, and provide life-cycle support.

Transportation is part of the physical distribution. The physical distribution components include transportation, warehousing, order processing, facility structure, and inventory. The major change in

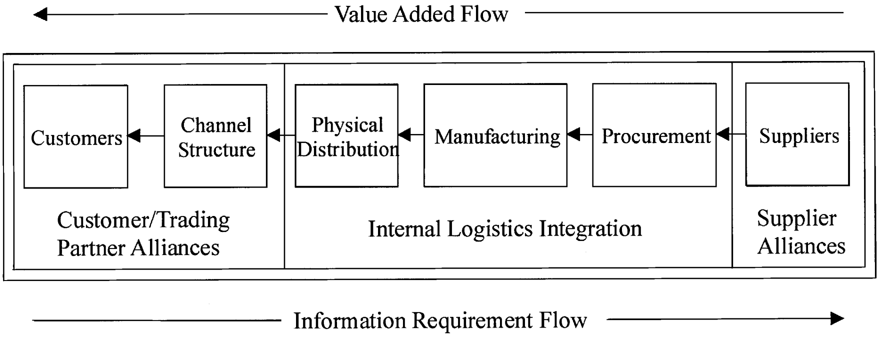

Value Added Flow
Customers
Channel Structure
Physical Distribution
Manufacturing
Procurement
Suppliers
Customer/Trading Partner Alliances
Internal Logistics Integration
Supplier Alliances
Information Requirement Flow

Figure 1 The Supply Chain Structure.

the supply chain in the last decade is information. It is information that allows transportation planners to reduce the costs in the supply chain in today's highly competitive environment. The integrated freight transportation systems that facilitate the movement of goods, information, and funds can target all areas of the supply chain. Every segment of the supply chain has a transportation need.

In an organization, transportation requirements may cover a wide range of territory and frequency characteristics. Decisions are usually made on the basis of cost as long as customer requirements are met. When making decisions that affect the organization's supply chain, it is important not to look at transportation alone or as an independent activity. Instead, transportation should be viewed in the context of the entire supply chain in order to make the most effective decisions (Figure 2).

The remainder of this chapter examines industrial engineering applications in transportation. We concentrate our attention in the transportation of goods. However, the techniques reviewed in the following sections can be applied to a variety of transportation problems. We do not provide a complete survey; instead, we present several representative applications in some detail.

7. TRANSPORTING GOODS

7.1. Cost, Time, and Quality Optimization in Transportation

In the transportation of goods, we often develop models that minimize total cost while maintaining acceptable levels of service. This cost may be a combination of the cost of operating a vehicle (fuel, maintenance, depreciation, etc.), the labor cost of the operator of the vehicle (driver, pilot, etc.), and possibly other fixed costs. In the transportation of small packages or less-than-truckload shipments, sorting cost is also considered, representing the cost of sorting and handling the packages to consolidate shipments. Cost may be considered indirectly, as in the case of minimizing empty vehicles or maximizing the load factor.

The time needed for transporting goods is another very important factor. Instead of minimizing general cost, transportation time may be minimized directly when time constraints are very tight or when the computation of cost is dominated by the labor cost. Whether the transportation time is

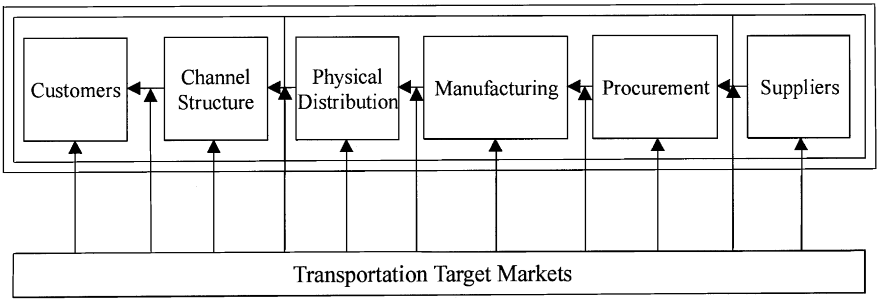

Customers
Channel Structure
Physical Distribution
Manufacturing
Procurement
Suppliers
Transportation Target Markets

Figure 2 Transportation Target Markets in the Supply Chain.

minimized directly or not, the model may include time constraints (e.g., time windows, total time of a driver's route) or the time element may be incorporated in the input data of the model (e.g., included in the underlying network).

There is a tradeoff between transportation cost and transportation time. In the last decades, shorter product cycles in manufacturing and notions like just-in-time inventory have resulted in an increasing demand for smaller transportation times at higher prices and higher transportation cost. The small-package transportation industry has responded with many "premium" services that guarantee short transportation times. Even when the transportation time is not guaranteed, it represents a primary element of quality of service along with other considerations such as minimization of lost and damaged goods, which are often handled indirectly or are external to the cost optimization models.

7.2. Integrating Customer Needs in Transportation Planning

Although we often try to minimize cost in transportation planning, what we really want to achieve is maximization of profit (revenue minus cost). Cost minimization assumes that demand is external to the model and unaffected by the solution obtained. Demand remains at assumed levels only if the obtained solution satisfies customer requirements and customer needs are integrated into the transportation planning process.

Excluding price and transportation time, many customer needs are not easily incorporated into a mathematical model. They may be included in the input data (e.g., different types of service offered) or considered when alternative solutions obtained by optimization are evaluated. Flexibility, a good set of transportation options, and good communication are of primary importance to customers, permitting them to effectively plan their own operations, pickups, and deliveries.

7.3. Forecasting in Transportation Planning

The development of effective transportation plans is highly dependent on our ability to forecast demand. Demand, in the case of the transportation of goods, refers to the expected number of shipments, the number of packages associated with a shipment, and the frequency with which such shipments occur. When developing transportation routes and driver schedules, demand levels are used as input and therefore need to be forecasted. Changes in demand can occur randomly or can follow seasonal patterns. In either case, if an accurate forecast is not produced, the transportation planning effort will yield less accurate results. These results have implications in the design of facilities (e.g., capacity), the acquisition of assets (e.g., delivery vehicles), and in the development of staffing plans (e.g., labor requirements). It is important to note that several factors affect demand in the transportation industry. Business cycles, business models, economic growth, the performance of the shipper's business, competition, advertising, sales, quality, cost, and reputation all have a direct impact on the demand for transportation services.

When developing a forecast, the planner must address some basic questions:

1. Does a relationship exist between the past and the future?
2. What will the forecast be used for?
3. What system is the forecast going to be applied to?
4. What is the size of the problem being addressed?
5. What are the units of measure?
6. Is the forecast for short-range, long-range, or medium-range planning purposes?

Once these questions have been addressed, the steps shown in Figure 3 guide the planner towards the development of a forecasting model that can be used in generating the forecast to support the transportation planning process.

Depending on the type of the transportation problem being solved (e.g., local pickup and delivery operations, large-scale network planning), the user may select different forecasting techniques and planning horizons. For long-range network planning in which the planner is determining the location of future distribution centers, long-range forecasts are required. Sales studies, demographic changes, and economic forecasts aid in the development of such forecasts. When developing aggregate plans in order to determine future staffing needs and potential facility expansion requirements, the planner develops medium-range forecasts covering planning horizons that range from one or two quarters to a year. Known techniques such as time series analysis and regression are usually applied to historical demand information to develop the forecast. Finally, in order to develop weekly and daily schedules and routes to satisfy demand in local pickup and delivery operations, the planner may develop daily, weekly, or monthly forecasts (short-range forecasts). Because demand fluctuations can have a direct impact on the effectiveness of pickup and delivery routes, the use of up-to-date information from shippers is critical. Techniques such as exponential smoothing and trend extrapolation facilitate this type of forecasting.

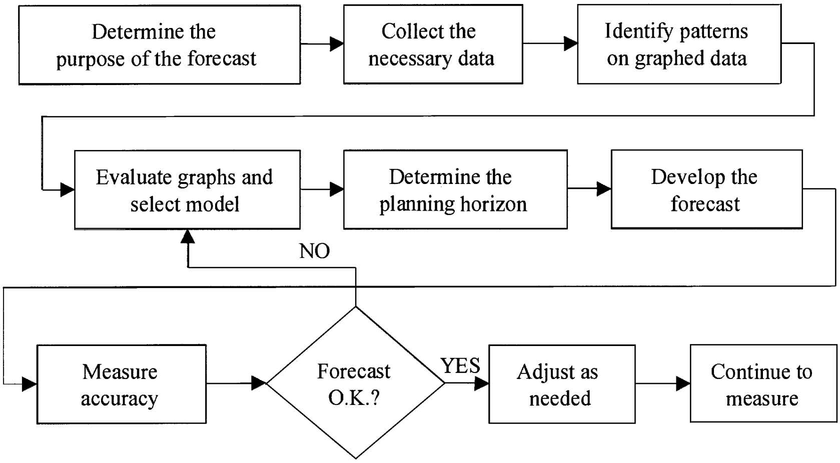

Determine the purpose of the forecast
Collect the necessary data
Identify patterns on graphed data
Evaluate graphs and select model
Determine the planning horizon
Develop the forecast
NO
Measure accuracy
Forecast O.K.?
YES
Adjust as needed
Continue to measure

Figure 3 Steps in Forecasting.

Forecasting is often considered an art. Because the inputs to any forecasting model are mostly historical and based on experience, the accuracy of such forecasts is based on the model selected and the frequency with which it is updated. Forecasting models are usually grouped into two categories: subjective and objective.

Subjective forecasting models are based on judgement and experience. Several techniques exist that attempt to make use of the "expert" 's knowledge to develop a forecast. These techniques include the Delphi method, jury of executives, and the use of sales force intelligence to develop the forecast.

Objective forecasting models are also known as quantitative models. The selection of a quantitative model is dependent on the pattern to be projected and the problem being addressed. There are two types of quantitative forecasting models: time series and explanatory models. In time series models, time is the independent variable and past relationships between time and demand are used to estimate what the demand will be in the future. Explanatory models, on the other hand, use other independent variables instead of or in addition to time. The variables used in the forecasting model are those that have shown a consistent relationship with demand.

When evaluating the impact of seasonal variability on the forecast, the planner has several indicators that can be used to refine the forecast. Leading and lagging economic indicators of the general business cycles can aid the forecaster in the refinement of plans. The Department of Commerce creates an index of well-known economic indicators. This index can be effectively applied to medium- and long-range forecasts to anticipate demand in the transportation of goods.

Several correlation coefficients can be calculated to determine how closely the forecasts correlate with actual demand. The sample mean forecast error (e.g., the square root of the sum of the squared forecast errors), which provides an approximation of the average forecast error of the forecasting model, can also be used. Choosing the best forecast technique requires an understanding of the particular forecasting problem and the characteristics of the available data. Ultimately, the planner's ability to use past and current information to forecast stops, delivery volume, and other key inputs to the transportation planning process will determine the quality and accuracy of the transportation plans developed. With changes in technology and with the increased availability and sharing of information between companies, we expect substantial improvements in the way forecasting is done in the transportation industry.

8. PICKUP AND DELIVERY

8.1. Pickup-and-Delivery Operations

Daily operations in parcel delivery include local pickup and delivery. Terminal facilities are assigned to service areas so that every package that has an origin or destination in the service area is handled through the assigned terminal. Each workday, a fleet of vehicles leaves a terminal (depot) carrying packages to be delivered to customers and returns carrying packages that have been picked up from customers. Most deliveries occur before the pickups, which take place late in the day. Each customer stop (either delivery or pickup) is characterized by a time window during which service must begin.

These windows are usually one-sided: a delivery must occur before a given time and a pickup after a given time. Some windows, though, may be wide open (e.g., a delivery must occur by the end of the workday) and other windows may be two-sided (a pickup must occur before an early closing time). Vehicle capacities are not a limitation in this problem; therefore, the fleet of vehicles may be considered homogenous if the difference in type of vehicle does not have a significant effect on cost. Maximum on-road time for the workday is an important limitation, where on-road time is defined as the total elapsed time from the moment a vehicle leaves the depot until it returns to the depot. The importance of on-road time derives from the fact that often drivers are paid overtime for hours of work beyond a certain length (e.g., 8 hours), and work rules prohibit work days beyond a certain length (e.g., 10 hours).

Efficient distribution implies good asset utilization. The primary assets in this problem are vehicles and drivers. Minimization of the number of drivers and of the total cost of the routes are frequently used objectives. Efficiency needs to be achieved while level of service is maintained; that is, service is provided to all customers in such a way as to satisfy all time-window constraints. This problem has a very short-term planning horizon. Since customer stops may differ from day to day and all the stops are known only a few hours before the actual pickup-and-delivery operation, the problem needs to be solved a few hours before implementation.

8.2. Modeling

The pickup-and-delivery problem can be modeled as a variation of the vehicle routing problem with time windows (VRPTW) and a single depot. Inputs for the VRPTW include matrices that specify the distance and travel time between every pair of customers (including the depot); service time and time window for each customer; maximum (or specified) number of drivers; starting time of the workday; and maximum on-road time of the workday. The maximum on-road time of the workday can be implemented as a time window on the depot, so that the pickup and delivery problem described above can be considered as a traveling salesman problem with time windows and multiple routes (m-TSPTW). However, the term *VRPTW* will be used in this section for this uncapacitated pickup-and-delivery problem.

Distances and travel times can be derived from the longitude and latitude of the customers and depot, assuming specific speed functions. Alternatively, distances and travel times can be obtained from an actual street network; this latter method provides greater accuracy. Longitude and latitude values are relatively easy to obtain; it is often considerably more difficult to acquire data for a street network.

The objective in the VRPTW is to minimize the number of drivers and/or minimize the total cost of the routes while satisfying all constraints. Cost is a function of total distance and on-road time. The output is a set of routes, each of which specifies a sequence of customers and starts and ends at the depot. Because of the short-term planning horizon, the drivers need to be notified for work before the problem is solved. In many cases, therefore, the number of drivers may be estimated and assumed given; the objective then becomes to minimize total cost.

The number of drivers is unlikely to be changed daily. However, the transportation of small packages is characterized by seasonality of demand. In the United States, demand increases by 40–50% during the months before Christmas. A pickup-and-delivery model can be used to obtain the minimum number of additional drivers that must be used to maintain level of service when demand increases.

The VRPTW is an extension of the traveling salesman problem with time windows (TSPTW), which in turn is an extension of the classical traveling salesman problem (TSP). The TSP and its variants have been studied extensively, and many algorithms have been developed based on different methodologies (Lawler et al. 1985). The TSP is NP-complete (Garey and Johnson 1979) and therefore is presumably intractable. For this reason, it is prohibitively expensive to solve large instances of the TSP optimally. Because the TSPTW (and the VRPTW with a maximum number of available drivers) are extensions of the TSP, these problems are NP-complete as well. In fact, for the TSPTW and VRPTW, not only is the problem of finding an optimal solution NP-complete; so is the problem of even finding a feasible solution (a solution that satisfies all time window constraints) (Savelsbergh 1985). For more on time constrained routing and scheduling problems, see Desrosiers et al. (1995).

Existing exact algorithms for the VRPTW have been reported to solve problems with up to 100 stops. However, the problems encountered in parcel delivery are often substantially larger than this. Moreover, these problems need to be solved quickly because of the short-term planning horizon. For these reasons, much of the work in this area has focused on the development of heuristic algorithms—algorithms that attempt to find good solutions instead of optimal solutions.

Heuristic algorithms for the TSPTW and VRPTW are divided into two general categories: route-construction heuristics and route-improvement heuristics. The first type of heuristic constructs a set of routes for a given set of customers. The second type of heuristic starts from an existing feasible solution (set of routes), and attempts to improve this solution. Composite procedures employ both types of heuristic, either by first constructing routes heuristically and then improving them or by applying route-improvement procedures to partially constructed routes at selected intervals during the route-construction process itself. Route-construction and route-improvement heuristics are discussed in more detail below.

8.3. Heuristic Construction Algorithms

There are two general strategies that a route-construction algorithm for the VRPTW can adopt. The first strategy is ''cluster first, route second,'' which first assigns stops to drivers, and then constructs a sequence for the stops assigned to each driver. The second strategy carries out the clustering and sequencing in parallel.

Because clustering is based primarily on purely spatial criteria, the cluster-first, route-second strategy is often more appropriate for the vehicle routing problem (VRP), where time windows are not present than for the VRPTW. However, cluster-first strategies can play a useful role in some instances of the VRPTW, as will be discussed later.

For now, we will confine our attention to algorithms that carry out clustering and sequencing in parallel. Within this general class of algorithms, there is a further distinction between sequential heuristics, which construct one route at a time until all customers are routed, and parallel heuristics, which construct multiple routes simultaneously. In each case, one proceeds by inserting one stop at a time into the emerging route or routes; the choice of which stop to insert and where to insert it are based on heuristic cost measures. Recent work suggests that parallel heuristics are more successful (Potvin and Rousseau 1993), in large part because they are less myopic than sequential approaches in deciding customer-routing assignments.

In the following discussion, we focus on heuristic insertion algorithms for the VRPTW. Solomon (1987) was the first to generalize a number of VRP route-construction heuristics to the VRPTW, and the heuristic insertion framework he developed has been adopted by a number of other researchers. Solomon himself used this framework as the basis for a sequential algorithm. In what follows, we briefly describe Solomon's sequential algorithm and then turn to extensions of the generic framework to parallel algorithms.

8.4. A Sequential Route-Construction Heuristic

This heuristic builds routes one at a time. As presented by Solomon (1987), sequential route construction proceeds as follows:

1. Select a ''seed'' for a new route.
2. If not all stops have been routed, select an unrouted stop and a position on the current route that have the best insertion cost. If a feasible insertion exists, make the insertion, else go to step 1.

In step 1, the heuristic selects the first stop of a new route (seed). There are various strategies for selecting a seed. One approach is to select the stop that is farthest from the depot; another is to select the stop that is most urgent in the sense that its time window has the earliest upper bound. In step 2, the heuristic determines the unrouted stop to be inserted next and the position at which it is to be inserted on the partially constructed route. In order to make this determination, it is necessary to compute, for each unrouted stop, the cost of every possible insertion point on the route.

Solomon introduced the following framework for defining insertion cost metrics. Let $(s_0, s_1, \ldots, s_m)$ be the current partial route, with s_0 and s_m representing the depot. For an unrouted stop u, $c_1(s_i, u, s_j)$ represents the cost of inserting u between consecutive stops s_i and s_j. If the insertion is not feasible, the cost is infinite. For a given unrouted stop u, the best insertion point $(i(u), j(u))$ is the one for which

$$c_1(i(u), u, j(u)) = \min_{p=1,\ldots,m} [c_1(s_{p-1}, u, s_p)]$$

The next stop to be inserted into the route is the one for which

$$c_1(i(u^*), u^*, j(u^*)) = \min_u [c_1(i(u), u, j(u))]$$

Stop u^* is then inserted in the route between $i(u^*)$ and $j(u^*)$. Possible definitions for the cost function c_1 are considered later.

8.5. A Parallel Route-Construction Heuristic

A parallel route-construction heuristic that extends Solomon's sequential algorithm is presented next; the presentation is based on ideas presented by Potvin and Rousseau (1993, 1995) and Russell (1995).

1. Run the sequential algorithm to obtain an estimate of the number of routes k. We can also retain the k seeds obtained by the sequential algorithm. Alternatively, once one has obtained an estimate of the number of routes, one can use some other heuristic to generate the seeds. A representative method is the seed-point-generation procedure of Fisher and Jaikumar (1981). The result of this process is a set of k routes that start from the depot, visit their respective seeds, and return to the depot.
2. If there are no unrouted stops, the procedure terminates. Otherwise, for each unrouted stop u and each partially constructed route $r = (s_0, \ldots, s_m)$, find the optimal insertion point $(p_r(u), q_r(u))$ for which

   $$c_1(p_r(u), u, q_r(u)) = \min_{j=1,\ldots,m} [c_1(s_{j-1}, u, s_j)]$$

   for some cost function c_1. For each unrouted u, its optimal insertion point $(p(u), q(u))$ is taken to be the best of its route-specific insertion points:

   $$c_1(p(u), u, q(u)) = \min_r [c_1(p_r(u), u, q_r(u))]$$

   Select for actual insertion the node u^* for which

   $$c_2(p(u^*), u^*, q(u^*)) = \text{optimal}_u [c_2(p(u), u, q(u)]$$

   for some cost function c_2 that may (but need not) be identical to c_1. If a feasible insertion exists, insert the stop u^* at its optimal insertion point and repeat step 2. Otherwise, go to step 3.
3. Increase the number of routes by one and go to step 2.

Possible definitions for c_1 and c_2 are presented below; again, the presentation here follows that of Potvin and Rousseau (1993).

We can measure the desirability of inserting u between i and j as a weighted combination of the increase in travel time and cost that would result from this insertion. Thus, let

d_{ij} = cost from stop i to stop j

$d(i,u,j) = d_{iu} + d_{uj} - d_{ij}$

e_j = current service start time at j

$e_{u,j}$ = new service start time at j, given that u is now on the route

$t(i,u,j) = e_{u,j} - e_j$

Then a possible measure of the cost associated with the proposed insertion is given by

$$c_1(i,u,j) = \alpha_1 d(i,u,j) + \alpha_2 t(i,u,j) \tag{1}$$

where α_1 and α_2 are constants satisfying

$$\alpha_1 + \alpha_2 = 1$$
$$\alpha_1, \alpha_2 \geq 0$$

One way of defining the measure c_2 is simply by setting $c_2 = c_1$. In this case, the optimal c_2-value is a minimum. An alternative approach is to define c_2 as a "maximum regret" measure.

A regret measure is a kind of "look-ahead" heuristic: it helps select the next move in a search procedure by heuristically quantifying the negative consequences that would ensue, if a given move were not selected. In the context of vehicle routing, a simple regret measure for selecting the next customer to be routed might proceed as follows. For each unrouted customer, the "regret" is the difference between the cost of the best feasible insertion point and the cost of the second-best insertion point; the next customer to be added to a route is one whose regret measure is maximal. But this is still a relatively shortsighted approach. One obtains better results by summing the differences between the best alternative and *all* other alternatives (Potvin and Rousseau 1993; Kontoravdis and Bard 1995). The regret measure that results from this idea has the form

$$c_2(u) = \Sigma_{r \neq r^*} [c_1(p_r(u), u, q_r(u)) - c_1(p_{r^*}(u), u, q_{r^*}(u))] \tag{2}$$

Underlying the use of this regret measure is a notion of urgency: the regret measure for a stop u is likely to be high if u has relatively few feasible or low-cost insertion points available. We would like

TABLE 1 Travel Times between Each Pair of Nodes Including the Depot (node 0)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
0	0	48	29	36	44	54	44	49	51	45	43	42	39	21	43	16	29	46	65
1	56	0	38	43	49	89	59	16	18	54	29	61	64	53	25	56	62	40	72
2	39	40	0	37	45	76	55	41	43	48	31	54	64	38	32	39	46	42	63
3	44	43	35	0	62	86	33	44	49	27	46	34	54	53	33	51	60	24	77
4	52	49	43	62	0	67	63	53	45	74	37	62	67	42	52	47	42	64	58
5	62	89	74	86	67	0	93	92	89	92	79	89	73	73	87	73	64	92	51
6	52	59	53	33	63	93	0	55	64	21	62	20	52	65	51	61	74	38	89
7	57	16	39	44	53	92	55	0	24	50	30	58	63	55	23	57	64	36	75
8	59	18	41	49	45	89	64	24	0	56	28	66	67	53	29	57	62	44	70
9	53	54	46	27	74	92	21	50	56	0	55	28	53	61	43	58	70	28	85
10	51	29	29	46	37	79	62	30	28	55	0	63	61	43	29	48	53	44	62
11	50	61	52	34	62	89	20	58	66	28	63	0	44	61	53	57	70	42	85
12	47	64	62	54	67	73	52	63	67	53	61	44	0	61	72	54	64	64	79
13	29	53	36	53	42	73	65	55	53	61	43	61	61	0	52	19	24	56	59
14	51	25	30	33	52	87	51	23	29	43	29	53	72	52	0	53	58	29	72
15	24	56	37	51	47	73	61	57	57	58	48	57	54	19	53	0	28	55	63
16	37	62	44	60	42	64	74	64	62	70	53	70	64	24	58	28	0	66	54
17	54	40	40	24	64	92	38	36	44	28	44	42	64	56	29	55	66	0	81
18	73	72	61	77	58	51	89	75	70	85	62	85	79	59	72	63	54	81	0

to route such a *u* relatively early to prevent its few good insertion points from being taken by other customers.

8.6. A Numerical Example

We illustrate the route-construction process by considering an instance of the VRPTW with 18 stops, together with the depot (node 0). Table 1 displays the time matrix for this problem, and Table 2 the time windows. All times are listed in hundredths of an hour; for example, 9:45 is represented as 975. The starting time from the depot is 875 for all drivers. Three routes are to be generated, using as seeds nodes 18, 6, and 4.

The parallel route-construction procedure initializes each route by inserting its seed node. Hence, this process creates the three one-stop routes, (0–18–0), (0–6–0), and (0–4–0). Next, for each un-

TABLE 2 Time Windows

Node	Earliest	Latest
0	875	1300
1	875	1200
2	875	1050
3	875	1200
4	875	1300
5	875	1300
6	875	1200
7	875	1050
8	875	1200
9	875	1200
10	875	1200
11	875	1050
12	875	1200
13	875	1200
14	875	1200
15	875	1200
16	875	1050
17	875	1200
18	875	1300

routed node, the best feasible insertion cost for each of the three current routes is computed from formula (1), where $\alpha_1 = 0$, $\alpha_2 = 1$, and time is used as the cost measure. This computation obtains for node 1 the best insertion costs (55 63 53) for the three routes. Hence, the best feasible insertion of node 1 is into route 3 ($r^* = 3$ in formula (2)). The regret measure for node 1 using formula (2) is calculated as follows:

$$c_2(1) = 55 - 53 + 63 - 53 + 53 - 53 = 12$$

The regret measures for the remaining unrouted nodes are computed similarly. The regret values for each node are shown in the following array, where *x* indicates a node that has already been included in a route.

$$\text{Regrets} = (12\ 16\ 52\ x\ 100\ x\ 3\ 23\ 96\ 29\ 86\ 21\ 31\ 1\ 24\ 50\ 48\ x)$$

At each iteration of the procedure, the next stop to be inserted into a route is a stop with the maximal regret measure. In the present case, node 5 has the maximal regret. Hence, the procedure inserts node 5 into its best feasible insertion point (first stop after the depot on route 1). After node 5 is inserted, the three routes under construction become (0–5–18–0), (0–6–0), and (0–4–0). The algorithm repeats this process until all nodes have been inserted.

The three routes returned by the procedure are shown in Table 3 and Figure 4.

8.7. Infeasible Problems

One of the inputs to a typical instance of the VRPTW is the maximum number of drivers available. In the case where the time windows are relatively tight, an algorithm may not be able to find a feasible solution (i.e., a solution in which all time windows are satisfied). The way one proceeds in this case depends on the nature of the application. In some applications, where there is a service guarantee and the same penalty applies no matter how close to meeting the service commitment a late delivery happens to be, we wish to minimize the number of missed time windows. In other applications, where the penalty is proportional to the lateness of a delivery, we wish to minimize,

TABLE 3 Solution Routes

Node	Time Window		Arrival Time
Route 1			
0	875	1300	875
5	875	1300	929
18	875	1300	980
16	875	1050	1034
13	875	1200	1058
15	875	1200	1077
0	875	1300	1101
Route 2			
0	875	1300	875
12	875	1200	914
11	875	1050	958
6	875	1200	978
9	875	1200	999
17	875	1200	1027
3	875	1200	1051
0	875	1300	1095
Route 3			
0	875	1300	875
2	875	1050	904
14	875	1200	936
7	875	1050	959
1	875	1200	975
8	875	1200	993
10	875	1200	1021
4	875	1300	1058
0	875	1300	1110

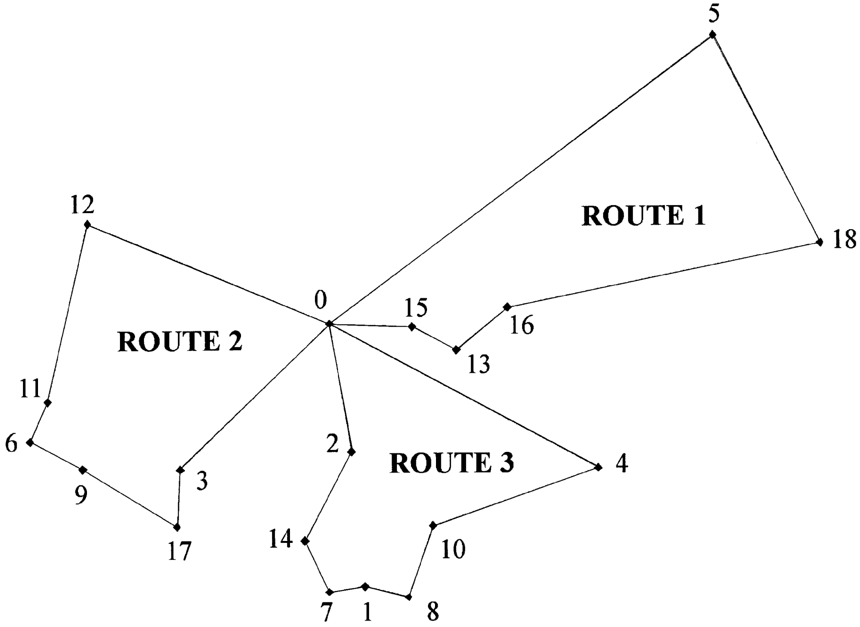


Figure 4 Example Solution.

over all stops, the total time by which the upper bounds of the time windows of stops are missed. In either case, we can still apply the heuristic insertion framework described above by adding an appropriate penalty function to the standard heuristic cost functions.

8.8. Work Breaks

The algorithms presented thus far provide a framework for solving a variety of routing and scheduling problems. However, the specific algorithms we have considered are applicable directly only to a small range of problems; they generally have to be extended in some way in order to deal with real-world applications. One important extension arises in routing applications in which the driver spends an entire day on the road. For such applications, we must take into account the necessity of scheduling lunch and other breaks in order to satisfy work rules and other constraints. Taking these factors into account complicates an already difficult problem.

We can characterize a break by specifying the following values: earliest start time for the break, latest start time for the break, and duration of the break. For example, we might specify that we are to include a 50-minute lunch break that begins any time between 11 am and 1 pm.

To accommodate breaks within the heuristic insertion framework, the most natural way to proceed is to represent the breaks themselves as nodes. Note that a break does come with a natural time window (earliest and latest start times) and an analogue of service time (duration of break). The one crucial missing item is location. In some cases, the location of breaks may be specified as part of the input to the algorithm. In general, however, the algorithm will have to select a location for each break, based on the locations of the nodes that represent actual stops.

For example, suppose we wish to schedule a break after stop a but before stop c. According to the heuristic insertion framework, this is just the problem of inserting the break node b between nodes a and c. One possible approach is simply to assign a's location to b. In this case, we calculate the arrival time at c as follows. Let $[e_x, l_x]$ be the time window for a given stop x, svcTime_x the service time at x, and t_{xy} the travel time from x to y. If x has been assigned to a route, we let arrival_x be the arrival time of the driver at location x. Then the time at which service begins at stop x is given by

$$\text{startTime}_x = \max(\text{arrival}_x, e_x)$$

In the case being considered here, where break node b is inserted between stops a and c, we have

$$\text{startTime}_c = \text{startTime}_a + \text{svcTime}_a + \text{svcTime}_b + t_{ac}$$

Recall that svcTime_b is just the duration of the break.

This is the basic picture, but there is at least one complication that we have to take into account. Suppose, as before, that we wish to insert break node *b* between *a* and *c*. Suppose also that the travel time from *a* to *c* is 30 minutes, that service is completed at *a* at 10:45, and that the time window for *b* is [11:00, 1:00]. Then, according to the simple scheme just described, we would have to wait for 15 minutes at location *a* before starting lunch at *a*!

To avoid this sort of awkward behavior, we would actually insert *b* at the first point on segment (*a*, *c*) such that there is no waiting time at *b*. If there is no such point, then we conclude that no lunch break can be inserted between stops *a* and *c*. For example, in the scenario just described, we would insert *b* halfway between *a* and *c*.

When a break has been fully represented this way as a node with time window, location, and service time, it is treated just like any other node within the heuristic insertion framework.

8.9. Route-Improvement Heuristics

Heuristic route-improvement procedures play a fundamental role in vehicle routing algorithms. Such procedures take as input a feasible solution consisting of a route or set of routes and seek to transform this initial solution into a lower-cost feasible solution.

One of the most successful route-improvement strategies has involved the use of edge-exchange heuristics. The edge-exchange technique was introduced by Lin and Kernighan (1973) as a local search strategy in the context of the conventional traveling salesman problem, but it has since been applied to a variety of related problems, including the vehicle routing problem.

For an individual route, a *k*-exchange involves replacing *k* edges currently on the route by *k* other edges. For example, a 2-exchange involves replacing two edges (say $(i, i + 1)$ and $(j, j + 1)$) by two other edges ((i, j) and $(i + 1, j + 1)$). Usually, all the available *k*-exchanges are examined and the best one is implemented. This is repeated as long as an improved solution is obtained. Since there are $\binom{n}{k}$ subsets of *k* edges in a cycle of *n* edges, the computational complexity of the edge-exchange method increases rapidly with *k*. Even one *k*-exchange requires $O(n^k)$ time, so attention is usually confined to the cases $k = 2$ and $k = 3$.

The idea of an edge exchange can be extended in a straightforward way to the case of pairs of routes. In this case, one exchanges entire paths between routes, where a path consists of a sequence of one or more nodes. For example, let $\text{routeA} = (a_1, a_2, \ldots, a_k)$, $\text{routeB} = (b_1, b_2, \ldots, b_n)$. Then, after exchanging paths between the two routes, the routes that result would be of the form, $\text{routeA} = (a_1, \ldots, a_i, b_{j+1}, \ldots, b_{j+r}, a_m, \ldots, a_k)$, and similarly for routeB.

As mentioned, the goal of a route-improvement heuristic is to reduce routing costs; typically this means that we wish to minimize route duration. However, when we are dealing with time windows, we must also verify that any proposed exchange retains time window feasibility. Techniques for efficient incorporation of time window constraints into edge-exchange improvement methods were developed by Savelsbergh (1985, 1990, 1992); see also Solomon et al. (1988).

The most recent work on route improvement has focused on the development of metaheuristics, which are heuristic strategies for guiding the behavior of more conventional search techniques, with a view towards avoiding local optima and thereby achieving higher-quality solutions. Metaheuristic techniques include tabu search, genetic algorithms, simulated annealing, and greedy randomized adaptive search (Glover 1989, 1990; Kontoravdis and Bard 1995; Potvin et al. 1996; Rochat and Taillard 1995; Taillard et al. 1997; Thangiah et al. 1995).

To illustrate the pertinent concepts, we focus here on tabu search. Tabu search helps overcome the problem of local optimality, which often arises when traditional deterministic optimization heuristics are applied. The problem of local optimality arises in the following way. A wide range of optimization techniques consists of a sequence of moves that lead from one trial solution to another. For example, in a vehicle-routing algorithm, a trial solution consists of a route for each vehicle; and a "move," in the route-improvement phase, might consist of some sort of interroute exchange. A deterministic algorithm of this general type selects a move that will most improve the current solution. Thus, such a procedure "climbs a hill" through the space of solutions until it cannot find a move that will improve the current solution any further. Unfortunately, while a solution discovered with this sort of hill-climbing approach cannot be improved through any local move, it may not represent a global optimum.

Tabu search provides a technique for exploring the solution space beyond points where traditional approaches become trapped at a local optimum. Tabu search does not supplant these traditional approaches. Instead, it is designed as a higher-level strategy that guides their application.

In its most basic form, the tabu method involves classifying certain moves as forbidden (or "tabu"); in particular, a move to any of the most recently generated solutions is classified as tabu.

(When we speak of "moves" here, we mean moves generated by some traditional search method such as edge exchange.) The tabu algorithm then selects the best move from the set of moves not classified as tabu, in order to drive the search into new regions. No concern is given to the fact that the best available moves may not improve the current solution. In particular, because a recently achieved local optimum will be classified as tabu, the method can drive the search down the hill away from this local optimum. The hope is that the expanded search will lead to an improved final solution. There is no guarantee that an improved solution will be found, but variations on the method just described have achieved considerable success and have become increasingly popular.

What has been described here is a very simple tabu search framework, which performs "bookkeeping" operations to ensure that the algorithm does not return to recently explored solutions. More sophisticated metaheuristic strategies more explicitly guide the direction of the search itself.

8.10. Preassigned Routes and Preassigned Territories

The heuristic route-construction methods described thus far are designed to minimize route cost while satisfying time window and other constraints. Route cost is typically a function of total on-road time and distance. In an approach to routing that seeks only to minimize travel time and distance, one typically follows a reoptimization strategy. That is, one is faced each day by a new problem, defined by the set of customers who actually require service that day; and one solves that problem by applying a heuristic route-construction algorithm. Because there is no concept, in such algorithms, of geographical area or regularity of service, the solution for one day will be independent of the solution for another. While these solutions may be very efficient from the standpoint of total cost and travel time, they may not be satisfactory for some applications. The reason for this shortcoming is that in some applications, there may be additional, hidden costs that the reoptimization strategy fails to take into account.

One such cost involves driver knowledge of the area in which he or she is providing service. A driver who is familiar with the road network and traffic conditions in his service area is likely to be more efficient than a driver for whom the delivery area is relatively new. A second, related cost involves the development of business relationships. For many service providers, an important goal is to achieve regularity and personalization of service by having the same driver visit a particular customer every time that customer requires service.

These considerations suggest that for some applications, it may be desirable to maintain some consistency from one day to the next in assigning customers to drivers. In this case, we need to devise some alternative to the reoptimization strategy.

One way of devising such an alternative is to interpret our problem as a probabilistic vehicle routing problem (PVRP) (Jaillet and Odoni 1988). According to this interpretation, we are given a service region and the set of all potential customers in that region. On a given day, only a subset of the customers will actually need service and the exact customer set for a given day cannot be predicted. However, we are also given a probability distribution, based on historical data, over the set of potential customers. The probability assigned to a potential customer represents the probability that that customer will require service on any given day.

One way of defining an objective function for this problem is provided by the a priori optimization strategy investigated by Jaillet (1988) and Bertsimas et al. (1990). In order to develop the ideas underlying this strategy, it is helpful to focus on the TSP. In the conventional TSP, we are given a complete graph $G = (V,E)$ and wish to find a lowest-cost tour that visits every node in V. We obtain the probabilistic TSP (PTSP) by supposing that, on any given day, the traveling salesman may have to visit only some subset S of the nodes in V. The probability that any v in V is present (must be visited) in a given problem instance is given by a probability function p over V. We identify a problem instance with the subset S of nodes that are present in that instance. Thus there are 2^n possible instances of the PTSP on G, where $n = |V|$.

According to the a priori optimization strategy, one finds a priori a tour through all n nodes of G. For any given instance of the problem, the $k \leq n$ nodes that are actually present are visited in the same order as they appear in the a priori tour; the $(n - k)$ missing nodes are simply skipped.

A natural measure of effectiveness for the a priori strategy is given by the notion of minimum length in the expected value sense. Thus, let τ be the a priori tour, and let L_τ be the length of τ. If the problem instance S occurs with probability $p(S)$ and requires covering a total length $L_\tau(S)$ according to the a priori tour, then it will receive a weight of $p(S)L_\tau(S)$ in the computation of expected length. Therefore, according to this construal of the PTSP, the objective is to find an a priori tour τ_0 through the n nodes of G, which minimizes the quantity

$$E[L_\tau] = \Sigma_{S \subseteq V}\, p(S) L_\tau(S)$$

We extend this model to the VRP by requiring that an a priori tour be constructed for each driver in such a way that all potential customers are assigned to some driver. Note that a set of a priori tours of this sort does provide the kind of regularity of service described above.

At first glance, the task of calculating $E[L_\tau]$ for a given a priori tour τ may appear problematic because the summation is over all 2^n subsets of V. However, Jaillet derived an efficient closed-form expression for $E[L_\tau]$, which requires only $O(n^2)$ time to compute (see discussion in Jaillet 1988).

Of course, being able to calculate $E[L_\tau]$ efficiently for a given a priori tour is very different from actually finding a tour that minimizes $E[L_\tau]$. Because the PTSP is at least as hard as the TSP, there is little hope of developing exact optimization methods that could solve more than modestly sized instances of the problem. Consequently, one must employ heuristics in order to develop practically viable PTSP algorithms. However, it has proven difficult to develop effective heuristic approaches to the PTSP, at least in part because the class of optimal solutions to the PTSP has properties very different from those associated with the conventional (Euclidean) TSP. For example, one of the fundamental properties of the Euclidean TSP is that an optimal solution cannot intersect itself; this property follows directly from the triangle inequality. In contrast, an optimal solution for the PTSP *can* intersect itself, even when the triangle inequality is satisfied (systematic treatments of the differences between the TSP and PTSP are presented by Jaillet [1988] and Jaillet and Odoni [1988]). One consequence of the fact that the PTSP has features very different from the TSP is that, in general, we cannot expect heuristic approaches designed specifically for the TSP to be extended successfully to the PTSP. In general, therefore, entirely new solution approaches are needed. One of the few promising approaches to the PTSP that has been discussed in the literature is based on the idea of "spacefilling curves" (Bartholdi and Platzman 1988). But the probabilistic versions of the TSP and VRP remain very much under the heading of research topics.

The problem becomes even more difficult when we turn to probabilistic versions of the VRPTW. For this problem, presumably, the a priori strategy would involve multiple objectives; in addition to minimizing the expected length of the a priori tour, we would also want to minimize the expected number of missed time windows. But it is difficult even to formulate this problem properly, much less solve it.

For this reason, when time windows are involved, we try in actual applications to capture some subset of the key features of the problem. In this case, instead of trying to construct a priori tours of the kind just described, we simply try to construct a solution in which a large percentage of repeat customers is handled by the same driver and each driver is relatively familiar with most of the territory to which he or she is assigned.

One way of achieving such a solution is to partition the service area into geographical regions and then always assign the same driver to the same region. In operational terms, a solution of this sort is one in which drivers have well-defined territories: a driver travels from the depot to his or her service area, carries out his or her work there, and then returns to the depot. This picture is to be contrasted with the one that emerges from the heuristic insertion framework, which usually results in routes that form a petal structure: in general, it creates routes in the form of intersecting loops, in which stops appear in a roughly uniform way.

To construct a solution in which drivers are assigned to territories in this way, it is appropriate to adopt the cluster-first, route-second strategy mentioned earlier. Thus, we proceed by first assigning stops to drivers, thereby partitioning the service area. We then construct a sequence for the stops assigned to each driver. This strategy represents a departure from the heuristic insertion framework, according to which clustering and routing proceed in parallel.

One way of implementing a cluster-first strategy is as follows. Suppose that, on a given day, we wish to construct k routes. We can proceed by (heuristically) solving a k-median problem (Daskin 1995); we then carry out the initial clustering by assigning each stop to its closest median. Having created the clusters, we then construct a route for each cluster individually by solving the resulting TSPTW. Unfortunately, it is often impossible to construct a feasible route for each cluster. When feasible routes cannot be constructed, we have to shift stops between clusters in such a way as to achieve feasibility while maintaining well-defined territories as much as possible.

If we begin by solving a new k-median problem each day, then we may succeed in producing well-defined territories for drivers; but we are unlikely to achieve consistency from one day to the next in assigning customers to drivers. A natural way of extending this approach to achieve consistency is to take historical information into account. Thus, we can solve a weighted k-median problem over the set of all locations that have required service over a specified number of past days, where the weight of a location is proportional to the number of times it has actually required service. We would then consistently use these medians as the basis for clustering. There may well be days in which the initial clusters thus constructed are significantly unbalanced with respect to numbers of customers per driver; in this case, we have to shift stops between clusters, as described above.

Whether the procedure just outlined can be applied successfully in the presence of time windows depends in part on the nature of those time windows. For some applications, time windows are one-sided, and there are only a few different service levels: for example, delivery by 9:00 am for express packages, delivery by 11:00 am for priority packages, and so on. For such applications, it is plausible to suppose that the sort of strategy just described, which emphasizes the spatial aspects of the problem, can be implemented successfully. If, on the other hand, there is a variety of different, overlapping, two-sided time windows, then we are faced with what is really a three-dimensional problem, with time as the third dimension. In such a case, where the temporal aspect of the problem predominates, the prospects for a successful cluster-first strategy of the type described here are considerably less promising.

An alternative heuristic strategy is to employ a sweep method for clustering, introduced by Gillet and Miller (1974); see also Fisher and Jaikumar (1981) and Hall et al. (1994). According to this strategy, one begins by dividing the service area into annuli centered at the depot, using some heuristic method for determining the radius of each annulus. For each annulus, one then sorts the set of customers by increasing polar angle and then builds routes sequentially, inserting stops in order into the current route.

As in the *k*-median method, the prospects for successful application, in the presence of time windows, of what is essentially a spatial method depend on the nature of those time windows. Again as in the *k*-median method, one would have to take historical data into account in order to achieve assignments of customers to drivers that are reasonably consistent over time.

8.11. Implementation Issues

It is evident, on the basis of the discussion in this section, that for vehicle-routing applications one may require a suite of algorithms, each of which best serves a somewhat different objective. Furthermore, to make the best use of these algorithms, it is important that the actual user be specially trained for the application.

Accuracy of the input data is very important in this problem (and in optimization problems in general). Because the objective is to minimize cost, the model is very sensitive to data that affect cost calculations. Because traffic conditions change with the time of day, weather conditions, and so on, travel times are generally estimated even when the best data are available. When longitude and latitude values are used as the basis for travel time estimates, a further loss of accuracy results. This indicates that the extra cost of getting provably optimal solutions may not be justified, and also that effort needs to be made to develop methodologies that are robust with respect to data inaccuracies. One must certainly take into account the possibility of such inaccuracies when one reports the output routes. It is especially important to do so if the intention is to provide precise directions to drivers who are inexperienced in their assigned areas. Obvious inaccuracies may generate distrust for the model by the users.

The pickup-and-delivery problem discussed in this section is static in the sense that all demand is known before any service begins. But one can also consider real time dispatching problems, in which demand is generated in real time and routing decisions are made after the driver has left the depot. Advances in communications technology make it feasible to implement real-time systems of this sort. They also make feasible the implementation of hybrid systems, in which drivers are given predetermined routes but these routes may be modified dynamically as new demand arises during the workday. Dynamic problems of this sort will require new algorithmic techniques to supplement those presented in this section.

9. LARGE-SCALE TRANSPORTATION NETWORK PLANNING

9.1. Line-Haul Movement of Packages

During the pickup and delivery operation of a parcel delivery company, packages are picked up and brought to a local terminal, as discussed in Section 8. These packages are sometimes transported to their destination terminal directly and delivered to the customers. In most cases, however, packages go though a series of transfer terminals where they get sorted and consolidated. Transportation between terminals is achieved using several modes of transportation. The most prevalent modes on the ground are highway and rail, using a fleet of tractors and trailers of different types. A fleet of aircraft is used for transporting by air.

A transfer terminal often shares the same building with one or more local terminals. Several sorting operations may be run daily in a transfer terminal and are repeated every day of the work week at the same times. A large transfer terminal may have four sorting operations at different times of the day or night. Packages get unloaded, sorted, and reloaded on trailers. Several types of trailers may be used with different capacities and other characteristics. One or more trailers are attached to a tractor and depart for the next terminal on their route. This may be another transfer terminal or their destination terminal, or a rail yard for the trailers to be loaded and transported by rail, or an air terminal, if the packages have to be transported by air. In this section, we will examine a model that determines the routes of packages from their origin to their destination terminals and the equipment used to transport them, only for packages that travel on the ground using two modes, highway or rail.

Package handling and transporting between terminals represents a large part of the operating costs for a package transportation company. A good operating plan that minimizes cost and/or improves service is crucial for efficient operations. In recent years, many new products and services have been introduced in response to customer demand. These changes as well as the changes in package volume require additions and modifications to the underlying transportation network to maintain its efficiency. A planning system that can evaluate alternatives in terms of their handling and transporting costs and their impact on current operations is very important. There is a tradeoff between the number of sorting operations a package goes through and the time needed to reach its destination. More sorting operations increase the handling cost, decrease the transportation cost because they consolidate loads better, but also increase the total time needed until a package is delivered to the customer. All these factors need to be taken into account in the design of a transportation network for routing packages.

9.2. A Network-Optimization Problem

The network-optimization problem for transporting packages on the ground can be defined as follows. Determine the routes of packages, mode of transportation (highway or rail), and types of equipment to minimize handling and transportation costs while the following constraints are met: service requirements are respected, the capacities of sorting operations are not surpassed, no sorting operation is underutilized, the number of doors of each facility available for loading trailers is considered, trailers balance by building, and all the packages to the same destination are routed along the same path.

To maintain level of service, packages must be transported between particular origin/destination (OD) pairs in a given time. The capacity of a sorting operation cannot be surpassed but also, if the number of packages through a sorting operation falls below a given value, the sorting operation needs to be closed down. The number of loading doors of each facility represents the maximum number of trailers that can be loaded simultaneously during a sorting operation. Note that two or more of the trailers loaded simultaneously may be going to the same terminal next on their route.

Trailers need to balance by building daily. This means that the trailers of each type that go into a building must also leave the building during the daily cycle. Balancing is achieved by introducing empty trailers into the system. There are several types of trailers that are used, with different capacities. Some types can be used both on highway and rail. Trailers can also be rented from the railroads, and these trailers may not need to balance. Each tractor can pull one, two, and, on rare occasions, three trailers, depending on the trailer types and highway restrictions. At a terminal, tractor–trailer combinations are broken up, the trailers unloaded, the packages sorted, the trailers loaded again, and the tractor and trailers reassembled to form new combinations. There is no constraint on the number of trailers of any type that are used.

An important operational constraint is that all the packages to the same destination must be routed along the same path. This implies that all packages with the same destination terminal follow the same path from the terminal where they first meet to their destination. This constraint results from the fact that sorting is done by destination. If sorting became totally automated, this constraint might be modified.

A network-optimization model can be used as a long-term planning tool. It can evaluate alternatives for locating new building facilities, opening or closing sorting operations, and changing an operating plan when new products are introduced or the number of packages in the system changes. Because any such changes require retraining of the people who manually sort the packages, the routing network in parcel delivery is not changed often or extensively. For a network-optimization model to be used as an intermediate-term planning tool to modify the actual routing network, the model needs to obtain incremental changes, that is, to obtain solutions as close as possible to the current operations.

9.3. Modeling

The network-optimization problem is formulated on a network consisting of nodes that represent sorting operations and directed links that represent movement of packages from one sorting operation to another. Both the highway and rail modes are represented by one link if the same time is needed to traverse the link by either mode. If a different time is needed by each mode, two different links are used, each representing one mode. In the latter case, all the packages need to travel by only one mode even if both modes are feasible so that all the packages that start together arrive at the same time to their destination. It is assumed that every day is repeated unchanged without any distortion of the pattern because of weekends. Since a sorting operation is repeated every day at the same time, one node represents a sorting operation for as many days as necessary for a complete cycle of operations to occur in the system. To avoid any confusion, we will use in the description of the network design problem the terms *origin* and *destination* or *OD pair* to refer to the origin and destination nodes of a package; we will use the terms *initial* and *final* nodes to refer to the origin and destination nodes of a directed link.

Each node is characterized by a daily starting and ending time of the sorting operation, a sorting capacity, a sorting cost, and the number of loading doors available. Each link is characterized by a travel time and distance between its initial and final nodes and the types of combinations that are permitted on the link with their costs and capacities. The combinations representing both highway and rail are associated with the same link when the travel time is the same for highway and rail. If the travel time is different and the link represents a particular mode, only the combinations of this mode are associated with the link. In addition, for each OD pair, the number of packages that need to be transported is given as well as the maximum time permitted from the moment a package starts at its origin until it reaches its destination.

A network-optimization system that solves the problem presented above utilizes large amounts of data and must have a reliable data interface in addition to the solver. The data interface extracts, verifies, and validates the data needed for the problem corresponding to the particular geographical area applicable to the problem. It then generates the network and, after the solver is applied, produces reports of the results. The solver consists of optimization algorithms that obtain solutions of the problem.

The network described above can be very large. To facilitate the application of the optimization algorithms, the network size is reduced using various aggregation rules. OD pairs may be consolidated by building or destination group, and local terminals in the same building may be combined into one node. If all the possible connections between each pair of sorting operations are added, the network becomes very dense. For this reason, only links that are likely to be used in the solution are included (e.g., direct links between sorting operations that are farther apart than a given distance are omitted).

The network design problem is formulated below as a mixed-integer programming problem (MIP) (Nemhauser and Wolsey 1988). A graph $G(N,E)$ is given with N a set of nodes and E a set of directed links. A node represents a sorting operation for each consecutive day of a whole cycle of operations. A directed link (i,j) represents the movement of packages from sorting operation i (initial node of the link) to sorting operation j (final node of the link). Because of the presence of two different links that are used in parallel between the same sorting operations representing two different modes with different travel times, link (i,j) may not be unique. To simplify the presentation, however, we disregard the presence of parallel links. The formulation can be easily extended to include this case because at most one of such parallel links can be used in the solution. Additional notation is introduced next.

9.3.1. Notation

Parameters

A = set of OD pairs = $\{(k,l) \mid k \in N$ is the origin and $l \in N$ is the destination, $k \neq l\}$
M = set of all trailer-combination types (a trailer combination consists of a tractor and one or more trailers)
M_{ij} = set of trailer-combination types of link $(i,j) \in E$; $M_{ij} \subseteq M$
Q = set of trailer types
F = set of trailer types that must balance (rented rail trailers may not need to balance); $F \subseteq Q$
B = $\{B_k \mid B_k$ is a building; $B_k \subseteq N\}$; $j \in B_k$ means that sorting operation $j \in N$ resides in building B_k
c_{ijm} = cost of moving trailer-combination type $m \in M_{ij}$ on link $(i,j) \in E$
d_{ikl} = cost of handling all the packages of OD pair $(k,l) \in A$ through node $i \in N$
k_q = capacity of trailer type $q \in Q$ in number of average-size packages
δ_{qm} = number of single trailers of type $q \in Q$ in trailer-combination type $m \in M$
h_{ij} = starting time of sorting operation j minus ending time of sorting operation i so that the packages sorted at $i \in N$ are transported on $(i,j) \in E$ and sorted next at $j \in N$
r_j = duration of sorting operation j
s_{kl} = starting time at origin $k \in N$ of the packages to be transported to destination $l \in N$
v_{kl} = number of average-size packages of OD pair $(k,l) \in A$
τ_{kl} = total time permitted for the packages of OD pair (k,l) to travel from k to l
u_i = capacity of sorting operation $i \in N$ (maximum number of packages that can go through node i)
g_i = minimum number of packages going through $i \in N$, if sorting operation i is open
f_i = maximum number of outgoing-trailer positions (loading doors) available at node $i \in N$
INF = a very large number

Decision variables

x_{ijm} = number of trailer combinations of type $m \in M_{ij}$ used on link $(i,j) \in E$
w_{ijq} = number of nonempty trailers of type $q \in Q$ on link $(i,j) \in E$
y_{ijkl} = 1 if link $(i,j) \in E$ is used to transport the packages of OD pair $(k,l) \in A$; 0 otherwise
z_i = 1 if sorting operation $i \in N$ is open; 0 otherwise
t_{ikl} = departure time of packages of OD pair $(k,l) \in A$ from node $i \in N$ (end time of sorting operation i).

The cost d_{ikl} is estimated as the average cost of a package through sorting operation $i \in N$ multiplied by the number of packages of OD pair $(k,l) \in A$. The cost c_{ijm} is estimated as a function of the distance of link $(i,j) \in E$, fuel prices, driver wages as well as depreciation and cost of the trailer-combination type $m \in M_{ij}$.

It is assumed that all the packages processed through a sorting operation are available at the beginning and leave at the end of the sorting operation. The time h_{ij} of link $(i,j) \in E$ is the difference between the starting time of the sorting operation j and the ending time of the sorting operation i with the appropriate time difference included for the packages to be transported on link (i,j) and be available at j on time. The time h_{ij} as defined above makes it unnecessary to consider the time windows of the sorting operations explicitly in the following formulation. The time h_{ij} is also equal to the travel time on link (i,j) plus the difference in time between the beginning of the sorting operation at j and the arrival of the packages at j, which corresponds to the wait time until the sorting operation j starts. The capacity of a sorting operation is represented by an upper bound that cannot be exceeded. A lower bound may also be used to force sorting operations to be closed if they are underutilized.

9.4. Network Design Formulation

$$\text{Min } Z(\mathbf{x},\mathbf{w},\mathbf{y},\mathbf{z},\mathbf{t}) = \Sigma_{j\in N}\, \Sigma_{(k,l)\in \mathrm{A}}\, d_{jkl}\, \Sigma_{i\in N}\, y_{ijkl} + \Sigma_{(i,j)\in E}\, \Sigma_{m\in M_{ij}}\, c_{ijm}\, x_{ijm} \tag{3}$$

subject to

$$\Sigma_{j\in N}\, y_{ijkl} - \Sigma_{p\in N}\, y_{pikl} = b \qquad \forall\ (k,l) \in A,\ i \in N \tag{4}$$

where $b = 1$ for $i = k$; $b = -1$ for $i = l$; $b = 0$ otherwise

$$t_{jkl} - t_{ikl} - \mathrm{INF}(y_{ijkl} - 1) \geq h_{ij} + r_j \qquad \forall\ i \in N,\ j \in N,\ (k,l) \in A \tag{5}$$

$$t_{kkl} = s_{kl} \qquad \forall\ (k,l) \in A \tag{6}$$

$$t_{lkl} \leq \tau_{kl} + s_{kl} + r_l \qquad \forall\ (k,l) \in A \tag{7}$$

$$\Sigma_{i\in N}\, \Sigma_{(k,l)\in A}\, v_{kl}\, y_{ijkl} \leq u_j \qquad \forall\ j \in N \tag{8}$$

$$\Sigma_{i\in N}\, \Sigma_{(k,l)\in A}\, v_{kl}\, y_{ijkl} - g_j\, z_j \geq 0 \quad \forall\ j \in N \tag{9}$$

$$z_j - y_{ijkl} \geq 0 \qquad \forall\ i \in N,\ j \in N,\ (k,l) \in A \tag{10}$$

$$y_{ijkl} + y_{ij'k'l} \leq 1 \quad \forall\ (i,j) \in E,\ (i,j') \in E,\ (k,l) \in A,\ (k',l) \in A,\ k < k',\ j \neq j' \tag{11}$$

$$\Sigma_{j\in B_k}\, \Sigma_{i\in N}\, \Sigma_{m\in M_{ij}}\, \delta_{qm}\, x_{ijm} - \Sigma_{j\in B_k}\, \Sigma_{p\in N}\, \Sigma_{m\in M_{ij}}\, \delta_{qm}\, x_{jpm} = 0 \qquad \forall\ B_k \in B,\ q \in F \tag{12}$$

$$\Sigma_{q\in Q}\, k_q\, w_{ijq} - \Sigma_{(k,l)\in A}\, v_{kl}\, y_{ijkl} \geq 0 \qquad \forall\ (i,j) \in E \tag{13}$$

$$\Sigma_{j\in N}\, \Sigma_{q\in Q}\, w_{ijq} \leq f_i \qquad \forall\ i \in N \tag{14}$$

$$\Sigma_{m\in M_{ij}}\, \delta_{qm}\, x_{ijm} - w_{ijq} \geq 0 \qquad \forall\ (i,j) \in E,\ q \in Q \tag{15}$$

$$x_{ijm} \geq 0 \text{ and integer} \qquad \forall\ i,j,m \tag{16}$$

$$w_{ijq} \geq 0 \text{ and integer} \qquad \forall\ i,j,q \tag{17}$$

$$y_{ijkl} = 0 \text{ or } 1 \qquad \forall i,j,k,l \tag{18}$$

$$z_i = 0 \text{ or } 1 \qquad \forall\ i \tag{19}$$

$$t_{jkl} \geq 0 \qquad \forall\ j,k,l \tag{20}$$

The objective function (3) minimizes the total cost of handling and transporting the packages of all the OD pairs. Constraints (4) are balancing constraints ensuring that the packages of an OD pair start from their origin, end at their destination, and, if they enter an intermediate node, also exit the node. Each constraint (5) computes the departure time of the packages of OD pair (k,l) from node j, using the departure time of the preceding node i. If link $(i,j) \in E$ is used to transport the packages of OD pair $(k,l) \in A$ ($y_{ijkl} = 1$), the constraint becomes $t_{jkl} - t_{ikl} \geq h_{ij} + r_j$. If link $(i,j) \in E$ is not used ($y_{ijkl} = 0$), the constraint is not binding. The starting time at each origin is set with constraints (6).

Constraints (7) ensure that the service requirements are met, that is, the movement of the packages of OD pair (k,l) from their origin k to their destination l takes no more than τ_{kl} time. These constraints also force constraints (5) to apply with equality if necessary.

Constraints (8) ensure that the capacities of the sorting operations are not violated, and constraints (9) prevent sorting operations that are open from being underutilized. Constraints (10) ensure that packages enter node j only if the sorting operation j is open. That is, if $y_{ijkl} = 1$ for any i, j, k, l then $z_j = 1$ from constraints (10). If $y_{ijkl} = 0$ for all i, j, k, l then $z_j = 0$ from constraints (9) and (10).

Constraints (11) ensure that all the packages going though sorting operation i and having the same destination l use the same network link to the next sorting operation on their path, that is, there are no split paths to the same destination. The only case that is permitted is $k \neq k', j = j'$. The case $k = k', j \neq j'$ is excluded by constraints (4) which ensure a unique path for each OD pair (k,l). The case $k \neq k', j \neq j'$ is excluded by constraints (11). Using $k < k'$ avoids repeating the constraints (11) twice.

Constraints (12) are balancing constraints ensuring that for each building B_k and for each trailer type $q \in F$ that must balance, the same number of trailers that enter building B_k also exit it. To achieve balancing of trailers, empty trailers may be introduced into the system by constraints (12). These constraints are more complicated than they would be if each trailer-combination type rather than each trailer type needed to balance. Constraints (13) are the volume constraints, ensuring that for each link (i,j) trailers with enough capacity are used to transport the number of packages on the link.

Constraints (14) ensure that the number of trailers that are filled during a sorting operation is no larger than the number of available loading doors. It is assumed that at most one trailer is filled from each door during a sorting operation. This is a good approximation although it may be conservative at times because occasionally it is possible to fill a trailer at a door and replace it with a new trailer during the same sorting operation. Constraints (15) ensure that the number of nonempty trailers of type $q \in Q$, on link $(i,j) \in E$ represented by w_{ijq} is no larger than the total number of trailers of type $q \in Q$ (empty and nonempty) forming the trailer combinations of type $m \in M_{ij}$ on link $(i,j) \in E$ represented by x_{ijm}. Note that w_{ijq} can have alternative solution values in the previous formulation apart from being exactly equal to the number of nonempty trailers. The value of w_{ijq} is always at least as large as the number of nonempty trailers from constraint (13). If empty trailers are introduced for balancing so that constraint (15) is not binding, w_{ijq} can have alternative solution values larger than the number of nonempty trailers, depending on how tight the corresponding constraint (14) is. Constraints (16) to (20) are the integrality and nonnegativity constraints.

For a large transportation company, the $G(N,E)$ network may have 40,000 links even when care is taken to keep it as sparse as possible. The number of binary variables y_{ijkl} of the above MIP formulation may be 14,000,000 or more. That is also the number of constraints (5), while the number of constraints (11) is much larger.

Formulation (3)–(20) is too large to be supplied to an MIP solver and solved directly. Instead, a heuristic solution for the network design problem can be obtained by solving sequentially two smaller problems that are embedded in formulation (3)–(20). These are the package-routing problem and the trailer-assignment problem. In the first step, the routes of the packages are obtained assuming a representative trailer type for highway and rail. This step produces the number of packages that are transported from the initial node to the final node of each link of the network. In the second step, given the number of packages and the types of permitted equipment on each link, the actual modes and trailer combinations that balance are obtained. These two problems are presented next.

9.5. Package-Routing Problem

The package-routing problem determines the routes of packages on the $G(N,E)$ network for each OD pair so that the service requirements are met, the capacities of the sorting operations are not exceeded, the sorting operations are not underutilized, the number of loading doors of the buildings is not exceeded, packages to a common destination are routed along the same path, trailers are not underutilized, and total cost is minimized. The package-routing problem does not determine the trailer combinations that are used, nor does it balance trailer types.

In the following, we still present a simplified formulation where we do not indicate mode, although we continue to use one link for both modes if the travel times are the same for both modes on the link and two different links if the travel times are different. We extract from formulation (3)–(20) constraints (4)–(8), (18), and (20), which involve only the y_{ijkl} and t_{jkl} variables, putting aside for the moment the constraints on the number of loading doors, underutilized sorting operations, split paths to common destination, and maximization of trailer utilization. Because of their very large number, constraints (11) are not included.

Because the objective function (3) involves variables other than y_{ijkl} and t_{jkl}, we replace it with a new objective function that uses an estimated cost parameter. First, we select a representative trailer-combination type for each mode and estimate the cost per package, per mode for each link by dividing the combination cost on the link by the combination capacity. For links that share both modes, the smaller cost of the two modes is used. Using the cost per package, the following cost parameter is computed:

d_{ijkl} = cost of transporting all the packages of OD pair $(k,l) \in A$ on link $(i,j) \in E$

Using this new cost, the objective function (3) is replaced by the following approximation:

$$\text{Min } Z(\mathbf{y},\mathbf{t}) = \Sigma_{j\in N}\,\Sigma_{(k,l)\in A}\, d_{jkl}\,\Sigma_{i\in N}\, y_{ijkl} + \Sigma_{(i,j)\in E}\,\Sigma_{(k,l)\in A}\, d_{ijkl}\, y_{ijkl} \tag{21}$$

Objective function (21) can be simplified if we combine the transporting cost of all the packages of OD pair (k,l) along link (i,j) and their handling cost through sorting operation j in the following cost parameter:

c_{ijkl} = cost of transporting all the packages of OD pair $(k,l) \in A$ on link $(i,j) \in E$ and handling them through node $j \in N$

The objective function becomes

$$\text{Min } Z(\mathbf{y},\mathbf{t}) = \Sigma_{(i,j)\in E}\,\Sigma_{(k,l)\in A}\, c_{ijkl}\, y_{ijkl} \tag{22}$$

The MIP formulation (22), (4)–(8), (18), and (20) of the package-routing problem is still a very large problem and difficult to solve directly. If the complicating constraints (8) are removed, the problem is solved relatively easily. We take advantage of this characteristic by using a Lagrangian relaxation approach (Ahuja et al. 1993; Fisher 1985; Geoffrion 1974). This is combined with search heuristics that also implement the constraints that are completely omitted in the formulation, which are the constraints on the number of loading doors, the lower bound constraints on the capacity of each sorting operation that is open, the split paths constraints, and an additional lower bound constraint on the number of packages on any link $(i,j) \in E$ that is actually used. The last constraint forces more efficient trailer utilization by ensuring that no link carries too few packages, which would result in the use of trailers with very low load factors in the trailer-assignment step.

The following Lagrangian dual problem is obtained by dualizing constraints (8) using their nonnegative Lagrange multipliers λ_j, $j \in N$.

9.6. Lagrangian Relaxation of the Package-Routing Problem

$$\begin{aligned}\text{Min } Z(\mathbf{y},\mathbf{t},\lambda) &= \Sigma_{(i,j)\in E}\,\Sigma_{(k,l)\in A}\, c_{ijkl}\, y_{ijkl} + \Sigma_{j\in N}\,\lambda_j\,(\Sigma_{i\in N}\,\Sigma_{(k,l)\in A}\, v_{kl}\, y_{ijkl} - u_j)\\ &= \Sigma_{(i,j)\in E}\,\Sigma_{(k,l)\in A}\,(c_{ijkl} + \lambda_j v_{kl})\, y_{ijkl} + \text{CONSTANT}\\ &= \Sigma_{(i,j)\in E}\,\Sigma_{(k,l)\in A}\,\bar{c}_{ijkl}\, y_{ijkl} + \text{CONSTANT}\end{aligned} \tag{23}$$

subject to

$$\Sigma_{j\in N}\, y_{ijkl} - \Sigma_{p\in N}\, y_{pikl} = b \qquad \forall\ (k,l) \in A,\ i \in N \tag{24}$$

where $b = 1$ for $i = k$; $b = -1$ for $i = l$; $b = 0$ otherwise

$$t_{jkl} - t_{ikl} - \text{INF}(y_{ijkl} - 1) \geq h_{ij} + r_j \qquad \forall\ i \in N, j \in N, (k,l) \in A \tag{25}$$

$$t_{kkl} = s_{kl} \quad \forall\ (k,l) \in A \tag{26}$$

$$t_{lkl} \leq \tau_{kl} + s_{kl} + r_l \quad \forall\ (k,l) \in A \tag{27}$$

$$y_{ijkl} = 0 \text{ or } 1 \qquad \forall\ i,j,k,l \tag{28}$$

$$t_{jkl} \geq 0 \quad \forall\ j,k,l \tag{29}$$

$$\lambda_j \geq 0 \qquad \forall\ j \tag{30}$$

The cost $\bar{c}_{ijkl} = c_{ijkl} + \lambda_j v_{kl}$ is the modified cost of transporting all the packages of OD pair $(k,l) \in$ A on link $(i,j) \in E$ and handling them through node $j \in N$, given the Lagrange multipliers λ_j, $j \in N$.

The Lagrangian dual problem (23)–(30) can be solved relatively easily because it decomposes into $|A|$ constrained shortest-path problems, one for each OD pair. A constrained shortest-path problem finds a path from an origin $k \in N$ to a destination $l \in N$ with the smallest cost that has total time no greater than τ_{kl}. Although the constrained shortest-path problem is NP-complete, it can be used as part of an algorithm for the package-routing problem because there are several algorithms that solve large enough problems in good computer running times (Ahuja et al. 1993; Desrosiers et al. 1995).

A heuristic algorithm for the package-routing problem is presented next, based on the previous analysis. It is applied to the network described before that includes only one link between a pair of sorting operations representing the cheaper mode if the times h_{ij} for both modes are the same, and two links, one for each mode, if the times are different. A more detailed (but still high-level) description of each step follows the algorithm.

9.7. Package-Routing Heuristic Algorithm

1. Find the OD pairs that have a unique feasible path from their origin to their destination using a *k*-shortest path algorithm. Route all the packages with unique shortest paths along the obtained paths, remove them from the package-routing problem, and update all the parameters.
2. Solve a constrained shortest path for each OD pair using the costs $\bar{c}_{ijkl} = c_{ijkl} + \lambda_j v_{kl}$, where the Lagrange multipliers are set to zero for the first iteration.
3. Reroute packages to (i) eliminate split paths from each sorting operation to a common destination, (ii) eliminate split paths between different modes that have different times on links so that only one mode is used, (iii) enforce the capacity constraint, and (iv) enforce the constraint on the number of loading doors for each sorting operation.
4. If the solution can be improved and the limit on the number of Lagrangian iterations is not reached, compute new costs $\bar{c}_{ijkl}$ using subgradient optimization or some other methodology and go to step 2.
5. Reroute packages from underutilized links to consolidate loads while preserving solution feasibility.
6. Reroute packages to close the underutilized sorting operation that is permitted to be closed and realizes the most savings. If a sorting operation is closed, disable the node representing it and start over from step 2.

In step 1, a *k*-shortest path algorithm (Minieka 1978) for $k = 2$ is used to obtain all the OD pairs that have unique feasible paths. The packages of these OD pairs are routed along their single paths and removed from the problem, updating all the parameters.

Steps 2–4 implement the Lagrangian relaxation algorithm. In step 2, a constrained shortest-path problem is solved for each OD pair using the costs $\bar{c} = c_{ijkl} + \lambda_j v_{kl}$. For the first iteration, the original link costs are used that result from setting the Lagrange multipliers to zero.

Step 3 takes care of the split-paths, capacity, and loading-door constraints. Each node of the network is examined if it satisfies the split-paths constraints (11). If multiple paths to the same destination exist starting from the node, the best path is selected based on several criteria and all the packages from the node to the common destination are routed along this path. Also, if packages are split between two parallel links of different modes, they are rerouted along only one parallel link. Each node is also examined to find whether it satisfies the capacity constraints (8). For each node that surpasses the permitted capacity, OD pairs that use the node are selected according to several criteria and their packages are removed until the capacities are satisfied. The removed packages are routed again sequentially using the constrained shortest-path algorithm on a network where the nodes that have reached their capacities are disabled. The split-paths constraints and the capacity constraints are examined repeatedly until they are satisfied or the algorithm indicates that it cannot find a feasible solution. Finally, each node is examined if it satisfies constraint (14) on the number of loading doors. If a node is found that violates the door constraint, packages are rerouted to sequentially reduce the number of doors but keep the capacity and split constraints valid.

In step 4, if the solution can be improved and more Lagrangian iterations are permitted, new Lagrange multipliers are computed using subgradient optimization (Ahuja et al. 1993; Crowder 1976) or some other methodology. These are used to obtain new costs $\bar{c}_{ijkl}$ and a new Lagrangian iteration starts at step 2. No more details are given here about the Lagrangian relaxation or the subgradient optimization algorithm. A Lagrangian relaxation approach is used to solve the trailer-assignment problem and a more detailed description of this optimization procedure is given in that section. When the Lagrangian iterations are completed, the solution satisfies the split-paths, capacity, and door constraints. If at any point the heuristic cannot obtain a solution that satisfies a constraint, it stops and indicates that it cannot find a feasible solution.

In step 5, links are found that carry too few packages for a complete load and the packages are rerouted so that the capacity, split-paths, and door constraints continue to be satisfied. These constraints are included to improve the results of the trailer-assignment problem that is solved after the package-routing problem and that represents the true cost of a network design solution.

Finally, in step 6, underutilized sorting operations are examined to implement constraints (9). The packages of underutilized sorting operations are rerouted and the underutilized sorting operation for which the total cost of the solution decreases most is eliminated. The node representing the eliminated sorting operation is disabled and the algorithm starts over from step 2.

The routing decisions obtained by the package-routing heuristic algorithm outlined above are used as input in the trailer-assignment problem that is described next.

9.8. Trailer-Assignment Problem

Given the number of packages on each link obtained by solving the package-routing problem, the trailer-assignment problem determines the number and type of trailer combinations on each link of the network $G(N,E)$ that have enough combined capacity to transport all the packages on the link, balance by trailer type for each building, and have the least cost.

The trailer-assignment problem is described next. To solve the problem more efficiently, the network $G(N,E)$ can be modified into the network $G'(N',E')$ as follows. Each node $i \in N'$ represents a building, that is, all the nodes representing sorting operations in the same building are collapsed into one node. All the links that carry packages in the solution of the package-routing problem are included. Among the links in $G(N,E)$ that do not carry packages, only those that may be used to carry empty combinations for balancing are included so that $E' \subseteq E$. In particular, among parallel links between buildings that do not carry packages, only one is retained in G'. Still, the network G' generally has several parallel links between each pair of nodes, in which case link (i,j) is not unique. To avoid complicating the formulation and because it is easy to extend it to include parallel links, we will not include indexing to indicate parallel links, exactly as we did in formulation (3)–(20).

The trailer-assignment problem is formulated below on the network $G'(N',E')$ using constraints (12), (13), and (16) as well as the objective function (3) with some modifications.

9.9. Trailer-Assignment Formulation

$$\text{Min } Z(\mathbf{x}) = \Sigma_{(i,j)\in E'} \Sigma_{m\in M_{ij}} c_{ijm}\, x_{ijm} \tag{31}$$

subject to

$$\Sigma_{i\in N'} \Sigma_{m\in M_{ij}} \delta_{qm}\, x_{ijm} - \Sigma_{p\in N'} \Sigma_{m\in M_{ij}} \delta_{qm}\, x_{jpm} = 0 \qquad \forall\, j \in N',\ q \in F \tag{32}$$

$$\Sigma_{m\in M_{ij}} k_m\, x_{ijm} \geq \overline{v}_{ij} \qquad \forall\, (i,j) \in E' \tag{33}$$

$$x_{ijm} \geq 0 \text{ and integer} \qquad \forall\, i,j,m \tag{34}$$

Objective function (31) is the same as objective function (3). It does not include the first component because it is a constant since the values of the y_{ijkl} variables are known from the solution of the package-routing problem. Constraints (32) are the same as constraints (12) except that the summation over buildings is now unnecessary because a node represents a building. Constraints (33) are similar to constraints (13) where, as in the objective function, the number of packages, $\overline{v}_{ij}$, for all OD pairs on link (i,j) is now a constant. The variables w_{ijq} are not needed anymore and only the variables x_{ijm} are used, representing combinations with both full and empty trailers. So the trailer capacities k_q for $q \in Q$ are replaced by the trailer-combination capacities k_m for $m \in M_{ij}$.

The integer-programming (IP) problem (31)–(34) is still difficult to solve. If constraints (32) are removed, the problem without balancing is easy to solve because it breaks into many very small problems, one for each link. To take advantage of this characteristic, Lagrangian relaxation (Ahuja et al. 1993; Fisher 1985; Geoffrion 1974) is used to solve problem (31)–(34), dualizing the balancing constraints (32). A Lagrange multiplier λ_{jq}, unrestricted in sign, is used for each balancing constraint and the following Lagrangian dual problem is obtained.

$$\begin{aligned} \text{Min } Z(\mathbf{x},\lambda) &= \Sigma_{(i,j)\in E'} \Sigma_{m\in M_{ij}} c_{ijm}\, x_{ijm} \\ &\quad + \Sigma_{j\in N'} \Sigma_{q\in F} \lambda_{jq} \left(\Sigma_{i\in N'} \Sigma_{m\in M_{ij}} \delta_{qm}\, x_{ijm} - \Sigma_{p\in N'} \Sigma_{m\in Mij} \delta_{qm}\, x_{jpm}\right) \\ &= \Sigma_{(i,j)\in E'} \Sigma_{m\in M_{ij}} \left(c_{ijm} + \Sigma_{q\in F} \delta_{qm} (\lambda_{jq} - \lambda_{iq})\right) x_{ijm} \\ &= \Sigma_{(i,j)\in E'} \Sigma_{m\in M_{ij}} \overline{c}_{ijm}\, x_{ijm} \end{aligned} \tag{35}$$

subject to

$$\Sigma_{m\in M_{ij}} k_m\, x_{ijm} \geq \overline{v}_{ij} \qquad \forall\, (i,j) \in E' \tag{36}$$

$$x_{ijm} \geq 0 \text{ and integer} \qquad \forall\, i,j,m \tag{37}$$

where $\bar{c}_{ijm} = c_{ijm} + \Sigma_{q \in F} \delta_{qm} (\lambda_{jq} - \lambda_{iq})$ is the modified cost of moving trailer-combination type $m \in M_{ij}$ on link $(i,j) \in E'$ given the vector $\boldsymbol{\lambda}$.

Problem (35)–(37) decomposes into $|E'|$ subproblems, one for each link $(i,j) \in E'$. Each one of the subproblems is a kind of integer reverse knapsack problem, similar to the integer knapsack problem (Martello and Toth 1990; Nemhauser and Wolsey 1988; Chvátal 1983) and can be solved by similar algorithms. Each subproblem is very small, having $|M_{ij}|$ variables for link $(i,j) \in E'$. The solution obtained by solving the Lagrangian dual (i.e., all the reverse knapsack problems) does not necessarily balance trailer combinations at each node even for optimal $\boldsymbol{\lambda}$ and is not generally feasible for the original problem (31)–(34). A feasible solution is obtained heuristically by solving sequentially one minimum-cost-flow problem (Ahuja et al. 1993) for each trailer type that needs to balance. Balancing is achieved by adding empty trailers or replacing one trailer type with another one of larger capacity that is not yet balanced or does not need to balance.

A Lagrangian relaxation heuristic algorithm that solves the Lagrangian dual problem (35)–(37) is presented next. It uses subgradient optimization to compute the Lagrange multipliers $\boldsymbol{\lambda}$.

9.10. Lagrangian Relaxation Algorithm for the Trailer-Assignment Problem

1. Set the Lagrange multipliers $\boldsymbol{\lambda}$ to zero in the Lagrangian dual problem and initialize $\bar{Z}$ (upper bound, best known feasible solution of the original problem) to a high value.
2. Solve the Lagrangian dual problem with the latest values of $\boldsymbol{\lambda}$ (by solving a set of integer reverse knapsack problems) to obtain the optimal objective function value, $Z^*(\mathbf{x}^*, \boldsymbol{\lambda})$, for the given $\boldsymbol{\lambda}$.
3. Apply a heuristic approach to obtain a feasible solution of the problem and update $\bar{Z}$.
4. If the gap between $\bar{Z}$ and $Z^*(\mathbf{x}^*, \boldsymbol{\lambda})$ is small or some other criterion is satisfied (e.g., a set number of iterations is reached or no more improvement is expected), stop.
5. Compute new values of the Lagrange multipliers $\boldsymbol{\lambda}$ using subgradient optimization and go to step 2.

Step 3 above may be implemented only occasionally instead of at every iteration. Improvement heuristics can also be used to modify the solution in several ways. They can be used to combine single trailers into more efficient trailer combinations. They can also be used to find any cycles of trailer types that are either empty or can be replaced by trailer types that do not need to balance. If a whole cycle of trailers that results in improved cost is modified, balancing of trailers is maintained. Improvement heuristics can also be used to replace whole cycles of trailers of excess capacity with trailers of smaller capacity if the total cost is decreased.

A subgradient optimization algorithm (Ahuja et al. 1993; Crowder 1976) is used in step 5 to compute an improved Lagrange multiplier vector and is described below.

9.11. Subgradient Optimization Algorithm

Given an initial Lagrange multiplier vector $\boldsymbol{\lambda}^0$, the subgradient optimization algorithm generates a sequence of vectors $\boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2, \ldots$ If $\boldsymbol{\lambda}^k$ is the Lagrange multiplier already obtained, $\boldsymbol{\lambda}^{k+1}$ is generated by the rule

$$t_k = \frac{\rho_k(\bar{Z} - Z^*(\mathbf{x}^*, \boldsymbol{\lambda}^k))}{\Sigma_j \Sigma_q (\Sigma_i \Sigma_m \delta_{qm} x^k_{ijm} - \Sigma_p \Sigma_m \delta_{qm} x^k_{jpm})^2} \tag{38}$$

$$\boldsymbol{\lambda}^{k+1}_{jq} = \lambda^k_{jq} + t_k (\Sigma_i \Sigma_m \delta_{qm} x^k_{ijm} - \Sigma_p \Sigma_m \delta_{qm} x^k_{jpm}) \qquad \forall\, j \in N',\, q \in F \tag{39}$$

where t_k is a positive scalar called the step size and ρ_k is a scalar that satisfies the condition $0 < \rho_k \leq 2$. The denominator of equation (38) is the square of the Euclidean norm of the subgradient vector corresponding to the optimal solution vector $\mathbf{x}^k$ of the relaxed problem at step k. Often a good rule for determining the sequence ρ_k is to set $\rho_0 = 2$ initially and then halve ρ_k whenever $Z^*(\mathbf{x}^*, \boldsymbol{\lambda}^k)$ has not increased in a specific number of iterations. The costs are calculated with the new values of $\boldsymbol{\lambda}$. If any negative costs are obtained, the value of ρ is halved and the values of $\boldsymbol{\lambda}$ recomputed until all the costs are nonnegative or ρ is too small to continue iterating.

A feasible solution is obtained in step 3, using a minimum-cost-flow algorithm (Ahuja et al. 1993) sequentially for each trailer type that needs to balance. First, for each building the excess or deficit of trailers of each type that has to balance is computed. Then a minimum-cost-flow algorithm is applied for each trailer type that obtains the optimal movements of trailers from each node of excess to each node of deficit. This may be the movement of a single empty trailer, the movement of an empty trailer that gets attached to an already used trailer to make up a permitted combination, or the movement of an already used trailer replacing another trailer type of smaller or equal capacity that is not yet balanced or does not need to balance.

Lagrangian relaxation is often used within a branch-and-bound procedure. The exact branch-and-bound algorithm has large computational cost; also, the trailer-assignment problem is only part of a heuristic algorithm for the network design problem. For these reasons, the Lagrangian relaxation algorithm is applied only once, at the root of the branch-and-bound tree, to find a heuristic solution to the trailer-assignment problem.

9.12. Extensions of the Network-Design Problem

If the results of the network-design problem are intended not only for long-term planning but to actually modify the routing network, the solution must represent an incremental change from the currently used solution. Otherwise, any savings from an improved solution may be lost in modifying the sorting operations to accommodate the solution changes. To this end, both the package-routing and the trailer-assignment problem can be modified relatively easily to handle presetting some of the variables. For the package-routing problem, this means presetting whole or partial paths of specific OD pairs. A solution is obtained by preprocessing the input data to eliminate or modify OD pairs that are completely or partially preset and updating the input data. Similarly, for the trailer-assignment problem, particular combinations on links may be preset. After appropriate variables are fixed, the Lagrangian relaxation algorithm optimizes the remaining variables.

The network-design problem presented considers only packages moving on the ground and chooses only one transportation mode when travel times differ between sorting operations. The network-design problem can be extended to include all modes of transportation and all types of products with different levels of service requirements. Different modes may have different costs and travel times between sorting operations, permitting parallel use of modes with different travel times along the same routes. This extension complicates considerably an already difficult problem and is not described here any further.

10. DRIVER SCHEDULING

10.1. Tractor-Trailer-Driver Schedules

This section examines one approach for solving the tractor-trailer-driver-scheduling problem for a package-transportation company. Tractor-trailer combinations transport packages between terminals of a package-transportation company, as discussed in Section 9. This involves movement of both equipment and drivers. While balancing is the only constraint for equipment routing, the movement of drivers is more complex. The daily schedule of a tractor-trailer driver starts at his or her base location (domicile) where he or she returns at the end of his workday. A driver schedule consists of one or more legs.

A leg is the smallest piece of work for a driver and consists of driving a tractor-trailer combination from one terminal to another or repositioning trailers inside a terminal. Each leg is characterized by an origin terminal and a destination terminal, which may coincide. There is an earliest availability time at the origin terminal, a latest required arrival time at the destination terminal, and a travel time (or repositioning time) associated with a leg. A tractor-trailer combination that is driven from the origin terminal to the destination terminal of a leg must start no earlier than the earliest availability time at the origin and arrive at the destination no later than the latest required arrival time. At the destination terminal of a leg, a driver may drop his or her current tractor-trailer combination and pick up a new one to continue work on the next leg of his or her daily schedule, take a break, or finish work for the day.

In this section, we assume that the legs are already determined and given. We want to generate driver schedules by combining legs. An acceptable schedule must meet specific work rules, which specify the minimum number of hours that a driver must be paid for a day's work (usually 8 hours) and the maximum length of a workday (usually 10 hours). In addition, a driver must return to his or her domicile every day and a workday must incorporate breaks of specified duration at specified times. For example, a lunch break must last 1 hour and be scheduled between 11:00 am and 2:00 pm.

Another consideration in the generation of driver schedules is the availability of packages for sorting within a terminal. During a sorting operation, loads should arrive at such a rate that the facility is not kept idle. If packages arrive late, the facility will be underutilized during the early stages of sorting while the facility capacity may be surpassed later. For this reason, the duration of each sorting operation is divided into equal time intervals, say, half-hour intervals. Driver schedules need to be generated in such a way that volume availability is satisfied, that is, a minimum number of packages arrives at each sorting facility by the end of each time interval.

The driver-scheduling problem has a short or intermediate planning horizon. It is usually solved regularly once or twice a year and the obtained schedules are bid by the drivers, based on seniority.

The problem may also be solved if changes in volume occur that render the existing schedules inadequate. If decisions about driver schedules are made at the local level and the schedules are bid separately by region, the problem is solved locally in a decentralized fashion.

10.2. Driver-Scheduling Problem

The problem of generating driver schedules can be defined as follows. Given a set of legs with time windows and travel times, generate driver schedules that assign work to drivers so that all the loads are moved within their time windows, the total work assigned to each driver meets given work rules, volume availability meets or exceeds sorting capacities, and the total cost of the schedules is as low as possible.

The problem is defined on an underlying network. The terminals are represented by nodes of the network and each leg is represented by an arc connecting the terminal of origin to the terminal of destination. Time windows on the nodes correspond to the earliest departure and latest arrival times at the terminals.

The cost of a schedule is a combination of both time and distance in addition to fixed costs because it is computed based on driver wages, cost of fuel, and vehicle depreciation. Feasible schedules can be generated using deadheading—that is, a driver can drive a tractor without a trailer to reposition himself or herself for the next leg. Tractor-only movements are used sparingly in a good set of schedules.

The driver-scheduling problem is formulated mathematically as a set-partitioning problem, a special type of integer-programming (IP) problem (Nemhauser and Wolsey 1988; Bradley et al. 1977).

10.2.1 Notation

Parameters

J = set of schedules (columns)
I_{leg} = set of legs (rows)
I_{dom} = set of domiciles (rows)
I_{sort} = set of sort intervals (rows)
c_j = cost of column j
a_{ij} = 1 if column j contains leg i; 0 otherwise
b_{ij} = 1 if column j has $i \in I_{\text{dom}}$ as its domicile; 0 otherwise
k_i^{lo} = minimum number of times domicile i must be used
k_i^{hi} = maximum number of times domicile i can be used
g_{ij} = number of packages contributed by column j to sort interval i
u_i = minimum number of packages required for sort interval i

Decision variables

x_j = 1 if column j is selected; 0 otherwise

10.3. Set-Partitioning Formulation with Side Constraints

$$\text{Min } \Sigma_{j\in J}\, c_j\, x_j \tag{40}$$

subject to

$$\Sigma_{j\in J}\, a_{ij}\, x_j = 1 \qquad \forall \text{ leg } i \in I_{\text{leg}} \qquad \text{(set-partitioning constraints)} \tag{41}$$

$$\Sigma_{j\in J}\, b_{ij}\, x_j \geq k_i^{\text{lo}} \qquad \forall \text{ domicile } i \in I_{\text{dom}} \qquad \text{(lower-domicile constraints)} \tag{42}$$

$$\Sigma_{j\in J}\, b_{ij}\, x_j \leq k_i^{\text{hi}} \qquad \forall \text{ domicile } i \in I_{\text{dom}} \qquad \text{(upper-domicile constraints)} \tag{43}$$

$$\Sigma_{j\in J}\, g_{ij}\, x_j \geq u_i \qquad \forall \text{ sort interval } i \in I_{\text{sort}} \qquad \text{(volume-availability constraints)} \tag{44}$$

$$x_j = 0 \text{ or } 1 \qquad \forall \text{ schedule } j \in J \qquad \text{(binary constraints)} \tag{45}$$

The objective function (40) minimizes the total cost of schedules. Constraints (41) are the set-partitioning constraints ensuring that each leg (row) is used by only one schedule (column). Constraints (42) and (43) are the domicile constraints and ensure that the lower and upper bounds for the selection of domiciles are met. According to work rules, each particular terminal must be used as a domicile a number of times within a given range. Constraints (44) are the volume-availability constraints and ensure that the required volume of packages is available for each sort interval. Constraints (45) are the binary constraints.

The presented formulation is a set-partitioning problem with additional side constraints. For a freight transportation company with 10,000 tractors and 15,000 drivers in the continental United States, the problem formulated above is too large to be solved for the whole country. Often, however, driver-scheduling decisions are made at the local level and the problem is broken naturally into smaller problems that are solved locally by each region.

It is sometimes difficult to obtain even a feasible solution of the IP problems formulated in (40)–(45). If the feasible region contains few feasible solutions or if there are errors in the data that make it difficult or impossible to find a feasible solution, the set-partitioning formulation (40)–(45) can be changed into a set-covering formulation (Nemhauser and Wolsey 1988; Bradley et al. 1977) with soft domicile and volume-availability constraints to help guide the cleaning of data and the solution process. Slack and surplus (auxiliary) variables are added to the equations and inequalities (41)–(45) and incorporated into the objective function (40) with very high costs. The following additional notation is introduced:

d_i = very high cost for auxiliary variable of row i.
s_i^+ = surplus variable for row i; if positive, it indicates the number of units that the original constraint is below its right-hand-side.
s_i^- = slack variable for row i; if positive, it indicates the number of units that the original constraint is above its right-hand-side.

10.4. Set-Covering Formulation with Soft Constraints

$$\text{Min } \Sigma_{j \in J} c_j x_j + \Sigma_{i \in I_{\text{leg}}} d_i s_i^- + \Sigma_{i \in I_{\text{dom}}} d_i s_i^+ + \Sigma_{i \in I_{\text{dom}}} d_i s_i^- + \Sigma_{i \in I_{\text{sort}}} d_i s_i^+ \tag{46}$$

subject to

$$\Sigma_{j \in J} a_{ij} x_j - s_i^- = 1 \qquad \forall \text{ leg } i \in I_{\text{leg}} \qquad \text{(set-covering constraints)} \tag{47}$$

$$\Sigma_{j \in J} b_{ij} x_j + s_i^+ \geq k_i^{\text{lo}} \qquad \forall \text{ domicile } i \in I_{\text{dom}} \qquad \text{(soft lower-domicile constraints)} \tag{48}$$

$$\Sigma_{j \in J} b_{ij} x_j - s_i^- \leq k_i^{\text{hi}} \qquad \forall \text{ domicile } i \in I_{\text{dom}} \qquad \text{(soft upper-domicile constraints)} \tag{49}$$

$$\Sigma_{j \in J} g_{ij} x_j + s_i^+ \geq u_i \qquad \forall \text{ sort interval } i \in I_{\text{sort}} \qquad \text{(soft volume-availability constraints)} \tag{50}$$

$$x_j = 0 \text{ or } 1 \qquad \forall \text{ schedule } j \in J \tag{51}$$

$$s_i^+ \geq 0 \qquad \forall \text{ row } i \tag{52}$$

$$s_i^- \geq 0 \qquad \forall \text{ row } i \tag{53}$$

The high value of the costs d_i of the slack and surplus variables prevents their inclusion in a solution with positive values, if this is possible. Constraints (47) resemble set-partitioning constraints but they can also be considered as set-covering constraints, that penalize overcoverage. If some slack or surplus variables have positive values in a solution, they may help identify reasons for not obtaining a true feasible solution or they may remain in the solution for as long as necessary in an iterative solution process that will be discussed later.

10.5. Column-Generation Methodology

Each one of the regional problem formulations of a large freight transportation company may have up to 2000 legs (2400 rows altogether) that can be combined into possibly billions of feasible schedules. Even if the binary constraints (51) are relaxed, it is impossible to solve the resulting linear program (LP) with all the columns included. Instead, a standard decomposition method for large LPs called column generation is used (Bradley et al. 1977; Chvátal, 1983).

When the binary constraints (51) are replaced with the following bounding constraints

$$0 \leq x_j \leq 1 \qquad \forall \text{ schedule } j \in J$$

a column generation approach refers to the resulting LP problem that includes all the possible columns as the master problem. The corresponding LP problem that includes only a subset of the columns is called the restricted master problem.

Solving the LP involves pricing out each column using the dual variables or shadow prices associated with the restricted master problem. Sometimes this pricing-out operation can be formulated as a known problem (e.g., a shortest-path problem) and is called a subproblem. The solution of the subproblem produces a column, not yet included in the restricted master problem, that prices out best. For a minimization problem, if the best new column obtained has negative reduced cost, its inclusion in the restricted master problem will improve the solution. Column generation iterates solving the subproblem, adding a new column with negative reduced costs to the restricted master problem, and solving the new restricted master problem until no more columns can be obtained with negative reduced costs. At that point the master problem has been solved optimally since all billions of possible schedules have been examined implicitly.

When the generation of columns cannot be structured as a known optimization problem, they must be generated explicitly. Usually, this approach results in solving the master problem heuristically because optimality is guaranteed only if all the feasible columns are examined. The schedules for the set-partitioning or set-covering problems presented above are obtained explicitly. The nature of the breaks imposed by work rules, the possible existence of additional local work rules in some regions, the presence of legs that carry priority loads and need to be included as early as possible in a schedule, and other local characteristics make it very difficult to generate schedules by solving a structured subproblem.

The schedules obtained from the solution of the LP problem by column generation may be fractional and therefore unacceptable. To obtain feasible schedules, an IP problem needs to be solved. An iterative heuristic approach for solving the driver-scheduling problem is presented next.

10.6. Iterative Process for Solving the Driver-Scheduling Problem with Column Generation

1. Start with a feasible solution. Set up an LP that consists of the columns of the feasible solution.
2. Solve the LP (restricted master problem). If the LP cost is low enough or a given number of iterations is reached, go to step 4.
3. Using the LP shadow prices, generate up to a given number of good new schedules and add them to the LP. If the number of columns exceeds a given maximum, remove columns with the worst reduced costs. Go to step 2.
4. Solve the IP.

In step 1, the algorithm starts with a feasible solution, which can be obtained from the currently used set of schedules. These schedules are broken apart to obtain the set of legs for the problem. In step 2, an LP problem is solved that consists of the current set of schedules and shadow prices are obtained for the rows. In step 3, the algorithm generates more schedules with reduced costs less than a set value using the shadow prices. The new schedules are added to the LP, omitting duplicate schedules. Because the LP is a minimization problem, columns with negative reduced costs will reduce the objective function value. More than one column is added to the LP at each iteration. For this reason, columns with low positive reduced costs may also be accepted. Such columns are also valuable in solving the IP problem in step 4. If the total number of columns in the restricted master problem exceeds a preset maximum number, the columns with the worst (highest) reduced costs are deleted. Steps 2 and 3 are repeated until an LP cost is obtained that is below a preset cutoff value or until a given number of iterations is reached. The resulting IP problem is solved in step 4.

In actual applications, the IP problem at step 4 of the previous algorithm obtained from column generation turned out to be a very difficult problem to solve with existing IP solvers. There were several reasons for this difficulty. The optimal LP solution at the beginning of step 4 included too many fractions, the problem structure exhibited massive degeneracy, and there were too many alternative optimal solutions. It was even difficult for IP solvers to obtain feasible solutions for large problems. Because of these difficulties, the following heuristic approach has been used that combines column generation with solution of the IP problem.

10.7. Integrated Iterative Process for Solving the Driver-Scheduling Problem

1. Start with a feasible solution. Set up an LP that consists of the columns of the feasible solution.
2. Solve the LP (restricted master problem). If the LP cost is low enough or a given number of iterations is reached, go to step 4.
3. Using the LP shadow prices, generate up to a given number of good new schedules and add them to the LP. If the number of columns exceeds a given maximum, remove columns with the worst reduced costs. Go to step 2.
4. Select a small number of columns with the highest fractional values, say 8. Using their legs as seeds, generate a small number of additional schedules, say 300, add to the LP, and solve it.
5. Select a small number of the highest fractional schedules, say 8. Restrict them to be integers and solve the resulting mixed-integer-programming (MIP) problem. If all variables in the solution of the current restricted master problem are integral, stop. Otherwise, set the selected columns to their integer solution values permanently, update the formulation, and go to step 2.

Steps 1, 2, and 3 are the same as before. In step 4, a preset number of columns with the highest fractional values are selected. The actual number used is set by experimentation. The legs making up these columns are used as seeds to generate a small number of additional columns that are added to the LP. The LP is solved and a small number of columns with the highest fractional values are selected and restricted to be integral. The resulting MIP problem is solved to optimality using an MIP solver. If all the columns of the restricted master problem have integral values in the solution, the algorithm stops. If some fractional values are still present in the solution, the selected columns to be integral are set permanently to their integer solution values and eliminated from the formulation and the column-generation phase starts again. In applications, up to about 40,000 columns were included in the restricted master problem during the iterative solution process.

10.8. Generation of Schedules

Approaches for generating new schedules are discussed in this section. Several techniques can be used to obtain new schedules explicitly, guided by the LP shadow prices. Each leg is associated with a shadow price in the LP solution, which indicates how expensive it is to schedule this leg. The generation of schedules focuses on the expensive legs to provide more alternative schedules that include them and drive the LP cost down. New schedules improve the LP solution if they have negative reduced costs. Usually the cutoff value is set higher than zero to include schedules with low positive costs that may combine well with the rest of them because columns are not added one at a time to the LP.

New schedules are generated probabilistically. Each available leg is assigned a probability of selection proportional to its shadow price. The list of legs is then shuffled using the assigned probabilities of selection as weights.* This means that legs with high shadow prices are more likely to be at the top of the shuffled list. Each leg in the shuffled list is used as a seed sequentially to generate up to a given number of legs.

Starting with a seed, schedules are generated using three different approaches: one based on depth-first search (Aho et al. 1983), a second approach that generates a given number of schedules in parallel, and a third method that recombines existing schedules to produce new and better ones. The schedules are generated by limited complete enumeration, that is, all the schedules that can be generated starting with a particular seed are generated, limited by an upper bound when too many combinations exist. Tractor-only movements for repositioning drivers are also used in the generation of feasible schedules. A partial schedule is accepted only if a complete feasible schedule can be generated from it, including appropriate work breaks. When a maximum total number of columns is obtained, the process stops.

The depth-first-search approach starts with a seed and adds legs sequentially until a complete schedule is obtained. Then a new schedule is started using either the same seed or the next seed in the list. All the feasible successors of a leg are shuffled based on their shadow prices as weights and used to obtain the next leg of a partial schedule. The parallel approach generates several schedules simultaneously by adding each one of its feasible successors to the current partial schedule.

The recombination approach identifies feasible schedules that are generated by removing one or more consecutive legs from one feasible schedule and replacing them with one or more legs from another schedule. Cycles of leg exchanges are then identified that produce new schedules of lower costs. In actual applications, the recombination approach tends to produce columns that drive the iterative solution process much more quickly toward a good overall result than when only columns obtained by the other two approaches are included.

10.9. Beyond Algorithms

Good algorithms that capture well the complexities of real-world problems are a big step towards achieving efficiency using optimization techniques. They are rarely, however, sufficient by themselves. The perfect algorithm is useless if it is not actually used, if it not used properly, or if the solution is not implemented. This is particularly important in the transportation of goods, which is a labor-intensive industry, and where the implementation of an optimization system may involve and affect a large number of people.

Often, a big challenge, beyond the development of good algorithms, is defining the correct problem to solve, finding the necessary data, setting up tools to extract the needed data, correcting and validating the data, having the model used correctly, and obtaining acceptance of the model and its results. A successful, decentralized application needs the involvement of the users, who must be able and willing to use it correctly and apply the results.

---

* This is exactly like regular shuffling except for the probabilities of selection that are not uniform. An ordered list of legs is obtained by randomly selecting one leg at a time from an original list of legs, based on the weights.

To obtain good results using the algorithms described in this chapter, correct data need to be used. Obtaining good, correct data is often a difficult, expensive, and time-consuming task. Errors in the data need to be identified and corrected easily for the system to succeed. The driver-scheduling problem is especially sensitive to the values of the time windows and cost input. An interface needs to be available that handles the cleaning of the data as well as the generation of the input to the algorithms in an appropriate format and that after a solution is obtained produces useful reports.

If a computer system is put in place for the first time to replace a manual system, the data needs of the computer system may be much larger than those of the manual system. Most optimization models obtain solutions by comparing explicitly or implicitly large numbers of alternatives for which data must be available. Manual systems, on the other hand, often obtain a new solution by modifying an existing one, that is, they examine very few alternatives and need fewer input data. The difference in input data also means that input-data needs for solving an existing application often have to be redefined for computer models. If good input data cannot be obtained for the computer model, its use cannot be expected to improve results.

The algorithms described previously are used to solve difficult MIP problems. Large computers and expensive commercial LP and IP solvers need to be used, which are usually available only in centralized locations. The interface that cleans the data and obtains the input can be implemented locally when the driver-scheduling system is applied separately by region.

A system like the one described has been deployed successfully by United Parcel Service using two different platforms. The data preparation is implemented on local computers available at all company locations. The optimization algorithms are solved remotely on large computers at another location using the company's intranet. The fact that two different platforms are used is hidden from the users, who are informed by e-mail about the status of a submitted job at various stages and get all the reports on the intranet. This kind of application is now possible because of the increase in computer power that permits the solution of difficult optimization problems in reasonable time and because of the evolution of the Internet that provides the tools supporting a remote implementation.

To improve the probability of success, user training is very important for both actual use of the system and interpretation of results. Optimization systems sometimes do not include some difficult characteristics of a real-life problem. In cases like this, interpretation of results, modification of the solution, or simply selection among alternative solutions is very important. A pilot implementation is also very helpful.

An important element for the success of the system is the existence of appropriate support to the users in applying the system, interpreting the results, and making appropriate decisions. This is especially important when the system results differ from current practices and need to be understood and their validity accepted or when they need to be slightly modified. A team of experts in both the computer system and the operational problem that is solved, who are accepted by the users and speak the same business language, needs to be available to assist the users, especially when the system is first deployed.

The optimization model described minimizes total cost of schedules, which often results in a solution using fewer drivers than those used in the current solution. How such a solution is implemented depends on company–labor agreements, which may differ among companies and regions. In the short term, available drivers beyond the number needed by the solution may be put on an ''on-call'' list and asked to come to work only if another driver is sick or a particular need arises.

It is generally very important to get the final users involved early on in the development process. In this way, the developer makes sure that the correct problem is solved and user needs are met as much as possible. Including the user in the development process early on increases the probability of acceptance of the model.

For the implementation of results, one other very significant factor for success concerns the reward structure of a company. If an optimization model minimizes cost but the user is not rewarded directly for implementing a solution that minimizes cost, the solution results are unlikely to be used. If a model minimizes the number of drivers but the decision maker is not rewarded for using fewer drivers, he is unlikely to jeopardize the goodwill of the people working with him by making a big effort to change the status quo.

11. QUALITY IN TRANSPORTATION

Companies that specialize in the transportation of goods must manage their costs, growth, and quality in order to remain competitive. However, without the appropriate measures and the systems to support performance measurement, it is practically impossible to manage any transportation system. In measuring quality in the freight-transportation industry several questions arise. First, of course, is the definition of quality itself. In this industry, quality is viewed differently at different steps in the transportation process. Shippers have different requirements from those of the receivers. Different internal processes have different views of quality and its measurement. However, we can summarize these requirements into five categories:

1. *Damages:* Was the shipment damaged in the process?
2. *On-time performance:* Were all service guarantees met?
3. *Accuracy:* Was the shipment delivered to the correct destination?
4. *Shipment integrity:* Were all the items in a shipment delivered together?
5. *Information integrity:* Was the information associated with a shipment available at all times?

The primary objective of the transportation-planning activity is to design processes that maintain high levels of performance in all five categories. Let's explore these requirements further:

- *Damages:* Of all the quality factors discussed, damages are perhaps one of the most important indicators of quality in both the receiver's and the shipper's view. When the freight-transportation company damages the merchandise being moved, the shipper, the receiver, and the transportation company itself are affected. Costs associated with insurance, returns, product replacement, and lost productivity are all a result of damaged goods. All transportation processes must be designed to prevent damaging the goods. Usually, every step in the transportation process has procedures to measure the number of damages created in any period of time. These procedures are used to establish accountability practices and to identify process-improvement opportunities.
- *On-time performance:* Freight-transportation companies compete on the basis of service performance and cost. In order to support the needs of the complex supply chains that exist today, high levels of on-time delivery reliability are expected from the transportation company. Several external and internal factors can have a direct impact on on-time delivery performance. External factors such as weather, traffic conditions, and subcontractor labor relations can have a direct impact on the ability of the transportation company to meet service commitments. With proper planning, transportation companies manage to minimize the impact of some of these factors. On the other hand, the one internal factor that will always have a negative impact on on-time delivery is lack of planning or, quite simply, poor planning. If the organization is not prepared to handle seasonal variations in pickup and delivery volumes or does not have contingency plans to deal with unexpected events, on-time delivery performance will be affected.
- *Accuracy:* Delivering to the correct destination is expected every time for every shipment. However, there are instances in which the transportation system fails to satisfy this basic requirement. Two major causes contribute to this type of service failure: missing or incorrect information and inadequate planning. For example, when the wrong address is attached to a shipment, the probability of making delivery mistakes increases substantially. Today, transportation companies offer a variety of services aimed at providing continuous shipment tracking and improved information quality. As indicated before, labor is usually the highest cost variable in the profitability equation of a transportation company. Labor is also the primary driver of quality in these organizations. When companies fail to develop staffing and training plans properly, delivery accuracy and reliability will be impacted.
- *Shipment integrity:* Receivers expect to receive all the items (e.g., packages) in a shipment on the same day and at the same time. It is the responsibility of the freight-transportation company to maintain the integrity of all shipments. The transportation system must be capable of using shipment information in its different processes to ensure the integrity of every shipment.
- *Information integrity:* As indicated earlier in this chapter, the information about a shipment is as important, in today's supply chains, as the movement of the shipment itself. Since shipment information is offered to shippers and receivers as a value-added service, the effectiveness with which this information is provided to them must be monitored and measured. Systems to effectively capture, store, and provide shipment information are critical in today's freight transportation business models. The freight-transportation industry will continue to be an information-based industry. Therefore, maintaining high levels of information accuracy and integrity will continue to be an important measure of performance.

Now that some basic measures have been defined, let's look into the quality-improvement process. Like other industries, freight-transportation companies make use of well-established quality-improvement techniques. Without a clear understanding of the facts and figures that affect the quality of the services offered, management cannot control and improve the processes involved. Remember that there are four continuous phases in the quality improvement process: Plan, Do, Check, and Act. Together, these phases are known as the Deming circle. Transportation processes must be designed and performance targets must be defined. As activities are completed throughout the different processes, regular checks must take place as the processes are monitored and controlled. The information gathered from these checks must be used to improve the process continuously.

12. TECHNOLOGY

12.1. Vehicle Routing

For solving transportation problems, the use of computers plays an important role in the development of models, schedules, network plans, and delivery and pickup routes. As described in previous sections, the complexity and magnitude of the transportation problems discussed in this chapter require extensive computation times. Making transportation decisions is a complex process. Dispatch managers in package-delivery companies must assess a variety of factors before making dispatch decisions. Some of those factors were discussed in previous sections. They include vehicle capacity, time windows, demand fluctuations, labor productivity, and the dynamic changes in pickup and delivery characteristics of customers, geographies, and traffic conditions.

Vehicle-routing problems fall into one of three basic segments: routing of service vehicles, passenger vehicles, and freight vehicles (Hall and Partyka 1997). Service vehicles are used to support jobs in the field and are generally used by service technicians. In this type of problem, the primary constraints are service time, time windows, and travel time. Because there is no major variation in the merchandise carried by any given vehicle, capacity constraints are not included in the formulation of service routes. Passenger-transportation services such as bus service face an additional constraint: capacity. The size of the vehicle determines the number of passengers that can be safely carried from one point to another. Freight vehicles are also constrained by their capacity. When completing pickups, a vehicle may run out of capacity at a certain customer location. At this point, the dispatcher must either dispatch another vehicle to the customer's location to complete service or ask the current driver to return to the depot, obtain an empty vehicle, and return to the customer's location to complete the service. It is clear from this example that the decision made by the dispatcher can have different cost and service implications and may affect more than one customer. A few years ago, the only way to make these dispatch decisions effectively was through human knowledge and experience. The ability of the dispatcher to make numerous decisions based on a few pieces of information could make or break the dispatching process. Today, things have changed. With the increasing implementation of technologies such as the geographic information systems (GIS) and global-positioning systems (GPS), dispatchers have new tools that automate and improve the vehicle-routing process.

Routing pickup-and-delivery vehicles does not end with the development of routes prior to the drivers' departure from the depot. Once drivers have left the depot, in-vehicle communications and route-information systems offer mechanisms not only to improve their performance but to meet on-demand customer requests. When dispatchers have the ability to communicate routing instructions and customer requests to drivers in the field, the opportunities for improving the overall efficiency of a dispatch plan increase substantially. However, the initial development of an efficient and effective dispatch plan is still critical.

Several software vendors have developed vehicle-routing software. In 1997, Hall and Partyka surveyed several vendors in order to compare the characteristics of their respective software systems. Table 4 presents an extract of this survey. The complete survey appeared in the June 1997 issue of *OR/MS Today.*

12.2. Information Gathering and Shipment Tracking

As indicated earlier in this chapter, the information associated with the goods being moved is as important as the transportation process itself. Today, transportation companies use a variety of tools to track, manage, and control the movement of goods from pickup to delivery. In addition, sophisticated electronic devices are being used by drivers not only to record the status of deliveries and pickups, but also to keep track of vehicle usage, time cards, and sales information.

12.3. New Trends: Intelligent Transportation Systems (ITS)

The transportation community has turned to the deployment of intelligent transportation systems (ITS) to increase the efficiency of existing highway, transit, and rail systems. One of the key variables in the vehicle-routing models described above is travel time. With the use of information from ITS, dispatchers can make better decisions. The U.S. Department of Transportation (DOT) has indicated that "ITS uses advanced electronics and information technologies to improve the performance of vehicles, highways, and transit systems. ITS provides a variety of products and services in metropolitan and rural areas."

As ITS evolves from pure research, limited prototyping, and pilot projects into routine usage, decision makers at "the corporate, state, regional, and local levels seek reliable information about the contribution that ITS products can make toward meeting the demand for safe and efficient movement of people and goods." Literature indicates that substantial benefits have already been realized in areas such as accident reduction, travel-time savings, customer service, roadway capacity, emission reduction, fuel consumption, and vehicle stops. Greater benefits are predicted with more extensive

TABLE 4 Routing Software Survey

Product	Publisher	Solvable Problem Size			Routing			GIS Product Interface	Special Features
		Number of Stops	Number of Vehicles	Number of Terminals	Real-Time Routing	Daily Routing	Route Planning		
GeoRoute	Kositzky & Associates, Inc.	4600	512	256	N	Y	Y	GeoWhiz	GeoRoute works for local delivery as well as over-the-road applications. Options for multidepot and redisptach are available.
GeoRoute 5	GIRO Enterprises, Inc.	Unlimited	Unlimited	Unlimited	N	Y	Y	ArcInfo, MapInfo	Software supports both point-to-point and street-by-street operations, as well as mixed requirements.
Load Manager	Roadnet Technologies, Inc.	N/A	N/A	N/A	N	N	N		
LoadExpress Plus	Information Software, Inc.	Unlimited	500	Unlimited	N	Y	Y	Proprietary, can work with data from all GIS systems	LoadExpress is a simple, powerful, and flexible choice for building and optimizing routes, scheduling deliveries, and analyzing distirbution patterns.
Manugistics Routing & Scheduling	Manugistics, Inc.	Unlimited	Unlimited	Unlimited	Y	Y	Y	Xeta, Rockwell, Cadec, Autoroach	Resource management—allows management of driver, tractor and trailer schedules, provides information on equipment requirements.
OVERS	Bender Management Consultants	10,000	1000	100	Y	Y	Y	Map Objects	Can optimize routes across multiple time periods, respect space/time constraints, optimize number and location of terminals and service areas.
RIMMS	Lightstone Group, Inc.	Unlimited	Unlimited	Unlimited	Y	Y	Y	ESRI shape files	Configurable by users across multiple industries, including both scheduling model and screen cosmetics. Interfaces via ODBC drivers.

ROADNET 5000	Roadnet Technologies, Inc.	Unlimited	Unlimited	Unlimited	N	Y	N	GDT maps	
RoadShow for Windows	ROADSHOW International, Inc.	8000	Unlimited	Unlimited	Y	Y	Y	GDT, Etak, MapInfo and proprietary software	ROADSHOW calculates cost-effective solutions based on actual costs incorporating specific information supplied by the user.
RoutePro	CAPS LOGISTICS	HW-based	HW-based	HW-based	Y	Y	Y	Etak, Horizons Technology, GDT, PCMiler	Customizable through fourth-generation macro language and ability to call functions from other languages.
RouteSmart Neighborhood	RouteSmart Technologies	Unlimited	Unlimited	1+ Intermediates	N	N	Y	GIS Plus (DOS version), ArcInfo	Meter-reading system, handles walking, driving, and combination routes.
RouteSmart Point-to-Point	RouteSmart Technologies	Unlimited	Unlimited	1+ Intermediates	N	Y	Y	ArcView version 3.0	
Routronics 2000	Carrier Logistics	Unlimited	Unlimited	Unlimited	N	Y	N	MapInfo	Routronics 2000 has been developed as a complete customer-service routing and dispatch system with interfaces to wireless communications.
SHIPCONS II	Insight, Inc.	Unlimited	Unlimited	Unlimited	N	Y	Y	GDT, MapInfo, Etak	Cost-based, integer optimization; user-configurable screens; Ad Hoc Report Writer; Digital Geography; Shipment Rater for TL and LTL carriers.
Taylor II	F&H Simulations, Inc.	1000	100	1000	N	N	Y		2 and 3D animation; Windows 95 and Windows NT; design of experiments; curve fitting of raw data (advanced statistics module).
Territory Planner	Roadnet Technologies	Unlimited	Unlimited	Unlimited	N	N	Y	GDT maps	
TESYS	Inform Software Corporation	3000	1000	500	Y	Y	N	CDPD/GPS/RF	
TransCAD	Caliper Corporation	Unlimited	Unlimited	Unlimited	Y	Y	Y	Maptitude, GIST, can work with data from all GIS systems	Toolkit of OR methods including min-cost network flow, transportation problem, and various optimal location methods.

Source: R. W. Hall and J. G. Partyka, “On the Road to Efficiency,” *OR/MS Today,* June 1997, pp. 38–47. Reprinted with permission.

deployment of more mature products. Freight-transportation companies face new constraints and challenges not only in meeting service commitments but in remaining competitive and cost effective while meeting governmental regulations. The use of ITS offers new opportunities to use information in the development of routes and schedules.

The ITS program is sponsored by the DOT through the ITS Joint Program Office (JPO), the Federal Highway Administration (FHWA), and the Federal Transit Administration (FTA).

ITS, formerly known as the intelligent vehicle highway systems (IVHS), were created after the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991 was established. ISTEA helped authorize larger spending for transit improvement. In January 1996, then Secretary of Transportation Frederico Peña launched Operation TimeSaver, which seeks to install a metropolitan intelligent transportation infrastructure in 75 major U.S. cities by 2005 to electronically link the individual intelligent transportation systems, sharing data so that better travel decisions can be made.

A projected $400 billion will be invested in ITS by the year 2011. Approximately 80% of that investment will come from the private sector in the form of consumer products and services.

The DOT has defined the following as the components of the ITS infrastructure:

- *Transit fleet management:* enables more efficient transit operations, using enhanced passenger information, automated data and fare collection, vehicle diagnostic systems, and vehicle positioning systems
- *Traveler information:* linked information network of comprehensive transportation data that directly receives transit and roadway monitoring and detection information from a variety of sources
- *Electronic fare payment:* uses multiuse traveler debit or credit cards that eliminate the need for customers to provide exact fare (change) or any cash during a transaction
- *Traffic signal control:* monitors traffic volume and automatically adjusts the signal patterns to optimize traffic flow, including signal coordination and prioritization
- *Freeway management:* provides transportation managers the capability to monitor traffic and environmental conditions on the freeway system, identify flow impediments, implement control and management strategies, and disseminate critical information to travelers
- *Incident management:* quickly identifies and responds to incidents (crashes, breakdowns, cargo spills) that occur on area freeways or major arteries
- *Electronic toll collection:* uses driver-payment cards or vehicle tags to decrease delays and increase roadway throughput
- *Highway–rail intersection safety systems:* coordinates train movements with traffic signals at railroad grade crossings and alerts drivers with in-vehicle warning systems of approaching trains
- *Emergency response:* focuses on safety, including giving emergency response providers the ability to pinpoint quickly the exact location of an incident, locating the nearest emergency vehicle, providing exact routing to the scene, and communicating from the scene to the hospital

The use of information from all of these system components will enhance the planner's ability in designing efficient transportation networks and delivery routes. In addition, as this information is communicated to the drivers, they will also have the capability of making better decisions that will enhance customer satisfaction and reduce overall costs.

For additional information, visit the DOT's website on ITS: http://www.its.dot.gov/.

Acknowledgments

The authors wish to thank Professor Mark S. Daskin of Northwestern University and Professor Ching-Chung Kuo of Pennsylvania State University for their constructive comments. They also thank their colleagues Ranga Nuggehalli, Doug Mohr, Hla Hla Sein, Mark Davidson, and Tai Kim for their assistance. They also thank Dr. Gerald Nadler of the University of Southern California for all his support.

REFERENCES

Aho, A. V., Hopcroft, J. E., and Ullman, J. D. (1983), *Data Structures and Algorithms*, Addison-Wesley, Reading, MA.

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993), *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall, Englewood Cliffs, NJ.

Bartholdi, J., and Platzman, L. (1988), "Heuristics Based on Spacefilling Curves for Combinatorial Problems in Euclidean Space," *Management Science*, Vol. 34, pp. 291–305.

Bertsimas, D., Jaillet, P., and Odoni, A. (1990), "A Priori Optimization," *Operations Research*, Vol. 38, pp. 1019–1033.

Bradley, S. P., Hax, A. C., and Magnanti, T. L. (1977), *Applied Mathematical Programming*, Addison-Wesley, Reading, MA.

Chvátal V. (1983), *Linear Programming*, W.H. Freeman, New York.

Crowder, H. (1976), "Computational Improvements for Subgradient Optimization," *Symposia Mathematica*, Vol. 19, pp. 357–372.

Daskin, M. S. (1995), *Network and Discrete Location: Models, Algorithms, and Applications*, John Wiley & Sons, New York.

Desrosiers, J., Dumas, Y., Solomon, M. M., and Soumis, F. (1995), "Time Constrained Routing and Scheduling," in *Network Routing*, Vol. 8 of *Handbooks in Operations Research and Management Science,* M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, Eds., North-Holland, Amsterdam, pp. 35–139.

Fisher, M. L. (1985), "An Applications Oriented Guide to Lagrangian Relaxation," *Interfaces*, Vol. 15, No. 2, pp. 10–21.

Fisher, M., and Jaikumar, R. (1981), "A Generalized Assignment Heuristic for Vehicle Routing," *Networks*, Vol. 11, pp. 109–124.

Garey, M., and Johnson, D. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York.

Geoffrion, A. M. (1974), "Lagrangian Relaxation for Integer Programming," *Mathematical Programming Study*, Vol. 2, pp. 82–114.

Gillet, B., and Miller, L. (1974), "A Heuristic Algorithm for the Vehicle Dispatching Problem," *Operations Research*, Vol. 22, pp. 340–349.

Glover, F. (1989), "Tabu Search—Part I," *ORSA Journal on Computing*, Vol. 1, pp. 190–206.

Glover, F. (1990), "Tabu Search—Part II," *ORSA Journal on Computing*, Vol. 2, pp. 4–32.

Hall, R. W., and Partyka, J. G. (1997), "On the Road to Efficiency," *OR/MS Today*, June, pp. 38–47.

Hall, R., Du, Y., and Lin, J. (1994), "Use of Continuous Approximations within Discrete Algorithms for Routing Vehicles: Experimental Results and Interpretation," *Networks*, Vol. 24, pp. 43–56.

Jaillet, P. (1988), "A Priori Solution of a Traveling Salesman Problem in Which a Random Subset of the Customers Are Visited," *Operations Research*, Vol. 36, pp. 929–936.

Jaillet, P., and Odoni, A. (1988), "The Probabilistic Vehcile Routing Problem," in *Vehicle Routing: Methods and Studies*, B. Golden and A. Assad, Eds., North-Holland, Amsterdam, pp. 293–318.

Kontoravdis, G., and Bard, J. (1995), "A GRASP for the Vehicle Routing Problem with Time Windows," *ORSA Journal on Computing*, Vol. 7, pp. 10–23.

Lawler, E., Lenstra, J., Rinnooy Kan, A., and Shmoys, D., Eds. (1985), *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*, John Wiley & Sons, New York.

Lin, S., and Kernighan, B. (1973), "An Effective Heuristic Algorithm for the Traveling Salesman Problem," *Operations Research*, Vol. 21, pp. 498–516.

Martello, S., and Toth, P. (1990), *Knapsack Problems: Algorithms and Computer Implementations*, John Wiley & Sons, New York.

Minieka, E. (1978), *Optimization Algorithms for Networks and Graphs*, Marcel Dekker, New York.

Nemhauser, G. L., and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization*, John Wiley & Sons, New York.

Potvin, J., and Rousseau, J. (1993), "A Parallel Route Building Algorithm for the Vehicle Routing and Scheduling Problem with Time Windows," *European Journal of Operational Research*, Vol. 66, pp. 331–340.

Potvin, J., and Rousseau, J. (1995), "An Exchange Heuristic for Routing Problems with Time Windows," *Journal of the Operational Research Society*, Vol. 46, pp. 1433–1446.

Potvin, J., Kervahut, T., Garcia, B., and Rousseau, J. (1996), "The Vehicle Routing Problem with Time Windows—Part I: Tabu Search," *INFORMS Journal on Computing*, Vol. 8, pp. 158–164.

Rochat, Y., and Taillard, E. (1995), "Probabilistic Diversification and Intensification in Local Search for Vehicle Routing," *Journal of Heuristics*, Vol. 1, pp. 147–167.

Russell, R. (1995), "Hybrid Heuristics for the Vehicle Routing Problem with Time Windows," *Transportation Science*, Vol. 29, pp. 156–166.

Savelsbergh, M. (1985), "Local Search in Routing Problems with Time Windows," *Annals of Operations Research*, Vol. 4, pp. 285–305.

Savelsbergh, M. (1990), "An Efficient Implementation of Local Search Algorithms for Constrained Routing Problems," *European Journal of Operational Research*, Vol. 47, pp. 75–85.

Savelsbergh, M. (1992), ‘‘The Vehicle Routing Problem with Time Windows: Minimizing Route Duration,’’ *ORSA Journal on Computing*, Vol. 4, pp. 146–154.

Solomon, M. (1987), ‘‘Algorithms for the Vehicle Routing and Scheduling Problems with Time Window Constraints,’’ *Operations Research*, Vol. 35, pp. 254–265.

Solomon, M., Baker, E., and Schaffer, J. (1988), ‘‘Vehicle Routing and Scheduling Problems with Time Window Constraints: Efficient Implementations of Solution Improvement Procedures,’’ in *Vehicle Routing: Methods and Studies*, B. Golden and A. Assad, Eds., North-Holland, Amsterdam, pp. 85–105.

Taillard, E., Badeau, P., Gendreau, M., Guertin, F., and Potvin, J. (1997), ‘‘A Tabu Search Heuristic for the Vehicle Routing Problem with Soft Time Windows,’’ *Transportation Science*, Vol. 31, pp. 170–186.

Thangiah, S., Osman, I., and Sun, T. (1995), ‘‘Metaheuristics for Vehicle Routing Problems with Time Windows,’’ Technical Report, Computer Science Department, Slippery Rock University, Slippery Rock, PA.

CHAPTER 31

Industrial Engineering Applications in Hotels and Restaurants

DOUGLAS C. NELSON
Purdue University

1. OVERVIEW **825**

2. DEVELOPING EFFICIENT WORK ENVIRONMENTS FOR HOTELS AND RESTAURANTS **826**
- 2.1. Overview 826
- 2.2. Design by Consensus 826
  - 2.2.1. Overview 826
  - 2.2.2. Relationship Charts 826
  - 2.2.3. Relationship Diagrams 829
  - 2.2.4. Designing for Supervision 829
  - 2.2.5. Designing for Efficient Utility Use 830
- 2.3. Evaluating the Efficiency of a Layout 830
  - 2.3.1. Overview 830
  - 2.3.2. Distance Charts 831
  - 2.3.3. Move Charts 831
  - 2.3.4. Travel Charts 831
  - 2.3.5. Evaluating the Charts 832
- 2.4. Kitchen Ergonomics 833
  - 2.4.1. Overview 833
  - 2.4.2. Table Heights 833
  - 2.4.3. Heights of Other Equipment 833
  - 2.4.4. Workstation Dimensions 834

3. CONTROLLING CAPITAL COSTS **834**
- 3.1. Overview 834
- 3.2. Value Engineering 834
- 3.3. Life-Cycle Costing 834
  - 3.3.1. Overview 834
  - 3.3.2. Information to Be Included in Life-Cycle Costing 835
  - 3.3.3. Converting Amounts to Present Value 835

4. SUMMARY **835**

REFERENCES **836**

1. OVERVIEW

In many aspects, the hospitality industry, which includes hotels and restaurants, is just like most other industries. In fact, components such as the kitchen and the laundry can be viewed as small factories. Several aspects, however, distinguish the hospitality industry from other industries. These differences center on the industry's products. Two of the most important product characteristics are the direct link between production and delivery and the perishable nature of the product (Nebel 1991). Unlike those in many industries, most of the hospitality industry's products must be consumed when and where they are produced. Therefore, in order to be competitive, hotels and restaurants must be able to produce products efficiently when and where consumers want them. The necessary levels of efficiency can be achieved through the application of industrial engineering techniques and principles. This chapter covers some of the more important applications that help hotels and restaurants maintain their competitiveness.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2. DEVELOPING EFFICIENT WORK ENVIRONMENTS FOR HOTELS AND RESTAURANTS

2.1. Overview

One area in which the application of industrial engineering techniques and principles has become increasingly important is improving worker efficiency. Between 1988 and 1997, the productivity, in output per hour, in food service kitchens decreased at an average annual rate of 0.6% per year (U.S. Bureau of the Census 1999).

Operators have tried many things in their attempts to slow and reverse this trend. Many operations have turned to technology and increased automation as a possible solution. While the incorporation of technology in the kitchen is important, Clark and Kirk (1997) failed to find a positive relationship between implementation of technology and productivity. Operations have also sought to improve productivity by purchasing labor in the form of convenience foods and increasing utilization of self-serve systems (Schechter 1997). Still other operations have sought to improve productivity though kitchen design. A study on trends in kitchen size, found that it has been steadily decreasing (Ghiselli et al. 1998). These smaller kitchens have reduced the distances workers must walk to prepare meals. If workers are walking less, they can become more productive (Liberson 1995). To develop smaller, more efficient kitchens, designers have used a variety of methods. One of the better ones is design by consensus (Avery 1985).

2.2. Design by Consensus

2.2.1. Overview

This technique uses the same standard relationship charts and diagrams as recommended for traditional design methods by some of the leading experts on food service layout and design (Almanza et al. 2000; Avery 1985; Kazarian 1989). However, there are major differences between design by consensus and some more traditional design methods. Traditional design methods have centered on management providing operational information to a kitchen design professional. The design professional takes that information and, based on experience and training, develops a layout for the facility. Design by consensus recognizes that workers who will be utilizing the facility have valuable information that can be used to improve the design. Design by consensus does not eliminate the need for the professional designer. Rather, it provides the designer with additional information that can lead to a more efficient, user-friendly design. Information is collected from workers by the use of relationship charts.

2.2.2. Relationship Charts

The first step to preparing a relationship chart is to identify all work centers in the facility. Once this has been accomplished, they are entered into a relationship chart as shown in Figure 1. The chart is now ready to distribute to the workers who will be using the facility. Along with the charts, the

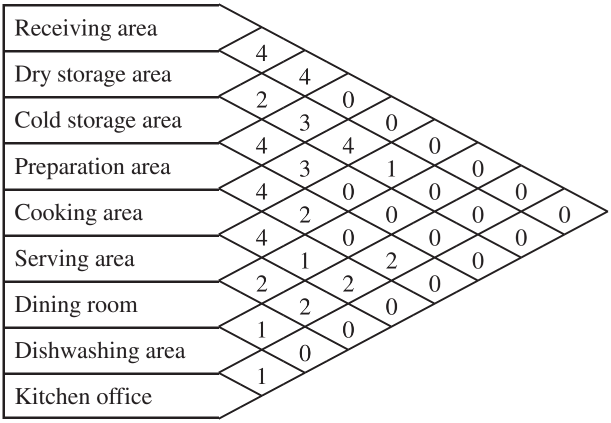


Figure 1 Relationship Chart for Work Centers within the Facility. (The ratings are anchored at 4, very important to be located close together, and 0, not important to be located close together.)

workers are provided instruction on how to complete the charts. They are instructed to rate the importance of placing each pair of work centers close together. The ratings can be on a seven-point scale, using numbers from zero to six (Kazarian 1989), or a five-point scale, using numbers from zero to four (Almanza et al. 2000; Avery 1985). While the seven-point scale allows for greater differentiation, the five-point scale is easier for nonprofessional designers to use. Because the workers are not professional designers, the five-point scale is often used.

The five-point scale is anchored at “not important to be located close together” (0) and “very important to be located close together” (4). To determine the level of importance, the workers must consider how often they move between the two work centers, the frequency, and other factors such as any load they will be carrying and time limitations to travel the distance between the work centers. Figure 1 shows an example of a completed relationship chart.

Once all workers have completed relationship charts, the charts are consolidated into a single chart. To ensure the lowest possible production costs, the highest-paid employees should engage in as little unproductive action as possible, such as walking between work centers. Therefore, when the charts are consolidated, weights are assigned to the different charts based on who submitted the chart. For example, the ratings of a relatively high-paid chef would be weighted higher than those of a lower-paid worker such as a dishwasher. This will lead to an arrangement that will help maximize the productivity of the highest-paid workers and as a result, lower production costs. Figure 2 shows an example of a consolidated relationship chart.

Once the consolidated relationship chart for the work centers has been completed, the designer can develop relationship charts for the equipment within the individual centers. As before, the first step is to determine all equipment that is to be located in the particular work center. The equipment is then listed on the relationship chart just as the work centers were listed on their relationship chart (Kazarian 1989).

While listing only equipment on the chart will help the designer determine the optimum layout for the particular work center, it ignores any relationships between equipment within the work center and other work centers. This is important because it is possible to have an efficient layout within a work center that is not the best layout for the facility. For example, Figure 2 shows that the most important links to the preparation work center are with the cooking work center, followed by the cold storage area and the serving area. If only the equipment is listed on the relationship diagram for the preparation area, then the designer will not know where in the preparation area the movement to the other work centers originates or terminates. Thus, while the layout developed for the preparation area might be the most efficient for that area, the designer has no way of knowing whether it is the optimum layout for the operation. To correct this deficiency, those work centers with the greatest link to the work center for which the relationship chart is being prepared should be included. As shown in Figure 3, the cooking, cold storage, and serving areas have been included in the relationship chart for the preparation area. It is important not to include more than the top two or three linked work centers in the chart. The addition of less important linked work centers to the chart will increase the complexity while adding very little useful information.

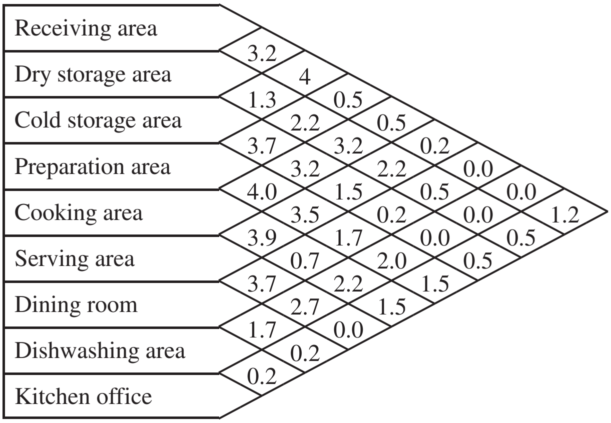


Figure 2 Relationship Chart for Consolidated Ratings for Work Centers within the Facility. (The ratings are anchored at 4, very important to be located close together, and 0, not important to be located close together.)

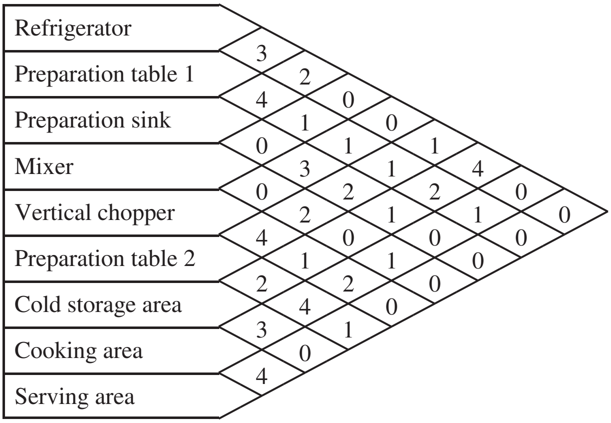


Figure 3 Relationship Chart for Equipment within the Preparation Work Center. (The ratings are anchored at 4, very important to be located close together, and 0, not important to be located close together.)

Using the information from Figure 3, the designer can easily see that the preparation area should be arranged so that preparation table 2 is located nearest the cooking area and the refrigerator is located nearest the cold storage area. The fact that there are no important links between the serving area and any of the equipment in the preparation area does not mean that the link between the two areas is not important, just that movement between the two areas is likely to be to or from any of the pieces of equipment in the work center.

While the consolidated relationship charts for the work centers within the facility and the equipment within the work centers provide valuable information, it is sometimes difficult to visualize the optimum layout in the numbers in the chart. This is particularly true if the relationship chart has a large number of work centers or pieces of equipment listed on it. Therefore, it is helpful to display the information in the form of a relationship diagram that highlights the most important links.

To prepare the information for use in the relationship diagrams, the numbers in the consolidated charts must be converted to whole numbers (Avery 1991). This is done by organizing the ratings for the individual links between work centers/equipment in descending order of importance. Only those links receiving a composite rating of 3.0 or higher are included in the ranking. The inclusion of less important links will clutter the relationship diagram and not provide the designer any additional information. Once the rank order of the importance of the links has been established, they are split into three or four groups based on the breaks in the data. Those links in the highest grouping are assigned an importance rank of four, those in the next highest grouping are assigned an importance rank of three, and so forth until the top four groupings have been assigned importance ranks. Table 1 shows the importance ranking for the work center links from Figure 2. The designer is now ready to prepare the relationship diagrams.

TABLE 1 Importance Rankings for the Top Links from Figure 2

Link	Consolidated Score	Importance Rank
Receiving and cold storage	4.0	4
Preparation and cooking	4.0	4
Cooking and serving	3.9	4
Cold storage and preparation	3.7	3
Serving and dining room	3.7	3
Preparation and serving	3.5	2
Receiving and dry storage	3.2	1
Dry storage and cooking	3.2	1
Cold storage and cooking	3.2	1

2.2.3. Relationship Diagrams

There are two different types of relationship diagrams: a bubble type and a layout type. The bubble type is done first. This relationship diagram, as shown in Figure 4, uses labeled circles connected by lines to help visualize the important links. The number of lines connecting the different circles corresponds to the importance rank of the links as determined in Table 1. The primary purpose of this type of relationship chart is to help determine which work center(s) should be centrally located, which can be done simply by counting the number of lines connected to the circle. The work center(s) with the greatest number of lines should be centrally located. For simpler drawings, this step can be eliminated, in which case the layout-type relationship chart is drawn directly from Table 1.

Once the work centers that must be centrally located have been determined, the layout-type relationship diagram is drawn. The work centers are drawn to scale inside the space that has been allocated for the operation. The first step in drawing this relationship diagram is to locate any work centers that are anchored to a specific part of the building. For example, the receiving area is generally located along an outside wall, out of public view. Once the centers that must be located in specific parts of the facility are drawn, the designer can begin inserting the remaining work centers, starting with those that must be centrally located. The last centers added to the layout are those centers with few, if any, links to the other work centers. At this point in the design process, the actual shape of each work center is not known. It is up to the designer to estimate the shape of each work center as it is drawn in the diagram. The result is that frequently the area for the last work center is not a functional shape, even though it is the correct size. For example, the remaining area for a 200-ft^2 office might be 2 ft wide and 100 ft long, or it could be split into two locations, neither of which would allow the construction of an acceptable office. If this happens, the layout-type relationship chart must be redrawn and the shapes and locations of some of the work centers modified to allow a functional office to be included. This process may need to be repeated several times to ensure that there is enough usable space for all work centers. An example of a completed layout-type relationship diagram is shown in Figure 5.

2.2.4. Designing for Supervision

Designing for ease of supervision as well as efficiency is one of the basic design principles (Birchfield 1988; Kazarian 1989). Relationship diagrams, however, do not take designing for supervision into account. For example, the location of kitchen office is essential for adequate supervision, but the information provided by the workers is not very helpful when it comes to locating the kitchen office. Because the office is not an area that workers visit frequently during production, it is likely to receive low importance ratings for being located near production areas, as can be clearly seen in Figures 1

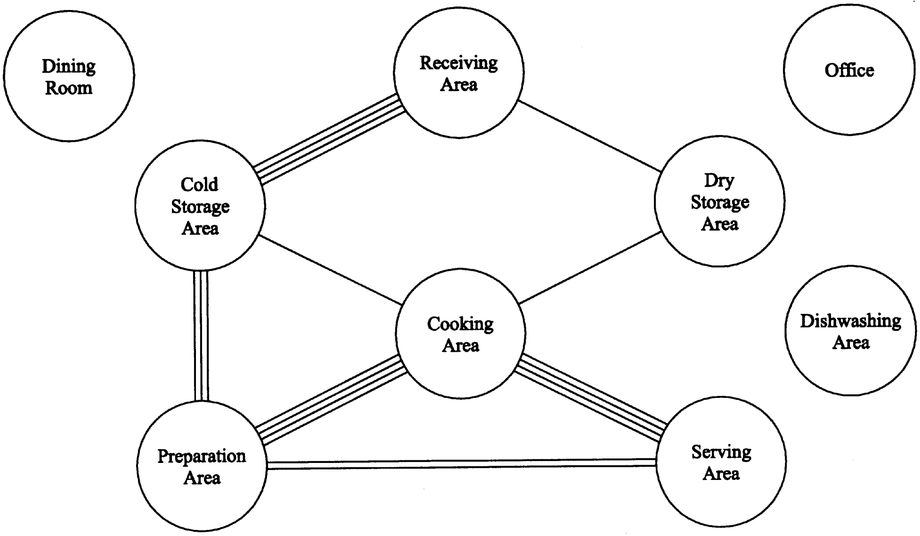


Figure 4 Bubble-Type Relationship Diagram Developed from the Information in Figure 3 and Table 1.

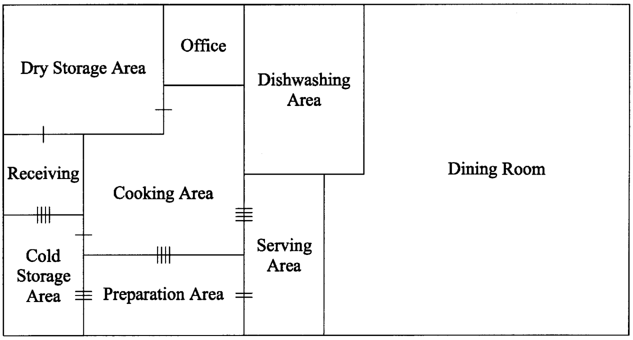


Figure 5 Layout-Type Relationship Diagram Developed from Figure 4.

and 2. Therefore, it is up to the designer to ensure that the kitchen office is properly placed to provide adequate supervision of the kitchen. The placement of the office is further complicated by other functions for which the office is used. Birchfield (1988) describes the need for the office to be accessible to customers needing to talk to management without having those customers walk through the kitchen to reach the office. Further, managers must also monitor the movement of food and supplies in and out of the building and storage areas. This places the office near the receiving area, which would cause problems for customers accessing the office. The final location of the office is often a compromise and depends on which function management views as its primary one.

2.2.5. Designing for Efficient Utility Use

Another important consideration when arranging the kitchen is efficient use of utilities. As with the office, the information provided by the relationship charts and diagrams will not adequately address this issue. Frequently, designing for efficient utility use conflicts with designing for maximum worker efficiency. Kazarian (1989) recommends that heating and cooling equipment be separated. While this is important for controlling energy use, it does not always provide for efficient production. Frequently, food is moved directly from a refrigerator to a piece of cooking equipment. Designing for efficient energy use would place the pieces of equipment some distance apart, while designing for efficient production would place them adjacent to each other. In this case, the designer must weigh the need for energy conservation against the need for worker efficiency and develop an arrangement that minimizes total production costs.

There are several other arrangement considerations that can help reduce energy consumption. One is locating steam- and hot water-generating equipment near the equipment that uses steam or hot water. Shortening the distance between the generation and use of steam or hot water will reduce energy loss from the pipes connecting the equipment. Another way to help reduce energy consumption is to consolidate as many pieces of heat-generating equipment under one exhaust hood as possible. This will reduce costs by reducing ventilation requirements. Finally, Avery (1985) recommends that compressors and condensers for large cooling equipment such as walk-in refrigerators and freezers be located so that the heat they generate can be easily exhausted, thereby reducing the cooling requirements for the kitchen. As with the example in the previous paragraph, arranging the kitchen to take advantage of these energy conservation techniques may conflict with arranging for maximum worker efficiency. It is up to the designer to develop a compromise arrangement that will minimize the total cost of operating the facility.

2.3. Evaluating the Efficiency of a Layout

2.3.1. Overview

Kazarian (1989) describes a cross-charting technique that uses distance, move, and travel charts to compare two or more layouts to determine which is the most efficient. This technique recognizes that if the time spent walking between equipment can be reduced, then the worker can spend more time being productive. Avery (1985) took the use of cross-charting one step further. He recognized that not all production workers are paid the same wage. It is possible to have an arrangement that is the most efficient based on total man-hours, but not the most efficient in total labor cost. Because of different training and skill requirements, it is possible for the highest-paid workers to make twice what the lowest-paid workers make. If the travel charts are adjusted for wages, then the comparison of arrangements is based on cost efficiency and not man-hour efficiency.

TABLE 2 Distance Chart[a]

	From				
To	Prep Table	Sink	Oven	Walk-in	Dishwasher
Prep table		2	3.5	8	11
Sink	2		3	6	10
Oven	3.5	3		6	18
Walk-in	8	6	6		18
Dishwasher	11	10	18	18	

[a] Distances are in meters.

2.3.2. *Distance Charts*

The first cross-chart that must be prepared is the distance chart. This chart contains the distance from each piece of equipment to all the other pieces of equipment in the work center. Table 2 shows an example of a completed distance chart. The pieces of equipment listed across the top of the chart are the starting points for the movements, and those listed down the side are the ending points. The order of the pieces of equipment in the list depends on the arrangement of the work center. Adjacent pieces of equipment should be listed next to each other in the chart. Listing the equipment in that order will make it easier to recognize ways to rearrange the work center so that it can be more productive. If the chart was prepared properly, then cells adjacent to the diagonal line will contain the smallest numbers and those farther away from the diagonal line will contain larger the numbers.

2.3.3. *Move Charts*

The move chart is a frequency chart that tracks the number of times a worker moves from one piece of equipment to another. Equipment is listed in these charts in the same order as in the distance chart. There are several ways to determine the frequencies of movement between the different pieces of equipment. The most accurate way is to observe the workers and record their movements. If this is not possible, a reasonable estimate can be developed using the menus and estimating the number of times workers will move between each pair of equipment. If the wages of the different classifications of workers using the facility are significantly different, then multiple move charts should be prepared, one for each classification. Table 3 shows examples of the move charts for two classifications of workers.

2.3.4. *Travel Charts*

After the distance and move charts are complete, they are used to generate travel charts. The individual cells in the travel charts are the products of the corresponding cells in the distance and move charts. The cells in the travel charts sum to the total distance traveled by the worker. Once the total distances traveled by each classification of workers are calculated, those numbers are weighted by

TABLE 3 Move Charts for Prep Cook and Dishwasher

Prep Cook	From				
To	Prep Table	Sink	Oven	Walk-in	Dishwasher
Prep table		40	40	15	35
Sink	30		10	45	15
Oven	40	10		10	0
Walk-in	40	20	10		0
Dishwasher	20	30	0	0	
Dishwasher	**From**				
To	Prep Table	Sink	Oven	Walk-in	Dishwasher
Prep table		0	0	0	60
Sink	0		0	0	0
Oven	0	0		0	0
Walk-in	0	0	0		0
Dishwasher	60	0	0	0	

their wages. Finally, a single sum is calculated for all the charts. That number can now be used to compare two or more arrangements. The arrangement that yields the smallest number is the most efficient in terms of labor costs. Table 4 show the travel charts for this example. The weighted sum for the prep cook's chart is 24,850 and the weighted sum of the dishwasher's chart is 9,240. Combining the two, the weighted sum for the operation is 34,090. The units are in currency times distance. The exact units used for currency and distances are not important, as long as they are consistent for each layout being evaluated.

2.3.5. Evaluating the Charts

The final step is to evaluate the charts to identify ways to improve the arrangement of equipment. There are basically two different ways to do this. One is to modify the arrangement, and the other is to modify the production procedures. For an efficient arrangement, the majority of the movement will be between adjacent equipment. Since the distance chart was set up so that adjacent equipment in the arrangement appeared in adjacent cells, the move chart can be used to identify problems with the arrangement. If the work center is properly arranged, the numbers in the cells along the diagonal line should be large and get progressively smaller the farther they are from the line. Rearranging the layout so that the largest numbers are adjacent to the line will improve the efficiency of the operation. In the prep cook's chart, one possible change in the arrangement that may reduce the total distance traveled is to move the walk-in closer to the sink than the oven is. The total number of moves between the prep table and the walk-in is 65 (45 from the walk-in to the sink and 20 from the sink to the walk-in), while the total number of moves between the walk-in and oven is only 20 (10 from the oven to the sink and 10 from the sink to the oven). Because moving one piece of equipment affects other pieces of equipment, the only way to be sure that the new arrangement is indeed better is to chart the new arrangement and compare the total weighted sums for both arrangements. When there is more than one travel chart because of different wage rates, both charts must be prepared for the new arrangement.

Be careful when making adjustments based on the move chart for the lowest-paid worker. While those adjustments may reduce the total man-hours need to perform the task, they may actually increase the total labor cost for the operation. For example, from the dishwasher's chart in Table 3, one possible change would be to move the dishwashing machine closer to the prep table. While this would reduce the travel of the lowest-paid worker, it may increase the travel of the highest-paid worker. Therefore, moving the dishwashing machine may increase labor dollars, making this an unwise adjustment to the arrangement.

Identifying procedural changes using these charts is not as straightforward as identifying arrangement changes. Basically, these charts can only be used to identify problems with procedures requiring further investigation. In the example in Table 3, there is a total of 100 trips to and from the dishwasher made by the prep cook and another 120 trips made by the dishwasher. The dishwasher's trips are all between the dishwasher and the prep table. This is an area requiring further investigation to see whether procedural changes are needed. The designer should seek to uncover the reason for the trips. It could be that the prep cook is bringing dirty dishes to the dishwasher and returning empty-handed and that the dishwasher is returning clean dishes to the prep table and returning empty-handed. If

TABLE 4 Travel Charts for Prep Cook and Dishwasher

Prep Cook	From				
To	Prep Table	Sink	Oven	Walk-in	Dishwasher
Prep table		80	140	120	385
Sink	60		30	270	150
Oven	140	30		60	0
Walk-in	320	120	60		0
Dishwasher	220	300	0	0	

Dishwasher	From				
To	Prep Table	Sink	Oven	Walk-in	Dishwasher
Prep table		0	0	0	660
Sink	0		0	0	0
Oven	0	0		0	0
Walk-in	0	0	0		0
Dishwasher	660	0	0	0	

this is the case, then the dishwasher may be able to drop off clean dishes and pick up the dirty dishes. Making this procedural change reduces the total trips between the two areas. Further savings are realized because all the trips are now made by the lowest-paid worker.

2.4. Kitchen Ergonomics

2.4.1. Overview

Another method that can be used to help improve the productivity in kitchens is the application of ergonomic principles. Commercial food preparation is a physically demanding occupation. Typically, employees must stand for long periods of time on hard surfaces. As their shifts proceed, workers often become fatigued and experience impaired perception and reduced physical performance, both of which will negatively impact their productivity (Almanza et al. 2000; Avery 1985). The application of sound ergonomic principles can delay the onset of fatigue, making the workers more productive over the course of the day. Something as simple as an incorrect table height can accelerate the onset of fatigue. If a worker is standing vertically, then between 20–27% of his or her energy expenditure can be used to perform work. If that same worker is stooping, then the percent of energy expenditure available to perform work drops to around 5% (Avery 1985). Providing workers the proper table height will allow them to have four to five times more energy to devote to production than they would have if they were stooping.

2.4.2. Table Heights

The correct table height depends on the type of work being performed and who is performing the work. For heavy work, the ideal height is the point where the wrist bends when the arms are hanging at the person's sides. This puts the table height for heavy work at 76–91 cm (34–36 in.), depending on the person (Almanza et al. 2000; Avery 1985). For lighter work, there seems to be some disagreement among the industry experts on what is the proper height in terms of body dimension. Katsigris and Thomas (1999) recommend that the surface height be 10 cm (4 in.) below the bend in the elbow. Avery (1985) recommends 2.5–7.6 cm (1 to 3 in.) below the bend in the elbow, and Kazarian (1989) recommends 7.6 cm (3 in.). Despite their differences on where the table should be located in reference to the worker, they all agree that the range is 94–99 cm (37–39 in.) for women and 99–104 cm (39–41 in.) for men.

The new smaller kitchens do not have enough room to have separate tables for different workers and different tasks. One way to overcome this problem and provide the optimum table height is to use adjustable tables. These are available, but their use is not yet widespread.

2.4.3. Heights of Other Equipment

While a significant amount of food preparation takes place at tables, it is also important that other kitchen equipment be at the proper height. Height is of particularly important for sinks, stacked equipment, storage equipment, and top-loading equipment. Most other equipment follows the same rules as tables when it comes to height.

According to Avery (1985), the height of the top of the sink should be 94–99 cm (37–39 in.) for women and 99–102 cm (39–40 in.) for men. The depth of the sink depends on its function. For standard sinks, the bottom of the sink should be at the tip of the worker's thumb when the arms are at the worker's sides. This places the bottom at approximately 69 cm (27 in.) for women and 74 cm (29 in.) for men. If the sink is to be used to soak pans, the bottom can be 15 cm (6 in.) lower, provided small items are not placed in the sink.

Avery (1985) also presents information on the importance of height for stacked items such as an oven. He recommends that deck ovens be stacked not more than two high and that the bottom oven be at least 51 cm (20 in.) off the ground. If three ovens are stacked and the first oven is at least 51 cm (20 in.) off the ground, there is a greater risk that workers will be burned trying to use the top oven. If the top oven is at an acceptable height, the bottom oven will be below 51 cm (20 in.), increasing the risk that a worker will be burned using that oven. Further, the worker will have to expend more energy stooping to place things in and remove them from the bottom oven. Stacking of other types of ovens is also not recommended. Conventional and convection ovens are too tall to stack without the risk of serious burns to those using them. It is far safer to put the oven on a stand so the opening is between the workers' waists and shoulders.

Restaurants often employ several different types of storage devices, from refrigeration and warming cabinets to plate and tray dollies to standard shelves. The proper height for storage of frequently used items is between the worker's waist and shoulders (Avery 1985). Placing items too high can lead to items being improperly placed on shelves, which can lead to falling objects and the related potential for damage and personal injury. Placing items too low forces workers to stoop, causing them to expend additional energy and putting a strain on their back muscles. In the process of designing small, efficient kitchens, designers have turned to undercounter storage units, such as refrigerators. If the use of undercounter storage units is necessary, the designers should consider using drawers instead of reach-ins because they require less bending and stooping to access products. Finally, if plate and tray dollies are used, they should be self-leveling, meaning that as one plate or tray is removed, the load on the springs at the bottom of the stack decreases, decreasing the compression of the spring and raising the stack. This keeps the items within easy reach of the workers.

With top-loading pieces of equipment such as steam-jacketed kettles, pots and pans, and mixers, it is important for the user to be able to reach comfortably and safely over the rim. This helps prevent burns, spills, and similar accidents when workers are adding ingredients or stirring the mixture. Avery (1985) recommends a rim height of no greater than 97 cm (38 in.) for steam-jacketed kettles, whenever possible. Operations that use large kettles must exceed that height to ensure that the draw-off is high enough to allow a pan to be placed beneath it. The same height recommendations for steam-jacketed kettles also apply to other top-loading equipment, such as stock pots and mixers.

2.4.4. Workstation Dimensions

In addition to proper surface height being maintained, the length and width of the workstations must also be addressed to help workers be productive. For a standing worker, the workstation should be arranged so that the majority of the tasks can be performed in a 46-cm (18-in.) arc centered on the worker (Avery 1985). Supplies needed to perform the tasks can be stored just outside that arc. Since people tend to spread out to fill their environment, it is important that table size be restricted to ensure efficient production. Almanza et al. (2000) recommend limiting the table size for a single work to 61–76 cm (24–30 in.) wide and 1.22–1.83 m (4–6 ft) long. The width of the table can be increased to 91 cm (36 in.) if the back of the table will be used for storage and to 107 cm (42 in.) if two workers will be using opposite sides of the table. If the two workers will be working side by side, then the table length should be 2.44–3.05 m (6–8 ft).

3. CONTROLLING CAPITAL COSTS

3.1. Overview

As with any industry, controlling capital costs is an important ingredient in developing and maintaining a successful operation. Earlier sections presented information on controlling production costs by increasing worker productivity. Equally important is the need to control capital cost. While capital costs for restaurant and hotels may not be as large as for heavy industry, they are significant when compared to the revenue generated by the respective operations. Therefore, it is important that operations not waste money on capital expenditures. By using value engineering and life-cycle costing, operators are able to control costs by making better capital-expenditure decisions.

3.2. Value Engineering

In its simplest terms, value engineering is the process of reviewing purchase decisions to determine whether they are cost effective. Value engineering seeks to determine whether the value added to the operation by the purchase provides the greatest possible return or whether there is a more cost-effective way to accomplish the same thing. When performing value engineering, it is important to have a good understanding of the operation and its goals. More than one concept has been diminished or destroyed by the improper application of value engineering. Deleting the required equipment to produce a restaurant's signature item in an attempt to cut cost has been a contributing factor in more than one business failure (Foster 1998). Foster also points out that operators should avoid doing last-minute value engineering. Last-minute budget cutting does not allow for proper evaluation of the options and can lead to the cutting of necessary pieces of equipment or options that will have to be added later at an increased cost. The problem for kitchen equipment is further compounded by a lack of understanding of the reasons behind the relatively high costs of commercial food-service equipment. A microwave oven designed to meet the sanitation and operational requirements of a commercial kitchen can cost $2000 or more. If the kitchen designer does not start the value-engineering process early and demonstrate the need for each piece of equipment and option, these items are likely to be the prime targets of last-minute value engineering. Foster recommends that designers start their own value-engineering process at the beginning of the design process. They should describe each piece of equipment and each option, noting its importance to the successful operation of the business. It is also helpful to obtain multiple bids for each item.

3.3. Life-Cycle Costing

3.3.1. Overview

In evaluating the different pieces of equipment and options, it is important to consider more than just the initial cost. Purchasing a holding cabinet made of anodized aluminum instead of stainless steel can significantly reduce the purchase cost. However, aluminum cabinets are not as durable and typically have shorter lives than stainless steel cabinets. Further, the salvage value of the steel cabinet is typically higher. If these other factors are not considered as part of the selection process, then the chance of selecting the most cost-efficient piece of equipment is greatly reduced (Borsenik and Stutts 1997). The challenge is to compare today's dollars with periodic payments and future dollars. Fortunately, relatively simple formulas can be used to bring all amounts back to present value so the different options can be properly evaluated.

3.3.2. Information to Be Included in Life-Cycle Costing

The process begins with collecting all pertinent information. The more information collected, the more accurate the prediction. Typically, information is collected on as many of the following items as are applicable.

1. *Initial costs:* This includes purchase price and installation costs.
2. *Periodic costs:* This includes energy usage, maintenance, supplies, and labor estimates. It can also include lease payments and service contracts.
3. *End of service costs/revenues:* This can be a revenue such as salvage or a cost such as a residual lease payment.
4. The organization's standard return on investment and anticipated inflation rate.

3.3.3. Converting Amounts to Present Value

The equation used to return all amounts to present value depends on when the amount is paid or received. First, it must be determined whether the amount is a cost or revenue. Since this is a costing method, costs are treated as positive numbers and revenues are treated as negative numbers. Once the sign of the amount has been determined, the next step is to determine whether it is a one-time payment/receipt or a periodic payment/receipt. Examples of one-time payments/receipts include purchase price, installation costs, and salvage value. Examples of periodic payments/receipts include rents, monthly charges such as energy cost, wages, and increased income directly attributed to using the equipment.

If the payment/receipt is a one-time event, then the next step is to determine when it occurs. If it occurs now, it is already in present value and no conversion is necessary. If, however, it occurs some time in the future, it must be converted to present value using the following equation from Newman (1980):

$$P = F(1 + i)^{-n} \tag{1}$$

where P = present value of the future amount (F)
i = interest rate
n = number of compounding periods

The length of the compounding periods must agree with the interest rate; that is, if the interest rate is an annual rate, the number of compounding periods must be expressed in years.

If the payment/receipt is periodic in nature, the following equation derived from information contained in Newman (1980) is used.

$$P = A(1 + i)^{-n} \frac{(1 - i)^n - (1 - \text{CPC})^n}{(i - \text{CPC})} \tag{2}$$

where A = periodic amount
CPC = compound price change or inflation rate

As before, the length of the compounding period must be the same for all terms. Equation (2) will work for all combinations of interest rate and CPC except those where the interest rate and CPC are equal. If the interest rate and the CPC are equal, Equation (2) is undefined. If this happens, the present value of the periodic amount can be found by multiplying the periodic amount by the number of compounding periods. Once all amounts have been converted to present value, the last step is to sum all payments/receipts. The equipment/option with the lowest sum is the most cost-effective alternative.

4. SUMMARY

The industrial engineering techniques discussed in this chapter are but a few of the techniques currently being used by hospitality operations. Successful food-service operations of the future will find ways to change the negative productivity trend that the industry has been experiencing for the last

20+ years. They will become more productive by making better use of existing resources and exploiting the benefits of new technology. The prudent application of the principles and techniques of industrial engineering will help make a positive productivity trend a reality.

REFERENCES

Almanza, B. A., Kotchevar, L. H., and Terrell, M. E. (2000), *Foodservice Planning: Layout, Design, and Equipment,* 4th Ed., Prentice Hall, Upper Saddle River, NJ.

Avery, A. C. (1985), *A Modern Guide to Foodservice Equipment,* Rev. Ed., Waveland Press, Prospect Heights, IL.

Borsenik, F. D., and Stutts, A. S. (1997), *The Management of Maintenance and Engineering Systems in the Hospitality Industry,* 4th Ed., John Wiley & Sons, New York.

Birchfield, J. C. (1988), *Design and Layout of Foodservice Facilities*, Van Nostrand Reinhold, New York.

Clark, J., and Kirk, D. (1997), ‘‘Relationships Between Labor Productivity and Factors of Production in Hospital and Hotel Foodservice Departments—Empirical Evidence of a Topology of Food Production Systems,’’ *Journal of Foodservice Systems*, Vol. 10, No. 1, pp. 23–39.

Foster, F. Jr., (1998), ‘‘Operations and Equipment: Reducing the Pain of ‘Value Engineering,’ ’’ *Nation's Restaurant News*, Vol. 20, No. 4, pp. 16–26.

Ghiselli, R., Almanza, B. A., and Ozaki, S. (1998), ‘‘Foodservice Design: Trends, Space Allocation, and Factors That Influence Kitchen Size,’’ *Journal of Foodservice Systems*, Vol. 10, No. 2, pp. 89–105.

Katsigris, C., and Thomas, C. (1999), *Design and Equipment for Restaurants and Foodservice: A Management View,* John Wiley & Sons, New York.

Kazarian, E. D. (1989), *Foodservice Facilities Planning*, Van Nostrand Reinhold, New York.

Liberson, J. (1995), ‘‘Food and Beverage: Cooking Up New Kitchens: Technology and Outsourcing Have Created Smaller, More Efficient Kitchens,’’ *Lodging*, Vol. 21, No. 2, pp. 69–72.

Nebel, D. C., III (1991), *Managing Hotels Effectively: Lessons From Outstanding General Managers*, Van Nostrand Reinhold, New York.

Newman, D. G. (1980), *Engineering Economic Analysis*, Engineering Press, San José, CA.

Schechter, M. (1997), ‘‘The Great Productivity Quest,’’ *Food Management*, Vol. 32, No. 1, pp. 46, 48, 52, 54.

U.S. Bureau of the Census (1999), *Statistical Abstract of the United States: 1999,* Washington, DC.

SECTION III

PERFORMANCE IMPROVEMENT MANAGEMENT

A. Organization and Work Design
B. Human Factors and Ergonomics

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

III.A

Organization and Work Design

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 32

Leadership, Motivation, and Strategic Human Resource Management

Taly Dvir
Tel Aviv University

Yair Berson
Polytechnic University

1. **INTRODUCTION** **842**
2. **LEADERSHIP AND MOTIVATION** **842**
    2.1. The Classic Paradigm of Leadership: A Transactional Approach 842
    2.2. The New Paradigm of Leadership: A Transformational Approach 843
    2.3. Differences in the Motivational Basis for Work between the Classic and the New Leadership Paradigms 845
        2.3.1. From a Calculative–Rational toward Emotional–Expressive Motivation to Work 845
        2.3.2. From an Individualistic-Oriented toward a Collectivistic-Oriented Motivation to Work 846
        2.3.3. From Extrinsic toward Intrinsic Motivation to Work 847
    2.4. The Full Range Leadership Model: An Integrative Framework 848
3. **LEADERSHIP OUTCOMES** **850**
    3.1. Performance 851
    3.2. Employee Development 852
        3.2.1. Employee Personal Development 853
        3.2.2. Development of Employee Attitudes toward the Leader 854
        3.2.3. Employee Group Development 855
4. **IMPLICATIONS FOR STRATEGIC HUMAN RESOURCE MANAGEMENT** **855**
    4.1. A Strategic Approach to Human Resource Management 856
    4.2. Recruiting: From Hiring to Socialization 856
        4.2.1. Implications for Selection 856
        4.2.2. Implications for Mentoring and Socialization 857
    4.3. From Performance Appraisal to Performance Management 858
        4.3.1. Implications for Feedback 858
        4.3.2. Implications for Alignment and Signaling of Strategic Goals 859
    4.4. From Training to Development 859
    4.5. Compensation: From Transactional to Intrinsic Reward Systems 861
    4.6. Involvement-Transformational vs. Inducement-Transactional HRM Systems: An Integrative Framework 862
5. **CONCLUSION** **863**

REFERENCES **864**

ADDITIONAL READING **867**

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

Most definitions of leadership reflect the assumption that it involves a social influence process whereby intentional influence is exerted by one person over other people to structure the activities and relationships in a group or organization (Yukl 1998). Leadership occurs when one group member modifies the motivation and competencies of others in the group (Bass 1990). Work motivation is the willingness of an individual to invest energy in productive activity. Thus, leadership and motivation are interwoven, inseparable concepts; a core outcome of effective leadership is a higher willingness on the part of the employees to invest energy in performing their tasks. A new genre of leadership and motivation theories has been shown to affect organizational effectiveness in ways that are quantitatively greater than, and qualitatively different from, the effects specified by previous theories (House and Shamir 1993; for a meta-analytic review see Lowe et al. 1996). These theories have led to new applications in human resource management. As shown in Figure 1, we first review the early vs. the most recent paradigms of leadership, namely the shift from transactional to transformational, charismatic, or visionary leadership. This shift is manifested by changes in the bases for work motivation from an emphasis on calculative, individualistic, extrinsic, short-term motivators toward more expressive, collectivistic, intrinsic, long-term motivators. Taken together, these new approches have been shown to positively impact organizational outcomes, including performance and employee development. The impact of the new leadership and motivation paradigm on organizational outcomes has implications for strategic human resource management, specifically recruiting, performance management, training and development, and compensation.

2. LEADERSHIP AND MOTIVATION

2.1. The Classic Paradigm of Leadership: A Transactional Approach

> I was his loyal friend, but no more. . . . I was so hurt when I discovered that in spite of my loyalty, he always preferred Arie Dery to me. I asked him: 'Bibi, why?' and he answered: "He brings me 10 seats in the parliament while you don't have such a bulk of voters behind you."
>
> —Itzhak Levi, former Minister of Education, on Benjamin Netanyahu, former Prime Minister of Israel

Many of the classic approaches to leadership concentrated on how to maintain or achieve results as expected or contracted between the leader and the employees. These transactional theories and practices viewed leadership in terms of contingent reinforcement, that is, as an exchange process in which employees are rewarded or avoid punishment for enhancing the accomplishment of agreed-upon objectives.

Figure 2 shows the process by which transactional leaders affect their employees' motivation and performance. Transactional leaders help employees recognize what the role and task requirements are to reach a desired outcome. The transactional leader helps clarify those requirements for employees, resulting in increased confidence that a certain level of effort will result in desired performance. By recognizing the needs of employees and clarifying how those needs can be met, the transactional leader enhances the employee's motivational level. In parallel, the transactional leader recognizes what the employee needs and clarifies for the employee how these needs will be fulfilled in exchange for the employee's satisfactory effort and performance. This makes the designated outcome of sufficient value to the employee to result in his or her effort to attain the outcome. The model is built upon the assumption that the employee has the capability to perform as required. Thus, the expected effort is translated into the expected performance (Bass 1985).

There is support in the literature for the effectiveness of transactional leadership. Using contingent reinforcement, leaders have been shown to increase employee performance and job satisfaction and to reduce job role uncertainty (Avolio and Bass 1988). For example, Bass and Avolio (1993) report results collected from 17 independent organizations indicating that the correlations between the contingent reward leadership style and employees' effectiveness and satisfaction typically ranged from 0.4 to 0.6, depending on whether it was promises or actual rewards. Similarly, Lowe et al. (1996), in their meta-analysis of 47 studies, report that the mean corrected correlation between contingent reward and effectiveness was 0.41.

Although potentially useful, transactional leadership has several serious limitations. First, the contingent rewarding, despite its popularity in organizations, appears underutilized. Time pressure, poor performance-appraisal systems, doubts about the fairness of the organizational reward system, or lack of managerial training cause employees not to see a direct relationship between how hard they work and the rewards they receive. Furthermore, reward is often given in the form of feedback from the superior, feedback that may also be counterproductive. What managers view as valued feedback is not always perceived as relevant by the employees and may be weighted less than feedback received from the job or coworkers. In addition, managers appear to avoid giving negative feedback to employees. They distort such feedback to protect employees from the truth. Second, transactional leadership may encourage a short-term approach toward attaining organizational objec-

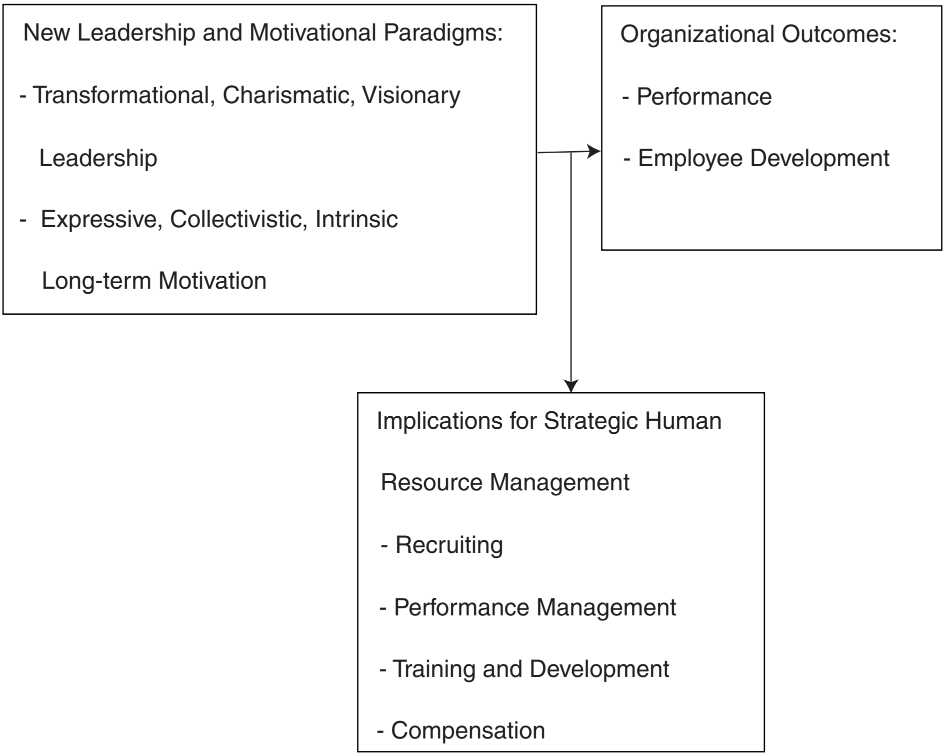

New Leadership and Motivational Paradigms:
- Transformational, Charismatic, Visionary Leadership
- Expressive, Collectivistic, Intrinsic Long-term Motivation
Organizational Outcomes:
- Performance
- Employee Development
Implications for Strategic Human Resource Management
- Recruiting
- Performance Management
- Training and Development
- Compensation

Figure 1 Linkages between Leadership and Motivational Paradigms, Organizational Outcomes, and Strategic Human Resource Management.

tives. A transactional leader who views leadership solely from a contingent reinforcement perspective may find that employees will circumvent a better way of doing things to maximize short-term gain to reach their goals in the most expeditious manner possible, regardless of the long-term implications. Third, employees do not always perceive contingent reinforcement as an effective motivator. The impact may be negative when the employees see contingent rewards given by the superior as manipulative (Avolio and Bass 1988).

Finally, one of the major criticisms directed toward this contract-based approach to leadership has been that it captures only a portion of the leader-follower relationships. Scholars like Hemphill (1960) and Stogdill (1963) (in Seltzer and Bass 1990) felt long ago that an expanded set of factors was necessary to fully describe what leaders do, especially leader–follower relationships that include motivation and performance that go beyond contracted performance. Bass and Avolio (1993) state that the ‘‘transaction’’ between a leader and a follower regarding the exchange of rewards for achieving agreed-upon goals cannot explain levels of effort and performance of followers who are inspired to the highest levels of achievement. There are many examples of leadership that do not conform to the ever-popular notion of a transaction between leader and follower. As Howell (1996) points out, the notion of leaders who manage meaning, infuse ideological values, construct lofty goals and visions, and inspire was missing entirely from the literature of leadership exchange. A new paradigm of leadership was needed to account for leaders who focus the attention of their employees on an idealized goal and inspire them to transcend themselves to achieve that goal. The actions of such leaders result in higher-order changes in employees and therefore in higher performance.

2.2. The New Paradigm of Leadership: A Transformational Approach

> Head in the clouds, feet on the ground, heart in the business.
>
> —Anita Roddick, founder of The Body Shop

Pat Summitt has been called the best basketball coach since the famed John Wooden. Just some

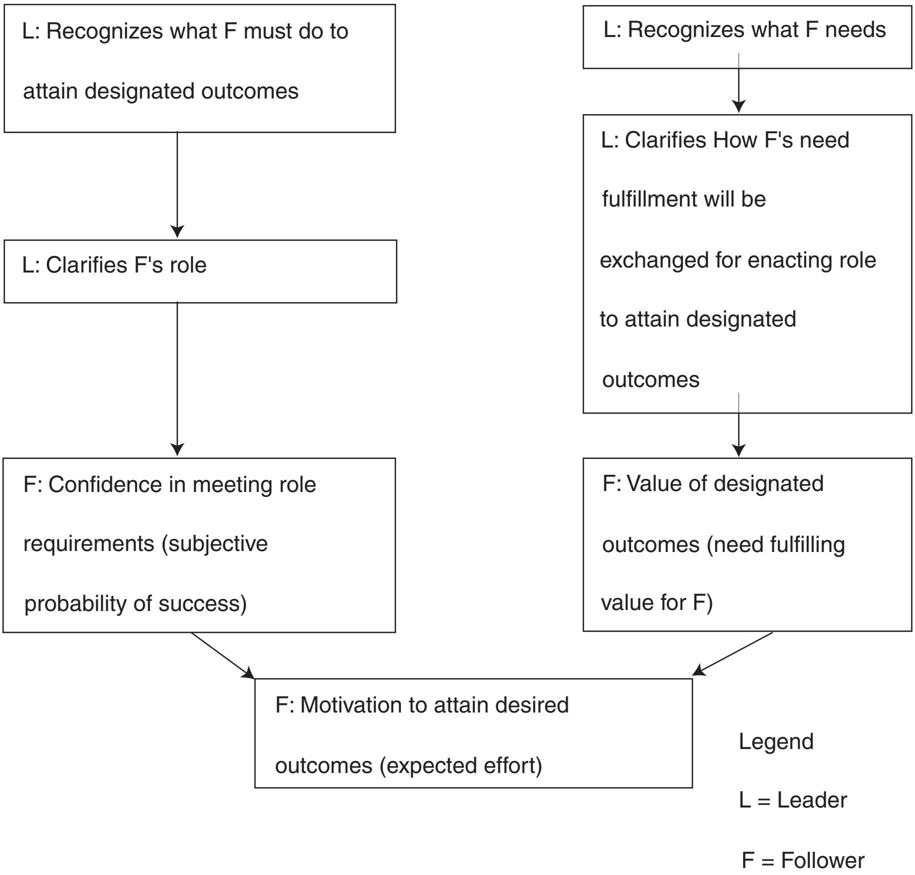

L: Recognizes what F must do to attain designated outcomes
L: Clarifies F's role
F: Confidence in meeting role requirements (subjective probability of success)
L: Recognizes what F needs
L: Clarifies How F's need fulfillment will be exchanged for enacting role to attain designated outcomes
F: Value of designated outcomes (need fulfilling value for F)
F: Motivation to attain desired outcomes (expected effort)
Legend
L = Leader
F = Follower

Figure 2 Transactional Leadership and Follower Motivation. (Adapted with the permission of The Free Press, a Division of Simon & Schuster, from *Leadership and Performance beyond Expectations,* by Bernard M. Bass, Copyright © 1985 by The Free Press.)

of her credentials include winning the first-ever Olympic gold medal in U.S. women's basketball history; six national championships; an .814 winning percentage, fifth among all coaches in college basketball history; and so many trips to basketball's Final Four that her record in all likelihood will never be equaled. Her 1997–1998 University of Tenessee team finished the regular season 30 and 0 and won its third consecutive national title. Beyond coaching records, however, what is Summitt like as a person? Among other things, when she walks into a room, her carriage is erect, her smile confident, her manner of speaking direct, and her gaze piercing. Her calendar documents her pressured schedule. But she is also a deeply caring person, appropriate for a farm daughter who grew up imitating her mother's selfless generosity in visiting the sick or taking a home-cooked dinner to anyone in need. Summitt clearly has many extraordinary qualities as a coach, but perhaps most striking is the nature of the relationship she develops with her players (Hughes et al. 1998). Pat Summitt does not rely only on her players' contracts to gain their highest levels of motivation and performance. The best transactional, contract-based leadership cannot assure such high performance on the part of the players. Summitt clearly has the additional qualities of what is referred to as transformational leadership.

Indeed, the past 15 years have seen some degree of convergence among organizational behavior scholars concerning a new genre of leadership theory, alternatively referred to as "transformational," "charismatic," and "visionary." Transformational leaders do more with colleagues and followers than set up simple transactions or agreements (Avolio 1999). Transformational leadership theory explains the unique connection between the leader and his or her followers that accounts for extraordinary performance and accomplishments for the larger group, unit, and organization (Yammarino and Dubinsky 1994), or, as stated by Boal and Bryson (1988, p. 11), transformational leaders ''lift ordinary people to extraordinary heights.'' Transformational leaders engage the ''full'' person with the purpose of developing followers into leaders (Burns 1978). In constrast to transactional leaders, who seek to satisfy the current needs of followers through transactions or exchanges via contingent reward behavior, transformational leaders (a) raise the follower level of awareness about the importance of achieving valued outcomes, a vision, and the required strategy; (b) get followers to transcend their own self-interests for the sake of the team, organization, or larger collectivity; and (c) expand the followers' portfolio of needs by raising their awareness to improve themselves and what they are attempting to accomplish. Transformational leaders encourage followers to develop and perform at levels above what they may have felt was possible, or beyond their own personal expectations (Bass and Avolio 1993).

Perhaps the major difference between the new paradigm of leadership and the earlier classic approaches mentioned above lies in the nature of the relationships that are built between leaders and employees, that is, the process by which the leaders energize the employees' effort to accomplish the goals, as well as in the type of goals set. According to the new genre of leadership theory, leaders transform the needs, values, preferences, and aspirations of the employees. The consequence of the leader's behaviors is emotional and motivational arousal of the employees, leading to self-reinforcement mechanisms that heighten their efforts and bring performance above and beyond the call of duty or contract (Avolio and Bass 1988; House and Shamir 1993).

2.3. Differences in the Motivational Basis for Work between the Classic and the New Leadership Paradigms

Individuals may approach social situations with a wide range of motivational orientations. The classic transactional and the new transformational paradigms of leadership differ in their assumptions about the core underlying bases for motivation to work. The main shifts in the motivational basis for work concentrate on three dimensions. The first deals with a transition from a calculative–rational toward an emotional–expressive motivation. The second concerns the shift from emphasizing individualistic-oriented motivation toward stressing the centrality of collectivistic-oriented motivation. The third dimension refers to the importance given to intrinsic motivators in addition to the traditional extrinsic incentives.

2.3.1. From a Calculative–Rational toward an Emotional–Expressive Motivation to Work

The transactional approach to leadership is rooted in behavioral and contingency theories. This transactional paradigm is manifested in the long fixation on the dimensions of consideration or people-orientation and initiating structure or task-orientation (review in Yukl 1998). Leadership behavior, as measured by the indexes of initiating structure and consideration, is largely oriented toward accomplishing the tasks at hand and maintaining good relationships with those working with the leader (Seltzer and Bass 1990). When the superior initiates structure, this enhances and reinforces the employees' expectancy that their efforts will succeed. Consideration by the superior is a desired benefit reinforcing employee performance (Bass 1985). This transactional approach was also carried into contingency theories such as the path–goal theory of leadership developed by House (1971). The path–goal theory took as its underlying axioms the propositions of Vroom's expectancy theory, which was the prevailing motivational theory at that time. According to expectancy theory, employees' efforts are seen to depend on their expectancy that their effort will result in their better performance, which in turn will result in valued outcomes for them (Yukl 1998). The path–goal theory to leadership asserted that ''the motivational function of the leader consists of increasing personal payoffs to subordinates for work-goal attainment and making the path to these payoffs easier travel by clarifying it, reducing roadblocks and pitfalls, and increasing the opportunities for personal satisfaction en route'' (House 1971, p. 324). The common denominator of the behavioral and contingency approaches to leadership is that the individual is assumed to be a rational maximizer of personal utility. They all explain work motivation in terms of a rational choice process in which a person decides how much effort to devote to the job at a given point of time. Written from a calculative-instrumental perspective, discussions and applications of work motivation have traditionally emphasized the reward structure, goal structure, and task design as the key factors in work motivation and deemphasized other sources of motivation. Classic theories assume that supervisors, managers, leaders, and their followers are able to calculate correctly or learn expected outcomes associated with the exercise of theoretically specified behaviors. These theories, then, make strong implicit rationality assumptions despite substantial empirical evidence that humans are subject to a myriad of cognitive biases and that emotions can be strong determinants of behavior (House 1995).

The new motivation and leadership paradigms recognize that not all of the relevant organizational behaviors can be explained on the basis of calculative considerations and that other considerations may enrich the explanation of human motivation (Shamir 1990). Transformational and charismatic leadership approaches assume that human beings are not only instrumental-calculative, pragmatic, and goal-oriented but are also self-expressive of feelings, values, and self-concepts. We are motivated to do things because it makes sense to do them from a rational-instrumental point of view, but also because by doing so we can discharge moral obligations or because through such a contribution we can establish and affirm a cherished identity for ourselves. In other words, because it is useful, but also because it is right, or because it ''feels right.'' Making the assumption that humans are self-expressive enables us to account for behaviors that do not contribute to the individual self-interest, the most extreme of which is self-sacrifice (House and Shamir, 1993; Shamir 1990; Shamir et al. 1993). Huy (1999), in discussing the emotional dynamics of organizational change, refers to the emotional role of charismatic and transformational leaders. According to Huy, at the organizational level, the emotional dynamic of encouragement refers to the organization's ability to instill hope among its members. Organizational hope can be defined as the wish that our future work situation will be better than the present one. Transformational leaders emotionally inspire followers through communication of vivid images that give flesh to a captivating vision so as to motivate them to pursue ambitious goals. The most important work for top managers is managing ideology and not strategy making. Transformational leaders can shape an ideological setting that encourages enthusiasm, nurtures courage, reveals opportunities, and therefore brings new hope and life into their organizations.

Thus, classic leadership theories addressed the instrumental aspects of motivation, whereas the new paradigm emphasizes expressive aspects as well. Such emotional appeals can be demonstrated by the words of Anita Roddick, founder and CEO of The Body Shop:

> Most businesses focus all the time on profits, profits, profits . . . I have to say I think that is deeply boring. I want to create an electricity and passion that bond people to the company. You can educate people by their passions . . . You have to find ways to grab their imagination. You want them to feel that they are doing something important . . . I'd never get that kind of motivation if we were just selling shampoo and body lotion. (Conger and Kanungo 1998, pp. 173–174)

2.3.2. From an Individualistic-Oriented toward a Collectivistic-Oriented Motivation to Work

Nearly all classic models of motivation in organizational behavior, in addition to being calculative, are also hedonistic. These classic approaches regard most organizational behavior as hedonistic and treat altruistic, prosocial, or cooperative behavior as some kind of a deviation on the part of the organizational members. This trend may be traced in part to the influence of the neoclassical paradigm in economics and psychology, which is based on an individualistic model of humans. American psychology, from which our work motivation theories have stemmed, has strongly emphasized this individualistic perspective (Shamir 1990).

The recent recognition that not all relevant work behaviors can be explained in terms of hedonistic considerations has led to the current interest in prosocial organizational behaviors, those that are formed with the intent of helping others or promoting others' welfare. However, between totally selfish work behavior and pure altruistic behaviors, many organizationally relevant actions are probably performed both for a person's own sake and for the sake of a collectivity such as a team, department, or organization. Individuals may approach social situations with a wide range of motivational orientations that are neither purely individualistic (concerned only with one's satisfaction) nor purely altruistic (concerned only with maximizing the other's satisfaction). Deutsch (1973, in Shamir, 1990) uses the term *collectivistic* to refer to a motivational orientation that contains a concern with both one's own satisfaction and others' welfare.

The importance of discussing the linkages between individual motivations and collective actions stems from the increasing recognition of the importance of cooperative behaviors for organizational effectiveness. During the 1990s, hundreds of American companies (e.g., Motorola, Cummins Engine, Ford Motor Co.) reorganized around teams to leverage the knowledge of all employees. Now it appears that the concept is going global, and recent research conducted in Western Europe has supported the wisdom of teams. For example, the Ritz-Carlton Hotel Co. has created ''self-directed work teams'' with the goal of improving quality and reducing costs. In the hotel's Tysons Corner, Virginia, facility, a pilot site for the company's program, the use of these teams has led to a decrease in turnover from 56 to 35%. At a cost of $4,000 to $5,000 to train each new employee, the savings were significant. At Air Products, a chemical manufacturer, one cross-functional team, working with suppliers, saved $2.5 million in one year. OshKosh B'Gosh has combined the use of work teams and advanced equipment. The company has been able to increase productivity, speed, and flexibility at its U.S. manufacturing locations, enabling it to maintain 13 of its 14 facilities in the United States, which made one of the few children's garment manufacturers able to do so (Gibson et al. 2000). This shift in organizational structure calls for higher cooperation in group situations. Cooperation, defined as the willful contribution of personal effort to the completion of interdependent jobs, is essential whenever people must coordinate activities among differentiated tasks. Individualism is the condition in which personal interests are accorded greater importance than are the needs of groups.

Collectivism occurs when the demands and interests of groups take precedence over the desires and needs of individuals. Collectivists look out for the well being of the groups to which they belong, even if such actions sometimes require that personal interests be disregarded. A collectivistic, as opposed to an individualistic orientation should influence personal tendencies to cooperate in group situations (Wagner 1995).

A core component incorporated in the new paradigm of leadership is that transformational leaders affect employees to transcend their own self-interests for the sake of the team, organization, or larger collectivity (Bass and Avolio 1990; Burns 1978). The effects of charismatic and transformational leaders on followers' relationships with the collective are achieved through social identification, that is, the perception of oneness with, or belonging to, some human aggregate (Shamir 1990). People who identify with a group or organization take pride in being part of it and regard membership as one of their most important social identities. High social identification may be associated with a collectivistic orientation in the sense that the group member is willing to contribute to the group even in the lack of personal benefits, places the needs of the group above individual needs, and sacrifices self-interest for the sake of the group (Shamir 1990; Shamir et al. 1998). A charismatic or transformational leader can increase social identification by providing the group with a unique identity distinguishing it from other groups. This can be achieved by the skillful use of slogans, symbols (e.g., flags, emblems, uniforms), rituals (e.g., singing the organizational song), and ceremonials. In addition, transformational and charismatic leaders raise the salience of the collective identity in followers' self-concepts by emphasizing shared values, making references to the history of the group or organization, and telling stories about past successes, heroic deeds of members, and symbolic actions by founders, former leaders, and former members of the group (Shamir et al. 1993). Frequent references by the leader to the collective identity and presentation of goals and tasks as consistent with that identity further bind followers' self-concepts to the shared values and identities and increase their social identification (Shamir et al. 1998).

Collectivistic appeals are illustrated by comments from Mary Kay Ash to her sales force at the company's annual convention:

> There was, however, one band of people that the Romans never conquered. These people were the followers of the great teacher from Bethlehem. Historians have long since discovered that one of the reasons for the sturdiness of this folk was their habit of meeting together weekly. They shared their difficulties and they stood side by side. Does this remind you of something? The way we stand side by side and share our knowledge as well as our difficulties with each other at our weekly unit meetings? . . . What a wonderful circle of friends we have. Perhaps it is one of the greatest fringe benefits of our company.
>
> —(Conger and Kanungo 1998, p. 160)

By implication, Mary Kay is saying that collective unity can work miracles in overcoming any odds.

2.3.3. From Extrinsic toward Intrinsic Motivation to Work

A classic categorization for work motivation is the intrinsic–extrinsic dichotomy. Extrinsic needs demand gratification by rewards that are external to the job itself. Extrinsic motivation derives from needs for pay, praise from superiors and peers, status and advancement, or physically comfortable working conditions. Intrinsic needs are satisfied when the activities that comprise the job are themselves a source of gratification. The desire for variety, for meaning and hope, and for challenging one's intellect in novel ways are examples of intrinsic motivation. When one engages in some activity for no apparent external reward, intrinsic motivation is likely at work. There is evidence that extrinsic rewards can weaken intrinsic motivation and that intrinsic rewards can weaken extrinsic motivation. However, intrinsic job satisfaction cannot be bought with money. Regardless of how satisfactory the financial rewards may be, most individuals will still want intrinsic gratification. Money is an important motivator because it is instrumental for the gratification of many human needs. Fulfilling basic existence needs, social needs, and growth needs can be satisfied with money or with things money can buy. However, once these needs are met, new growth needs are likely to emerge that cannot be gratified with money. Pay, of course, never becomes redundant; rather, intrinsic motivators to ensure growth-stimulating challenges must supplement it (Eden and Globerson 1992).

Classic motivational and leadership theories have emphasized both extrinsic and intrinsic rewards. However, transformational leadership theory goes beyond the rewards-for-performance formula as the basis for work motivation and focuses on higher-order intrinsic motivators. The motivation for development and performance of employees working with a transformational leader is driven by internal and higher-order needs, in contrast to the external rewards that motivate employees of transactional leaders (Bass and Avolio 1990). The transformational leader gains heightened effort from employees as a consequence of their self-reinforcement from doing the task. To an extent, transformational leadership can be viewed as a special case of transactional leadership with respect to exchanging effort for rewards. In the case of transformational leadership, the rewarding is internal (Avolio and Bass 1988).

The intrinsic basis for motivation enhanced by transformational and charismatic leaders is emphasized in Shamir et al.'s (1993) self-concept theory. Self-concepts are composed, in part, of identities. The higher the salience of an identity within the self-concept, the greater its motivational significance. Charismatic leaders achieve their transformational effects through implicating the self-concept of followers by the following four core processes:

1. By increasing the intrinsic value of effort, that is, increasing followers' intrinsic motivation by emphasizing the symbolic and expressive aspects of the effort, the fact that the effort itself reflects important values.
2. By empowering the followers not only by raising their specific self-efficacy perceptions, but also by raising their generalized sense of self-esteem, self-worth, and collective efficacy.
3. By increasing the intrinsic value of goal accomplishment, that is, by presenting goals in terms of the value they represent. Doing so makes action oriented toward the accomplishment of these goals more meaningful to the follower in the sense of being consistent with his or her self-concept.
4. By increasing followers' personal or moral commitment. This kind of commitment is a motivational disposition to continue a relationship, role, or course of action and invest effort regardless of the balance of external costs and benefits and their immediate gratifying properties.

2.4. The Full Range Leadership Model: An Integrative Framework

Bass and Avolio (1994) propose an integrative framework that includes transformational, transactional, and nontransactional leadership. According to Bass and Avolio's "full range leadership model," leadership behaviors form a continuum in terms of activity and effectiveness. Transformational leadership behaviors are at the higher end of the range and are described as more active–proactive and effective than either transactional or nontransactional leadership. Transformational leadership includes four components.

Charismatic leadership or idealized influence is defined with respect to both the leader's behavior and employee attributions about the leader. Idealized leaders consider the needs of others over their own personal needs, share risks with employees, are consistent and trustworthy rather than arbitrary, demonstrate high moral standards, avoid the use of power for personal gain, and set extremely challenging goals for themselves and their employees. Taken together, these behaviors make the leader a role model for his or her employees; the leader is admired, respected, trusted, and ultimately identified with over time. Jim Dawson, as the president of Zebco, the world's largest fishing tackle company, undertook a series of initiatives that would symbolically end the class differences between the workforce and management. Among other actions, he asked his management team to be role models in their own arrival times at work. One executive explained Dawson's thinking:

> What we realized was that it wasn't just Jim but we managers who had to make sure that people could see there were no double standards. As a manager, I often work long hours. . . Now the clerical and line staff don't see that we are here until 7 because they are here only until 4:45 p.m. As managers, we don't allow ourselves to use that as an excuse. . . Then it appears to the workforce that you have a double standard. An hourly person cannot come in at 7:05 a.m. He has got to be here at 7:00 a.m. So we have to be here at the same time.
>
> (Conger and Kanungo 1998, pp. 136–137)

Similarly, Lee Iacocca reduced his salary to $1 in his first year at Chrysler because he believed that leadership means personal example.

Inspirational motivation involves motivating employees by providing deeper meaning and challenges in their work. Such leaders energize their employees' desire to work cooperatively to contribute to the collective mission of their group. The motivational role of vision is achieved through the selection of goals meaningful to followers. For example, Mary Kay Ash articulates her company's mission as enhancing women's role in the world and empowering them to become more self-confident. This is a highly appealing message to an organization made up of women for whom selling Mary Kay Cosmetics may be their first working career. The importance of giving meaning to work is also illustrated in the speech of Orit Gadiesh, the vice chairman of Bain and Co., a well-known Boston management consulting firm. After a period of crisis, the financial picture of the company had turned around, but Gadiesh sensed that the organization had lost pride in itself. She chose the company's annual meeting in August 1992 to convey in words what she felt the organization had to realize, a "pride turnaround." Gadiesh said, "We've turned around financially, and we've turned around the business. . . . Now it's time to turn around what they [competitors] really fear, what they have always envied us for, what made them most uneasy—as crazy as this sounds. It's time to turn around our collective pride in what we do!" (Conger and Kanungo 1998, p. 182).

Intellectual stimulation entails the leader's and the employees' questioning of assumptions, reframing problems, and thinking about concepts using novel approaches or techniques. For instance, Richard Branson, the charismatic British entrepreneur who built the diversified, multibillion dollar business of Virgin, sets an example of an innovative and unconventional approach to business. Time after time, Branson has found opportunities in established industries by going against conventions. One of his earliest successful initiatives was in discount stores. When approximately 20 years old, he observed that despite the abolition of a government policy allowing manufacturers and suppliers to "recommend" prices to retail outlets (which had ensured high prices for records), music records remained overpriced. He decided to challenge norms around pricing by offering records through the mail, discounting them some 15% from retail stores. When he entered the airline business in 1984, he distinguished himself from competitors in various ways from lower fares to in-flight masseurs, fashion shows to musicians, and motorcycle transportation to the London airport. Anita Roddick of The Body Shop also expresses a hard-line stance against the traditional practices of the cosmetics industry: "It turned out that my instinctive trading values were dramatically opposed to the standard practices in the cosmetic industry. I look at what they are doing and walk in the opposite direction" (Conger and Kanungo 1998, p. 185).

Individualized consideration represents the leader's effort to understand and appreciate the different needs and viewpoints of employees while attempting continuously to develop employee potential. A soldier in the Israel Defense Forces (IDF) describes his platoon leader's individualized treatment

> On the way back, there was a very steep slope. I finished it with no air in my lungs. We still had 400 meters to run. I was last and I couldn't walk anymore. Then, K., the platoon commander came and said "run." I can't walk and he is talking to me about running. Ten meters before the end he said "now sprint." I did it and got to the camp. I felt great. I felt I had done it. Suddenly I discovered that K. was able to get me to do things that I never thought I could do.
>
> (Landau and Zakay 1994, pp. 32–33)

Donald Burr, the former CEO of People Express Airlines, who was deeply interested in creating a humane organization and was guided by a fundamental belief in people, summarizes his view: "The people dimension is the value added to the commodity. Many investors still don't fully appreciate this point" (Conger and Kanungo 1998, p. 144).

Conversely, transactional leaders exert influence on employees by setting goals, clarifying desired outcomes, providing feedback, and exchanging rewards for accomplishments. Transactional leadership includes three components. With the *contingent reinforcement* style, the leader assigns or secures agreements on what needs to be done and promises or actually rewards others in exchange for satisfactorily carrying out the assignments. Alternatively, using the *management-by-exception—active* style, the leader arranges actively to monitor deviations from standards, mistakes, and errors in the employees' assignments and takes corrective action as necessary. With the *management-by-exception—passive* style, the leader waits for problems to emerge and then takes corrective action. Nontransactional leadership is represented by the *laissez-faire* style, in which the leader avoids intervention, transactions, agreements, or setting expectations with employees (Bass 1996; Bass and Avolio 1993).

One of the fundamental propositions in Bass and Avolio's model is that transformational leadership augments transactional leadership in predicting leader effectiveness. Transformational leadership cannot be effective if it stands alone. Supporting most transformational leaders is their ability to manage effectively, or transact with followers, the day-to-day mundane events that clog most leaders' agenda. Without transactional leadership skills, even the most awe-inspiring transformational leaders may fail to accomplish their intended missions (Avolio and Bass 1988). Indeed, several studies (Hater and Bass 1988; Howell and Frost 1989; Seltzer and Bass 1990) have confirmed this "augmentation hypothesis" and shown that transformational or charismatic leadership explained additional variance in leader effectiveness beyond either transactional leadership or leadership based on initiating structure and consideration. Sometimes, with highly transformational or charismatic leaders, there is a need to find a complementary top manager with a more transactional-instrumental orientation. For example, Colleen Barrett, the executive vice president of Southwest, is the managerial complement to Herb Kelleher's charismatic, loosely organized style. She is a stickler for detail and provides the organizational counterweight to Kelleher's sometimes chaotic style (Conger and Kanungo 1998).

Thus, according to the full range leadership model, every leader displays each style to some degree. Three-dimensional optimal and suboptimal profiles are shown in Figure 3. The depth *frequency* dimension represents how often a leader displays a style of leadership. The horizontal *active* dimension represents the assumptions of the model according to which the laissez-faire style is the most passive style, whereas transactional leadership incorporates more active styles and transformational leadership is proactive. The vertical *effectiveness* dimension is based on empirical results that

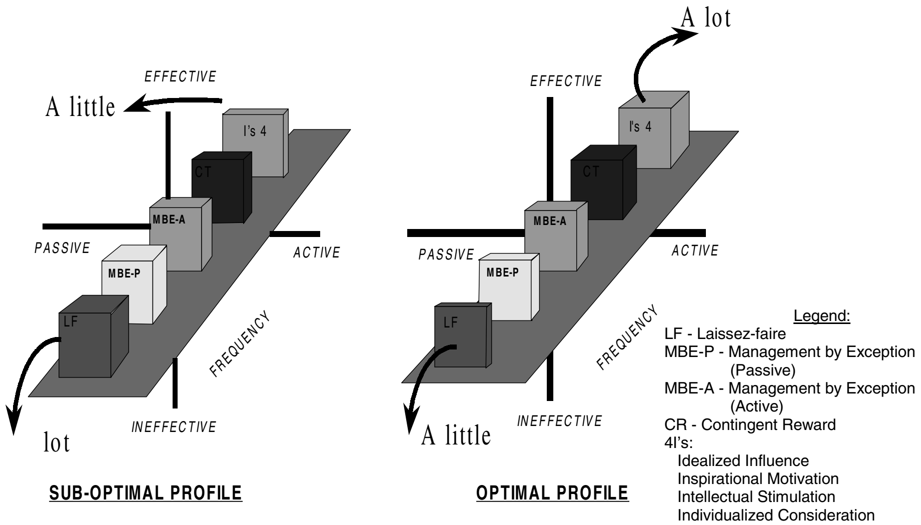

A little
EFFECTIVE
I's 4
CT
MBE-A
PASSIVE
ACTIVE
MBE-P
LF
FREQUENCY
INEFFECTIVE
lot
SUB-OPTIMAL PROFILE
A lot
EFFECTIVE
I's 4
CT
MBE-A
PASSIVE
ACTIVE
MBE-P
LF
FREQUENCY
INEFFECTIVE
A little
OPTIMAL PROFILE
Legend:
LF - Laissez-faire
MBE-P - Management by Exception (Passive)
MBE-A - Management by Exception (Active)
CR - Contingent Reward
4I's:
Idealized Influence
Inspirational Motivation
Intellectual Stimulation
Individualized Consideration

Figure 3 Full Range Leadership Model: Optimal and Suboptimal Profiles as Created by the Frequency of Displaying the Leadership Styles' Behaviors. (Adapted from B. J. Avolio, *Full Leadership Development: Building the Vital Forces in Organizations,* p. 53, copyright © 1999 by Sage Publications, Inc., reprinted by permission of Sage Publications, Inc.)

have shown active transactional and proactive transformational leadership to be far more effective than other styles of leadership. The frequency dimension manifests the difference between optimal and suboptimal profile. An optimal managerial profile is created by more frequent use of transformational and active transactional leadership styles along with less frequent use of passive transactional and laissez-faire leadership. The opposite directions of frequency create a suboptimal profile.

3. LEADERSHIP OUTCOMES

> My message was, and will always be, "There are no limits to the maximum."
> —T., an exemplary infantry platoon commander in the Israel Defense Forces

"Transformational leaders encourage follower to both *develop* and *perform* at levels above what they may have felt was possible, or beyond their own personal expectations" (Bass and Avolio 1990, p. 234, emphasis in original). Thus, achieving certain levels of performance as well as development become a targeted outcome. The linkages between transformational leadership, employee development, and performance are presented in Figure 4.

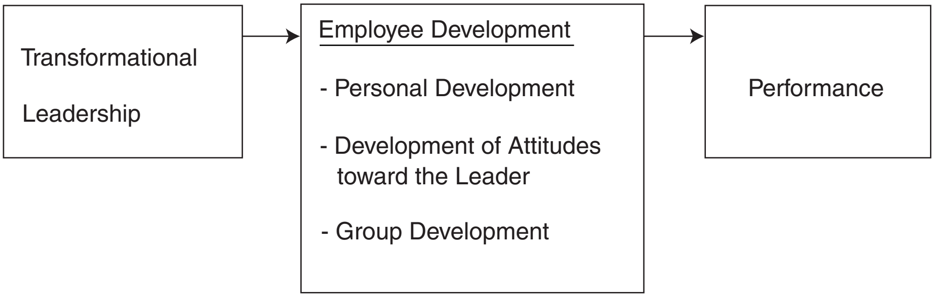

Transformational
Leadership
Employee Development
- Personal Development
- Development of Attitudes toward the Leader
- Group Development
Performance

Figure 4 Linkages among Transformational Leadership, Employee Development, and Performance.

3.1. Performance

Numerous studies have examined the relationships between various components of the full range leadership model and performance outcomes. Overall, these studies confirmed the positive impact of transformational leadership on perceived effectiveness as well as on hard measures of performance.

The most commonly studied perceived outcomes were perceived effectiveness of the leader and the work unit, extra effort, and satisfaction with the leader. In the majority of the studies, both transformational/transactional leadership and the outcome variables were evaluated by the same source, usually the immediate followers. For example, Hater and Bass (1988) found positive high correlations between transformational leadership and perceived effectiveness and satisfaction, both rated by the leaders' immediate followers, whereas the correlations between the transactional factors and followers' ratings were low to moderate. Seltzer and Bass (1990) studied 138 followers and 55 managers and confirmed that transformational leadership adds to transactional leadership (as represented by the dimensions of initiating structure and consideration) in explaining the variance of followers' satisfaction and ratings of leader effectiveness. Several studies have expanded the spectrum of outcome variables. Thus, for example, Bycio et al. (1995) conducted a study among 1376 nurses who evaluated both their superior's leadership style and the outcome variables. They found that in addition to the strong positive relationships between transformational leadership and followers' extra effort, satisfaction with the leader, and perceived leader effectiveness, transformational leadership also had positive correlations with followers' organizational commitment and negative correlations with followers' intentions to leave the profession and the job. Podsakoff et al.'s (1990) study among 988 exempt employees of a large petrochemical company confirmed that "core" transformational leadership behaviors and individualized support had positive effects on followers' trust and satisfaction with the leader. Bass and Avolio (1993) summarize the findings of this type from studies conducted in industrial, military, educational, and religious organizations. The pattern of relationships was consistent across 17 independent samples and pointed to a clear hierarchy. The correlations with perceived effectiveness and satisfaction with the leader typically ranged from 0.60 to 0.80 for transformational leadership, from 0.40 to 0.60 for contingent reward, from −0.30 to 0.30 for management-by-exception, and from −0.30 to −0.60 for laissez-faire leadership.

Fewer studies have collected survey measures of both leadership and outcomes from multiple sources. In these studies, different sources subjectively evaluated transformational/transactional leadership and organizational outcomes. The most frequent additional source used was the evaluation of managers' performance by their superiors. For example, Keller (1992) studied 66 project groups containing professional employees from three industrial research and development organizations. He showed that transformational leadership, as rated by immediate followers, was more predictive of project quality, as evaluated by the management, in research vs. development teams. Yammarino and Dubinsky (1994) collected leadership and perceived effectiveness data from various sources, including 105 salespersons and their 33 sales supervisors, in a $1 billion multinational medical products firm. They confirmed an individual-level positive relationship between transformational leadership, as perceived by followers, and extra effort and perceived performance of followers, as perceived by their superiors.

A growing number of studies have examined the impact of transformational leadership on objective performance measures. For example, Onnen (1987, in Bass and Avolio 1993) reports that transformational leadership of Methodist ministers, as evaluated by their district members, was positively related to Sunday church attendance and to growth in church membership. Avolio et al. (1988) found that transformational leadership and active transactional leadership were positively related to financial measures and productivity among MBA students engaged in a complex, semester-long competitive business simulation. Bryant (1990, in Bass and Avolio 1993) confirms that nursing supervisors who were rated by their followers as more transformational managed units with lower turnover rates. Howell and Avolio (1993) found positive relationships between transformational leadership and objective unit performance over a one-year interval among managers representing the top four levels of management in a large Canadian financial institution. German bank unit performance over longer vs. shorter time periods was higher in banks led by leaders who were rated by their employees as more transformational (Geyer and Steyrer 1998 in Avolio 1999).

However, "even when predictions regarding objective performance outcomes support the model, we are still faced with plausible alternative cause-and-effect relationships" (Bass and Avolio 1993, p. 69). Therefore, to establish causal effects, several experiments (Barling et al. 1996; Crookall 1989; Dvir et al. in press; Howell and Frost 1989; Kirkpatrick and Locke 1996; Sosik et al. 1997) were conducted either in the field or in the laboratory. Overall, these experiments confirmed the causal impact of transformational or charismatic leadership on performance outcomes. Such an experimental design can confirm that the direction of causal flow is indeed from transformational leadership to the hypothesized performance outcomes as opposed to instances where enhanced follower performance cause the higher transformational leadership ratings. Howell and Frost (1989) found that experimentally induced charismatic leadership positively affected task performance, task adjustment, and adjustment to the leader and the group. Kirkpatrick and Locke (1996) manipulated three core components common to charismatic and transformational leadership in a laboratory simulation among 282 students. They found that a vision of high quality significantly affected several attitudes, such as trust in the leader, and congruence between participants' beliefs and values and those communicated through the vision. Vision implementation affected performance quality and quantity. Barling et al. (1996), in a field experiment among 20 branch managers in a large Canadian bank, confirmed the positive impact of transformational leadership training on employees' organizational commitment and on two aspects of branch-level financial performance. Dvir et al. (in press) conducted a longitudinal randomized field experiment among military leaders and their followers. They confirm the positive impact of transformational leadership, enhanced through a training intervention, on direct followers' personal development and on indirect followers' objective performance.

Two meta-analytic studies have been conducted recently. Lowe et al. (1996) found that significantly higher relationships were observed between transformational scales and effectiveness than between transactional scales and effectiveness across 47 samples. This pattern held up across two levels of leadership and with both hard (number of units) and soft (performance appraisals) measures of performance (see Figure 5). Coleman et al. (1995) found that, across 27 studies, the transformational leadership styles were more strongly related to performance than the transactional styles. The average relationship across studies for the transformational leadership factors and performance ranged from 0.45 to 0.60; for transactional leadership, 0.44; for management-by-exception—active, 0.22; for management-by-exception—passive, 0.13; and for laissez-faire, −0.28.

To sum up, there is sufficient empirical evidence to conclude that transformational leadership has a positive impact on both perceived and objective performance and that this impact is stronger than the effects of transactional leadership.

3.2. Employee Development

Rather than solely focusing on the exchange with an eye toward performance, transformational leaders concentrate on developing employees to their full potential. Indeed, one of the most prominent aspects of transformational leadership compared to transactional leadership concerns employee developmental processes (Avolio and Gibbons 1988). The transformational leader evaluates the potential of all employees in terms of their being able to fulfill current commitments and future positions with even greater responsibilities. As a result, employees are expected to be more prepared to take on the responsibilities of the leader's position, and to be "transformed," as Burns (1978) originally argued, from followers into leaders. In contrast, working with the transactional leader, employees are expected to achieve agreed upon objectives but are not expected to undergo a developmental process whereby they will assume greater responsibility for developing and leading themselves and others. Follower

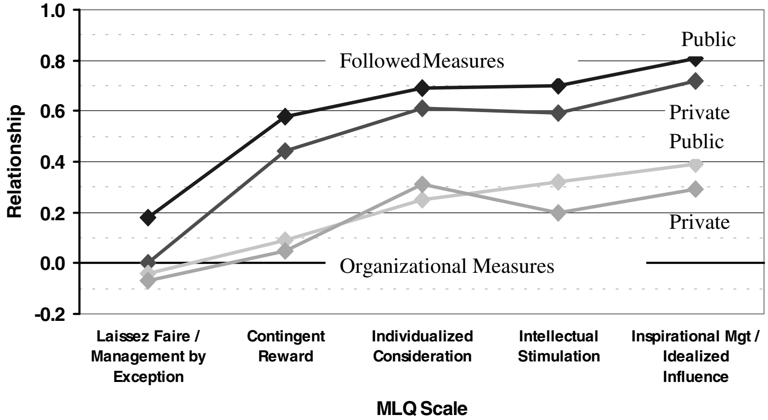


Figure 5 The Relationships between the Full Range Leadership Model's Styles and Leader Effectiveness in Public and Private Organizations in Lowe et al.'s (1996) Meta-analysis of 47 Samples. (MLQ = multifactor leadership questionnaire, which measures transformational, transactional, and nontransactional leadership; Relationship = correlations between leadership style and leader effectiveness; Follower measures = subordinate perceptions of leader effectiveness; organizational measures = quasi-institutional measures of leader effectiveness including both hard measures [e.g., profit] and soft measures [e.g., supervisory performance appraisals].)

developmental variables were classified here into the three broad categories of personal development, development of attitudes toward the leader, and group development.

3.2.1. Employee Personal Development

Most general references in the literature to follower development actually refer to an intrapersonal, developmental change that followers of transformational leaders presumably undergo; the leader transforms followers by serving as a coach, a teacher, and a mentor. Dvir et al. (in press) propose a conceptual framework for examining developmental aspects of transformational leadership theory on the basis of the sporadic theoretical discussions and empirical examinations in the literature. Their framework for follower personal development includes the three domains of motivation, morality, and empowerment. Dvir et al. found evidence of the impact of transformational leaders upon their direct followers' development in at least one measure of each category.

Motivation The emphasis in transformational leadership theory on employee motivational development can be traced back to Burns's (1978) references to two continuums of follower development. The first dealt with motivation and was based on Maslow's (1954) hierarchy of needs. Maslow conceived a hierarchy of human needs according to the following order: physiological, security, social, self-esteem, and self-actualization needs. Self-actualization, the realization of one's potential to become what one has the capacity to become, is at the highest level of need. According to Maslow, only upon satisfaction of the lower level needs does the motivation to satisfy a higher level need arise, and self-actualization needs are infinite. Burns (1978) proposes that transforming compared to transactional leaders motivate followers to achieve the highest possible level of need satisfaction such that the primary motive becomes the desire to satisfy *self-actualization needs* rather than lower needs. Based on Burns, Bass (1985, 1996) suggests that transformational leaders elevate the level of needs along Maslow's hierarchy or expand the employees' portfolio of needs. Unlike transactional leaders, who concentrate on fulfilling current employee needs, transformational leaders arouse needs that may have lain dormant.

Another motivational aspect associated with transformational leadership is the emphasis on employees' *extra effort*. Bass (1985) originally posited extra effort as a manifestation of employee motivation. He claimed that employees' extra efforts show how highly a leader motivates them to perform beyond expectations. Thus, it can be concluded that the emphasis on satisfying self-actualization needs reflects the type of need underlying the employees' motivation, whereas extra effort reflects the level of their motivation. Shamir et al. (1993), in discussing the motivational effects of charismatic leadership, propose that charismatic leaders increase employee effort by increasing the intrinsic value of their efforts through links to valued aspects of the employee's self-concept. Forming these linkages to the employee's self-concept helps to harness the motivational forces of self-expression, self-consistency, self-esteem, and self-worth. Extra effort is one of the most widely confirmed correlates of transformational leadership. Dvir et al. (in press) conducted a randomized field experiment in a military organization and confirmed the positive impact of transformational leadership, enhanced through training, on direct followers' perceived extra effort.

Morality Burns's (1978) second developmental continuum relates to the employees' moral development, based on Kohlberg's (1973) theory. Burns summarizes that "transforming leadership ultimately becomes moral in that it raises the level of human conduct and ethical aspiration of both leader and led, and thus it has a transforming effect on both" (p. 20). Bass (1985) purposely omitted the moral element from his original theory, apart from emphasizing the importance of collective interests. Recent articles have decried the omission of the moral element from transformational and charismatic leadership theories, and have called for its inclusion as part of the developmental process undergone by employees of transformational leaders. Bass (1996) came to agree with Burns that to be transformational, a leader must be morally uplifting. One of the difficulties in applying this part of the theory is that, according to Kohlberg, moving from one moral stage to the next may take years, a time span too long to wait for results to appear in typical leadership studies. Therefore, moral development in the short-term can be represented by the employees' *internalization of organizational moral values* (Dvir et al. in press).

The transition from a desire to satisfy solely personal interests to a desire to satisfy the broader collective interests is part of moral development, according to Kohlberg (1973). Bass (1985) places special emphasis on this aspect of moral development and suggests that transformational leaders get their employees to transcend their own self-interest for the sake of the team or organization. In Shamir's (1991) self-concept theory, the *collectivistic orientation* of the employees is one of the central transformational effects of charismatic leaders. Dvir et al. (in press) confirm the causal impact of transformational leaders on their direct followers' collectivistic orientation.

Empowerment Transformational leadership theory, in contrast to early charismatic leadership theories (e.g., House 1977), never assumed that leadership relationships are based on regression or weakness on the part of the employees. On the contrary, transformational leadership theory has consistently emphasized employees' development toward autonomy and empowerment over automatic followership. Neocharismatic leadership theories have ceased conceptualizing the employees as weak (House 1995; Shamir 1991).

The literature has referred to critical-independent thinking as one of the essential empowerment-related processes undergone by employees of transformational leaders. Bass and Avolio (1990) state that transformational leaders enhance employees' capacities to think on their own, develop new ideas, and question operating rules and systems that no longer serve the organization's purpose or mission. Avolio and Gibbons (1988) posit that a primary goal of transformational leadership is to develop employee self-management and self-development. Shamir's (1991) self-concept theory emphasized the transformational effects of charismatic leaders on employee autonomy, independence, and empowerment. A critical-independent employee as an outcome of transformational leadership also accords Kelley's (1992) conceptualization regarding styles of followership perceived within organizations as good followership. Kelley's respondents described the best followers as those who "think for themselves," "give constructive criticism," "are their own person," and "are innovative and creative." The worst followers were characterized as "must be told what to do" and "don't think." In between were the typical followers who "take direction" and "don't challenge the leader or group." Although Kelley labels this dimension critical-independent thinking, he actually refers to thinking and action, or in his words, "to become a full contributor, you need to cultivate independent, critical thinking and develop the courage to *exercise* it" (p. 114, emphasis added). Thus, it is more appropriate to label this developmental outcome of transformational leadership *critical-independent approach*. Howell and Avolio (1989) confirm the hypothesized relationships between transformational leadership and employee innovation, risk-taking, and creativity among Canadian executives. Dvir et al. (in press) found a positive effect of transformational leadership on direct followers' critical-independent approach toward the self, peers, superiors, and the organizational system.

Kelley's (1992) review of best, worst, and typical follower characteristics reveals a second dimension, namely active engagement in the task. The best followers "take initiative," "participate actively," "are self-starters," and "go above and beyond the job." The worst ones are "passive," "lazy," "need prodding," and "dodge responsibility." The typical followers basically "shift with the wind." Active engagement in the task involves the investment of energy into the followership role, usually toward a mutually shared dream or goal. Thus, *active engagement in the task* is defined as the energy invested in the employee role as expressed by high levels of activity, initiative, and responsibility taken by the employee, and is expected to be positively affected by transformational leadership.

Transformational and charismatic leadership theorists have emphasized the impact of the leader on employees' self-perceptions. Shamir et al. (1993) specify increased employee self-efficacy as one of the transformational effects of charismatic leaders. According to Conger and Kanungo (1988), self-efficacy can be tied to charismatic leadership through empowerment. Avolio and Gibbons (1988, p. 298) stated that "a significant part of developing or transforming followers is developing their feelings of self-efficacy." Eden (1988) argues about the difference between general and specific self-efficacy (GSE and SSE, respectively). GSE is relatively a stable trait and is resistant to change, whereas SSE represents a more labile state. Direct followers of transformational leaders were found to have higher levels of SSE (Dvir et al. in press). Based on the importance given to self-efficacy as an empowerment-related outcome of transformational and charismatic leadership, we posited *specific self-efficacy* as a malleable developmental outcome that is enhanced among employees of transformational leaders.

3.2.2. Development of Employee Attitudes toward the Leader

According to Yukl (1998), transformational leaders influence followers by arousing strong emotions and identification with the leader. The notion of follower development going in the direction of strong emotional attachment or favorable attitudes toward the leader is deeply rooted in charismatic and transformational leadership theory. For example, House (1977) asserts that followers in charismatic relationships feel affection toward the leader. According to Conger and Kanungo (1987), charismatic leaders are more likable and honorable than noncharismatic leaders. Shamir et al. (1993) suggest that personal commitment to the leader is part of the transformational effects of charismatic leaders. Favorable attitudes toward the leader have been the most commonly studied outcomes of transformational leadership. There is a considerable amount of evidence to conclude that transformational leadership enhances employees' *satisfaction with the leader* and *perceived effectiveness of the leader* (Bass and Avolio 1990).

The full range leadership model (Avolio and Bass 1988) suggests that having a shared sense of purpose and common norms between follower and leader should assist transformational leaders in successfully completing and communicating their mission. Furthermore, if norms are not common to both leader and follower, then the first step toward transforming followers may involve the leader's unfreezing followers' attitudes and values and then gaining conformity with his or her own value system. Many writers have emphasized the importance of the similarity between leaders and followers. For example, House and Baetz (1979, in Conger and Kanungo 1987) postulate that the followers of charismatic leaders are characterized by the similarity of their beliefs to those of the leader. The self-concept theory (Shamir 1991; Shamir et al. 1993) posits that the values and identities of followers within charismatic relationships are congruent with the leaders' vision. According to this theory, one of the two classes of charismatic leaders' behaviors is frame alignment, which refers to the linkage of follower and leader interpretive orientations so that the followers' interests, values, and beliefs and the leaders' activities, goals, and ideology become congruent and complementary. Some level of demonstrated and moral superiority is likely to be necessary for the leader to become a representative character, but the leader should also emphasize his or her similarity to the followers on certain dimensions. The main processes of psychological attachment through which the leader's behaviors of frame alignment influence their followers in the charismatic relationships are personal identification and value internalization. Personal identification refers to the followers' desire to emulate or vicariously gain the qualities of the leader. Such identification increases with the extent to which the leader represents desirable identities, values, and attributes in his or her behaviors. Value internalization refers to the incorporation of the leader's or group's values within the self as guiding principles. The self-concept theory stresses that personal identification is consistent with the theory, but a stronger emphasis is put on value internalization.

Several empirical studies have been conducted on the relationships between *leader–follower similarity or value congruence* and various outcome variables. Turban and Jones (1988) found that superior–follower similarity was related to followers' perceived performance. Megalino et al. (1989) showed that workers were more satisfied and committed when their values were congruent with the values of their superior. Enz (1988) found that perceived value congruity between department members and top managers, examined from the perspectives of both groups, accounted for unique variance in departmental power. Kirkpatrick and Locke (1996) found in a laboratory experiment that articulation of a vision emphasizing quality by the leader positively affected the followers' perceived congruence with the beliefs and values communicated through the vision.

3.2.3. Employee Group Development

Several models have explicitly considered leadership to be one determinant of team effectiveness. In addition, recent references to transformational and charismatic leadership have begun to emphasize the importance of group-level outcomes (Avolio 1999; Shamir et al. 1998). For example, Guzzo et al. (1993) suggest that transactional leadership might have indirect effects on group potency and effectiveness, whereas transformational leadership has both indirect and direct effects. Klein and House (1995) recently extended charismatic theory to the group level by conceiving groups as constellations of dyadic relationships. In their extension, groups and organizations vary along the dimension of homogeneity or heterogeneity of group members' charismatic relationships with the leader. Shamir (1990) suggests that leaders can enhance collectivistic motivation, that is, the individual contributions to collective work efforts, through calculative considerations, the internalization of values and personal norms, or the maintenance and affirmation of relevant identities. In his self-concept theory, Shamir (1991) posits collectivistic orientation as an important transformational effect on followers of charismatic leaders. In spite of the emphasis on follower group development in recent frameworks of transformational leadership, very few empirical investigations have connected transformational and transactional leader behaviors and group-level outcomes. Sivasubramaniam et al. (1997, in Avolio 1999) found that transformational leadership directly predicted the performance of work groups while also predicting performance indirectly through levels of group potency. In a longitudinal laboratory experiment, Sosik et al. (1997) largely confirm that transformational leadership affected group potency and group effectiveness more strongly than transactional leadership. Shamir et al. (1998) confirm that the leaders' emphasis on the collective identity of the unit was associated with the strength of the unit culture, as expressed in the existence of unique symbols and artifacts, and with unit viability as reflected by its discipline, morale, and cooperation. More research is needed on the effects of transformational leaders on follower group development.

4. IMPLICATIONS FOR STRATEGIC HUMAN RESOURCE MANAGEMENT

This section will introduce central implications of the implementation of the new genre of leadership theories in organizations. Our goal is to focus on the implications of leadership research in the framework of the major functions of human resource management (HRM). As with other aspects of organizational behavior, leadership and motivation research has contributed to the development of sophisticated HRM tools and techniques that have been shown to improve organizational effectiveness (e.g., Jackson and Schuler 1995). The major functions of HRM, discussed below, are recruiting, performance appraisal, training, and compensation. For each function, we provide a background based on a strategic approach to HRM. We then describe the utility of the above research findings, regarding leadership and motivation, to strategic HRM and provide examples.

4.1. A Strategic Approach to Human Resource Management

As traditional sources of competitive advantage, such as technology and economies of scale, provide less competitive leverage now than in the past, core capabilities and competence, derived from how people are managed have become more important. A strategic perspective of human resource management is critical for creating and sustaining human resource-based competitive advantage (Sivasubramaniam and Ratnam 1998). A strategic approach to HRM implies a focus on "planned HRM deployments and activities intended to enable the form to achieve its goals" (Wright and McMahan 1992, p. 298). Strategic HRM involves designing and implementing a set of internally consistent policies and practices that ensure that a firm's human capital (employees' collective knowledge, skills, and abilities) contributes to the achievement of its business objectives. Internal consistency is achieved by creating vertical and horizontal fit. Vertical fit involves the alignment of HRM practices and strategic management, and horizontal fit refers to the alignment of the different HRM functions (Wright and Snell 1998).

Strategic HRM is concerned with recognizing the impact of the outside environment, competition, and labor markets, emphasizes choice and decision making, and has a long-term focus integrated in the overall corporate strategy. For instance, companies like General Motors that emphasize a retrenchment or cost reduction strategy will have an HRM strategy of wage reduction and job redesign, while the growth corporate strategy of Intel requires that it will aggressively recruit new employees and rise wages (Anthony et al. 1998). Similarly, cost-reduction policies may require that leaders emphasize equity and contingent reward, while growth strategy should enhance transformational leadership. Cost-reduction and growth strategies, as well as other strategies, have clear implications for how companies should recruit employees, evaluate their performance, and design training and compensation systems.

Considerable evidence has shown the performance implications of HRM. These studies indicate that high-performing systems are characterized by careful selection, focus on a broad range of skills, employ multisource appraisals focused on employee development, employ teams as a fundamental building block of the organization's work systems, provide facilitative supervision, include human resource considerations in strategic decision making, and link rewards to performance that focuses the employees on the long-term goals of the firm. These findings have been fairly consistent across industries and cultures (Sivasubramaniam and Ratnam 1998).

4.2. Recruiting: From Hiring to Socialization

Recruiting, the process by which organizations locate and attract individuals to fill job vacancies, should be matched with company strategy and values as well as with external concerns, such as the state of the labor market. Recruiting includes both pre- and post-hiring goals. Among the pre-hiring goals are attracting large pools of well-qualified applicants. Post-hiring goals include socializing employees to be high performers and creating a climate to motivate them to stay with the organization (Fisher et al. 1996). Another major decision in the recruiting philosophy of a firm is whether to rely on hiring outside candidates or to promote from within the organization.

The above decisions and goals have implications for how organizations can benefit from recent research on leadership and motivation. Organizations that recruit from their internal labor market and accept candidates for long-term employment may seek candidates with leadership skills who can be potential managers. Such organization may use mentoring programs as a way of enhancing employee socialization. Other organizations may benefit from using leadership skills as part of the selection procedures they employ, though the use of leadership characteristics as a criterion for selection is highly controversial (e.g., Lowery 1995).

4.2.1. Implications for Selection

While good leadership qualities could be a criterion for selecting employees at all levels, if used at all, it is more common in managerial selection. Managerial selection is a particularly difficult task because there are many different ways to be a successful manager. Managing requires a wide range of skills, one of which could be leadership. Leadership effectiveness was found to relate to common selection criteria, such as practical intelligence, reasoning, creativity, and achievement motivation (Mumford et al. 1993). However, many HRM professionals are reluctant to use leadership as a criterion for selection, citing reasons, such as lack of definition, complexity of the construct, and difficulty in validating selection tools (Lowery 1995). These challenges pose a threat to the validity of the selection process. An invalid selection system may be illegal, as well as strategically harmful for the organization. Consequently, our discussion of leadership as a selection tool is mainly exploratory.

The most common use of leadership criteria in selection is as part of an assessment center (e.g., Russell 1990). Assessment centers are work samples of the job for which candidates are selected (mostly candidates for managerial positions). They can last from one day to one week and have multiple means of assessment, multiple candidates, and multiple assessors. Assessment centers are considered relatively valid predictors of both short- and long-term success in managerial positions (e.g., Russell 1987).

Many assessment centers include leadership evaluation in the form of leaderless group discussions, where candidates are evaluated on small-group management activities (Lowery 1995). In the simple form of leaderless group discussion, the initially leaderless group is assigned a problem to discuss to reach a group decision. Assessors judge who emerges as a leader, initiatives displayed by each participant, and other aspects of interpersonal performance (Bass 1990). Assessors are required to identify behaviors, such as gatekeeping, facilitate support, and evaluate the candidate's ability to accomplish the ''tasks.'' Although assessment centers are generally more suitable for evaluating candidates for promotion, they are also used with candidates for entry (Lowery 1995). In Lowery's review of the applicability of measuring leadership in an assessment center, he argues that we still do not know how to assess leadership as an independent performance dimension in assessment centers.

Recognizing the limitations of assessment centers, Russell (1990) suggested using biodata to select top corporate leaders. According to Kunhert and Lewis (1987), biodata captures early life events, which are hypothesized to impact leadership performance in adulthood. They argued that transformational leaders derive meaning from personal experience in a significantly different way than transactional leaders based on early development. Identifying the life experiences that lead to the development of transformational leaders may contribute to selecting such leaders based on their biodata reports. Indeed, Russell (1990) obtained correlations between prior life experiences and contribution to fiscal performance of candidates for top managerial positions at a Fortune 50 firm. According to Russell, transformational leadership measures may be a good source of anecdotal behavioral examples that correlate with biodata and can serve as tools for executive selection.

More recent attempts to utilize findings from the new genre of leadership theory in selection use both biodata and assessment center procedures. Avolio and colleagues from South Africa have developed a preliminary tool to assess the leadership potential of candidates to the South African police force. This tool is based on the multifactor leadership questionnaire (MLQ) developed by Bass and Avolio (1996) to measure the full range leadership styles. The assessment center also includes a biodata survey that has been used in previous research (Avolio 1994). Similarly, researchers from the Center for Creative Leadership examined the process of executive selection and identified that success in the top three jobs in large organizations was predicted mostly by ''relationships with subordinates'' and ''expressed affection'' (Kouzes 1999).

The above are examples of early application of the new genre of leadership theory in selection. However, since selection procedures require serious legal considerations, the application of findings from recent research to selection is relatively slow. However, recent leadership approaches may provide insights for other recruiting-related functions, such as mentoring.

4.2.2. Implications for Mentoring and Socialization

A strategic HRM approach requires that the staffing process not end once applicants are hired or promoted. To retain and maximize human resources, organizations must emphasize socializing newly hired/promoted employees. A thorough and systematic approach to socializing new employees is necessary if they are to become effective workers. Good socialization is essential to ensure employee understanding of the company's mission and vision. It includes a realistic job preview (RJP) that allows the creation of appropriate expectations for the job (Breaugh 1983). RJP includes presenting realistic information about job demands, the organization's expectations, and the work environment. At a later stage, when employees begin to feel part of the organization, a mentoring program, in which an established worker serves as an adviser to the new employee, may help in the socialization process. Mentoring programs are especially necessary in environments of high turnover, such as high technology (Messmer 1998).

Mentoring programs have both short- and long-term benefits. In addition to developing the skills, knowledge, and leadership abilities of new and seasoned professionals, these programs strengthen employee commitment (Messmer 1998). Mentoring programs have been successful in developing new talent, as well as recruiting junior level staff and bringing them up to speed on policies and procedures. Mentors provide more than company orientation; they become trusted advisors on issues ranging from career development to corporate culture. At a time when many executives are managing information instead of people, mentoring may offer a welcome opportunity to maintain the kinds of critical interpersonal skills that can further a manager's career. Although the mentors in most mentoring programs are not the supervisors of the mentees, the characteristics of good mentors are very similar to the characteristics of good leaders, especially transformational leaders (Bass 1985).

Bass (1985) argues that effective transformational leaders emphasize individualized consideration that involves the leader's service as a counselor for their proteges. Such leaders have more knowledge and experience and the required status to develop their proteges. Johnson (1980) found that two thirds of 122 recently promoted employees had mentors. Bass (1985) claims that since mentors are seen as authorities in the system, their reassurance helps mentees be more ready, willing and able to cooperate in joint efforts. Moreover, mentoring is an efficient method for leadership development because followers are more likely to model their leadership style on that of their mentor, as compared with a manager who is not seen as a mentor (Weiss 1978). To be effective, mentors need to be successful, competent, and considerate (Bass 1985). Companies like Bank of America and Pricewaterhouse-Coopers' employ mentoring mostly to increase retention. In competitive labor markets, such as the high-technology industry, increasing retention is a major HR goal. In PricewaterhouseCoopers', mentoring has led to an increase in female retention (Messmer 1998). Clearly, transformational leadership, and individualized consideration in particular, can provide a strategic advantage to organizations that employ mentoring programs.

The development of mentoring programs should include a review of needs for mentoring, a buy-in from senior management, and constant updating of all managers involved with the mentee. Such programs should also include a selection of mentors based on the above principles, a careful match of participants, and company events to demonstrate the importance of the mentoring program for the organization (Messmer 1998). Top managers, such as the long-time CEO of the Baton Rouge Refinery of Exxon, are seen as more effective if they develop their followers. In the case of Exxon, these followers were later promoted more than their mentor.

We chose to demonstrate the contribution of effective leadership to recruiting using the examples of selection and mentoring. However, leadership and motivation are important parts of other recruiting processes, such as in interviewer skills and assessment of candidates' motivation. We chose those aspects that we feel benefit especially from the new genre of leadership theories. We continue to use this approach with the HRM function of performance management.

4.3. From Performance Appraisal to Performance Management

Performance appraisal is the process by which an employee's contribution to the organization during a specific period of time is assessed. Performance feedback provides the employees with information on how well they performed. If used inappropriately, performance appraisal can be disastrous and lead to a decrease in employee motivation and performance. Performance appraisal has a significant strategic role in HRM. It provides strategic advantage by allowing the organization to monitor both its individual employees and the extent to which organizational goals are met. From a strategic HRM perspective, performance appraisal involves more than assessment or measurement. Rather, it is a method for performance management, including defining performance, measuring it, and providing feedback and coaching to employees (Fisher et al. 1996). In sum, performance management has three major strategic functions. First, it signals to employees which behaviors are consistent with organizational strategy. Second, if used appropriately, it is a useful method for organizational assessment. Finally, it is a feedback mechanism that promotes employee development. The movement from a focus on appraisal to performance management allows for better application of the new genre of leadership and motivation theory.

While the traditional performance-appraisal systems emphasized uniform, control-oriented, narrow-focus appraisal procedures that involved supervisory input alone, strategic performance management emphasizes customization, multipurpose, and multiple raters, focusing on development rather than past performance. Similarly, whereas traditional approaches to leadership and motivation (e.g., House 1971) emphasized control and exchange based on extrinsic motivators, more recent approaches emphasize development and influence through intrinsic motivators (e.g., Bass 1985; Shamir et al. 1993). Indeed, even during the 1980s and the beginning of the 1990s, employees and managers viewed performance appraisal negatively. Longnecker and Gioia (1996, in Vicere and Fulmer 1996) found that out of 400 managers, only one quarter were satisfied with appraisal systems and the higher a manager rose in the organization, the less likely he or she was to receive quality feedback. Nevertheless, they indicated that when used as a strategic tool for executives, systematic feedback was crucial for them. Many of them were also willing to be evaluated personally. According to Vicere and Fulmer (1996), performance-management processes are highly linked with leadership development. We chose to focus on the contributions of leadership research to two aspects of performance management: feedback from multiple raters (360°) and alignment of strategic goals.

4.3.1. Implications for Feedback

As we mentioned above, strategic performance-management systems involve more than the traditional supervisory evaluation. Current appraisal systems often include peer ratings, follower ratings, and self-ratings. Each source seems to contribute differently and their combination (a 360° approach) is used mainly for developmental reasons (Waldman et al. 1998). About 12% of American organizations are using full 360° programs for reasons such as management development, employee involvement, communication, and culture change (Waldman et al. 1998). Advocates of 360° systems believe that they contribute to leadership development by allowing managers to compare their self-perceptions with those of their employees (Yammarino and Atwater 1997).

The gaps between self and other ratings were found to predict the perceived performance of the leader. Agreement between self and other reports on *high* ratings predicted high performance (Yammarino and Atwater 1997). Gaps between self and other ratings may also relate to the leadership style of the targeted manager. Transformational executives who had ratings congruent with those of their followers were seen as highly considerate, emphasized meetings with their employees, and associated individual employee goals with organizational goals.

Nontransformational leaders, who had little congruence with their followers' ratings, were seen as detached and as having little or no impact (Berson and Avolio 1999). These findings may contribute to a better understanding of the gaps between raters and provide insight into the type of training and development that can be given to managers following performance appraisal. This is especially important given recent criticism of the use of 360° in performance management. Waldman et al. (1998) argued that 360° feedback, in addition to being a high cost, has become a fad rather than a systematically and appropriately used tool. Rather, 360° systems should be tailored to organizational needs.

4.3.2. Implications for Alignment and Signaling of Strategic Goals

A major strategic advantage of performance management is that it is a method by which managers signal to employees the mission of the unit or the organization. Managers can use performance management to align their employees' behaviors with organizational goals or strategy. Employees want to be rewarded and will engage in behaviors that their supervisors emphasize. We expect transactional leaders to use performance appraisal as a mechanism of contingent reward (Bass 1985; Bass and Avolio 1994). Such leaders will mostly benefit from performance appraisal in the short term. In order to achieve a long-term alignment with organizational goals, supervisors need to have followers who identify with them. Transformational or charismatic leaders may be able to use performance management to align followers with organizational goals and inspire and motivate them to achieve these goals. These leaders will be better communicators of organizational goals. Indeed, effective appraisal depends on the supervisor's ability to communicate organizational goals.

In a study of a large telecommunication company, Berson and Avolio (2000) found that employees of transformational leaders viewed their managers as more open and informal and as better conveyers and receivers of information than employees of nontransformational leaders. Moreover, employees of effective communicators were more aware of strategic organizational goals than employees who reported to less effective communicators. These findings hint that supervisors who are transformational leaders will be able to utilize performance-management systems to communicate strategic goals more effectively than transactional supervisors for whom performance management may serve as a contract with employees. In addition to better communication of organizational goals, charismatic leaders may also use performance-management systems to sense the environment. Such leaders can use performance management to assess the capabilities of the organization and determine quality goals (Conger and Kanungo 1998).

In summary, evidence from studies in the new genre of leadership theories offers insights regarding the strong impact of transformational or charismatic leaders on their followers. We believe that performance management is an excellent tool for such leaders to communicate strategic goals and sense followers' needs for personal development.

4.4. From Training to Development

With the rapid developments in technology and the changing nature of jobs, the provision of training and development exercises has become a major strategic objective for the organization. Organizations provide training for many reasons, from orientation to the organization and improving performance on the job to preparation for future promotion. Like other HRM functions, training should be aligned with the strategic goals of the organization. Organizations that rely on a highly committed, stable workforce will have to invest more in individuals than will organizations that employ unskilled temporary employees (Fisher et al. 1996).

When a company changes strategy, as Xerox did when it became "the Document Company," it begins intensive training to equip employees with the new skills. Xerox spent $7 million on developing a training center. In a world of networking and alliances between organizations, Xerox needs this center not only to train its employees but to train suppliers, customers, and other constituencies. Xerox, like other employers, has a long-term perspective on its workforce and is using training to increase employee retention. More and more professionals seek jobs that will provide them with the opportunity for professional development. Many engineers who work in high-technology industries have multiple job opportunities and are constantly in the job market (Messmer 1998). Consequently, organizations that wish to decrease turnover move from emphasizing short-term training for specific purposes to long-term development.

We believe that the shift to development calls for more leadership of the new genre type. While the short-term emphasis on training, where employees are trained to learn a new skill that they need for their current task or job, required mainly task-oriented transactional leadership, the move to long-term development entails transformational leadership. The need for more transformational or charismatic leaders in organizations is a consequence of several trends that reflect the shift from training to development. First, in addition to sending employee to off-the-job training workshops, more companies emphasize individual coaching. Coaching, like other types of on-the-job training, ensures maximized transfer between what was learned in the training and the actual job (Ford 1997). On-the-job training also allows for maximum trainee motivation because trainees see this type of training as highly relevant.

Second, as managerial jobs become more complex, there is an increased need for managerial training. Most typologies of managerial tasks emphasize that managers need to have leadership skills. Specifically, skills such as providing feedback, communication, motivation, aligning individual goals with organizational goals, and having good relationships with colleagues and followers should be the focus of development (Kraut et al. 1989).

Finally, as discussed above regarding socializing employees, coaching and mentoring programs have become more popular, sometimes as spontaneous mentor–protégé relationship and sometimes as an organized mentoring program (Messmer 1998).

These shifts in training, together with the emphasis on long-term development, highlight the importance of developing transformational leadership in the organization. Transformational leaders develop significant relationships with their followers and thus have a chance to be seen as coaches and to provide practical on-the-job training. These leaders are also intellectually stimulating and can constantly challenge employees and develop them. Personal development involves individualized consideration, another major style of transformational leaders. More specifically, individualized consideration consists of mentoring (Bass 1985).

Furthermore, both intellectual stimulation and individualized consideration can be cultivated and nurtured not only at the individual level but also at the team and organizational levels (Avolio and Bass 1995). Organizations can create cultures that encourage coaching and development while providing consideration and recognition of individual needs. Paul Galvin, a CEO of Motorola, created a culture, based on his leadership style, where risk-taking is advocated and seen as the most secure approach (Avolio 1999). Employees are encouraged to be creative and develop new products.

Intellectually stimulating leaders also empower their followers (Bass 1985). Empowerment is an effective method of training because employees learn more tasks and are responsible for problem solving. It is especially important in innovative environments where employees have to learn to take risks. Organizations that employ empowering techniques try to applaud both successful and unsuccessful risk-taking behavior. For example, Lockheed Martin takes a team approach to empowerment. Employees who have knowledge regarding an aspect of a project are encouraged to speak up (Wolff 1997, in Anthony et al. 1998). In this way, employees get to experience managerial roles.

Several studies using the transformational leadership paradigm have demonstrated the relationships between transformational leadership and empowerment. Masi (1994, in Avolio 1999) reported positive correlations between transformational leadership and empowering organizational culture norms. Moreover, transformational leadership at the top may filter to lower levels of the organization. Bass and Avolio (1994) suggested the "falling dominoes effect," where leaders who report to transformational leaders tend to be transformational as well. Bryce (1988, in Avolio 1999) found support for this effect in a study of Japanese senior company leaders. Finally, Spreitzer and Jansaz (1998, in Avolio 1999) found that empowered managers are seen as more intellectually stimulating and charismatic than nonempowered managers. These findings demonstrate that by being transformational, executives can create a culture that facilitates professional development for their employees. Even when their managers are empowered by other managers, followers seem to attribute charisma and intellectual stimulation to the empowered leader.

In addition to the contribution of transformational leaders to employee and managerial development, organizations can use transformational leadership as a model for training leadership. There seems to be agreement that leadership is to some extent innate but to some extent trainable (e.g., Avolio 1999). From the models of the new genre of leadership, training using the full range leadership model has been the most extensive. Crookall (1989) compared the effectiveness of training Canadian prison shop supervisors in transformational leadership and in situational leadership using a control group that did not receive any training. Crookall concluded that the transformational leadership workshop had a significant positive impact on two of the four performance measures subjectively evaluated by superiors, and the situational leadership training had significant positive influence on one of the four subjective performance measures. There was no change in the perceived performance of the untrained group. In addition, a significant positive impact was found in both training groups on turnover, work habits, and managers' evaluations regarding the personal growth of prisoners. Significant improvement regarding respect for superiors, professional skills, and good citizenship as evaluated by the manager, were found only for the transformational leadership workshop.

A more recent study (Barling et al. 1996) demonstrated that bank branch managers who received transformational leadership training had a more positive impact on follower commitment and unit financial performance than did those who did not receive training. Finally, Dvir et al. (in press) conducted a comprehensive randomized field experiment comparing the impact of military leaders who received training in transformational leadership vs. leaders who participated in an eclectic leadership training program. Indirect followers of leaders who participated in the transformational training demonstrated superior performance on a battery of objective measures. Direct followers of leaders who got the transformational training showed higher levels of personal development over time. The above findings, together with the extensive training conducted with the transformational leadership model in a wide range of organizations (Avolio 1999), suggest that organizations can benefit from the advantages of the new-genre leaders by training managers at all levels to become transformational (Bass 1990).

4.5. Compensation: From Transactional to Intrinsic Reward Systems

We conclude our discussion of HRM functions with compensation systems. Our goal is to demonstrate that although compensation systems seem to be a classic example of contractual or transactional conduct between employees and the organization, transformational leaders can utilize this strong rewarding agent to motivate employees to perform beyond expectations (Bass 1985).

Organizations expect employees to provide their services and in return to be compensated. Compensation exists to help employees fulfill their own needs. Consequently, classic compensation systems are based on pure exchange or contractual relationships between employees and the organization. In effect, innovative organizations use compensation systems as a strategic tool. Compensation, like performance-management systems, can help signal to employees the major objectives of the organization (e.g., customer focus), to attract and retain employees, encourage skill development, motivate employees to perform effectively, and help shape organizational culture (Fisher et al. 1996). Compensation is considered the most important HRM function and is seen as crucial by all employees. It includes different aspects of pay and benefits. Although compensation systems are supposed to motivate employees, there is high variability in their effectiveness.

HRM professionals and researchers have been using motivation theory to examine compensation systems. Most of these theories view compensation as a transaction and its components as extrinsic motivators. However, these theories tend to explain why employees do not just work for money (Fisher et al. 1996). For example, according to equity theory, individuals determine whether they are being fairly treated by comparing their input/outcome ratio to the input/outcome ratio of someone else. However, individuals may be motivated to work by factors beyond the direct input/output ratio of their peers. For example, the success of the software company SAS, which is growing at an annual rate of more than 25%, cannot be explained using equity models. Although most software companies pay their employees with bonuses and stock options, SAS offers an intellectually rewarding environment and a family corporate culture. SAS has a 4% turnover rate and is the largest privately owned software company (Pfeffer 1998). According to equity theory, there is little equity between SAS employees and other software employees in similar positions. However, based on turnover rates, it seems that SAS employees are quite motivated. Theories that make motivation contingent on compensation are, according to Pfeffer (1998), a myth that has little or no supporting evidence.

The intellectually engaging work and friendly environment as well as the ability to work using state-of-the-art technology (Pfeffer 1998) that SAS provides its employees can explain their extra effort according to the new genre of leadership theories. SAS employees and other employees do indeed expect monetary rewards, but their extra effort relates to intrinsic factors such as leadership and organizational culture. As Bass and Avolio (1994) suggest, this is an example of how transformational leadership augments transactional or contingent reward behaviors.

Transformational leaders have employees who trust them and are committed to work for more then short-term goals. Transactional leaders, who rely on contingent reward and management by exception, teach employees that they should work only when rewarded or when they are closely monitored. Such leaders signal to their followers that they do not trust them. Consequently, these employees will be less motivated than employees who feel that their supervisors do trust them. Indeed, companies like Southwest Airlines recognize these issues and highlight managing through trust and respect rather than exceptional compensation packages (Bass 1985; Pfeffer 1998).

Moreover, transformational leaders provide a personal example to employees. Executives who get paid over 80 times more than their employees do not signal to employees that pay is motivating. For example, Whole Food Markets pays nobody in the company more than 8 times the average company salary. Similarly, when the CEO of Southwest Airlines asked pilots to freeze their salaries for another year, he froze his own salary for four years. On the other hand, paying executives bonuses when employees are being laid off, as was done by GM in the 1980s, sends the wrong message to employees (Pfeffer 1998). Motivating employees using compensation has to be aligned with other organizational practices, including exemplification of top managers. By providing an example, executives send a message that they share a common fate and that the organization emphasizes a culture of cooperation and teamwork rather than competition.

Compensation systems often highlight individual incentives (Fisher et al. 1996). Among these incentives are commissions, bonuses, and merit pay. Although these methods lead to employee satisfaction, they have little relationship with organizational outcomes that can be measured only at the unit or the organizational levels. Indeed, when organizations want to encourage teamwork, rewarding individuals may be extremely harmful. They end up with a compensation system that undermines teamwork and focuses on short-term goals. Such compensation systems send a mixed message to employees regarding organizational goals. As the director of corporate industrial relations at Xerox said, ‘‘if managers seeking to improve performance or solve organizational problems use compensation as the only lever, they will get two results: nothing will happen, and they will spend a lot of money. That's because people want more out of their jobs than just money'' (Pfeffer 1998). Transformational leaders appeal to collective needs of employees and help align their individual goals with organizational goals (Shamir et al. 1993). Organizations need to emphasize collective rewards, such as profit-sharing and gain-sharing plans, where employees share organizational gains. These methods may motivate teamwork to a certain extent, but effective teamwork will still depend on leaders and the culture or environment that they create. Rather, transformational leaders may be able to use compensation and its components to signal what is important and thus shape company culture.

In summary, the traditional notion that compensation systems are effective motivators is based on a transactional contractual approach to employee–management relations. More recent approaches to compensation emphasize rewarding at least at the team if not the organizational level and posit that monetary rewards should accompany intrinsic rewards, such as trust and recognition. These recent approaches also emphasize decentralizing pay decisions (Gomez-Mejia et al. 1998). We argue here that these new compensation procedures can be better applied by transformational or charismatic leaders. Applying the full range leadership model, we argue that such leaders may also exhibit contingent reward and management by exception to support the basic compensation contract. However, aligned with transformational leadership, the above compensation procedures can lead to maximum motivation.

4.6. Involvement-Transformational vs. Inducement-Transactional HRM Systems: An Integrative Framework

One common characteristic of most extant research is the use of the ‘‘parts view'' of HRM. Typically, researchers identify dimensions of HRM to identify the dimensions that were significantly related to performance. New approaches to HRM argue that to be considered strategic, the set of practices should be part of a system, interrelated, interdependent, and mutually reinforcing. HRM systems are best viewed as configurations, systems of mutually reinforcing and interdependent set of practices that provide the overarching framework for strategic actions. The configurational approach to HRM has gained support in recent empirical studies (review in Sivasubramaniam and Ratnam 1998). Dyer and Holder (1988) have proposed three clusters of human resource strategies, labeled inducement, investment, and involvement.

Inducement human resource strategy is based on the concept of motivation through rewards and punishment. Companies following this philosophy tend to emphasize pay and benefits, perquisites, or assignment to nontraditional work environments as inducement to perform and remain with the firm. In addition, such firms demand high levels of reliable role behaviors, define jobs narrowly, hire on the basis of meeting minimum qualifications, and provide directive supervision. The inducement strategy is most closely linked to transactional leadership (Kroeck 1994).

Investment human resource strategy is built around extensive training and development. Companies adhering to this strategy place a premium on the long-range education of employees and expect employees to exercise a fair amount of initiative and creativity in carrying out their tasks. Dominant corporate values are personal growth, respect, equity, justice, and security, not autonomy and empowerment. Due to the developmental aspects of this strategy, some (e.g., Kroeck 1994) link it to the individualized consideration style in the full range leadership model. However, because this strategy often takes on a paternalistic approach to management, and because of the lack of emphasis on autonomy and empowerment, core elements in the relationships between transformational leaders and their followers, we will refer to this strategy as paternalistic.

Involvement human resource strategy is built around creating a very high level of employee commitment. Employees are motivated by the stimulus of autonomy, responsibility, and variety and of being able to see their contribution to the final product or service. The employees at these companies are expected to exercise considerable initiative and creativity as well as display a high degree of flexibility to adapt to rapid change. Team-based work systems are the building blocks of the organization, and supervision is facilitative and minimal. The involvement strategy calls for all four styles of transformational leadership (Kroeck 1994).

It was found that in general, firms following an involvement strategy outperformed firms pursuing inducement strategy. For example, Sivasubramaniam, et al. (1997, in Avolio 1999) conducted a study among 50 Indian firms and found higher correlations between involvement-transformational human

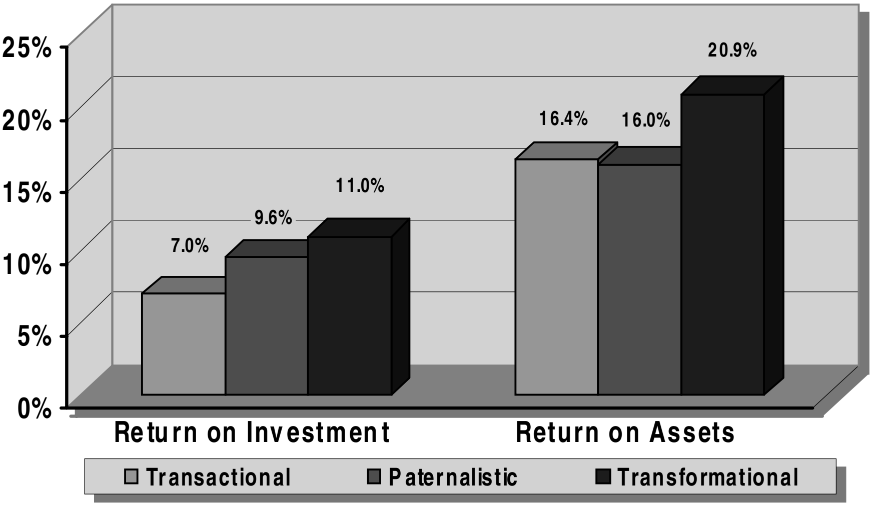


Figure 6 Relationships between Type of Human Resource Management System and Financial Performance among 50 Indian Firms According to Sivasubramaniam, et al.'s (1997) Findings.

resource strategy and the firm's return on investment and return on assets compared to the inducement-transactional and investment-paternalistic strategies (see Figure 6). Sivasubramaniam and Ratnam (1998) found that firms pursuing an involvement-transformational human resource strategy outperformed the other firms in actual accounting returns. Specifically, adopting an involvement human resource strategy translated into nearly a 4.5% edge in return on investment. For a typical firm in this study, this would mean an additional profit of nearly $400,000 per year, a sustainable advantage accrued due to superior human resource practices. It should be noted that these results were obtained after controlling for past year's profitability and hence cannot be dismissed as possibly due to profitable firms having the resources to invest in innovative human resource practices. These firms also reported lower levels of employee turnover as compared to their closest competitors, as well as superior human resource performance. In contrast, inducement-transactional type firms outperformed other firms in labor productivity because their primary focus is on efficiency.

5. CONCLUSION

An overwhelming body of evidence supports the notion that transformational leadership exists at many hierarchical levels, in a variety of organizational settings, and across many cultures. This new genre of leadership theories holds different assumptions on human motivation than traditional leadership theories. The emphasis has shifted from calculative, individualistic, short-term, extrinsic motivation to work toward more expressive, collectivistic, long-term, intrinsic motivators. Numerous studies have supported the greater effectiveness of transformational leadership compared to transactional and nontransactional leadership in enhancing employees' development and performance.

The advantages of transformational leadership have contributed to practical applications in the major HRM functions. Beginning with recruiting, transformational leaders emphasize long-term socialization rather than focusing on hiring alone. Moreover, the commitment of such leaders to employee growth transcends performance appraisal to become developmentally oriented performance-management systems. Furthermore, such systems can be used as a basis for programs that view employee skills and potential contribution rather than training for specific tasks. Finally, while transactional leaders rely mostly on monetary rewards that build contractual relations with employees, transformational leaders augment these rewards with intrinsic motivators that lead to building a highly committed workforce.

As the market conditions get tougher and competitors attempt to duplicate every single source of advantage, what may be sustainable and inimitable by competition is the human resource-based advantage (Sivasubramaniam and Ratnam 1998). Therefore, an integrated, interdependent, and mutually reinforcing HRM system that represents the new leadership and motivational paradigms may contribute to organizational effectiveness.

REFERENCES

Anthony, W. P., Perrewe, P. L., and Kacmar, K. M. (1998), *Human Resource Management: A Strategic Approach,* Dryden, Fort Worth, TX.

Avolio, B. J. (1994), "The 'Natural': Some Antecedents to Transformational Leadership," *International Journal of Public Administration,* Vol. 17, pp. 1559–1580.

Avolio, B. J. (1999), *Full Leadership Development: Building the Vital Forces in Organizations,* Sage, Thousand Oaks, CA.

Avolio, B. J., and Bass B. M. (1988), "Transformational Leadership, Charisma and Beyond," in *Emerging Leadership Vistas,* J. G. Hunt, H. R. Baliga, H. P. Dachler, and C. A. Schriesheim, Eds., Heath, Lexington, MA.

Avolio, B. J., and Bass. B. M. (1995), "Individual Consideration Viewed at Multiple Levels of Analysis: A Multi-Level Framework for Examining the Diffusion of Transformational Leadership," *Leadership Quarterly,* Vol. 6, pp. 199–218.

Avolio, B. J., and Gibbons, T. C. (1988), "Developing Transformational Leaders: A Life Span Approach," in *Charismatic Leadership: The Elusive Factor in Organizational Effectiveness,* J. A. Conger and R. N. Kanungo, Eds., Jossey-Bass San Francisco.

Avolio, B. J., Waldman, D. A., and Einstein, W. O. (1988), "Transformational Leadership in a Management Simulation: Impacting the Bottom Line," *Group and Organization Studies,* Vol. 13, pp. 59–80.

Barling, J., Weber, T., and Kelloway, K. E. (1996), "Effects of Transformational Leadership Training on Attitudinal and Financial Outcomes: A Field Experiment," *Journal of Applied Psychology,* Vol. 81, pp. 827–832.

Bass B. M. (1985), *Leadership and Performance Beyond Expectations,* Free Press, New York.

Bass B. M. (1990), *Bass and Stogdill's Handbook of Leadership: Theory, Research and Management Applications,* Free Press, New York.

Bass B. M. (1996), *A New Paradigm of Leadership: An Inquiry into Transformational Leadership,* Army Institute for the Behavioral and Social Sciences, Alexandria, VA.

Bass B. M., and Avolio, B. J. (1990), "The Implications of Transactional and Transformational Leadership for Individual, Team, and Organizational Development," *Research in Organizational Change and Development,* Vol. 4, pp. 231–272.

Bass B. M., and Avolio, B. J. (1993), "Transformational Leadership: A Response to Critiques," in *Leadership Theory and Research: Perspectives and Directions,* M. M. Chemers and R. Ayman, Eds., Academic Press, San Diego.

Bass B. M., and Avolio, B. J. (1994), "Introduction," in B. M. Bass and B. J. Avolio, Eds., *Improving Organizational Effectiveness Through Transformational Leadership,* Sage, Thousand Oaks, CA.

Bass B. M., and Avolio, B. J. (1996), *Manual for the Multifactor Leadership Questionnaire,* Mind Garden, Palo Alto, CA.

Berson, Y., and Avolio, B. J. (1999), "The Utility of Triangulating Multiple Methods for the Examination of the Level(s) of Analysis of Leadership Constructs," Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology (Atlanta, April).

Berson, Y., and Avolio, B. J. (2000), "An Exploration of Critical Links between Transformational and Strategic Leadership," Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology (New Orleans, April).

Boal, K. B., and Bryson J. M. (1988), "Charismatic Leadership: A Phenomenological and Structural Approach," in Emerging Leadership Vistas, J. G. Hunt, B. R. Baliga, H. P. Dachler, and C. A. Schriesheim, Eds., Lexington Books, Lexington, MA.

Breaugh, J. A. (1983), "Realistic Job Previews: A Critical Appraisal and Future Research Directions, Academy of Management Review, Vol. 8, pp. 612–620.

Burns, J. M. (1978), *Leadership,* Harper and Row, New York.

Bycio, P., Hackett, R. D., and Allen, J. S. (1995), "Further Assessments of Bass's (1985) Conceptualization of Transactional and Transformational Leadership," *Journal of Applied Psychology,* Vol. 80, pp. 468–478.

Coleman, E. P., Patterson, E., Fuller, B., Hester, K., and Stringer, D. Y. (1995), "A Meta-Analytic Examination of Leadership Style and Selected Follower Compliance Outcomes," University of Alabama.

Conger, J. A., and Kanungo, R. N. (1987), "Toward a Behavioral Theory of Charismatic Leadership in Organizational Settings," *Academy of Management Review,* Vol. 12, pp. 637–647.

Conger, J. A., and Kanungo, R. N. (1988), "Behavioral Dimensions of Charismatic Leadership," in *Charismatic Leadership: The Elusive Factor in Organizational Effectiveness,* J. A. Conger and R. N. Kanungo, Eds., Jossey-Bass San Francisco.

Conger, J. A., and Kanungo, R. N. (1998), *Charismatic Leadership in Organizations,* Sage, Thousand Oaks, CA.

Crookall, P. (1989), "Leadership in Prison Industry," Doctoral dissertation, University of Western Ontario, London, ON, Canada.

Dvir, T., Eden, D., Avolio, B. J., and Shamir, B. (in press), "Impact of Transformational Leadership on Follower Development and Performance: A Field Experiment," *Academy of Management Journal.*

Dyer, L., and Holder, G. W. (1988), "A Strategic Perspective of Human Resource Management," in *Human Resource Management: Evolving Roles and Responsibilities,* L. Dyer, Ed., BNA, Washington, DC, pp. –46.

Eden, D. (1988), "Pygmalion, Goal Setting, and Expectancy: Compatible Ways to Raise Productivity," *Academy of Management Review,* Vol. 13, pp. 639–652.

Eden, D., and Globerson, S. (1992), "Financial and Nonfinancial Motivation," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York.

Enz, C. A. (1988), "The Role of Value Congruity in Intraorganizational Power," *Administrative Science Quarterly,* Vol. 33, pp. 284–304.

Fisher, C. D., Schoenfeldt, L. F., and Shaw, J. B. (1996), *Human Resource Management,* 3rd Ed., Houghton Mifflin, Boston.

Ford, J. K. (1997), "Advances in Training Research and Practice: An Historical Perspective," in *Improving Training Effectiveness in Work Organizations,* K. J. Ford, S. W. J. Kozlowski, K. Kraiger, E. Salas, and M. S. Teachout, Eds., Lawrence Erlbaum, Mahwah, NJ.

Gibson, J. L., Ivancevich, J. M., and Donnelly, J. H. (2000), *Organizations: Behavior, Structure, Processes,* 10th Ed., Irwin/McGraw-Hill, Boston.

Gomez-Mejia, L. R., Balkin, D. B., and Cardy, R. L. (1998), *Managing Human Resources,* Prentice Hall, Upper Saddle River, NJ.

Guzzo, R. A., Yost, P. R., Campbell, R. J., and Shea, G. P. (1993), "Potency in Groups: Articulating a Construct," *British Journal of Social Psychology,* Vol. 32, pp. 87–106.

Hater, J. J., and Bass B. M. (1988), "Superiors" Evaluations and Subordinates' Perceptions of Transformational and Transactional Leadership," *Journal of Applied Psychology,* Vol. 73, pp. 695–702.

House, R. J. (1971), "A Path Goal Theory of Leader Effectiveness," *Administrative Science Quarterly,* Vol. 16, pp. 321–338.

House, R. J. (1995), "Leadership in the Twenty-First Century," in *The Changing Nature of Work,* A. Howard, Ed., Jossey-Bass, San Francisco.

House, R. J., and Shamir, B. (1993), "Toward the Integration of Transformational, Charismatic, and Visionary Theories," in *Leadership Theory and Research: Perspectives and Directions,* M. M. Chemers and R. Ayman, Eds., Academic Press, San Diego, CA.

Howell, J. M. (1996), *Organization Contexts, Charismatic and Exchange Leadership*. Unpublished manuscript, University of Western Ontario, London, ON.

Howell, J. M., and Avolio, B. J. (1989), "*Transformational versus Transactional Leaders: How They Impart Innovation, Risk-Taking, Organizational Structure and Performance,*" Paper presented at the Annual Meeting of the Academy of Management (Washington, DC, August).

Howell, J. M., and Frost, P. J. (1989), "A Laboratory Study of Charismatic Leadership," *Organizational Behavior and Human Decision Processes,* Vol. 43, pp. 243–269.

Hughes, R. L., Ginnett, R. C., and Curphy, G. J. (1998), *Leadership: Enhancing the Lessons of Experience,* Irwin, Chicago.

Jackson, S. E., and Schuler, R. S. (1995), "Understanding Human Resource Management in the Context of Organizations and Their Environments," *Annual Review of Psychology,* Vol. 46, pp. 237–264.

Johnson, M. C. (1980), "Speaking from Experience: Mentors—the Key to Development and Growth," *Training and Development Journal,* Vol. 34, pp. 55–57.

Keller, R. T. (1992), "Transformational Leadership and the Performance of Research and Development Project Groups," *Journal of Management,* Vol. 18, No. 3, pp. 489–501.

Kelley, R. E. (1992), *The Power of Followership,* Currency Doubleday, New York.

Kirkpatrick, S. A., and Locke, E. A. (1996), "Direct and Indirect Effects of Three Core Charismatic Leadership Components on Performance and Attitudes," *Journal of Applied Psychology,* Vol. 81, pp. 36–51.

Klein, K. J., and House, R. J. (1995), "On Fire: Charismatic Leadership and Levels of Analysis," *Leadership Quarterly,* Vol. 6, pp. 183–198.

Kohlberg, L. (1973), "Stages and Aging in Moral Development: Some Speculations," *Gerontologist,* Vol. 13, No. 4, pp. 497–502.

Kouzes, J. M. (1999), "Getting to the Heart of Leadership," *Journal for Quality and Participation,* Vol. 22, p. 64.

Kraut, A. I., Pedigo, P. R., McKenna, D. D., and Dunnette, M. D. (1989), "The Roles of the Manager: What's Really Important in Different Management Jobs," *Academy of Management Executive,* Vol. 3, pp. 286–292.

Kroeck, G. K. (1994), "Corporate Reorganization and Transformations in Human Resource Management," in *Improving Organizational Effectiveness Through Transformational Leadership,* B. M. Bass and B. J. Avolio, Eds., Sage, Thousand Oaks, CA.

Kunhert, K. W., and Lewis, P. (1987), "Transactional and Transformational Leadership: A Constructive/Developmental Analysis," *Academy of Management Review,* Vol. 12, pp. 648–657.

Landau, O., and Zakay, E. (1994), *Ahead and with Them: The Stories of Exemplary Platoon Commanders in The Israel Defense Forces,* IDF School for Leadership Development, Netania, Israel (in Hebrew).

Lowe, K. B., Kroeck. K. G., and Sivasubramaniam, N. (1996), "Effectiveness Correlates of Transformational and Transactional Leadership: A Meta-Analytic Review of the MLQ Literature," *Leadership Quarterly,* Vol. 7, pp. 385–425.

Lowery, P. E. (1995), "The Assessment Center Process: Assessing Leadership in the Public Sector," *Public Personnel Management,* Vol. 24, pp. 443–450.

Maslow, A. H. (1954), *Motivation and Personality,* Harper, New York.

Megalino, B. M., Ravlin, E. C., and Adkins, C. L. (1989), "A Work Values Approach to Corporate Culture: A Field Test of the Value Congruence and Its Relationship to Individual Outcomes," *Journal of Applied Psychology,* Vol. 74, pp. 424–432.

Messmer, M. (1998), "Mentoring: Building Your Company's Intellectual Capital," *HR Focus,* Vol. 75, pp. 11–12.

Mumford, M. D., O'Connor, J., Clifton, T. C., Connelly, M. S., and Zaccaro, S. J. (1993), "Background Data Constructs as Predictors of Leadership Behavior," *Human Performance,* Vol. 6, pp. 151–195.

Pfeffer, J. (1998), "Six Dangerous Myths about Pay," *Harvard Business Review,* Vol. 76, pp. 108–119.

Podsakoff, P. M., Mackenzie, S. B., Moorman, R. H., and Fetter, R. (1990), "Transformational Leader Behaviors and Their Effects on Followers' Trust in Leader, Satisfaction, and Organizational Citizenship Behaviors," *Leadership Quarterly,* Vol. 1, pp. 107–142.

Russell, C. J. (1987), "Person Characteristics versus Role Congruency Explanations for Assessment Center Ratings," *Academy of Management Journal,* Vol. 30, pp. 817–826.

Russell, C. J. (1990), "Selecting Top Corporate Leaders: An Example of Biographical Information," *Journal of Management,* Vol. 16, pp. 73–86.

Seltzer, J., and Bass B. M. (1990), "Transformational Leadership: Beyond Initiation and Consideration," *Journal of Management,* Vol. 16, No. 4, pp. 693–703.

Shamir, B. (1990), "Calculations, Values, and Identities: The Sources of Collectivistic Work Motivation," *Human Relations,* Vol. 43, No. 4, pp. 313–332.

Shamir, B. (1991), "The Charismatic Relationship: Alternative Explanations and Predictions," *Leadership Quarterly,* Vol. 2, pp. 81–104.

Shamir, B., House, R. J., and Arthur, M. B. (1993), "The Motivational Effects of Charismatic Leadership: A Self-Concept Based Theory," *Organization Science,* Vol. 4, No. 2, pp. 1–17.

Shamir, B., Zakai, E., Breinin, E., and Popper, M. (1998), "Correlates of Charismatic Leader Behavior in Military Units: Individual Attitudes, Unit Characteristics and Performance Evaluations," *Academy of Management Journal,* Vol. 41, pp. 387–409.

Sivasubramaniam, N., and Ratnam, C. S. V. (1998), "Human Resource Management and Firm Performance: The Indian Experience," Paper presented at the Annual Meeting of The Academy of Management (San Diego, August).

Sosik, J. J., Avolio, B. J., and Kahai, S. S. (1997), "Effects of Leadership Style and Anonymity on Group Potency and Effectiveness in a Group Decision Support System Environment," *Journal of Applied Psychology,* Vol. 82, pp. 89–103.

Vicere, A. A., and Fulmer, R. M. (1996), *Leadership by Design,* Harvard Business School Press, Boston.

Wagner, J. A. (1995), ‘‘Studies of Individualism–Collectivism: Effects on Cooperation in Groups,’’ *Academy of Management Journal,* Vol. 38, pp. 152–172.

Waldman, D. A., Atwater, L. E., and Antonioni, D. (1998), ‘‘Has 360 Degree Feedback Gone Amok?,’’ *Academy of Management Executive,* Vol. 12, pp. 86–96.

Weiss, H. M. (1978), ‘‘Social Learning of Work Values in Organizations,’’ *Journal of Applied Psychology,* Vol. 63, pp. 711–718.

Wright, P. M., and McMahan, G. C. (1992), ‘‘Theoretical Perspectives for Strategic Human Resource Management,’’ *Journal of Management,* Vol. 18, pp. 295–320.

Wright, P. M., and Snell, S. A. (1998), ‘‘Toward a Unifying Framework for Improving Fit and Flexibility in Strategic Human Resource Management,’’ *Academy of Management Review,* Vol. 23, pp. 756–772.

Yammarino, F. J., and Atwater, L. E. (1997), ‘‘Do Managers See Themselves as Others See Them? Implications for Self–Other Rating Agreement for Human Resource Management,’’ *Organizational Dynamics,* Vol. 25, pp. 35–44.

Yammarino, F. J., and Dubinsky, A. J. (1994), ‘‘Transformational Leadership Theory: Using Levels of Analysis to Determine Boundary Conditions,’’ *Personnel Psychology,* Vol. 47, pp. 787–811.

Yukl, G. A. (1998), *Leadership in Organizations,* Prentice Hall, Englewood Cliffs, NJ.

ADDITIONAL READING

House, R. J., ‘‘A 1976 Theory of Charismatic Leadership,’’ in *Leadership: The Cutting Edge,* J. G. Hunt and L. L. Larson, Eds., Southern Illinois University Press, Carbondale, IL, 1977.

Howell, J. M., and Avolio, B. J., ‘‘Transformational Leadership, Transactional Leadership, Locus of Control, and Support for Innovation: Key Predictors of Consolidated Business-Unit Performance,’’ *Journal of Applied Psychology,* Vol. 78, pp. 891–902, 1993.

Huy, Q. H., ‘‘Emotional Capability, Emotional Intelligence, and Radical Change,’’ *Academy of Management Review,* Vol. 24, pp. 325–345, 1999.

Turban, D. B., and Jones, A. P.,‘‘Superior–Subordinate Similarity: Types, Effects, and Mechanisms,’’ *Journal of Applied Psychology,* Vol. 73, pp. 228–234, 1988.

CHAPTER 33

Job and Team Design

Gina J. Medsker
Human Resources Research Organization

Michael A. Campion
Purdue University

1. **INTRODUCTION** **869**
    - 1.1. Job Design 869
    - 1.2. Team Design 870
2. **JOB DESIGN** **870**
    - 2.1. Mechanistic Job Design 870
        - 2.1.1. Historical Background 870
        - 2.1.2. Design Recommendations 874
        - 2.1.3. Advantages and Disadvantages 874
    - 2.2. Motivational Job Design 874
        - 2.2.1. Historical Background 874
        - 2.2.2. Design Recommendations 875
        - 2.2.3. Advantages and Disadvantages 875
    - 2.3. Perceptual/Motor Job Design 875
        - 2.3.1. Historical Background 875
        - 2.3.2. Design Recommendations 875
        - 2.3.3. Advantages and Disadvantages 876
    - 2.4. Biological Job Design 876
        - 2.4.1. Historical Background 876
        - 2.4.2. Design Recommendations 876
        - 2.4.3. Advantages and Disadvantages 876
    - 2.5. Conflicts and Trade-offs among Approaches 876
3. **TEAM DESIGN** **877**
    - 3.1. Historical Background 877
    - 3.2. Design Recommendations 878
        - 3.2.1. Input Factors 878
        - 3.2.2. Process Factors 879
        - 3.2.3. Output Factors 880
    - 3.3. Advantages and Disadvantages 880
4. **IMPLEMENTATION ADVICE FOR JOB AND TEAM DESIGN** **882**
    - 4.1. When to Consider Design and Redesign of Work 882
    - 4.2. Procedures to Follow 884
        - 4.2.1. Procedures for the Initial Design of Jobs or Teams 884
        - 4.2.2. Procedures for Redesigning Existing Jobs or Teams 885
    - 4.3 Methods for Combining Tasks 885
    - 4.4. Individual Differences among Workers 886
    - 4.5. Some Basic Decisions 888
    - 4.6. Overcoming Resistance to Change 889
5. **MEASUREMENT AND EVALUATION OF JOB AND TEAM DESIGN** **889**
    - 5.1. Using Questionnaires to Evaluate Job and Team Design 889
    - 5.2. Choosing Sources of Data 892
    - 5.3. Evaluating Long-Term Effects and Potential Biases 892
    - 5.4. Example of an Evaluation of a Job Design 893
    - 5.5. Example of an Evaluation of a Team Design 893

REFERENCES **894**

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

1.1. Job Design

Job design is one of those aspects of managing organizations that is so commonplace that it often goes unnoticed. Most people realize the importance of job design when an organization is being built and production processes are being developed. Some even recognize the importance of job design when changes are taking place in organizational structures and processes. However, few people realize that job design may be affected when organizations grow, retrench, or reorganize, when managers use their discretion in the assignment of miscellaneous tasks on a day-to-day basis, or when the people in the jobs or their managers change. Fewer yet realize that job design change can be used as an intervention to enhance important organizational goals.

Many different aspects of an organization influence job design, including an organization's structure, technology, processes, and environment. A discussion of these influences is beyond the scope of this chapter, but they are dealt with in other sources (e.g., Davis 1982; Davis and Wacker 1982). These influences impose constraints on how jobs are designed. However, considerable discretion often exists in the design of jobs in many organizational situations. The job (defined as a set of tasks performed by a worker) is a convenient unit of analysis in developing new organizations or changing existing ones.

Several distinctions are useful to clarify the terminology in this chapter. One distinction is between a task and a job. A task is a set of actions performed by a worker who transforms inputs into outputs through the use of tools, equipment, or work aids. The actions of the task may be physical, mental, or interpersonal. On the other hand, a job is an aggregation of tasks assigned to a worker (Gael 1983; U.S. Department of Labor 1972). When the same set of tasks is performed by more than one worker, those workers are said to have the same job.

It is often useful to distinguish among positions, jobs, and occupations. A position is the set of tasks performed by one worker (e.g., the specific industrial engineering position held by employee X). A job is a set of similar positions (e.g., the industrial engineering positions in manufacturing in a particular company). The tasks performed in a given position are usually a combination of tasks that are common to all workers in that job and of tasks that are unique to that position. The unique tasks are sometimes a function of organizational requirements (e.g., different product or equipment) and sometimes a function of the disposition of the particular worker (e.g., different strengths or interests). An occupation is a collection of similar jobs (e.g., all industrial engineering jobs across companies). Job design usually focuses, by definition, on the job level. Differences in design between positions are assumed to be small and are often ignored. This may not be the case in all situations, however. And there can be great differences in design across jobs within an occupation.

Among the most prolific writers on job design in the industrial engineering literature over the last 35 years has been Louis Davis and his associates (Davis 1957; Davis et al. 1955; Davis and Taylor 1979; Davis and Valfer 1965; Davis and Wacker 1982, 1987). As he and his colleagues point out, many of the personnel and productivity problems in industry may be the direct result of the design of jobs. Job design can have a strong influence on a broad range of important efficiency and human resource outcomes:

- Productivity
- Quality
- Job satisfaction
- Training time
- Intrinsic work motivation
- Staffing
- Error rates
- Accident rates
- Mental fatigue
- Physical fatigue
- Stress
- Mental ability requirements
- Physical ability requirements
- Job involvement
- Absenteeism
- Medical incidents
- Turnover
- Compensation rates

As indicated by many of these outcomes, job-design decisions can influence other human resource systems. For example, training programs may need to be developed, revised, or eliminated. Hiring standards may need to be developed or changed. Compensation levels may need to be increased or decreased. Performance appraisal can be affected due to changed responsibilities. Promotion, transfer, and other employee-movement systems may also be influenced. Thus, aspects of many human resource programs may be dictated by initial job-design decisions or may need to be reconsidered following job redesign. In fact, human resource outcomes may constitute the goals of the design or redesign project. Research supporting these outcomes is referenced below during the description of the approaches.

Unfortunately, many people mistakenly view the design of jobs as technologically determined, fixed, and inalterable. However, job designs are actually social inventions that reflect the values of the era in which they were constructed. These values include the economic goal of minimizing immediate costs (Davis et al. 1955; Taylor 1979) and the theories of human motivation that inspire work designers (Steers and Mowday 1977; Warr and Wall 1975). These values, and the designs they influence, are not immutable but subject to change and modification (Campion and Thayer 1985).

The question is, what is the best way to design a job? In fact, there is no single best way. There are actually several major approaches to job design. Each derives from a different discipline and reflects different theoretical orientations and values. This chapter describes the various approaches and their advantages and disadvantages. It highlights the trade-offs and compromises that must be made in choosing among these approaches. This chapter provides tools and procedures for developing and assessing jobs in all varieties of organizations.

1.2. Team Design

This chapter also compares the design of jobs for individuals working independently to the design of work for teams of individuals working interdependently. The major approaches to job design usually focus on designing jobs for individual workers. In recent years, design of work around groups or teams, rather than at the level of the individual worker, has become more popular (Goodman et al. 1987; Guzzo and Shea 1992; Hollenbeck et al. 1995; Tannenbaum et al. 1996). New manufacturing systems and advancements in understanding of team processes have encouraged the use of team approaches (Gallagher and Knight 1986; Majchrzak 1988).

In designing jobs for teams, one assigns a task or set of tasks to a group of workers rather than to an individual. The team is then considered to be the primary unit of performance. Objectives and rewards focus on team, rather than individual, behavior. Team members may be performing the same tasks simultaneously or they may break tasks into subtasks to be performed by different team members. Subtasks may be assigned on the basis of expertise or interest, or team members may rotate from one subtask to another to provide variety and cross-train members to increase their breadth of skills and flexibility (Campion et al. 1994).

The size, complexity, or skill requirements of some tasks seem to naturally fit team job design, but in many cases there may be a considerable degree of choice regarding whether to design work around individuals or teams. In such situations, job designers should consider the advantages and disadvantages of the different design approaches in light of the organization's goals, policies, technologies, and constraints (Campion et al. 1993, 1996). The relative advantages and disadvantages of designing work for individuals and for teams are discussed in this chapter, and advice for implementing and evaluating the different work-design approaches is presented.

2. JOB DESIGN

This chapter is based on an interdisciplinary perspective on job design. That is, several approaches to job design are considered, regardless of the scientific disciplines from which they came. Interdisciplinary research on job design has shown that several different approaches to job design exist; that each is oriented toward a particular subset of outcomes for organizations and employees; that each has costs as well as benefits; and that trade-offs are required when designing jobs in most practical situations (Campion 1988, 1989; Campion and Berger 1990; Campion and McClelland 1991, 1993; Campion and Thayer 1985). The four major approaches to job design are reviewed below in terms of their historical development, design recommendations, and benefits and costs. Table 1 summarizes the approaches, while Table 2 provides detail on specific recommendations.

2.1. Mechanistic Job Design

2.1.1. Historical Background

The historical roots of job design can be traced back to the concept of the division of labor, which was very important to early thinking on the economies of manufacturing (Babbage 1835; Smith 1981). The division of labor led to job designs characterized by specialization and simplification. Jobs designed in this fashion had many advantages, including reduced learning time, reduced time

TABLE 1 Interdisciplinary Approaches to Job Design and Human Resource Benefits and Costs

APPROACH/Discipline Base (example references)	Illustrative Recommendations	Illustrative Benefits	Illustrative Costs
MECHANISTIC/Classic Industrial Engineering (Gilbreth 1911; Niebel 1988; Taylor 1911)	Increase specialization simplification repetition automation Decrease spare time	Decrease in training staffing difficulty making errors mental overload and fatigue mental skills and abilities compensation	Increase in absenteeism boredom Decrease in satisfaction motivation
MOTIVATIONAL/Organizational Psychology (Hackman and Oldham 1980; Herzberg 1966)	Increase variety autonomy significance skill usage participation feedback recognition growth achievement	Increase in satisfaction motivation involvement performance customer service catching errors Decrease in absenteeism turnover	Increase in training time/cost staffing difficulty making errors mental overload stress mental skills and abilities compensation
PERCEPTUAL-MOTOR/Experimental Psychology, Human Factors (Salvendy 1987; Sanders and McCormick 1987)	Increase lighting quality display and control quality user-friendly equipment Decrease information-processing requirements	Decrease in making errors accidents mental overload stress training time/cost staffing difficulty compensation mental skills and abilities	Increase in boredom Decrease in satisfaction
BIOLOGICAL/Physiology, Biomechanics, Ergonomics (Astrand and Rodahl 1977; Grandjean, 1980; Tichauer, 1978)	Increase seating comfort postural comfort Decrease strength requirements endurance requirements environmental stressors	Decrease in physical abilities physical fatigue aches and pains medical incidents	Increase in financial cost inactivity

Advantages and disadvantages are based on findings in previous interdisciplinary research (Campion 1988, 1989; Campion and Berger 1990; Campion and McClelland 1991, 1993; Campion and Thayer 1985).

TABLE 2 Multimethod Job Design Questionnaire (MJDQ)

(Specific Recommendations from Each Job Design Approach)

Instructions: Indicate the extent to which each statement is descriptive of the job, using the scale below. Circle answers to the right of each statement. Scores for each approach are calculated by averaging applicable items.

Please use the following scale:

(5) Strongly agree
(4) Agree
(3) Neither agree nor disagree
(2) Disagree
(1) Strongly disagree
( ) Leave blank if do not know or not applicable

Mechanistic Approach

1. *Job specialization:* The job is highly specialized in terms of purpose, tasks, or activities.	1	2	3	4	5
2. *Specialization of tools and procedures:* The tools, procedures, materials, etc. used on this job are highly specialized in terms of purpose.	1	2	3	4	5
3. *Task simplification:* The tasks are simple and uncomplicated.	1	2	3	4	5
4. *Single activities:* The job requires you to do only one task or activity at a time.	1	2	3	4	5
5. *Skill simplification:* The job requires relatively little skill and training time.	1	2	3	4	5
6. *Repetition:* The job requires performing the same activity(s) repeatedly.	1	2	3	4	5
7. *Spare time:* There is very little spare time between activities on this job.	1	2	3	4	5
8. *Automation:* Many of the activities of this job are automated or assisted by automation.	1	2	3	4	5

Motivational Approach

9. *Autonomy:* The job allows freedom, independence, or discretion in work scheduling, sequence, methods, procedures, quality control, or other decision making.	1	2	3	4	5
10. *Intrinsic job feedback:* The work activities themselves provide direct and clear information as to the effectiveness (e.g., quality and quantity) of job performance.	1	2	3	4	5
11. *Extrinsic job feedback:* Other people in the organization, such as managers and coworkers, provide information as to the effectiveness (e.g., quality and quantity) of job performance.	1	2	3	4	5
12. *Social interaction:* The job provides for positive social interaction such as team work or coworker assistance.	1	2	3	4	5
13. *Task/goal clarity:* The job duties, requirements, and goals are clear and specific.	1	2	3	4	5
14. *Task variety:* The job has a variety of duties, tasks, and activities.	1	2	3	4	5
15. *Task identity:* The job requires completion of a whole and identifiable piece of work. It gives you a chance to do an entire piece of work from beginning to end.	1	2	3	4	5
16. *Ability/skill-level requirements:* The job requires a high level of knowledge, skills, and abilities.	1	2	3	4	5
17. *Ability/skill variety:* The job requires a variety of knowledge, skills, and abilities.	1	2	3	4	5
18. *Task significance:* The job is significant and important compared with other jobs in the organization.	1	2	3	4	5
19. *Growth/learning:* The job allows opportunities for learning and growth in competence and proficiency.	1	2	3	4	5
20. *Promotion:* There are opportunities for advancement to higher level jobs.	1	2	3	4	5
21. *Achievement:* The job provides for feelings of achievement and task accomplishment.	1	2	3	4	5
22. *Participation:* The job allows participation in work-related decision making.	1	2	3	4	5
23. *Communication:* The job has access to relevant communication channels and information flows.	1	2	3	4	5

TABLE 2 (*Continued*)

24. *Pay adequacy:* The pay on this job is adequate compared with the job requirements and with the pay in similar jobs.	1	2	3	4	5
25. *Recognition:* The job provides acknowledgement and recognition from others.	1	2	3	4	5
26. *Job security:* People on this job have high job security.	1	2	3	4	5
Perceptual/Motor Approach					
27. *Lighting:* The lighting in the workplace is adequate and free from glare.	1	2	3	4	5
28. *Displays:* The displays, gauges, meters, and computerized equipment on this job are easy to read and understand.	1	2	3	4	5
29. *Programs:* The programs in the computerized equipment on this job are easy to learn and use.	1	2	3	4	5
30. *Other equipment:* The other equipment (all types) used on this job is easy to learn and use.	1	2	3	4	5
31. *Printed job materials:* The printed materials used on this job are easy to read and interpret.	1	2	3	4	5
32. *Workplace layout:* The workplace is laid out such that you can see and hear well to perform the job.	1	2	3	4	5
33. *Information-input requirements:* The amount of information you must attend to in order to perform this job is fairly minimal.	1	2	3	4	5
34. *Information-output requirements:* The amount of information you must output on this job, in terms of both action and communication, is fairly minimal.	1	2	3	4	5
35. *Information-processing requirements:* The amount of information you must process, in terms of thinking and problem solving, is fairly minimal.	1	2	3	4	5
36. *Memory requirements:* The amount of information you must remember on this job is fairly minimal.	1	2	3	4	5
37. *Stress:* There is relatively little stress on this job.	1	2	3	4	5
Biological Approach					
38. *Strength:* The job requires fairly little muscular strength.	1	2	3	4	5
39. *Lifting:* The job requires fairly little lifting, and/or the lifting is of very light weights.	1	2	3	4	5
40. *Endurance:* The job requires fairly little muscular endurance.	1	2	3	4	5
41. *Seating:* The seating arrangements on the job are adequate (e.g., ample opportunities to sit, comfortable chairs, good postural support, etc.).	1	2	3	4	5
42. *Size differences:* The workplace allows for all size differences between people in terms of clearance, reach, eye height, leg room, etc.	1	2	3	4	5
43. *Wrist movement:* The job allows the wrists to remain straight without excessive movement.	1	2	3	4	5
44. *Noise:* The workplace is free from excessive noise.	1	2	3	4	5
45. *Climate:* The climate at the workplace is comfortable in terms of temperature and humidity, and it is free of excessive dust and fumes.	1	2	3	4	5
46. *Work breaks:* There is adequate time for work breaks given the demands of the job.	1	2	3	4	5
47. *Shift work:* The job does not require shift work or excessive overtime.	1	2	3	4	5
For jobs with little physical activity due to single workstation add:					
48. *Exercise opportunities:* During the day, there are enough opportunities to get up from the workstation and walk around.	1	2	3	4	5
49. *Constraint:* While at the workstation, the worker is not constrained to a single position.	1	2	3	4	5
50. *Furniture:* At the workstation, the worker can adjust or arrange the furniture to be comfortable (e.g., adequate legroom, footrests if needed, proper keyboard or work surface height, etc.).	1	2	3	4	5

Adapted from Campion et al. (1993). See reference and related research (Campion et al. 1996) for reliability and validity information. Scores for each preference/tolerance are calculated by averaging applicable items.

for changing tasks or tools, increased proficiency from the repetition of the same tasks, and the development of special-purpose tools and equipment.

A very influential person for this early perspective on job design was Frederick Taylor (Taylor 1911; Hammond 1971), who expounded the principles of scientific management, which encouraged the study of jobs to determine the ‘‘one best way’’ to perform each task. Movements of skilled workers were studied using a stopwatch and simple analysis. The best and quickest methods and tools were selected, and all workers were trained to perform the job in the same manner. Standards were developed, and incentive pay was tied to the standard performance levels. Gilbreth was also a key founder of this job-design approach (Gilbreth 1911). Through the use of time and motion study, he tried to eliminate wasted movements in work by the appropriate design of equipment and placement of tools and materials.

Surveys of industrial job designers indicate that this mechanistic approach to job design, characterized by specialization, simplification, and time study, was the prevailing practice throughout the last century (Davis et al. 1955; Davis and Valfer 1965). These characteristics are also the primary focus of many modern-day writers on job-design (Barnes 1980; Niebel 1988; Mundel 1985; also see Chapter 38). The discipline base is indicated as ‘‘classic’’ industrial engineering in Table 1. Modern-day industrial engineers may practice a variety of approaches to job design, however.

2.1.2. Design Recommendations

Table 2 provides a brief list of statements that describe the essential recommendations of the mechanistic approach. In essence, jobs should be studied to determine the most efficient work methods and techniques. The total work in an area (e.g., department) should be broken down into highly specialized jobs that are assigned to different employees. The tasks should be simplified so that skill requirements are minimized. There should also be repetition in order to gain improvement from practice. Idle time should be minimized. Finally, activities should be automated or assisted by automation to the extent possible and economically feasible.

2.1.3. Advantages and Disadvantages

The goal of this approach is to maximize efficiency, in terms of both productivity and the utilization of human resources. Table 1 summarizes some of the human resource advantages and disadvantages that have been observed in previous research. Jobs designed according to the mechanistic approach are easier and less expensive to staff. Training times are reduced. Compensation requirements may be less because skill and responsibility are reduced. And because mental demands are less, errors may be less common.

The mechanistic approach also has disadvantages. Too much of the mechanistic approach may result in jobs that are so simple and routine that employees experience less job satisfaction and motivation. Overly mechanistic work can lead to health problems from the physical wear that can result from highly repetitive and machine-paced work.

2.2. Motivational Job Design

2.2.1. Historical Background

Encouraged by the human relations movement of the 1930s (Mayo 1933; Hoppock 1935), people began to point out the unintended drawbacks of the overapplication of the mechanistic design philosophy in terms of worker attitudes and health (Argyris 1964; Blauner 1964; Likert 1961). Overly specialized and simplified jobs were found to lead to dissatisfaction (Caplan et al. 1975; Karasek 1979; Kornhauser 1965; Shepard 1970) and to adverse physiological consequences for workers (Frankenhaeuser 1977; Johansson et al. 1978). Jobs on assembly lines and other machine-paced work were especially troublesome in this regard (Salvendy and Smith 1981; Walker and Guest 1952). These trends led to an increasing awareness of the psychological needs of employees.

The first efforts to enhance the meaningfulness of jobs simply involved the exact opposite of specialization. It was recommended that tasks be added to jobs, either at the same level of responsibility (i.e., job enlargement) or at a higher level (i.e., job enrichment) (Ford 1969; Herzberg 1966). This job-design trend expanded into a pursuit of identifying and validating the characteristics of jobs that make them motivating and satisfying (Griffin 1982; Hackman and Lawler 1971; Hackman and Oldham 1980; Turner and Lawrence 1965). This approach to job design considers the psychological theories of work motivation (Mitchell 1976; Steers and Mowday 1977; Vroom 1964), thus this ‘‘motivational’’ approach to job design draws primarily from organizational psychology as a discipline base.

A related trend following later in time but somewhat comparable in content is the sociotechnical approach (Cherns 1976; Emory and Trist 1960; Rousseau 1977). It focuses not only on the work, but also on the technology itself. Interest is less on the job per se and more on roles and systems. The goal, and key concept, is the joint optimization of both the social and technical systems. Although this approach differs somewhat in that it also gives consideration to the technical system, it is similar in that it draws on the same psychological job characteristics that affect satisfaction and motivation.

2.2.2. Design Recommendations

Table 2 provides a list of statements that describe the recommendations from the motivational approach. It suggests that jobs should allow the worker some autonomy to make decisions about how and when tasks are to be done. The worker should feel that the work is important to the overall mission of the organization or department. This is often done by allowing the worker to perform a larger unit of work or to perform an entire piece of work from beginning to end. Feedback on job performance should be given to the worker from the task itself, as well as from the supervisor and others. The worker should be able to use a variety of skills and have opportunities to learn new skills and personally grow on the job. Aside from these characteristics that make jobs meaningful from a task-oriented perspective, this approach also considers the social or people-interaction aspects of the job. That is, jobs should have opportunities for participation, communication, and recognition. Finally, other human resource systems should contribute to the motivating atmosphere, such as adequate pay, promotion, and job-security systems.

2.2.3. Advantages and Disadvantages

The goal of this approach is to enhance the psychological meaningfulness of the jobs, thus influencing a variety of attitudinal and behavioral outcomes. Table 1 summarizes some of the human resource benefits and costs from previous research. Jobs designed according to the motivational approach have more satisfied, motivated, and involved employees. Furthermore, job performance may be higher and absenteeism lower. Customer service may even be improved, in part because employees may take more pride in the work, and in part because employees can catch their own errors by performing a larger part of the work.

In terms of disadvantages, jobs that are too high on the motivational approach may have longer training times and be more difficult and expensive to staff because of their greater skill and ability requirements. Higher skill and responsibility may in turn require higher compensation. Overly motivating jobs may be so stimulating that workers become predisposed to mental overload, fatigue, errors, and occupational stress.

2.3. Perceptual/Motor Job Design

2.3.1. Historical Background

This approach draws on a scientific discipline that goes by many names, including human factors, human factors engineering, human engineering, man–machine systems engineering, and engineering psychology. As a field, it developed from a number of other disciplines, primarily experimental psychology, but also industrial engineering (Meister 1971; Meister and Rabideau 1965). Within experimental psychology, job-design recommendations draw heavily from knowledge of human skilled performance (Welford 1976) and the analysis of humans as information processors (Fogel 1976). The main concern of this approach is with the efficient and safe utilization of humans in human–machine systems, with emphasis on the selection, design, and arrangement of system components so as to take account of both people's capabilities and limitations (Pearson 1971). It is more concerned with equipment than is the motivational approach and more concerned with people's abilities than is the mechanistic approach.

The perceptual/motor approach received public attention through the Three Mile Island incident, after which it was concluded that the control room-operator job in the nuclear power plant may have created too many demands on the operator in an emergency situation, thus creating a predisposition to errors of judgment (Campion and Thayer 1987). Federal government regulations issued since that time require that nuclear power plants consider the "human factors" in their design (NRC 1981). The primary emphasis suggested by the title of these regulations is on perceptual and motor abilities of people. This approach is the most prolific with respect to recommendations for proper job design, with the availability of many handbooks giving specific advice for all types of equipment, facilities, and layouts (Salvendy 1987; Sanders and McCormick 1987; Van Cott and Kinkade 1972; Woodson 1981).

2.3.2. Design Recommendations

Table 2 provides a list of statements describing some of the most important recommendations of the perceptual/motor approach. They refer either to equipment and environments on the one hand or information processing requirements on the other. Their thrust is to take into consideration the mental capabilities and limitations of people, such that the attention and concentration requirements of the job do not exceed the abilities of the least-capable potential worker. The focus is on the limits of the least-capable worker because this approach is concerned with the effectiveness of the total system, which is no better than its weakest link. Jobs should be designed to limit the amount of information workers must pay attention to, remember, and think about. Lighting levels should be appropriate, displays and controls should be logical and clear, workplaces should be well laid out and safe, and equipment should be easy to use (i.e., user friendly).

2.3.3. Advantages and Disadvantages

The goals of this approach are to enhance the reliability and safety of the system and to gain positive user reactions. Table 1 summarizes some of the human resource advantages and disadvantages found in previous research. Jobs designed according to the perceptual/motor approach have lower likelihoods of errors and accidents. Employees may be less stressed and mentally fatigued because of the reduced mental demands of the job. Like the mechanistic approach, it reduces the mental ability requirements of the job. Thus, it may also enhance some human resource efficiencies, such as reduced training times and staffing difficulties.

On the other hand, costs to the perceptual/motor approach may result if it is excessively applied. In particular, less satisfaction, less motivation, and more boredom may result because the jobs provide inadequate mental stimulation. This problem is exacerbated by the fact that the least-capable potential workerplaces the limits on the mental requirements of the job.

2.4. Biological Job Design

2.4.1. Historical Background

This approach and the perceptual/motor approach share a joint concern for proper person–machine fit. The primary difference is that this approach is more oriented toward biological considerations of job design and stems from such disciplines as work physiology (Astrand and Rodahl 1977), biomechanics (i.e., the study of body movements) (Tichaeur 1978), and anthropometry (i.e., the study of body sizes) (Hertzberg 1972). Like the perceptual/motor approach, the biological approach is concerned with the design of equipment and workplaces as well as the design of tasks (Grandjean 1980).

2.4.2. Design Recommendations

Table 2 provides a list of important recommendations from the biological approach. This approach tries to design jobs to reduce physical demands, and especially to avoid exceeding people's physical capabilities and limitations. Jobs should not have excessive strength and lifting requirements, and again the capabilities of the least-physically able potential worker set the maximum level. Chairs should be designed so that good postural support is provided. Excessive wrist movement should be reduced by redesigning tasks and equipment. Noise, temperature, and atmosphere should be controlled within reasonable limits. Proper work/rest schedules should be provided so that employees can recuperate from the physical demands of the job.

2.4.3. Advantages and Disadvantages

The goals of this approach are to maintain the comfort and physical well being of the employees. Table 1 summarizes some of the human resource advantages and disadvantages observed in the research. Jobs designed according to the biological approach require less physical effort, result in less fatigue, and create fewer injuries and aches and pains than jobs low on this approach. Occupational injuries and illnesses, such as lower back pain and carpal tunnel syndrome, are fewer on well-designed jobs. There may even be lower absenteeism and higher job satisfaction on jobs that are not physically arduous.

A direct cost of this approach may be the expense of changes in equipment or job environments needed to implement the recommendations. At the extreme, there may be other costs. For example, it is possible to design jobs with so few physical demands that the workers become drowsy or lethargic, thus reducing their performance or encouraging them to leave their workplace. Clearly, extremes of physical activity and inactivity should be avoided, and there may even be an optimal level of physical activity for various employee groups (e.g., male, female, young, old).

2.5. Conflicts and Trade-offs among Approaches

Although one should strive to construct jobs that are well designed on all the approaches, it is clear that there are some direct conflicts in design philosophies. As Table 1 illustrates, the benefits of some approaches are the costs of others. No one approach can satisfy all outcomes. As noted above, the greatest potential conflicts are between the motivational approach on the one hand and the mechanistic and perceptual/motor approaches on the other. They produce nearly opposite outcomes. The mechanistic and perceptual/motor approaches recommend designing jobs that are simple, easy to learn, safe, and reliable, with minimal mental demands on workers. The motivational approach encourages more complicated, challenging, and stimulating jobs, with greater mental demands.

Because of these conflicts, trade-offs and compromises may be necessary in many practical situations. The major trade-offs will be in terms of the mental demands of jobs created by the alternative design strategies. Making the job more mentally demanding increases the likelihood of achieving the workers' goals of satisfaction and motivation. On the other hand, making the job less mentally demanding increases the chances of reaching the organization's goals of reduced training and staffing costs and errors. Which trade-offs will be made depends on which types of outcomes one wants to maximize. In most situations, probably a compromise strategy may be optimal.

Trade-offs may not be needed in all situations, however. Jobs can often be improved on one approach while still maintaining their quality on other approaches. For example, in a recent redesign study, the motivational approach was applied to a group of clerical jobs to improve employee satisfaction and customer service (Campion and McClelland 1991). The expected benefits occurred along with some expected costs (e.g., increased training and compensation requirements), but not all potential costs occurred (e.g., efficiency did not decrease).

One strategy for minimizing trade-offs is to avoid design decisions that influence the mental demands of jobs. An example of this strategy is to enhance motivational design by focusing on the social aspects (e.g., social interaction, communication, participation, recognition, feedback, etc.). These design features can be increased without incurring the costs of increased mental demands. Moreover, many of these design features are under the direct control of those who manage the job.

The independence of the biological approach provides another opportunity to improve design without incurring trade-offs with the other approaches. One can reduce physical demands without influencing the mental demands of a job. Of course, the cost of equipment may need to be considered.

Finally, the adverse effects of trade-offs can often be reduced by avoiding designs that are extremely high or low on any of the approaches. Alternatively, one might require minimally acceptable levels on each approach. Knowing all the approaches to job design and their corresponding outcomes will help one make more intelligent job-design decisions and avoid unanticipated consequences.

3. TEAM-DESIGN

3.1. Historical Background

The major approaches to job design, as discussed in Section 2, typically focus on designing jobs for individual workers; however, it is also possible to design jobs around work teams. In designing jobs for teams, one assigns a task or set of tasks to a group of workers rather than an individual and considers the group to be the unit of performance. Objectives and rewards focus on group, not individual, behavior. Depending on the nature of the tasks, a team's workers may be performing the same tasks simultaneously or they may break tasks into subtasks to be performed by individuals within the team. Subtasks could be assigned on the basis of expertise or interest, or team members might rotate from one subtask to another to provide job variety and increase the breadth of skills and flexibility in the workforce.

Some tasks, because of their size or complexity or for other reasons, seem to fit naturally into a team job design, whereas others may seem to be appropriate only at the individual job level. In many cases, though, there may be a considerable degree of choice regarding whether to organize work around teams or individuals. In such situations, the engineer should consider the advantages and disadvantages of the use of teams as the unit for job design with respect to an organization's goals, policies, technologies, and constraints.

Team-based approaches to organizing work have become very popular in the last two decades in the United States (Goodman et al. 1987; Guzzo and Shea 1992; Hollenbeck et al. 1995; Tannenbaum et al. 1996). Theoretical treatments of team effectiveness have predominantly used input–process–output (IPO) models, as popularized by such authors as McGrath (1964), as frameworks to discuss team design and effectiveness (Guzzo and Shea 1992). Many variations on the IPO model have been presented in the literature over the years (e.g., Denison et al. 1996; Gladstein 1984; Sundstrom et al. 1990).

Social psychologists have studied groups and teams for several decades, mostly in laboratory settings. They have identified problems such as social loafing or free riding, groupthink, decision-making biases, and process losses and inhibitions that operate in groups (Diehl and Strobe 1987; Harkins 1987; Janis 1972; McGrath 1984; Paulus 1998; Steiner 1972; Zajonc 1965). Some empirical field studies have found that the use of teams does not necessarily result in positive outcomes (e.g., Katz et al. 1987; Tannenbaum et al. 1996), while others have shown positive effects from the implementation of teams (e.g., Banker et al. 1996; Campion et al. 1993, 1996; Cordery et al. 1991). Given that so many organizations are transitioning to team-based work design, it is imperative that the design and implementation of teams be based on the increasing knowledge from team-design research.

3.2. Design Recommendations

Design recommendations are organized around the IPO model of work team effectiveness shown in Figure 1. The variables in the model are briefly described below. A more detailed explanation of each variable is contained in Figure 1.

3.2.1. Input Factors

Inputs are the design ingredients that predispose team effectiveness. There are at least four basic types of inputs needed to ensure that teams are optimally designed:

1. Design the jobs to be motivating and satisfying. The characteristics of jobs that make them motivating in a team setting are basically the same as those that make them motivating in an individual setting. Some of the key characteristics applicable in teams are listed below and described in Figure 1 in more detail. They can be used to evaluate or design the jobs in your client's organization.
    - **(a)** Allow the team adequate self-management.
    - **(b)** Encourage participation among all members.
    - **(c)** Encourage task variety; all members should perform varied team tasks.
    - **(d)** Ensure that tasks are viewed by members as important.
    - **(e)** Allow the team to perform a whole piece of work.
    - **(f)** Make sure the team has a clear goal or mission.
2. Make the jobs within the team interdependent. Teams are often formed by combining interdependent jobs. In other cases, the jobs can be made to be interdependent to make them appropriate for teams. For example, reorganizing a company around its business processes normally requires making the work interdependent. Listed below (and in Figure 1) are several ways jobs can be interdependent.

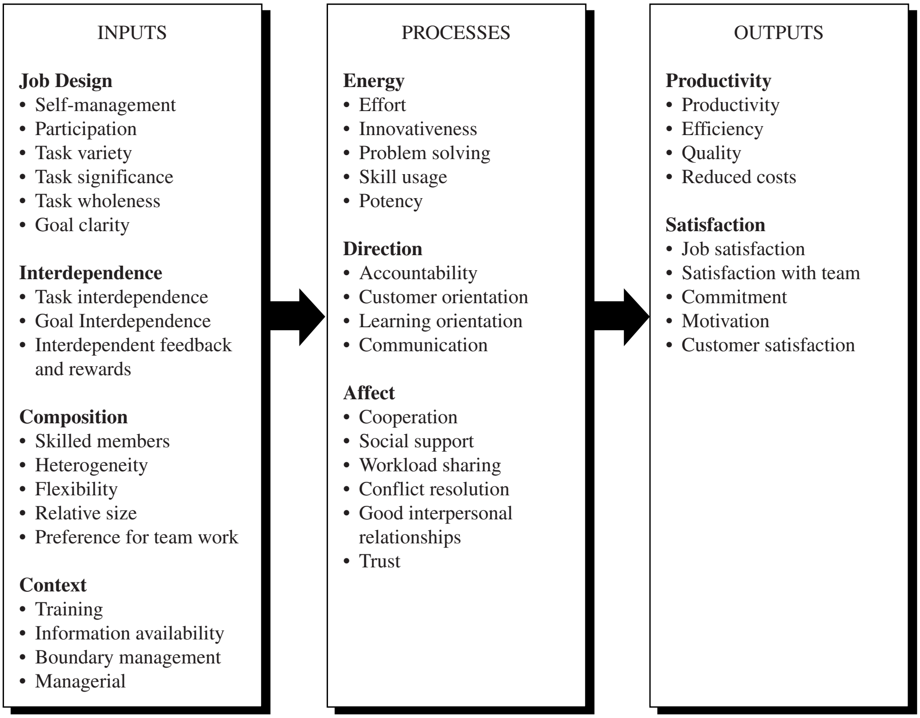

INPUTS
Job Design
• Self-management
• Participation
• Task variety
• Task significance
• Task wholeness
• Goal clarity
Interdependence
• Task interdependence
• Goal Interdependence
• Interdependent feedback and rewards
Composition
• Skilled members
• Heterogeneity
• Flexibility
• Relative size
• Preference for team work
Context
• Training
• Information availability
• Boundary management
• Managerial
PROCESSES
Energy
• Effort
• Innovativeness
• Problem solving
• Skill usage
• Potency
Direction
• Accountability
• Customer orientation
• Learning orientation
• Communication
Affect
• Cooperation
• Social support
• Workload sharing
• Conflict resolution
• Good interpersonal relationships
• Trust
OUTPUTS
Productivity
• Productivity
• Efficiency
• Quality
• Reduced costs
Satisfaction
• Job satisfaction
• Satisfaction with team
• Commitment
• Motivation
• Customer satisfaction

Figure 1 Model of Work Team Effectiveness.

   - **(a)** Make tasks interdependent; jobs must be linked or teams are not needed.
   - **(b)** Set interdependent goals for team members.
   - **(c)** Have interdependent feedback and rewards for team members, with both linked to team performance.
3. Compose the team of the right people. Properly selecting people for a team may be even more important than for individual jobs because a poor selection decision can affect the performance of the entire team. Staffing teams may also be more complex. Some of the key variables to consider are as follows.
   - **(a)** Ensure that team members have a high level of skills.
   - **(b)** Have varied team membership, especially in terms of complementary skills.
   - **(c)** Staff the team with people who are flexible in terms of job assignments.
   - **(d)** Ensure appropriate team size; just large enough (but not too large) for the work involved.
   - **(e)** Select employees with a preference for working in teams.
4. Arrange the context to support teamwork. In order for a new work arrangement like teams to be effective, the organization must be arranged to enable or facilitate them. Following are several key examples.
   - **(a)** Train the work teams in both team skills and task skills.
   - **(b)** Make all necessary information available to the team.
   - **(c)** Ensure proper boundary management by facilitating communication and cooperation between teams.
   - **(d)** Provide managerial support for the team.

3.2.2. Process Factors

Process factors are intermediate indicators of effectiveness. (Note that ‘‘team process,’’ which refers to how the team operates, is different than ‘‘business process,’’ which refers to how the work flows through the organization to get accomplished.) There are at least three categories of team process indicators of effectiveness. Although these are intermediate outcomes, they can also be influenced through the design of the team or through encouragement, as indicated below.

1. Encourage a high-energy environment. This is the first indication that the team may be properly designed. It refers not just to working harder, but to heightened levels of interest and enthusiasm. Even if this is not forthcoming immediately, it can be encouraged in many ways, such as the following (also see Figure 1).
   - **(a)** Create or encourage a high-effort norm within the team.
   - **(b)** Create an environment that encourages innovation within the team.
   - **(c)** Facilitate problem solving within the team.
   - **(d)** Create opportunities for skill usage.
   - **(e)** Encourage team spirit (potency).
2. Ensure that the team is properly directed. The higher energy of the team must be properly directed if it is going to enhance the attainment of organizational goals. There are a number of indicators that a team is properly directed. Key examples follow below. Each can also be encouraged by the management above the team or the consultant.
   - **(a)** Encourage a sense of accountability and responsibility within the team.
   - **(b)** Create or encourage a customer orientation in the team.
   - **(c)** Create a learning orientation in the team.
   - **(d)** Facilitate communication within the team.
3. Encourage proper affective and interpersonal behavior on the team. Not only is this important from a quality of work life point of view, but the long-term viability of the team depends on the members’ willingness to work with each other in the future. Many key indicators of proper affect within the team should be present or encouraged, such as the following:
   - **(a)** Facilitate cooperation within the team.
   - **(b)** Encourage social support among team members.
   - **(c)** Ensure the workload is shared appropriately among team members.
   - **(d)** Facilitate the prompt resolution of conflict within the team.
   - **(e)** Encourage good interpersonal relationships within the team.
   - **(f)** Encourage trust within the team.

3.2.3. Output Factors

Outputs are the ultimate criteria of team effectiveness. The most important outcome is whether the teams enhance the business process. There are two basic categories of outputs. Both are central to the definition of team effectiveness as well as to effective business processes.

1. Effective teams are productive and efficient, and they may also improve quality and reduce costs, as explained in the first part of this chapter.
2. Effective teams are satisfying. This includes not only job satisfaction, but motivated and committed employees. Satisfaction also applies to the customers of the team's products or services. These outcomes were also explained in more detail in the first part of the chapter.

3.3. Advantages and Disadvantages

Work teams can offer several advantages over the use of individuals working separately. Table 3 lists some of these advantages. To begin with, teams can be designed so that members bring a combination of different knowledge, skills, and abilities (KSAs) to bear on a task. Team members can improve their KSAs by working with those who have different KSAs (McGrath 1984), and cross-training on different tasks can occur as a part of the natural workflow. As workers become capable of performing different subtasks, the workforce becomes more flexible. Members can also provide performance feedback to one another that they can use to adjust and improve their work behavior. Creating teams whose members have different KSAs provides an opportunity for synergistic combinations of ideas and abilities that might not be discovered by individuals working alone. Heterogeneity of abilities and personalities has been found to have a generally positive effect on team performance, especially when task requirements are diverse (Goodman et al. 1986; Shaw 1983).

Other advantages include social facilitation and support. Facilitation refers to the fact that the presence of others can be psychologically stimulating. Research has shown that such stimulation can have a positive effect on performance when the task is well learned (Zajonc 1965) and when other team members are perceived as potentially evaluating the performer (Harkins 1987; Porter et al. 1987). With routine jobs, this arousal effect may counteract boredom and performance decrements (Cartwright 1968). Social support can be particularly important when teams face difficult or unpopular decisions. It can also be important in groups such as military squads and medical teams for helping workers deal with difficult emotional and psychological aspects of tasks they perform.

Another advantage of teams is that they may increase the information exchanged between members. Communication can be increased through proximity and the sharing of tasks (McGrath 1984). Intrateam cooperation may also be improved because of team-level goals, evaluation, and rewards (Deutsch 1949; Leventhal 1976). Team rewards can be helpful in situations where it is difficult or impossible to measure individual performance or where workers mistrust supervisors' assessments of performance (Milkovich and Newman 1996).

Increased cooperation and communication within teams can be particularly useful when workers' jobs are highly coupled. There are at least three basic types of coupling, or sequencing of work: pooled, sequential, and reciprocal. In pooled coupling, members share common resources but are otherwise independent. In sequential coupling, members work in a series. Workers whose tasks come later in the process must depend on the performance of workers whose tasks come earlier. In reciprocal coupling, workers feed their work back and forth among themselves. Members receive both inputs and outputs from other members (Thompson 1967; Mintzberg 1979). Team job design would be especially useful for workflows that have sequential or reciprocal coupling.

Many of the advantages of work teams depend on how teams are designed and supported by their organization. The nature of team tasks and their degree of control can vary. According to much of the theory behind team job design, which is primarily from the motivational approach, decision making and responsibility should be pushed down to the team members (Hackman 1987). By pushing decision making down to the team and requiring consensus, the organization should find greater acceptance, understanding, and ownership of decisions among workers (Porter et al. 1987). The increased autonomy resulting from making work decisions should be both satisfying and motivating for teams (Hackman 1987).

The motivational approach would also suggest that the set of tasks assigned to a team should provide a whole and meaningful piece of work (i.e., have task identity) (Hackman 1987). This allows team members to see how their work contributes to a whole product or process, which might not be possible for individuals working alone. This can give workers a better idea of the significance of their work and create greater identification with a finished product or service. If team workers rotate among a variety of subtasks and cross-train on different operations, workers should also perceive greater variety in the work. Autonomy, identity, significance, variety, and feedback are all characteristics of jobs that have been found to enhance motivation. Finally, teams can be beneficial to the organization if team members develop a feeling of commitment and loyalty to their team (Cartwright 1968).

Thus, designing work around teams can provide several advantages to organizations and their workers. Unfortunately, there are also some disadvantages to using work teams. Whether or not teams are beneficial can depend on several factors particular to the composition, structure, and environment of teams and the nature of their tasks. Table 3 lists some of the possible disadvantages of designing work around teams.

For example, some individuals may dislike teamwork and may not have the necessary interpersonal skills or desire to work in a team. In addition, individuals may experience less autonomy and less personal identification when working on a team task than on an individual task. Designing work around teams does not guarantee individual team members greater variety, significance, and identity. If members within the team do not rotate among tasks or if some team members are assigned exclusively to less desirable tasks, not all members will benefit from team job-design. Members can still have fractionated, demotivating jobs. How one organizes work within the team is important in determining the effects of team job design.

Teamwork can also be incompatible with cultural norms. The United States has a very individualistic culture (Hofstede 1980). In addition, organizational norms, practices, and labor–management relations may be incompatible with team job design, making its use more difficult.

Some of the advantages of team design can create other disadvantages. Although team rewards can spur greater cooperation and reduce competition *within* a team, they may cause greater competition and reduced communication *between* teams. If members identify too strongly with the team, they may fail to recognize when behaviors that benefit the team detract from organizational goals. Competition between teams can be motivating up to a point, after which it can create conflicts that are detrimental to productivity.

Increased communication within teams may not always be task relevant. Teams may spend work time socializing. Team decision making can take longer than individual decision making, and reaching a consensus can be time consuming. The need for coordination within teams takes time and faulty coordination can create problems.

Team processes can also inhibit decision making and creativity. When teams become highly cohesive they may become so alike in their views that they develop ‘‘groupthink’’ (Janis 1972; Paulus 1998). When groupthink occurs, teams tend to underestimate their competition, fail to adequately critique fellow team members' suggestions, fail to survey and appraise alternatives adequately, and fail to work out contingency plans. In addition, team pressures can distort judgments. Decisions may be based more on the persuasive abilities of dominant individuals or the power of majorities than on the quality of information and decisions. Research has found a tendency for group judgments to be more extreme than the average of individual members' predecision judgments (Isenberg 1986; McGrath 1984; Pruitt 1971). This may aid reaching a consensus, but it may be detrimental if judgments are poor.

Although evidence shows that highly cohesive groups are more satisfied with the group, high cohesiveness is not necessarily related to high productivity. Whether cohesiveness is related to performance depends on group norms and goals. If a group's norm is to be productive, cohesiveness

TABLE 3 Advantages and Disadvantages of Work Teams

Advantages	Disadvantages
Group members learn from one another	Lack of compatability of some individuals with group work
Possibility of greater workforce flexibility with cross-training	Additional need to select workers to fit group as well as job
Opportunity for synergistic combinations of ideas and abilities	Possibility some members will experience less-motivating jobs
New approaches to tasks may be discovered	Possible incompatibility with cultural, organizational, or labor–management norms
Social facilitation and arousal	Increased competition and conflict between groups
Social support for difficult tasks	More time consuming due to socializing coordination losses, and need for consensus
Increased communication and information exchange between team members	Inhibition of creativity and decision-making processes; possibility of groupthink
Greater cooperation among team members	Less powerful evaluation and rewards; social loafing or free-riding may occur
Beneficial for interdependent workflows	Less flexibility in cases of replacement, turnover, or transfer
Greater acceptance and understanding of decisions when team makes decisions	
Greater autonomy, variety, identity, significance, and feedback possible for workers	
Commitment to the group may stimulate performance and attendance	

TABLE 4 When to Design Jobs around Work Teams

1. Do the tasks require a variety of knowledge, skills, abilities such that combining individuals with different backgrounds would make a difference in performance?
2. Is cross-training desired? Would breadth of skills and work force flexibility be essential to the organization?
3. Could increased arousal, motivation, and effort to perform make a difference in effectiveness?
4. Can social support help workers deal with job stresses?
5. Could increased communication and information exchange improve performance rather than interfere?
6. Could increased cooperation aid performance?
7. Are individual evaluation and rewards difficult or impossible to make or are they mistrusted by workers?
8. Could common measures of performance be developed and used?
9. Would workers' tasks be highly interdependent?
10. Is it technologically possible to group tasks in a meaningful and efficient way?
11. Would individuals be willing to work in groups?
12. Does the labor force have the interpersonal skills needed to work in groups?
13. Would group members have the capacity and willingness to be trained in interpersonal and technical skills required for group work?
14. Would group work be compatible with cultural norms, organizational policies, and leadership styles?
15. Would labor–management relations be favorable to group job design?
16. Would the amount of time taken to reach decisions, consensus, and coordination not be detrimental to performance?
17. Can turnover be kept to a minimum?
18. Can groups be defined as a meaningful unit of the organization with identifiable inputs, outputs, and buffer areas that give them a separate identity from other groups?
19. Would members share common resources, facilities, or equipment?
20. Would top management support group job design?

Affirmative answers support the use of team work design.

will enhance productivity; however, if the norm is not one of commitment to productivity, cohesiveness can have a negative influence (Zajonc 1965; Stogdill 1972).

The use of teams and team-level rewards can also decrease the motivating power of evaluation and reward systems. If team members are not evaluated for their individual performance, do not believe that their output can be distinguished from the team's, or do not perceive a link between their own performance and their outcomes, free-riding or social loafing (Albanese and Van Fleet 1985; Cartwright 1968; Latane et al. 1979) can occur. In such situations, teams do not perform up to the potential expected from combining individual efforts.

Finally, teams may be less flexible in some respects because they are more difficult to move or transfer as a unit than individuals (Sundstrom et al. 1990). Turnover, replacements, and employee transfers may disrupt teams. And members may not readily accept new members.

Thus, whether work teams are advantageous or not depends to a great extent on the composition, structure, reward systems, environment, and task of the team. Table 4 presents questions that can help determine whether work should be designed around teams rather than individuals. The greater the number of questions answered in the affirmative, the more likely teams are to succeed and be beneficial. If one chooses to design work around teams, suggestions for designing effective work teams and avoiding problems are presented below.

4. IMPLEMENTATION ADVICE FOR JOB AND TEAM DESIGN

4.1. When to Consider Design and Redesign of Work

There are at least eight situations when design or redesign of work should be considered.

1. When starting up or building a new plant or work unit. This is the most obvious application of job design.
2. During innovation or technological change. Continual innovation and technological change are important for survival in most organizations. These changes in procedures and equipment mean there are changes in job design. This is not unique to manufacturing jobs. The introduction of

electronic equipment is changing many office jobs. Proper consideration of job design is needed to ensure that the innovations are successful.

3. When markets, products, or strategies change. Modern consumer and industrial markets change rapidly. To keep up with changing demands, organizations must often change marketing strategies and product line mixes. Such changes can affect many jobs throughout an organization and require redesign. For example, salespersons' territories, product line responsibilities, and compensation packages may need to be modified to reflect changing strategies. Production workers' jobs may also require redesign as styles and quantities of products change.
4. During reorganization. Reorganizations of management hierarchies and organizational units frequently mean changes in job assignments or responsibilities of many employees due to the creation and elimination of jobs. In order to ensure a successful reorganization, principles of proper job design must be considered.
5. During growth or downsizing. As an organization grows, new jobs are formed. These jobs are often designed haphazardly, reflecting a collection of tasks other employees do not have time to do. Likewise, during downsizing, jobs are eliminated and some tasks are included in other jobs. This can lead to unfavorable changes in the designs of the remaining jobs.
6. When jobs are needed for special positions or persons. Even existing organizations create new positions. Also, new persons may be hired to fill positions that have different backgrounds, skills, and capabilities than former employees. Both these situations may create a need to reevaluate job-design. For example, hiring handicapped workers may require that managers redesign jobs for them. Frequently, special jobs are also designed for newcomers to the organization, for special administrative assistants, or for temporary assignments.
7. When the workforce or labor markets change. Changing demographics, education levels, and economic conditions affecting employment levels can cause changes in the quality and size of the organization's labor force and labor markets from which the organization hires new workers. Jobs may need to be redesigned to fit a workforce whose average education level has increased over time, or physically demanding jobs may need to be redesigned to accommodate increasing numbers of older or female workers.
8. When there are performance, safety, or satisfaction problems. It is quite common for the employee to be blamed when there are problems with performance, safety, or satisfaction. In many of these instances, the job design is at least partly to blame. Several examples may illustrate this. Human error was named as the cause of the nearly catastrophic nuclear power plant incident at Three Mile Island noted previously, but the design of the operator's job might have been the actual cause. In a study of a wood products company (Campion and Thayer 1985), one sawmill job with multiple employees involved pulling two-by-fours off a moving belt and placing them in racks. The employees were described as lazy and apathetic. But examination of the job from a motivational design point of view revealed that it lacked variety and any significant skill requirements. It was unimportant, repetitive, and boring. It is no surprise that the employees were not motivated or satisfied. In that same study, a plywood plant required an employee to align strips of wood on a moving belt just before they entered the dryer. One time when dryer utilization was not up to standard, the supervisor concluded that the incumbent was negligent and gave her a written reprimand. But the job was very poorly designed from a biological perspective. The employee had to operate a foot pedal while standing and thus spent all day with most of the body weight on one foot. She also had to bend over constantly while extending her arms to adjust the strips of wood, resulting in biomechanical stresses on the back, arms, and legs. Everyone hated the job, yet the employee was blamed.

As a final example, the authors discovered that a personnel-recruiter job in a large company was in need of improved mechanistic design. The job involved running the engineering co-op program that consisted of hundreds of engineering students coming to the company and returning to school each semester. The recruiter had to match employees' interests with managers' needs, monitor everyone's unique schedule, keep abreast of the requirements of different schools, administer salary plans and travel reimbursement, and coordinate hire and termination dates. The job was completely beyond the capability of any recruiter. The solution involved having a team of industrial engineers study the job and apply the mechanistic approach to simplify tasks and streamline procedures.

It is clear that some types of jobs are naturally predisposed to be well designed on some job-design approaches and poorly designed on others. It may be in these latter regards that the greatest opportunities exist to benefit from job redesign. For example, many factory, service, and otherwise low-skilled jobs lend themselves well to mechanistic design, but the ideas of specialization and simplification of tasks and skill requirements can be applied to other jobs in order to reduce staffing difficulties and training requirements. Jobs can often be too complex or too large for employees, leading to poor performance or excessive overtime. This is common with professional and managerial jobs, as was illustrated in the recruiter example above. Professional jobs are usually only evaluated in terms of the motivational approach to job design, but often they can be greatly improved by mechanistic design principles. Finally, if workload in an area temporarily rises without a corresponding increase in staffing levels, the mechanistic approach may be applied to the jobs to enhance efficiency.

Most managerial, professional, and skilled jobs are fairly motivational by their nature. Factory, service, and low-skilled jobs tend naturally not to be motivational. The latter clearly represent the most obvious examples of needed applications of the motivational approach. But there are many jobs in every occupational group, and aspects of almost every job, where motivational features are low. Application of motivational job-design is often limited only by the creativity of the designer.

Jobs involving the operation of complex machinery (e.g., aircraft, construction, and factory control rooms) are primary applications of the perceptual/motor approach. Likewise, many product-inspection and equipment-monitoring jobs can tax attention and concentration capabilities of workers. But jobs in many other occupations may also impose excessive attention and concentration requirements. For example, some managerial, administrative, professional, and sales jobs can be excessively demanding on the information-processing capabilities of workers, thus causing errors and stress. And nearly all jobs have periods of overload. Perceptual/motor design principles can often be applied to reduce these demands of jobs.

Traditional heavy industries (e.g., coal, steel, oil, construction, and forestry) represent the most obvious applications of the biological approach. Similarly, this approach also applies to many jobs that are common to most industries (e.g., production, maintenance) because there is some physical demands component. Biological design principles can be applied to physically demanding jobs so that women can better perform them (e.g., lighter tools with smaller hand grips). But there may also be applications to less physically demanding jobs. For example, seating, size differences, and posture are important to consider in the design of many office jobs, especially those with computer terminals. This approach can apply to many light-assembly positions that require excessive wrist movements that can eventually lead to the wrist ailment carpal tunnel syndrome. It should be kept in mind, however, that jobs designed with too little physical activity (i.e., movement restricted due to single position or workstation) should be avoided. Likewise, jobs that require excessive travel should be avoided because they can lead to poor eating and sleeping patterns.

4.2. Procedures to Follow

There are at least several general guiding philosophies that are helpful when designing or redesigning jobs:

1. As noted previously, designs are not fixed, unalterable, or dictated by the technology. There is at least some discretion in the design of all jobs and substantial discretion in most jobs.
2. There is no single best design for a given job, there are simply better and worse designs depending on one's job-design perspective.
3. Job design is iterative and evolutionary. It should continue to change and improve over time.
4. When possible, participation of the workers affected generally improves the quality of the resulting design and acceptance of suggested changes.
5. Related to number 4, the process aspects of the project are very important to success. That is, how the project is conducted is important in terms of involvement of all the parties of interest, consideration of alternative motivations, awareness of territorial boundaries, and so on.

As noted previously, surveys of industrial job designers have consistently indicated that the mechanistic approach represents the dominant theme of job design (Davis et al. 1955; Taylor 1979). Other approaches to job design, such as the motivational approach, have not been given as much explicit consideration. This is not surprising because the surveys only included job designers trained in engineering-related disciplines, such as industrial engineers and systems analysts. It is not necessarily certain that other specialists or line managers would adopt the same philosophies. Nevertheless, there is evidence that even fairly naive job designers (i.e., college students taking management classes) also seem to adopt the mechanistic approach in job-design simulations. That is, their strategies for grouping tasks were primarily similarity of functions or activities, and also similarity of skills, education, difficulty, equipment, procedures, or location (Campion and Stevens 1989). Even though the mechanistic approach may be the most natural and intuitive, this research has also revealed that people can be trained to apply all four of the approaches to job design.

4.2.1. Procedures for the Initial Design of Jobs or Teams

In consideration of process aspects of conducting a design project, Davis and Wacker (1982) have suggested a strategy consisting of four steps:

1. Form a steering committee. The steering committee usually consists of a group of high-level executives that have a direct stake in the new jobs. The purpose of this committee is fourfold: (a) to bring into focus the objective of the project, (b) to provide resources and support for the project, (c) to help gain the cooperation of all the parties affected by the project, and (d) to oversee and guide the project.
2. Form a design task force. The task force may include engineers, managers, job-design experts, architects, specialists, and others with knowledge or responsibility relevant to the project. The purpose of the task force is to gather data, generate and evaluate design alternatives, and help implement recommended designs.
3. Develop a philosophy statement. The first goal of the task force is to develop a philosophy statement to guide the many decisions that will be involved in the project. The philosophy statement is developed with considerable input from the steering committee and may include such factors as the purposes of the project, the strategic goals of the organization, assumptions about workers and the nature of work, process considerations, and so on.
4. Proceed in an evolutionary manner. The essential point here is that jobs should not be overspecified. With considerable input from eventual jobholders, the designs of the jobs will continue to change and improve over time.

4.2.2. Procedures for Redesigning Existing Jobs or Teams

According to Davis and Wacker (1982), the process of redesigning existing jobs is much the same as that of designing original jobs with two additions. First, the existing job incumbents must be involved. Second, more attention needs to be given to implementation issues. Most importantly, those involved in the implementation must feel ownership of the change. They should believe that the redesign represents their own interests and efforts. This is important not only so that they will be emotionally committed to the change and willing to put in the effort to make it happen, but also so that they will understand the details of the redesign so as to reduce inherent uncertainty.

Along with steps related to the process issues discussed above and in the previous subsection, a redesign project would also include the following five steps:

1. Measure the design of the existing job or team. A questionnaire methodology may be used as well as other analysis tools such as job analysis, time and motion study, and variance analysis. The goal is to gain a measure of the job as it currently exists.
2. Diagnose potential job- and team-design problems. Based partly on the measures collected in step 1, the job is analyzed for potential problems. The job/team-design task force and employee involvement are particularly important at this step. Focused group meetings are often a useful vehicle for identifying and evaluating potential problems.
3. Determine job- and team-design changes. These changes will be guided by the goals of the project, the problems identified in step 3, and one or more of the theoretical approaches to job and team design. Often several potential changes are generated and evaluated. Evaluation of alternative changes may consist of a consideration of the costs and benefits identified in previous research (see Table 1) and the opinions of engineers, managers, and employees. This may be the point when trade-offs become the most apparent.
4. Make the job- and team-design changes. Implementation plans should be developed in detail along with back-up plans in case there are a few difficulties with the new design. Communication and training are keys to successful implementation. Consideration might also be given to pilot testing the changes before widespread implementation is undertaken.
5. Conduct a follow-up evaluation of the new design. Evaluating the new design after implementation is probably the most neglected component of the process in most applications. Part of the evaluation might include the collection of job and team-design measurements on the redesigned job or team using the same instruments as in step 1. Evaluation may also be conducted on the outcomes from the redesign, such as employee satisfaction, error rates, and training times (e.g., Table 1). And it should be noted that some of the effects of job and team design are not always easy to demonstrate. Scientifically valid evaluations require experimental research strategies with control groups. Such studies may not always be possible in ongoing organizations, but often quasiexperimental and other field research designs are possible (Cook and Campbell 1979). Finally, the need for iterations and fine adjustments is identified through the follow-up evaluation.

4.3. Methods for Combining Tasks

In many cases, designing jobs or teams is largely a function of combining tasks. Generally speaking, most writing on job design has focused on espousing overall design philosophies or on identifying those dimensions of jobs (once the jobs exist) that relate to important outcomes, but little research has focused on how tasks should be combined to form jobs in the first place. Some guidance can be gained by extrapolating from the specific design recommendations in Table 2. For example, variety in the motivational approach can be increased by simply combining different tasks into the same job.

Conversely, specialization from the mechanistic approach can be increased by including only very similar tasks in the same job. It is also possible when designing jobs to first generate alternative combinations of tasks, then evaluate them using the design approaches in Table 2.

A small amount of research within the motivational approach has focused explicitly on predicting the relationships between combinations of tasks and the design of resulting jobs (Wong 1989; Wong and Campion 1991). This research suggests that the motivational quality of a job is a function of three task-level variables.

1. Task design. The higher the motivational quality of the individual tasks, the higher the motivational quality of the job. Table 2 can be used to evaluate the individual tasks, then motivational scores for the individual tasks can be summed together. Summing is recommended rather than averaging because it includes a consideration of the number of tasks (Globerson and Crossman 1976). That is, both the motivational quality of the tasks and the number of tasks are important in determining the motivational quality of a job.
2. Task interdependence. Interdependence among the tasks has been shown to have an inverted-U relationship with the motivational quality of a job. That is, task interdependence is positively related to motivational value up to some moderate point; beyond that point, increasing interdependence leads to lower motivational value. Thus, when tasks are being combined to form motivational jobs, the total amount of interdependence among the tasks should be kept at a moderate level. Both complete independence among the tasks and excessively high interdependence should be avoided. Table 5 contains the dimensions of task interdependence and provides a questionnaire that can be used to measure interdependence. Table 5 can be used to judge the interdependence of each pair of tasks being evaluated for inclusion into a particular job.
3. Task similarity. Some degree of similarity among tasks may be the oldest rule of job-design (as discussed previously) and seems to have little influence on the motivational quality of the job. But beyond a moderate level, it tends to decrease the motivational value. Thus, when motivational jobs are being designed, high levels of similarity should be avoided. Similarity at the task pair level can be judged in much the same manner as interdependence by using a subset of the dimensions in Table 5 (see the note).

Davis and Wacker (1982 1987) have provided a list of criteria for grouping tasks into jobs. Part of their list is reproduced below. There are two points to notice. First, the list represents a collection of criteria from both the motivational approach to job-design (e.g., 1, 5, 9) as well as the mechanistic approach (e.g., 2, 8). Second, many of the recommendations could be applied to designing work for teams, as well as individual jobs.

1. Each set of tasks is a meaningful unit of the organization.
2. Task sets are separated by stable buffer areas.
3. Each task set has definite, identifiable inputs and outputs.
4. Each task set has associated with it definite criteria for performance evaluation.
5. Timely feedback about output states and feedforward about input states are available.
6. Each task set has resources to measure and control variances that occur within its area of responsibility.
7. Tasks are grouped around mutual cause–effect relationships.
8. Tasks are grouped around common skills, knowledge, or data bases.
9. Task groups incorporate opportunities for skill acquisition relevant to career advancement.

4.4. Individual Differences Among Workers

A common observation made by engineers and managers is that not all employees respond the same way to the same job. Some people on a given job have high satisfaction, while others on the very same job have low satisfaction. Some people seem to like all jobs, while others dislike every job. Clearly, there are individual differences in how people respond to their work.

There has been a considerable amount of research looking at individual differences in reaction to the motivational approach to job design. It has been found that some people respond more positively (e.g., are more satisfied) than others to highly motivational work. These differences were initially considered to be reflections of underlying work ethic (Hulin and Blood 1968), but later were viewed more generally as differences in needs for personal growth and development (Hackman and Oldham 1980).

TABLE 5 Dimensions of Task Interdependence

Instructions: Indicate the extent to which each statement is descriptive of the pair of tasks using the scale below. Circle answers to the right of each statement. Scores are calculated by averaging applicable items.

Please use the following scale:
(5) Strongly agree
(4) Agree
(3) Neither agree nor disagree
(2) Disagree
(1) Strongly disagree
( ) Leave blank if do not know or not applicable

Inputs of the Tasks					
1. *Materials/supplies:* One task obtains, stores, or prepares the materials or supplies to perform the other task.	1	2	3	4	5
2. *Information:* One task obtains or generates information for the other task.	1	2	3	4	5
3. *Product/service:* One task stores, implements, or handles the products or services produced by the other task.	1	2	3	4	5
Processes of the Task					
4. *Input–output* relationship: The products (or outputs) of one task are the supplies (or inputs) necessary to perform the other task.	1	2	3	4	5
5. *Method and procedure:* One task plans the procedures or work methods for the other task.	1	2	3	4	5
6. *Scheduling:* One task schedules the activities of the other task.	1	2	3	4	5
7. *Supervision:* One task reviews or checks the quality of products or services produced by the other task.	1	2	3	4	5
8. *Sequencing:* One task needs to be performed before the other task.	1	2	3	4	5
9. *Time sharing:* Some of the work activities of the two tasks must be performed at the same time.	1	2	3	4	5
10. *Support service:* The purpose of one task is to support or otherwise help the other task be performed.	1	2	3	4	5
11. *Tools/equipment:* One task produces or maintains the tools or equipment used by the other task.	1	2	3	4	5
Outputs of the Tasks					
12. *Goal:* One task can only be accomplished when the other task is properly performed.	1	2	3	4	5
13. *Performance:* How well one task is performed has a great impact on how well the other task can be performed.	1	2	3	4	5
14. *Quality:* The quality of the product or service produced by one task depends on how well the other task is performed.					

Adapted from Wong and Campion (1991). The task similarity measure contains 10 comparable items (excluding items 4, 6, 8, 9, and 14 and including an item on customer/client). Copyright © 1991 by the American Psychological Association. Adapted with permission.

Using the broader notion of preferences/tolerances for types of work, the consideration of individual differences has been expanded to all four approaches to job design (Campion 1988; Campion and McClelland 1991). Table 6 provides a set of rating scales that can be used with job incumbents to determine their preferences/tolerances. These scales can be administered in the same manner as the questionnaire measures of job design discussed previously.

Although a consideration of employee differences is strongly encouraged, in many situations there are limits to which such differences can be accommodated. As examples, many jobs have to be designed for groups of people that may differ in their preferences/tolerances, often jobs need to be designed without knowledge of the future workers, and the workers on a job may change over time. Fortunately, even though the cumulative evidence is that individual differences moderate reactions to the motivational approach (Loher et al. 1985) the differences are of degree, not direction. In other words, some people respond more positively than others to motivational work, but very few respond negatively. It is likely that this also applies to the other approaches to job design.

TABLE 6 Preferences/Tolerances for Types of Work

Instructions: Indicate the extent to which each statement is descriptive of the job incumbent's preferences and tolerances for types of work on the scale below. Circle answers to the right of each statement. Scores are calculated by averaging applicable items.

Please use the following scale:
(5) Strongly agree
(4) Agree
(3) Neither agree nor disagree
(2) Disagree
(1) Strongly disagree
( ) Leave blank if do not know or not applicable

Preferences/Tolerances for Mechanistic Design					
I have a high tolerance for routine work.	1	2	3	4	5
I prefer to work on one task at a time.	1	2	3	4	5
I have a high tolerance for repetitive work.	1	2	3	4	5
I prefer work that is easy to learn.	1	2	3	4	5
Preferences/Tolerances for Motivational Design					
I prefer highly challenging work that taxes my skills and abilities.	1	2	3	4	5
I have a high tolerance for mentally demanding work.	1	2	3	4	5
I prefer work that gives a great amount of feedback as to how I am doing.	1	2	3	4	5
I prefer work that regularly requires the learning of new skills.	1	2	3	4	5
I prefer work that requires me to develop my own methods, procedures, goals, and schedules.	1	2	3	4	5
I prefer work that has a great amount of variety in duties and responsibilities	1	2	3	4	5
Preferences/Tolerances for Perceptual/Motor Design					
I prefer work that is very fast paced and stimulating.	1	2	3	4	5
I have a high tolerance for stressful work.	1	2	3	4	5
I have a high tolerance for complicated work.	1	2	3	4	5
I have a high tolerance for work where there are frequently too many things to do at one time.	1	2	3	4	5
Preferences/Tolerances for Biological Design					
I have a high tolerance for physically demanding work.	1	2	3	4	5
I have a fairly high tolerance for hot, noisy, or dirty work.	1	2	3	4	5
I prefer work that gives me some physical exercise.	1	2	3	4	5
I prefer work that gives me some opportunities to use my muscles.	1	2	3	4	5

Adapted from Campion (1988). See reference for reliability and validity information. Interpretations differ slightly across the scales. For the mechanistic and motivational designs, higher scores suggest more favorable reactions from incumbents to well designed jobs. For the perceptual/motor and biological approaches, higher scores suggest less unfavorable reactions from incumbents to poorly designed jobs. Copyright © 1988 by the American Psychological Association. Adapted with permission.

4.5. Some Basic Decisions

Hackman and Oldham (1980) have provided five strategic choices that relate to implementing job redesign. They note that little research exists indicating the exact consequences of each choice and that correct choices may differ by organization. The basic decisions are given below:

1. Individual vs. group designs for work. A key initial decision is to either enrich individual jobs or create self-managing work teams. This also includes consideration of whether any redesign should be undertaken and its likelihood of success.
2. Theory-based vs. intuitive changes. This choice was basically defined as the motivational (theory) approach vs. no particular (atheoretical) approach. In the present chapter, this choice may be better framed as choosing among the four approaches to job design. However, as argued earlier, consideration of only one approach may lead to some costs or additional benefits being ignored.
3. Tailored vs. broadside installation. The choice here is between tailoring the changes to the individual employee or making the changes for all employees in a given job.
4. Participative vs. top-down change processes. The most common orientation, and that of this chapter, is that participative is best. However, there are costs to participation, including the

time commitment involved and the fact that incumbents may lack needed broader knowledge of the business.

5. Consultation vs. collaboration with stakeholders. The effects of job-design changes often extend far beyond the individual incumbent and department. For example, the output from the job may be an input to another job elsewhere in the organization, and the presence of a union always constitutes another interested party. Depending on many considerations, participation of stakeholders may range from no involvement to consultation to full collaboration.

4.6. Overcoming Resistance to Change

Resistance to change can be a problem in any project involving major change. Failure rates of implementations demonstrate a need to give more attention to the human aspects of change projects. It has been estimated that between 50% and 75% of newly implemented manufacturing technologies in the United States have failed, with a disregard for human and organizational issues considered to be a bigger cause of failure than technical problems (Majchrzak 1988; Turnage 1990). The number one obstacle to implementation was considered to be resistance to change (Hyer 1984).

Guidelines for reducing resistance to change include the following (Gallagher and Knight 1986; Majchrzak 1988; Turnage 1990):

1. Involve workers in planning the change. Workers should be informed in advance of changes and involved in the process of diagnosing problems and developing solutions because resistance is reduced when workers participate and feel the project is their own.
2. Top management should visibly support the change. When workers feel managers are not committed, they are less likely to take a project seriously.
3. Create change that is consistent with workers' needs and existing values. Resistance is less if change is seen to reduce burdens, offer interesting experience, and not threaten workers' autonomy or security. Workers need to see advantages to them of their involvement in the change.
4. Create an environment of open, supportive communication. If participants experience support and trust, there will be less resistance. Misunderstandings and conflicts should be expected as natural to the innovation process. Adequate provision should be made for clarification and communication.
5. Allow for flexibility. Resistance is reduced if a project is open to revision and reconsideration based on experience.

5. MEASUREMENT AND EVALUATION OF JOB AND TEAM DESIGN

5.1. Using Questionnaires to Evaluate Job and Team Design

One easy and versatile way to measure job design is by using questionnaires or checklists. Job design can then be defined from an operational point of view as "a set of dimensions of jobs that can be used to describe all jobs, regardless of job content, which influence a wide range of benefits and costs for both the organization and the employee." This method of measuring job-design is also highlighted because it has been used widely in research on job design, especially on the motivational approach.

Several questionnaires exist for measuring the motivational approach to job design (Hackman and Oldham 1980; Sims et al. 1976). Only one questionnaire has been developed that measures all four approaches to job design. A version of that questionnaire is presented in Table 2. It is called the multimethod job-design questionnaire (MJDQ) because of its interdisciplinary emphasis. It yields an evaluation of the quality of a job's design based on each of the four approaches. Table 2 also includes a rating scale so that it can simply be copied and used without being retyped.

Table 7 presents a scale that can be used to measure team-design characteristics. It can be used to evaluate input and process characteristics of teams. Background information and examples of the use of this measure can be found by Campion et al. (1993 1996).

Questionnaires may be used in several different contexts:

1. When designing new jobs. When a job does not yet exist, the questionnaire is used to evaluate proposed job descriptions, workstations, equipment, and so on. In this role, it often serves as a simple design checklist.
2. When redesigning existing jobs. When a job exists, there is a much greater wealth of information. Questionnaires can be completed by incumbents, managers, and engineers. Questionnaires can be used to measure job design before and after changes are made and to evaluate proposed changes.

TABLE 7 Team-Design Measure

Instructions: This questionnaire contains statements about your team and how your team functions as a group. Please indicate the extent to which each statement describes your team by circling a number to the right of each statement.

Please use the following scale:
(5) Strongly agree
(4) Agree
(3) Neither agree nor disagree
(2) Disagree
(1) Strongly disagree
( ) Leave blank if do not know or not applicable

Self-Management

1. The members of my team are responsible for determining the methods, procedures, and schedules with which the work gets done. 1 2 3 4 5
2. My team, rather than my manager, decides who does what tasks within the team. 1 2 3 4 5
3. Most work-related decisions are made by the members of my team rather than by my manager. 1 2 3 4 5

Participation

4. As a member of a team, I have a real say in how the team carries out its work. 1 2 3 4 5
5. Most members of my team get a chance to participate in decision making. 1 2 3 4 5
6. My team is designed to let everyone participate in decision making. 1 2 3 4 5

Task Variety

7. Most members of my team get a chance to learn the different tasks the team performs. 1 2 3 4 5
8. Most everyone on my team gets a chance to do the more interesting tasks. 1 2 3 4 5
9. Task assignments often change from day to day to meet the workload needs of the team. 1 2 3 4 5

Task Significance (Importance)

10. The work performed by my team is important to the customers in my area. 1 2 3 4 5
11. My team makes an important contribution to serving the company's customers. 1 2 3 4 5
12. My team helps me feel that my work is important to the company, 1 2 3 4 5

Task Identity (Mission)

13. The team concept allows all the work on a given product to be completed by the same set of people. 1 2 3 4 5
14. My team is responsible for all aspects of a product for its area. 1 2 3 4 5
15. My team is responsible for its own unique area or segment of the business. 1 2 3 4 5

Task Interdependence

16. I cannot accomplish my tasks without information or materials from other members of my team. 1 2 3 4 5
17. Other members of my team depend on me for information or materials needed to perform their tasks. 1 2 3 4 5
18. Within my team, jobs performed by team members are related to one another. 1 2 3 4 5
19. My work goals come directly from the goals of my team. 1 2 3 4 5
20. My work activities on any given day are determined by my team's goals for that day. 1 2 3 4 5
21. I do very few activities on my job that are not related to the goals of my team. 1 2 3 4 5

Interdependent Feedback and Rewards

22. Feedback about how well I am doing my job comes primarily from information about how well the entire team is doing. 1 2 3 4 5
23. My performance evaluation is strongly influenced by how well my team performs. 1 2 3 4 5
24. Many rewards from my job (pay, promotion, etc.) are determined in large part by my contributions as a team member. 1 2 3 4 5

TABLE 7 **(*Continued*)**

Heterogeneity (Membership)					
25. The members of my team vary widely in their areas of expertise.	1	2	3	4	5
26. The members of my team have a variety of different backgrounds and experience.	1	2	3	4	5
27. The members of my team have skills and abilities that complement each other.	1	2	3	4	5
Flexibility					
28. Most members of my team know each other's jobs.	1	2	3	4	5
29. It is easy for the members of my team to fill in for one another.	1	2	3	4	5
30. My team is very flexible in terms of membership.	1	2	3	4	5
Relative Size					
31. The number of people in my team is too small for the work to be accomplished. (Reverse score)	1	2	3	4	5
Preference for Team Work					
32. If given the choice, I would prefer to work as part of a team than work alone.	1	2	3	4	5
33. I find that working as a member of a team increased my ability to perform effectively.	1	2	3	4	5
34. I generally prefer to work as part of a team.	1	2	3	4	5
Training					
35. The company provides adequate technical training for my team.	1	2	3	4	5
36. The company provides adequate quality and customer service training for my team.	1	2	3	4	5
37. The company provides adequate team skills training for my team (communication, organization, interpersonal relationships, etc.).	1	2	3	4	5
Managerial Support					
38. Higher management in the company supports the concept of teams.	1	2	3	4	5
39. My manager supports the concept of teams.	1	2	3	4	5
Communication/Cooperation between Work Groups					
40. I frequently talk to other people in the company besides the people on my team.	1	2	3	4	5
41. There is little competition between my team and other teams in the company.	1	2	3	4	5
42. Teams in the company cooperate to get the work done.	1	2	3	4	5
Potency (Team Spirit)					
43. Members of my team have great confidence that the team can perform effectively.	1	2	3	4	5
44. My team can take on nearly any task and complete it.	1	2	3	4	5
45. My team has a lot of team spirit.	1	2	3	4	5
Social Support					
46. Being in my team gives me the opportunity to provide support to other team members.	1	2	3	4	5
47. My team increases my opportunities for positive social interaction.	1	2	3	4	5
48. Members of my team help each other out at work when needed.	1	2	3	4	5
Workload Sharing					
49. Everyone on my team does their fair share of the work.	1	2	3	4	5
50. No one in my team depends on other team members to do their work for them.	1	2	3	4	5
51. Nearly all the members of my team contribute equally to the work.	1	2	3	4	5
Communication/Cooperation with the Work Group					
52. Members of my team are very willing to share information with other team members about our work.	1	2	3	4	5
53. Teams enhance the communications among people working on the same product.	1	2	3	4	5
54. Members of my team cooperate to get the work done.	1	2	3	4	5

Adapted from Campion et al. (1993). Scores for each preference/tolerance are calculated by averaging applicable items. Adapted with permission of Personnel Psychology, Inc.

3. When diagnosing problem jobs. When problems occur, regardless of the apparent source of the problem, the job-design questionnaire can be used as a diagnostic device to determine whether any problems exist with the design of the jobs.

The administration of questionnaires can be conducted in a variety of ways. Employees can complete them individually at their convenience at their workstation or some other designated area, or they can complete them in a group setting. Group settings allow greater standardization of instructions and provide the opportunity to answer questions and clarify ambiguities. Managers and engineers can also complete the questionnaires either individually or in a group session. Engineers and analysts usually find that observation of the job site, examination of the equipment and procedures, and discussions with any incumbents or managers are important methods of gaining information on the job before completing the questionnaires.

Scoring for each job-design approach is usually accomplished by simply averaging the applicable items. Then the scores from different incumbents, managers, or engineers are combined by averaging (Campion 1988; Campion and McClelland 1991). The implicit assumption is that slight differences among respondents are to be expected because of legitimate differences in viewpoint. However, the absolute differences in scores should be examined on an item-by-item basis, and large discrepancies (e.g., more than one point) should be discussed to clarify possible differences in interpretation. It is often useful to discuss each item until a consensus group rating is reached.

The higher the score on a particular job-design scale, the better the quality of the design of the job based on that approach. Likewise, the higher the score on a particular item, the better the design of the job on that dimension. How high a score is needed or necessary cannot be stated in isolation. Some jobs are naturally higher or lower on the various approaches as described previously, and there may be limits to the potential of some jobs. The scores have most value in comparing jobs or alternative job designs rather than evaluating the absolute level of the quality of job design. However, a simple rule of thumb is that if the score for an approach is smaller than three, the job is poorly designed on that approach and should be reconsidered. Even if the average score on an approach is greater than three, examine any individual item scores that are at two or one.

5.2. Choosing Sources of Data

1. Incumbents. Incumbents are probably the best source of information if there is an existing job. In the area of job analysis, incumbents are considered subject matter experts on the content of their jobs. Also, having input into the job design can enhance the likelihood that suggested changes will be accepted. Involvement in such work-related decisions can enhance feelings of participation, thus increasing motivational job design in itself (see item 22 of the motivational scale in Table 2). One should include a large number of incumbents for each job because there can be slight differences in perceptions of the same job due to individual differences. Evidence suggests that one should include all incumbents or at least 10 incumbents for each job (Campion 1988; Campion and McClelland 1991).
2. Managers or supervisors. First-level managers or supervisors may be the next-most knowledgeable persons about an existing job. They may also provide information on jobs under development if they have insight into the jobs through involvement in the development process. Differences in perceptions of the same job among managers should be smaller than among incumbents, but slight differences will exist and multiple managers should be used. Evidence suggests that one should include all managers with knowledge of the job or at least three to five managers for each job (Campion 1988; Campion and McClelland 1991).
3. Engineers or analysts. Engineers, if the job has not been developed yet, may be the only source of information because they are the only ones with insight into what the job will eventually look like. But also for existing jobs, an outside perspective by an engineer, analyst, or consultant may provide a more objective viewpoint. Again, there can be small differences among engineers, so at least two to five should evaluate each job (Campion and Thayer 1985; Campion and McClelland 1991).

5.3. Evaluating Long-Term Effects and Potential Biases

It is important to recognize that some of the effects of job design may not be immediate, others may not be long lasting, and still others may not be obvious. The research has not tended to address these issues directly. In fact, these effects are offered here as potential explanations for some of the inconsistent findings in the literature. The purpose is to simply put the reader on the alert for the possibility of these effects.

Initially when jobs are designed and employees are new, or right after jobs are redesigned, there may be a short-term period of positive attitudes (often called a "honeymoon effect"). As the legendary Hawthorne studies indicated, often changes in jobs or increased attention given to workers tends to create novel stimulation and positive attitudes (Mayo 1933). Such transitory elevations in affect should not be mistaken for long-term improvements in satisfaction, as they may wear off over time. In fact, with time the employees may realize that the job is now more important or bigger and should require higher compensation (Campion and Berger 1990). These are only examples to illustrate how dissipating and lagged effects might occur.

Likely candidates for costs that may lag in time include compensation, as noted. Stress and fatigue may also take a while to build up if a job's mental demands have been increased excessively, and boredom may take a while to set in after a job's mental demands have been overly decreased. In terms of lagged benefits, productivity and quality are likely to improve with practice and learning on the new job. And some benefits, like reduced turnover, simply take a period of time to estimate accurately.

Benefits that may potentially dissipate with time include satisfaction, especially if the elevated satisfaction is a function of novelty rather than basic changes to the motivating value of the job. Short-term increases in productivity due to heightened effort rather than better design may not last over time. Costs that may dissipate include the training requirements and staffing difficulties. Once the jobs are staffed and everyone is trained, these costs disappear until turnover occurs. So these costs will not go away completely, but they may be less after initial start-up. Dissipating heightened satisfaction but long-term increases in productivity were observed in a recent motivational job-redesign study (Griffin 1989).

Another potential effect that may confuse the proper evaluation of the benefits and costs of job-design is spillover. Laboratory research has shown that job satisfaction can bias employees' perceptions of the motivational value of their jobs (O'Reilly et al. 1980). Likewise, the level of morale in the organization can have a spillover effect onto employees' perceptions of job design in applied settings. If morale is particularly high, it may have an elevating effect on how employees view their jobs; conversely, low morale may have a depressing effect on employees' views. The term *morale* refers to the general level of job satisfaction across employees, and it may be a function of many factors, including management, working conditions, and wages. Another factor included that has an especially strong effect on employee reactions to job-design changes is employment security. Obviously, employee enthusiasm for job-design changes will be negative if they view them as potentially decreasing their job security, and every effort should be made to eliminate these fears. The best method of addressing these effects is to be attentive to their potential existence and conduct longitudinal evaluations of job design.

5.4. Example of an Evaluation of a Job Design

One study is briefly described here as an illustration of a job-redesign project (Campion and McClelland 1991). It best illustrates the evaluation component of redesign and the value of considering both potential benefits and costs, rather than the implementation and process components of redesign. The setting was a large financial services company. The unit under study processed the paperwork in support of other units that sold the company's products. Jobs were designed in a mechanistic manner in that separate employees prepared, sorted, coded, computer keyed, and performed other specific functions on the paper flow.

The organization viewed the jobs as perhaps too mechanistically designed. Guided by the motivational approach, the project intended to enlarge jobs by combining existing jobs. In so doing, the organization hoped to attain three objectives. First, larger jobs might enhance motivation and satisfaction of employees. Second, larger jobs might increase incumbent feelings of ownership of the work, thus increasing customer service. Third, management recognized that there might be potential costs of enlarged jobs in terms of lost efficiency, and thus every attempt was made to maintain (i.e., avoid decreased) productivity.

As indicated by the third objective, the study considered the consequences of the redesign in terms of all approaches to job design. It was anticipated that the project would increase motivational consequences, decrease mechanistic and perceptual/motor consequences, and have no effect on biological consequences (Table 1).

The evaluation consisted of collecting detailed data on job design and a broad spectrum of potential benefits and costs of enlarged jobs. The research strategy involved comparing several varieties of enlarged jobs with each other and with unenlarged jobs. Questionnaire data were collected and focus group meetings were conducted with incumbents, managers, and analysts. The study was repeated at five different geographic sites.

Results indicated that enlarged jobs had the benefits of more employee satisfaction, less boredom, better quality, and better customer service; but they also had the costs of slightly higher training, skill, and compensation requirements. Another finding was that all the potential costs of enlarging jobs were not observed, suggesting that redesign can lead to benefits without incurring every cost in a one-to-one fashion. Finally, the study revealed several improvements to the enlarged jobs.

5.5. Example of an Evaluation of a Team Design

This illustration demonstrates the use of multiple sources of data and multiple types of team-effectiveness outcomes. The setting was the same financial services company as in the job-design evaluation above. Questionnaires based on Table 7 were administered to 391 clerical employees and 70 managers on 80 teams (Campion et al. 1993) and to 357 professional workers on 60 teams (Campion et al. 1996) to measure teams' design characteristics. Thus, two sources of data were used, both team members and managers, to measure the team-design characteristics.

In both studies, effectiveness outcomes included the organization's satisfaction survey, which had been administered at a different time than the team-design characteristics questionnaire, and managers' judgments of team effectiveness. In the first study, several months of records of teams' productivity were also used to measure effectiveness. In the second study, employees' judgments of team effectiveness, managers' judgments of team effectiveness measured three months after the first managers' judgements measure, and the average of team members' most recent performance ratings were also used as outcome measures.

Results indicated that all of the team-design characteristics had positive relationships with at least some of the outcomes. Relationships were strongest for process characteristics. Results also indicated that when teams were well designed according to the team-design approach, they were higher on both employee satisfaction and team-effectiveness ratings.

One final cautionary note regarding evaluation. Different sources (e.g., incumbents, managers) provide different perspectives and should always be included. Collecting data from a single source could lead one to draw different conclusions about a project than if one obtains a broader picture of results by using multiple sources of data.

REFERENCES

Albanese, R. and Van Fleet, D. D. (1985), "Rational Behavior in Groups: The Free-Riding Tendency," *Academy of Management Review,* Vol. 10, pp. 244–255.

Argyris, C. (1964), *Integrating the Individual and the Organization,* John Wiley & Sons, New York.

Astrand, P. O., and Rodahl, K. (1977), *Textbook of Work Physiology: Physiological Bases of Exercise,* 2nd ed., McGraw-Hill, New York.

Babbage, C. (1835), *On the Economy of Machinery and Manufacturers,* 4th Ed., in *Design of Jobs,* 2nd Ed.), L. E. Davis and J. C. Taylor, Eds., Goodyear, Santa Monica, CA, pp. 3–5.

Banker, R. D., Field, J. M., Schroeder, R. G., and Sinha, K. K. (1996), "Impact of Work Teams on Manufacturing Performance: A Longitudinal Study," *Academy of Management,* Vol. 39, pp. 867–890.

Barnes, R. M. (1980), *Motion and Time Study: Design and Measurement of Work,* 7th Ed., John Wiley & Sons, New York.

Blauner, R. (1964), *Alienation and Freedom,* University of Chicago Press, Chicago.

Campion, M. A. (1988), "Interdisciplinary Approaches to Job Design: A Constructive Replication with Extensions," *Journal of Applied Psychology,* Vol. 73, pp. 467–481.

Campion, M. A. (1989), "Ability Requirement Implications of Job Design: An Interdisciplinary Perspective," *Personnel Psychology,* Vol. 42, pp. 1–24.

Campion, M. A., and Berger, C. J. (1990), "Conceptual Integration and Empirical Test of Job Design and Compensation Relationships," *Personnel Psychology,* Vol. 43, pp. 525–554.

Campion, M. A., and McClelland, C. L. (1991), "Interdisciplinary Examination of the Costs and Benefits of Enlarged Jobs: A Job Design Quasi-Experiment,*" Journal of Applied Psychology,* Vol. 76, pp. 186–198.

Campion, M. A., and McClelland, C. L. (1993), "Follow-Up and Extension of the Interdisciplinary Costs and Benefits of Enlarged Jobs," *Journal of Applied Psychology,* Vol. 78, pp. 339–351.

Campion, M. A., and Stevens, M. J. (1989), "A Laboratory Investigation of How People Design Jobs: Naive Predispositions and the Influence of Training," in *Academy of Management Best Papers Proceedings* (Washington, DC), Academy of Management, Briarcliff Manor, NY, pp. 261–264.

Campion, M. A., and Thayer, P. W. (1985), "Development and Field Evaluation of an Interdisciplinary Measure of Job Design," *Journal of Applied Psychology,* Vol. 70, pp. 29–43.

Campion, M. A., and Thayer, P. W. (1987), "Job Design: Approaches, Outcomes, and Trade-offs," *Organizational Dynamics,* Vol. 15, No. 3, pp. 66–79.

Campion, M. A., Medsker, G. J., and Higgs, A. C. (1993), "Relations between Work Group Characteristics and Effectiveness: Implications for Designing Effective Work Groups," *Personnel Psychology,* Vol. 46, pp. 823–850.

Campion, M. A., Cheraskin, L., and Stevens, M. J. (1994), "Job Rotation and Career Development: Career-Related Antecedents and Outcomes of Job Rotation," *Academy of Management Journal,* Vol. 37, pp. 1518–1542.

Campion, M. A., Papper, E. M., and Medsker, G. J. (1996), "Relations between Work Team Characteristics and Effectiveness: A Replication and Extension," *Personnel Psychology,* Vol. 49, pp. 429–452.

Caplan, R. D., Cobb, S., French, J. R. P., Van Harrison, R., and Pinneau, S. R. (1975*), Job Demands and Worker Health: Main Effects and Occupational Differences,* HEW Publication No. (NIOSH) 75-160, U.S. Government Printing Office, Washington, DC.

Cartwright, D. (1968), "The Nature of Group Cohesiveness," in *Group Dynamics: Research and Theory,* 3rd Ed., D. Cartwright and A. Zander, Eds., Harper & Row, New York, pp. 91–109.

Cherns, A. (1976), "The Principles of Socio-technical Design," *Human Relations,* Vol. 29, pp. 783–792.

Cook, T. D., and Campbell, D. T. (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings,* Rand-McNally, Chicago.

Cordery, J. L., Mueller, W. S., and Smith, L. M. (1991), "Attitudinal and Behavioral Effects of Autonomous Group Working: A Longitudinal Field Study," *Academy of Management,* Vol. 34, pp. 464–476.

Davis, L. E. (1957), "Toward a Theory of Job Design," *Journal of Industrial Engineering,* Vol. 8, pp. 305–309.

Davis, L. E. (1982), "Organization Design," in *Handbook of Industrial Engineering,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 2.2.1–2.2.29

Davis, L. E., and Taylor, J. C. (1979), *Design of Jobs,* 2nd Ed., Goodyear, Santa Monica, CA.

Davis, L. E., and Valfer, E. S. (1965), "Intervening Responses to Changes in Supervisor Job Design," *Occupational Psychology,* Vol. 39, pp. 171–189.

Davis, L. E., and Wacker, G. L. (1982), "Job Design," in *Handbook of Industrial Engineering,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 2.5.1–2.5.31.

Davis, L. E., and Wacker, G. L. (1987), "Job Design," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 431–452.

Davis, L. E., Canter, R. R., and Hoffman, J. (1955), "Current Job Design Criteria," *Journal of Industrial Engineering,* Vol. 6, No. 2, pp. 5–8, 21–23.

Denison, D. R., Hart, S. L., and Kahn, J. A. (1996), "From Chimneys to Cross-Functional Teams: Developing and Validating a Diagnostic Model," *Academy of Management Journal,* Vol. 39, pp. 1005–1023.

Deutsch, M. (1949), "An Experimental Study of the Effects of Cooperation and Competition upon Group Process," *Human Relations,* Vol. 2, pp. 199–231.

Diehl, M., and Stroebe, W. (1987), "Productivity Loss in Brainstorming Groups: Toward the Solution of a Riddle," *Journal of Personality and Social Psychology,* Vol. 53, pp. 497–509.

Emory, F. E., and Trist, E. L. (1960), "Socio-technical Systems," in *Management Sciences, Models, and Techniques,* C. W. Churchman and M. Verhulst, Eds., Vol. 2, Pergamon Press, London, pp. 83–97.

Fogel, L. J. (1976), *Human Information Processing,* Prentice Hall, Englewood Cliffs, NJ.

Ford, R. N. (1969), *Motivation through the Work Itself,* American Management Association, New York.

Frankenhaeuser, M. (1977), "Job Demands, Health, and Well-Being," *Journal of Psychosomatic Research,* Vol. 21, pp. 313–321.

Gael, S. (1983), *Job Analysis: A Guide to Assessing Work Activities,* Jossey-Bass, San Francisco.

Gallagher, C. C., and Knight, W. A. (1986), *Group Technology Production Methods in Manufacture,* Ellis Horwood, Chichester.

Gilbreth, F. B. (1911), *Motion Study: A Method for Increasing the Efficiency of the Workman,* Van Nostrand, New York.

Gladstein, D. L. (1984), "Groups in Context: A Model of Task Group Effectiveness," *Administrative Science Quarterly,* Vol. 29, pp. 499–517.

Globerson, S., and Crossman, E. R. F. W. (1976), "Nonrepetitive Time: An Objective Index of Job Variety," *Organizational Behavior and Human Performance,* Vol. 17, pp. 231–240.

Goodman, P. S., Ravlin, E. C., and Argote, L. (1986), "Current Thinking about Groups: Setting the Stage for New Ideas," in *Designing Effective Work Groups,* P. S. Goodman & Associates, Eds., Jossey-Bass, San Francisco, pp. 1–33.

Goodman, P. S., Ravlin, E. C., and Schminke, M. (1987), "Understanding Groups in Organizations," in *Research in Organizational Behavior,* B. M. Staw and L. L. Cummings, Eds., Vol. 9, JAI Press, Greenwich, CT, pp. 121–173.

Grandjean, E. (1980), *Fitting the Task to the Man: An Ergonomic Approach,* Taylor & Francis, London.

Griffin, R. W. (1982), *Task Design: An Integrative Approach,* Scott-Foresman, Glenview, IL.

Griffin, R. W. (1989), "Work Redesign Effects on Employee Attitudes and Behavior: A Long-Term Field Experiment," in *Academy of Management Best Papers Proceedings* (Washington, DC), Academy of Management, Briarcliff Manor, NY, pp. 214–219.

Guzzo, R. A., and Shea, G. P. (1992), "Group Performance and Intergroup Relations in Organizations," in *Handbook of Industrial and Organizational Psychology,* 2nd Ed., Vol. 3, M. D. Dunnette and L. M. Hough, Eds., Consulting Psychologists Press, Palo Alto, CA, pp. 269–313.

Hackman, J. R. (1987), "The Design of Work Teams," in *Handbook of Organizational Behavior,* J. Lorsch, Ed., Prentice Hall, Englewood Cliffs, NJ, pp. 315–342.

Hackman, J. R. and Lawler, E. E. (1971), "Employee Reactions to Job Characteristics," *Journal of Applied Psychology,* Vol. 55, pp. 259–286.

Hackman, J. R., and Oldham, G. R. (1980), *Work Redesign,* Addison-Wesley, Reading, MA.

Hammond, R. W. (1971), "The History and Development of Industrial Engineering," in *Industrial Engineering Handbook,* 3rd Ed., H. B. Maynard, Ed., McGraw-Hill, New York.

Harkins, S. G. (1987), "Social Loafing and Social Facilitation," *Journal of Experimental Social Psychology,* Vol. 23, pp. 1–18.

Hertzberg, H. T. H. (1972), "Engineering Anthropology," in *Human Engineering Guide to Equipment Design,* Rev. Ed., H. P. Van Cott and R. G. Kinkade, Eds., U. S. Government Printing Office, Washington, DC, pp. 467–584.

Herzberg, F. (1966), *Work and the Nature of Man,* World, Cleveland.

Hofstede, G. (1980), *Culture's Consequences,* Sage, Beverly Hills, CA.

Hollenbeck, J. R., Ilgen, D. R., Tuttle, D. B., and Sego, D. J. (1995), "Team Performance on Monitoring Tasks: An Examination of Decision Errors in Contexts Requiring Sustained Attention," *Journal of Applied Psychology,* Vol. 80, pp. 685–696.

Hoppock, R. (1935), *Job Satisfaction,* Harper & Row, New York.

Hulin, C. L. and Blood, M. R. (1968), "Job Enlargement, Individual Differences, and Worker Responses," *Psychological Bulletin,* Vol. 69, pp. 41–55.

Hyer, N. L. (1984), "Management's Guide to Group Technology," in *Group Technology at Work,* N. L. Hyer, Ed., Society of Manufacturing Engineers, Dearborn, MI, pp. 21–27.

Isenberg, D. J. (1986), "Group Polarization: A Critical Review and Meta-Analysis," *Journal of Personality and Social Psychology,* Vol. 50, pp. 1141–1151.

Janis, I. L. (1972), *Victims of Groupthink,* Houghton-Mifflin, Boston.

Johansson, G., Aronsson, G., and Lindstrom, B. O. (1978), "Social Psychological and Neuroendocrine Stress Reactions in Highly Mechanized Work," *Ergonomics,* Vol. 21, pp. 583–599.

Karasek, R. A. (1979), "Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign," *Administrative Science Quarterly,* Vol. 24, pp. 285–308.

Katz, H. C., Kochan, T. A., and Keefe, J. H. (1987), "Industrial Relations and Productivity in the U. S. Automobile Industry," *Brookings Papers on Economic Activity,* Vol. 3, pp. 685–727.

Kornhauser, A. (1965), *Mental Health of the Industrial Worker: A Detroit Study,* John Wiley & Sons, New York.

Latane, B., Williams, K., and Harkins, S. (1979), "Many Hands Make Light the Work: The Causes and Consequences of Social Loafing," *Journal of Personality and Social Psychology,* Vol. 37, pp. 822–832.

Leventhal, G. S. (1976), "The Distribution of Rewards and Resources in Groups and Organizations," in *Advances in Experimental Social Psychology,* Vol. 9, L. Berkowitz and E. Walster, Eds., Academic Press, New York, pp. 91–131.

Likert, R. (1961), *New Patterns of Management,* McGraw-Hill, New York.

Loher, B. T., Noe, R. A., Moeller, N. L., and Fitzgerald, M. P. (1985), "A Meta-Analysis of the Relation Between Job Characteristics and Job Satisfaction," *Journal of Applied Psychology,* Vol. 70 pp. 280–289.

Majchrzak, A. (1988), *The Human Side of Factory Automation,* Jossey-Bass, San Francisco.

Mayo, E. (1933), *The Human Problems of an Industrial Civilization,* Macmillan, New York.

McGrath, J. E. (1964), *Social Psychology: A Brief Introduction,* Holt, Rinehart & Winston, New York.

McGrath, J. E. (1984), *Groups: Interaction and Performance,* Prentice Hall, Englewood Cliffs, NJ.

Meister, D. (1971), *Human Factors: Theory and Practice,* John Wiley & Sons, New York.

Meister, D., and Rabideau, G. F. (1965), *Human Factors Evaluation in System Development,* John Wiley & Sons, New York.

Milkovich, G. T., and Newman, J. M. (1996), *Compensation,* 5th ed., Irwin, Chicago.

Mintzberg, H. (1979), *The Structuring of Organizations: A Synthesis of the Research,* Prentice Hall, Englewood Cliffs, NJ.

Mitchell, T. R. (1976), "Applied Principles in Motivation Theory," in *Personal Goals in Work Design,* P. Warr, Ed., John Wiley & Sons, New York, pp. 163–171.

Mundel, M. E. (1985), *Motion and Time Study: Improving Productivity,* 6th ed., Prentice Hall, Englewood Cliffs, NJ.

Niebel, B. W. (1988), *Motion and Time Study,* 8th Ed., Richard D. Irwin, Homewood, IL.

Nuclear Regulatory Commission (NRC) (1981), *Guidelines for Control Room Design Reviews* NUREG-0700, NRC, Washington, DC.

O'Reilly, C., Parlette, G., and Bloom, J. (1980), "Perceptual Measures of Task Characteristics: The Biasing Effects of Differing Frames of Reference and Job Attitudes," *Academy of Management Journal,* Vol. 23, pp. 118–131.

Paulus, P. B. (1998), "Developing Consensus about Groupthink after All These Years," *Organizational Behavior and Human Decision Processes,* Vol. 73, Nos. 2–3, pp. 362–374.

Pearson, R. G. (1971), "Human Factors Engineering," in *Industrial Engineering Handbook,* 3rd Ed., H. B. Maynard, Ed., McGraw-Hill, New York.

Porter, L. W., Lawler, E. E., and Hackman, J. R. (1987), "Ways Groups Influence Individual Work Effectiveness," in *Motivation and Work Behavior,* 4th ed., R. M. Steers and L. W. Porter, Eds., McGraw-Hill, New York, pp. 271–279.

Pruitt, D. G. (1971), "Choice Shifts in Group Discussion: An Introductory Review," *Journal of Personality and Social Psychology,* Vol. 20, pp. 339–360.

Rousseau, D. M. (1977), "Technological Differences in Job Characteristics, Employee Satisfaction, and Motivation: A Synthesis of Job Design Research and Socio-technical Systems Theory," *Organizational Behavior and Human Performance,* Vol. 19, pp. 18–42.

Salvendy, G., Ed. (1987), *Handbook of Human Factors,* John Wiley & Sons, New York, 1987.

Salvendy, G., and Smith, M. J., Eds. (1981), *Machine Pacing and Occupational Stress,* Taylor & Francis, London.

Sanders, M. S., and McCormick, E. J. (1987), *Human Factors in Engineering and Design,* 6th ed., McGraw-Hill, New York.

Shaw, M. E. (1983), "Group Composition," in *Small Groups and Social Interaction,* Vol. 1, H. H. Blumberg, A. P. Hare, V. Kent, and M. Davies, Eds., John Wiley & Sons, New York, pp. 89–96.

Shepard, J. M. (1970), "Functional Specialization and Work Attitudes," *Industrial Relations,* Vol. 23, pp. 185–194.

Sims, H. P., Szilagyi, A. D., and Keller, R. T. (1976), "The Measurement of Job Characteristics," *Academy of Management Journal,* Vol. 19, pp. 195–212.

Smith, A. (1981), *An Inquiry into the Nature and Causes of the Wealth of Nations,* R. H. Campbell and A. S. Skinner, Eds., Liberty Classics, Indianapolis, IN.

Steers, R. M., and Mowday, R. T. (1977), "The Motivational Properties of Tasks," *Academy of Management Review,* Vol. 2, pp. 645–658.

Steiner, I. D. (1972), *Group Processes and Productivity,* Academic Press, New York.

Stogdill, R. M. (1972), "Group Productivity, Drive, and Cohesiveness," *Organizational Behavior and Human Performance,* Vol. 8, pp. 26–43.

Sundstrom, E., De Meuse, K. P., and Futrell, D. (1990), "Work Teams: Applications and Effectiveness," *American Psychologist,* Vol. 45, pp. 120–133.

Tannenbaum, S. I., Salas, E., and Cannon-Bowers, J. A. (1996), "Promoting Team Effectiveness," in *Handbook of Work Group Psychology,* M. A. West, Ed., John Wiley & Sons, New York, pp. 503–529.

Taylor, F. W. (1911), *The Principles of Scientific Management,* W. W. Norton, New York, 1911.

Taylor, J. C. (1979), "Job Design Criteria Twenty Years Later," in *Design of Jobs,* 2nd ed., L. E. Davis and J. C. Taylor, Eds., John Wiley & Sons, New York, pp. 54–63.

Thompson, J. D. (1967), *Organizations in Action,* McGraw-Hill, New York.

Tichauer, E. R. (1978), *The Biomechanical Basis of Ergonomics: Anatomy Applied to the Design of Work Situations,* John Wiley & Sons, New York.

Turnage, J. J. (1990), ‘‘The Challenge of New Workplace Technology for Psychology,’’ *American Psychologist,* Vol. 45, pp. 171–178.

Turner, A. N., and Lawrence, P. R. (1965), *Industrial Jobs and the Worker: An Investigation of Response to Task Attributes,* Harvard Graduate School of Business Administration, Boston.

U. S. Department of Labor (1972), *Handbook for Analyzing Jobs,* U.S. Government Printing Office, Washington, DC.

Van Cott, H. P., and Kinkade, R. G., Eds. (1972), *Human Engineering Guide to Equipment Design,* Rev. Ed., U.S. Government Printing Office, Washington, DC.

Vroom, V. H. (1964), *Work and Motivation,* John Wiley & Sons, New York.

Walker, C. R., and Guest, R. H. (1952), *The Man on the Assembly Line,* Harvard University Press, Cambridge, MA.

Warr, P., and Wall, T. (1975), *Work and Well-Being,* Penguin, Maryland.

Welford, A. T. (1976), *Skilled Performance: Perceptual and Motor Skills,* Scott-Foresman, Glenview, IL.

Wong, C. S. (1989), ‘‘Task Interdependence: The Link Between Task Design and Job-design,’’ Ph.D. dissertation, Purdue University, West Lafayette, IN.

Wong, C. S., and Campion, M. A. (1991), ‘‘Development and Test of a Task Level Model of Job-design,’’ *Journal of Applied Psychology,* Vol. 76, pp. 825–837.

Woodson, W. E. (1981), *Human Factors Design Handbook,* McGraw-Hill, New York.

Zajonc, R. B. (1965), ‘‘Social Facilitation,’’ *Science,* Vol. 149, pp. 269–274.

CHAPTER 34

Job Evaluation in Organizations

John M. Hannon
Jerry M. Newman
State University of New York-Buffalo

George T. Milkovich
Cornell University

James T. Brakefield
Western Illinois University

1. INTRODUCTION 900
- 1.1. The Influence of Society and Its Values on Job Evaluation 901
- 1.2. The Influence of Individuals on Job Evaluation 901

2. TRADITIONAL JOB VALUATION 902
- 2.1. Ranking Method 902
- 2.2. Classification Method 903
- 2.3. Factor Comparison Method 903
  - 2.3.1. Conduct Job Analysis 904
  - 2.3.2. Select Benchmark Jobs 904
  - 2.3.3. Rank Benchmark Jobs on Each Factor 904
  - 2.3.4. Allocate Benchmark Wages Across Factors 904
  - 2.3.5. Compare Factor and Wage-Allocation Ranks 906
  - 2.3.6. Construct Job Comparison Scale 906
  - 2.3.7. Apply the Scale 906
- 2.4. Point Method 907
  - 2.4.1. Conduct Job Analysis 907
  - 2.4.2. Choose Compensable Factors 907
  - 2.4.3. Establish Factor Scales 908
  - 2.4.4. Establish Factor Weights 908
  - 2.4.5. Evaluate Jobs 909
- 2.5. Single Factor Systems 910

3. OTHER METHODS OF VALUING JOBS 910
- 3.1. Market-Based Pay Systems 910
- 3.2. Knowledge-Based Pay Systems 911
- 3.3. Skill-Based Pay Systems 911

4. CREATING THE JOB-EVALUATION SYSTEM 911
- 4.1. Implementing Job Evaluation 911
- 4.2. Specifying the Macro Objectives of Job Evaluation 911
- 4.3. Specifying the Micro Objectives of Job Evaluation 912
- 4.4. Choosing the Job-Evaluation Method 912
- 4.5. Deciding Who Will Participate in the Job-Evaluation Process 912
  - 4.5.1. Compensation/Job-Evaluation Committees 913
  - 4.5.2. Employee–Manager Participation 913
  - 4.5.3. Unions 913

5. MAINTAINING THE JOB-EVALUATION SYSTEM 913
- 5.1. Handling Appeals and Reviews 913
- 5.2. Training Job Evaluators 913
- 5.3. Approving and Communicating the Results of the Job-Evaluation Process 913
- 5.4. Using Information Technology in the Job-Evaluation Process 914

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

5.5. Future Trends in the Job-Evaluation Process 914

6. EVALUATING THE JOB-EVALUATION SYSTEM 914

6.1. Reliability: Consistent Results 914

6.2. Validity: Legitimate Results 915

6.2.1. Hit rates: Agreement with Predetermined Benchmark Structures 915

6.2.2. Convergence of Results 915

6.3. Utility: Cost-Efficient Results 916

6.4. Nondiscriminatory: Legally Defensible Results 916

6.5. Acceptability: Sensible Results 917

7. SUMMARY 917

REFERENCES 917

1. INTRODUCTION

Why is it that Sam Jones, engineer, makes more money than Ann Banks, who is also an engineer in the same company? Is this an example of sex discrimination in wages? What if we were also to report that Ann Banks makes more money in her engineering job than Ted Adams, an entry-level programmer? Would this lessen your suspicions about the wage-setting practices of our fictitious company? If your response is one of uncertainty, then you probably recognize that several factors need to be considered in determining wages for individuals. First, any wages paid to employees should satisfy an internal consistency criterion. Jobs *inside an organization* are compared to a set of standards and each other to determine their relative contributions to the organization's objectives. To satisfy employee expectations about fairness, more valuable jobs should receive higher "scores" in the comparison process. In our example above, internal consistency triggers the question: How does the work of an engineer compare with that of an entry-level computer programmer? The second wage-determination factor is external competitiveness. Wages for jobs inside an organization should be compared against *wages outside the organization paid by competitors.* How much do other employers pay engineers, and how much do we wish to pay our engineers in comparison to what other employers would pay them? Finally, wages are also a function of the distinctive contributions that individual employees make on their jobs. The level of individual contributions depends on an assessment of *performance and/or seniority of people doing the same job or possessing the same job skills.* Before we jump to the conclusion that Sam Jones should not be making more than Ann Banks because they both are engineers, we must first assess whether their individual contributions have been identical. The pay differential may be warranted if Sam consistently performs better than Ann or if he has more seniority.

Of these three factors affecting wages, this chapter concentrates on only one: the process of determining internal consistency. Specifically, we focus on ways that organizations compare jobs in terms of their relative contributions to the goals of the firm. To the extent this process of ensuring internal consistency is successful, several positive outcomes can be expected. Research suggests that internal consistency may improve both employee satisfaction and performance (Lawler 1986). Alternatively, a lack of internal consistency can lead to turnover, grievances and decreased motivation (Livernash 1957). Without a fair structure, employees may resent the employer, resist change, become depressed, and "lack that zest and enthusiasm which makes for high efficiency and personal satisfaction in work" (Jacques 1961).

The first stage in determining the relative worth of jobs is to assess what the content of these jobs is! This process, as described elsewhere in this Handbook, is called job analysis. A job analyst is charged with the responsibility of acquiring valid (relevant) and reliable (consistent) information about the contents and requirements of jobs. The information obtained through job analysis is usually codified and documented in a job description. It provides a foundation for various human resource management functions, such as establishing selection criteria, setting performance standards, and determining compensation. For our purposes here, the most important function of job analysis is to provide input information into determining the relative worth of jobs within an organization. This process of systematically comparing the contents and requirements of jobs to determine their relative worth (rank ordering) within the organization is called job evaluation. One of the outcomes of this evaluation process is usually a hierarchy of jobs arranged from most valuable to least valuable.

The resulting job structure can be used as a guide in setting pay rates. For the rates to be equitable, jobs that are higher in the structure should be paid more than jobs that are lower in the job structure. This is an important point! Even though this chapter focuses primarily on the ways that organizations determine the relative value (i.e., compared to each other) of jobs, at some point a comparison must be made to external market wages. This external comparison may be the source of an important conflict. Occasionally, jobs that are similar in worth to the organization may be dissimilar in price in the labor market! Suppose, for example, that for a particular organization "skill" and "effort" are judged by top management to be equally important in achieving corporate objectives. Some jobs in that organization may require more skill than effort and other jobs may require more effort than skill.

These jobs will nonetheless be valued similarly in the job structure. The market rates for these jobs, however, may be quite different. Other organizations may not value skill and effort equally. Or perhaps market supply is lower and market demand is higher for people capable of performing the skilled jobs, resulting in higher market wages for the ‘‘skill’’ jobs relative to the ‘‘effort’’ jobs. Thus, for the organization to attract the most qualified workers, it may have to offer wages that are higher than it would offer on the basis of internal consistency alone.

The balance between internal consistency and external competitiveness is a key issue in any employer’s compensation strategy. One firm may emphasize an integrated approach to all human resource management, and internal consistency of pay would be part of that strategy. If so, there would be a relatively close correspondence between its job structure and its pay structure. Another firm may emphasize the relationship between its pay level and pay levels in the labor market. In this firm, there may not be as close a correspondence between the company’s job structure, as originally determined through job evaluation, and its pay structure. Indeed, as we shall discover, the firm may not even systematically develop a job structure through job evaluation, choosing rather to adopt the external market’s evaluation of jobs (i.e., adopt wholesale the market rate without considering internal worth).

This tension between value as assessed within an organization and value as assessed by competitors in the external labor market is but one of several conflicts that may arise in deciding on wages for jobs. Indeed, other ‘‘actors’’ have also influenced the wage-determination process.

1.1. The Influence of Society and Its Values on Job Evaluation

In some societies, at different times through history, egalitarian value systems have been adopted by entire countries. An egalitarian philosophy implies a belief that all workers should be treated equally (Matthew 20.1–16). To some extent, this philosophy underlies the job-evaluation process in those remaining countries that can be classified as communist or socialist. Although some differentials do exist across different jobs, the size of these differentials is much smaller than if this societal influence were not present. Given the recent movement toward capitalism around the world, it is evident that an egalitarian policy may not continue to exert a strong influence over the valuation of jobs.

A second example of societal impacts on wage determination is illustrated by the ‘‘just wage’’ doctrine (Cartter 1959). In the 13th century, skilled artisans and craftsmen began to prosper at the expense of nobles and landowners by selling goods and services to the highest bidders. The church and state reacted by proclaiming a schedule of ‘‘just wages’’ that tended to reflect that society’s class structure and that were consistent with the prevailing notion of birthrights. In essence, the policy explicitly denied economic factors as appropriate determinants of pay.

The proliferation of computers and accompanying information explosion in the recent past has forever changed the way work is done. Not surprisingly, countless companies (like Bayer) have been forced to make ‘‘retain, reject, or redesign’’ decisions about their job-evaluation systems. Most have chosen the redesign option in order to keep the values that have made them so successful but incorporate their new perspectives regarding employee autonomy, teamwork, responsibility, and the like (Laabs 1997). Sometimes referred to as competencies or value driver, job characteristics such as leadership required and customer impact are beginning to form the basis for a whole new set of compensable factors (Kanin-Lovers et al. 1995; McLagan 1997).

1.2. The Influence of Individuals on Job Evaluation

Normally great pains are taken to ensure that position evaluation is kept entirely independent from person evaluation (i.e., job evaluation is kept distinct from performance evaluation, which involves the evaluation of individuals as they perform jobs). Seasoned job evaluators counsel novices to determine the worth of a job independent of its incumbent. The focus should always be on the work, not the worker. After all, a job is relatively stable, whereas the person holding that job may change regularly. For the purposes of determining job worth, individuals are viewed as interchangeable. To deal with the distinction between job and person value, organizations traditionally have set upper and lower limits on job worth (called pay grade minimums and pay grade maximums) and allowed salary to fluctuate within that grade as a function of individual performance or worth.

For certain jobs, though, the worth of the job is inextricably linked to the incumbent performing the job (Pierson 1983). This exception is particularly evident for managerial and executive positions. The person’s unique abilities and knowledge may shape the job. For these jobs, the relative importance of the individual occupying the job leads to increased emphasis on personal attributes in job valuation. The top jobs in almost any organization seem to be designed more around the talents and experience of the individuals involved than around any rigidly defined duties and responsibilities. For professional workers, too, the nature of their work and the knowledge they bring to the task may make it difficult to distinguish job worth from individual worth. Thus, for professionals such as scientists or engineers, pay may reflect individual attributes, accomplishments, or credentials (i.e., a B.S. in Chemistry, a Ph.D. in Engineering).

2. TRADITIONAL JOB EVALUATION

The traditional way to value jobs involves a mix of internal organizational factors as well as external market conditions in setting pay rates. Various job-evaluation techniques have evolved different strategies for incorporating both of these essential influences into the wage-setting process.

In spite of the long-standing existence and recent expansion of some alternative individual (such as commissions and bonuses), market-based (free agent auctions), and parsimonious (delayering and broadbanding) compensation schemes, formal job evaluation continues to stand the test of time. Like the employment interview, which has been criticized harshly but still is most useful, job evaluation has been accused of being "a barrier to excellence" and "an institutional myth" (Emerson 1991; Quaid 1993). Nevertheless, it, too, remains as an essential building block for human resource management. In fact, over 70% of the organizations in this country are estimated to use job evaluation (Bureau of National Affairs 1976).

As noted in the following sections, for both the ranking method and the factor comparison method, external and internal factors are incorporated throughout the job-evaluation process. In the classification method and the point method, internal factors and external factors are considered separately at first and are later reconciled with each other. In the point method, for example, point totals denoting relative internal worth can be reconciled with market data through statistical procedures such as regression analysis.

Determining which of the job-evaluation processes (outlined in the pages that follow) provides the best fit for a given organization depends on numerous considerations. One may be more appropriate than the other, but there is no one best scheme (Fowler 1996).

2.1. Ranking Method

Ranking simply involves ordering the job descriptions from highest to lowest based on a predetermined definition of value or contribution. Three ways of ranking are usually considered: simple ranking, alternation ranking, and paired comparison ranking. Simple ranking requires that evaluators order or rank jobs according to their overall value to the organization. Alternation ranking involves ordering the job descriptions alternately at each extreme (e.g., as shown in Figure 1).

Agreement is reached among evaluators on which job is the most valuable, then the least valuable. Job evaluators alternate between the next most valued and next-least valued, and so on, until all the jobs have been ordered. For example, evaluators agreed that the job of master welder was the most valued of the six jobs listed above and receiving clerk was the least valued. Then they selected most and least valued jobs from the four remaining titles on the list. After this, a final determination would be made between the last two jobs.

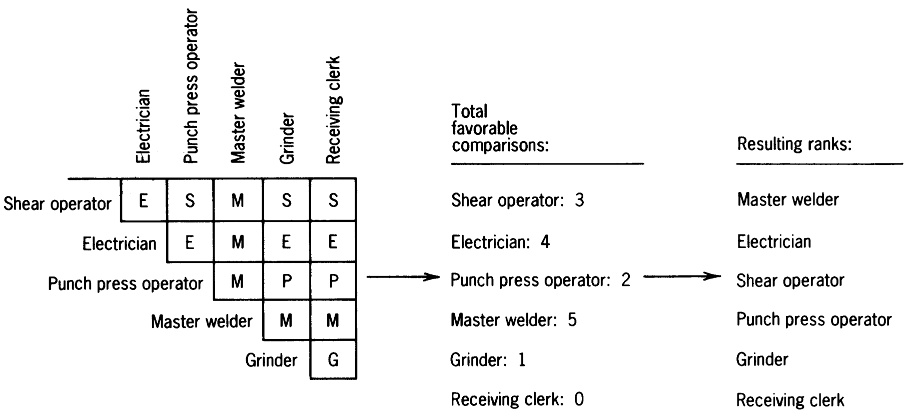
The paired comparison method involves comparing all possible pairs of jobs under study. A simple way to do paired comparison is to set up a matrix, as shown in Figure 2.

The higher-ranked job is entered in the cell. For example, of the shear operator and the electrician, the electrician is ranked higher. Of the shear operator and the punch press operator, the shear operator is ranked higher. When all comparisons have been completed, the job with the highest tally of "most valuable" rankings (the biggest winner) becomes the highest-ranked job, and so on. Some evidence suggests that the alternation ranking and paired comparison methods are more reliable (produce similar results more consistently) than simple ranking (Chesler 1948).

Caution is required if ranking is chosen. The criteria or factors on which the jobs are ranked are usually so poorly defined (if they are specified at all) that the evaluations become subjective opinions

Jobs		*Rank*
Number	***Title***	***Most valued***
1	Shear operator	Master welder
2	Electrician	Electrician
3	Punch press operator	
4	Master welder	
5	Grinder	
6	Receiving clerk	Receiving clerk
		Least valued

Figure 1 Alternation Ranking.



	Electrician	Punch press operator	Master welder	Grinder	Receiving clerk
Shear operator	E	S	M	S	S
Electrician		E	M	E	E
Punch press operator			M	P	P
Master welder				M	M
Grinder					G

Total favorable comparisons:	Resulting ranks:
Shear operator: 3	Master welder
Electrician: 4	Electrician
Punch press operator: 2	Shear operator
Master welder: 5	Punch press operator
Grinder: 1	Grinder
Receiving clerk: 0	Receiving clerk

Figure 2 Paired Comparison Ranking. (From Milkovich and Newman 1993)

that are difficult, if not impossible, to explain and justify in work-related terms. Further, evaluators using this method must be knowledgeable about every single job under study. And as the organization changes, it is difficult to retain command of all this job information. Even if such a person exists, the sheer number of rankings to complete becomes onerous, if not impossible. For example, using the paired comparison process where 50 jobs are involved requires $(n)(n - 1)/2 = 1225$ comparisons. Some organizations try to overcome these difficulties by ranking jobs within single departments and merging the results. However, without greater specification of the factors on which the rankings are based, merging ranks is a major problem.

2.2. Classification Method

The classification method has been in use for over 100 years. It originated as a technique to reform abuses in hiring and paying government workers. Variations of the classification method are still widely used by public-sector employers. The basic procedure is simple: create a set of job categories and sort jobs into them. The categories should be conceived such that jobs that fall into the same category are more similar to each other than to any jobs in other categories. Then, for pay purposes, jobs are treated equally within each category and are treated differently across categories.

Each category is defined by a *class description*. For example, the federal government classification method describes grade 1 as all classes of positions the duties of which are to be performed under immediate supervision, with little or no latitude for the exercise of independent judgment, (1) the simplest routine work in office, business, or fiscal operations, or (2) elementary work of a subordinate technical character in a professional, scientific, or technical field. These class descriptions should be detailed enough to differentiate jobs but general enough to make it fairly easy to slot jobs. While detailed class descriptions make some evaluations more consistent, they can limit the variety of jobs that can readily be classified. It would be difficult, for example, to slot clerical jobs into classes created with sales jobs in mind.

Job classes can be made more concrete by anchoring them with benchmark jobs. For a job to be used as a benchmark, it must be commonly known, relatively stable in content, and perceived to be paid fairly. Where feasible, there should be at least one benchmark job for each job class.

The appropriate number of job classes depends on the diversity of jobs and on promotion paths. A common rule of thumb is 7 to 14 classes (Belcher 1974). Some argue for having many classes, saying that employees favor frequent advancement to higher grades. Today, however, prevailing opinion argues for having fewer classes, saying that it reduces needless bureaucracy.

A problem with the classification method is that it provides incentive for incumbents to "aggrandize" a job title to get it into a higher classification. This may seem appropriate to a manager whose immediate concern is to secure a pay raise for a subordinate; but others may see it as underhanded, and it may even lead to a pay discrimination lawsuit.

2.3. Factor Comparison Method

In the factor comparison method, jobs are evaluated based on two criteria: (1) a set of compensable factors and (2) wages for a select group of benchmark jobs. The two criteria are combined to form a job-comparison scale, which is then applied to nonbenchmark jobs. Unfortunately, the method's complexity often limits its usefulness (Benge et al. 1941). A simplified explanation of this method would include the following steps:

2.3.1. Conduct Job Analysis

As with all job-evaluation methods, information about the jobs must be collected and job descriptions prepared. The Factor Comparison Method differs, however, in that it requires that jobs be analyzed and described in terms of the compensable factors used in the plan. The originators of the method, Benge et al. (1941), prescribed five factors: mental requirements, skill requirements, physical factors, responsibility, and working conditions. They considered these factors to be universal (applicable to all jobs in all organizations) but allowed some latitude in the specific definition of each factor among organizations.

2.3.2. Select Benchmark Jobs

The selection of benchmark jobs is critical because the entire method is based on them. Benchmark jobs (also called key jobs) serve as reference points. The exact number of benchmarks required varies; some rules of thumb have been suggested (15 to 25), but the number depends on the range and diversity of the work to be evaluated.

2.3.3. Rank Benchmark Jobs on Each Factor

Each benchmark job is ranked on each compensable factor. In Table 1, a job family consisting of six jobs is first ranked on mental requirements (rank of 1 is highest), then on experience/skills, and so on.

This approach differs from the straight ranking plan in that each job is ranked on each factor rather than as a whole job.

2.3.4. Allocate Benchmark Wages across Factors

Once each benchmark job is ranked on each factor, the next step is to allocate the current wages paid for each benchmark job among the compensable factors. Essentially, this is done by deciding how much of the wage rate for each benchmark job is associated with mental demands, how much with physical requirements, and so on, across all the compensable factors. This is done for each benchmark job and is usually based on the judgment of a compensation committee. For example, in Table 2, of the $5.80 per hour paid to the punch press operator, the committee had decided that $0.80 of it is attributable to the job's mental requirements, another $0.80 is attributable to the job's experience/skill requirements, $2.40 is attributable to the job's physical requirements, $1.10 is attributable to the job's supervisory requirements, and $0.70 is attributable to the job's other responsibilities. The total $5.80 is thus allocated among the compensable factors. This process is repeated for each of the benchmark jobs.

After the wage for each job is allocated among that job's compensable factors, the dollar amounts for each factor are ranked. The job that has the highest wage allocation for mental requirements is ranked 1 on that factor, next highest is 2, and so on. Separate rankings are done for the wage allocated to each compensable factor. In Table 3, the parts-inspector position has more of its wages allocated to mental demands than does any other job and so it receives the highest rank for that factor.

There are now two sets of rankings. The first ranking is based on comparisons of each benchmark job on each compensable factor. It reflects the relative presence of each factor among the benchmark jobs. The second ranking is based on the proportion of each job's wages that is attributed to each factor. The next step is to see how well the two rankings agree.

TABLE 1 Factor Comparison Method: Ranking Benchmark Jobs by Compensable Factors[a]

Benchmark Jobs	Mental Requirements	Experience/ Skills	Physical Factors	Supervision	Other Responsibilities
A. Punch press operator	6	5	2	4	4
B. Parts attendant	5	3	3	6	1
C. Riveter	4	6	1	1	3
D. Truck operator	3	1	6	5	6
E. Machine operator	2	2	4	2	5
F. Parts inspector	1	3	5	3	2

[a] Rank of 1 is high.

Source: Milkovich and Newman 1993.

TABLE 2 Factor Comparison Method: Allocation of Benchmark Job Wages across Factors

			Factors							
Benchmark Jobs	Current Wage Rate ($/hr)		Mental Requirements $		Experience/ Skills $		Physical Factors $		Supervision $		Other Responsibilities $
A. Punch press operator	5.80	=	0.80	+	0.80	+	2.40	+	1.10	+	0.70
B. Parts attendent	9.60	=	2.15	+	2.35	+	1.90	+	0.60	+	2.60
C. Riveter	13.30	=	2.50	+	3.10	+	2.45	+	4.50	+	0.75
D. Truck operator	8.50	=	3.40	+	3.20	+	0.60	+	0.80	+	0.50
E. Machine operator	11.80	=	3.60	+	2.90	+	1.75	+	2.90	+	0.65
F. Parts inspector	11.40	=	4.50	+	2.20	+	1.20	+	2.50	+	1.10

Source: Milkovich and Newman 1993.

TABLE 3 Ranking Wage Allocations[a]

	Factors									
	Mental Requirements		Experience/Skills		Physical Factors		Supervision		Other Responsibilities	
Benchmark Jobs	$	Rank	$	Rank	$	Rank	$	Rank	$	Rank
A. Punch press operator	0.80	6	0.80	6	2.40	2	1.10	4	0.70	4
B. Parts attendant	2.15	5	2.35	4	1.90	3	0.60	6	2.60	1
C. Riveter	2.50	4	3.10	2	2.45	1	4.50	1	0.75	3
D. Truck operator	3.40	3	3.20	1	0.60	6	0.80	5	0.50	6
E. Machine operator	3.60	2	2.90	3	1.75	4	2.90	2	0.65	5
F. Parts inspector	4.50	1	2.20	5	1.20	5	2.50	3	1.10	2

Source: Milkovich and Newman 1993.

2.3.5. Compare Factor and Wage-Allocation Ranks

The two rankings are judgments based on comparisons of compensable factors and wage distributions. They agree when each benchmark is assigned the same location in both ranks. If there is disagreement, the rationale for the wage allocations and factor rankings is reexamined. Both are judgments, so some slight tuning or adjustments may bring the rankings into line. The comparison of the two rankings is simply a cross-checking of judgments. If agreement cannot be achieved, then the job is no longer considered a benchmark and is removed.

2.3.6. Construct Job Comparison Scale

Constructing a job-comparison scale involves slotting benchmark jobs into a scale for each factor based on the amount of pay assigned to each factor. Such a scale is illustrated in Figure 3. Under mental requirements, the punch press operator is slotted at $0.80, the parts attendant at $2.15, and so on. These slottings correspond to the wage allocations shown in Figure 3.

2.3.7. Apply the Scale

The job-comparison scale is the mechanism used to evaluate the remaining jobs. All the nonbenchmark jobs are now slotted into the scales under each factor at the dollar value thought to be appro-

$ Value	Mental requirements	Experience/ skills	Physical demands	Supervision	Other responsibilities
.00					
.20					Truck operator
.40			Truck operator	Parts attendant	Machine operator
.60					Punch press operator
.80	Punch press operator	Punch press operator		Truck operator	Riveter
1.00			STOCKER		STOCKER
.20			Parts inspector	Punch press operator	Parts inspector
.40	STOCKER			STOCKER	
.60				Parts inspector	
.80			Machine operator Parts attendant		
2.00					
.20	Parts attendant	Parts inspector Parts attendant			
.40			Punch press operator		
.60	Riveter	STOCKER	Riveter		
.80		Machine operator		Machine operator	Parts attendant
3.00		Riveter			
.20		Truck operator			
.40	Truck operator				
.60	Machine operator				
.80					
4.00					
.20					
.40	Parts inspector			Riveter	
.60					
.80					
5.00					

Figure 3 Job Comparison Scale. (From Milkovich and Newman 1993)

priate. This is done by comparing the factors in the job descriptions of nonbenchmark jobs with the factors in the reference points. Consider the position of parts stocker, a nonbenchmark job. The evaluator reads the stocker job description, examines the first compensable factor on the job comparison scale (mental requirements), and locates two benchmark jobs between which the mental requirements of the stocker job rank. After examining the job descriptions for punch press operator and parts attendant the stocker job might be judged to require greater mental demands than those required for the punch press operator but less than those for the parts attendant and might be slotted at a rate of $1.40 for mental requirements. The final worth of each job is derived from a summation of the dollars allocated to the job across all compensable factors.

Historically, only about 10% of employers using formal job evaluations have used the factor comparison approach (Nash and Carroll 1975). The method is complex and difficult to explain, particularly to employees who are dissatisfied with the final ranking their job achieves. In addition, as the agreed-upon wage rates of the benchmark jobs change, the relationships among the jobs may change, and the allocation of the wages among the factors must be readjusted. So continuous updating is required.

In spite of these difficulties, the factor comparison approach represents a significant improvement over simple ranking and classification. First, the criteria for evaluating jobs (i.e., the compensable factors) are agreed upon and made explicit. Second, the use of existing wage rates of benchmark jobs as one of the criteria for designing and explaining the pay structure is unique. In a sense, factor comparison more systematically links external market forces with internal, work-related factors. Finally, in the factor comparison approach, we see the use of a scale of degrees of worth (dollars) for each compensable factor in the job-comparison scale.

These three features—defining compensable factors, scaling the factors, and linking an agreed-upon wage structure with the compensable factors—are also the basic building blocks on which point plans are based.

2.4. Point Method

Like factor comparison, designing a point system is rather complex and often requires outside assistance by consultants. But once designed, the plan is relatively simple to understand and administer, which accounts for its widespread use. Indeed, it is the system used by the vast majority of companies in this country (Milkovich and Newman 1993).

Point methods have three common characteristics: (1) compensable factors, with (2) numerically scaled factor degrees to distinguish different levels within a factor, and (3) weights reflecting the relative importance of each factor.

With the point method, as with all job-evaluation plans, the first step is job analysis. The next steps are to choose the factors, scale them, establish the factor weights, and then evaluate jobs.

2.4.1. Conduct Job Analysis

Information about the jobs to be evaluated is the cornerstone of all job evaluation. While ideally, all jobs will be analyzed, the relevant work content—the behaviors, tasks performed, abilities/skills required, and so on—of a representative sample of jobs forms the basis for deriving compensable factors.

2.4.2. Choose Compensable Factors

Compensable factors play a pivotal role in the point method. In choosing factors, an organization must decide: “What factors are valued in our jobs? What factors will be paid for in the work we do?” Compensable factors should possess the following characteristics:

Work Related They must be demonstrably derived from the actual work performed in the organization. Some form of documentation (i.e., job descriptions, job analysis, employee and/or supervisory interviews) must support the factors. Factors that are embedded in a work-related logic can help withstand a variety of challenges to the pay structure. For example, managers often argue that the salaries of their subordinates are too low in comparison to other employees or that the salary offered to a job candidate is too low for the job. Union members may question their leaders about why one job is paid differently from another. Allegations of illegal pay discrimination may be raised. Line managers, union leaders, and compensation specialists must be able to explain differences in pay among jobs. Differences in factors that are work related help provide that rationale. Properly selected factors may even diminish the likelihood of these challenges arising.

Business Related Compensable factors need to be consistent with the organization’s culture and values, its business directions, and the nature of the work. Changes in the organization or its business strategies may necessitate changing factors. While major changes in organizations are not daily occurrences, when they do occur, the factors need to be reexamined to ensure that they are consistent with the new circumstances.

Acceptable to the Parties Acceptance of the pay structure by managers and employees is critical. This is also true for the compensable factors used to slot jobs into the pay structure. To achieve acceptance of the factors, all the relevant parties' viewpoints need to be considered.

Discriminable In addition to being work related, business related, and acceptable, compensable factors should have the ability to differentiate among jobs. As part of differentiating among jobs, each factor must be unique from other factors. If two factors overlap in what they assess in jobs, then that area of overlap will contribute disproportionately to total job points, which may bias the results. Factor definitions must also possess clarity of terminology so that all concerned can understand and relate to them.

There are two basic ways to select and define factors: Adapt factors from an existing standard plan or custom design a plan. In practice, most applications fall between these two. Standard plans often are adjusted to meet the unique needs of a particular organization, and many custom-designed plans rely heavily on existing factors. Although a wide variety of factors are used in conventional, standard plans, they tend to fall into four generic groups: skills required, effort required, responsibility, and working conditions. These four were used originally in the National Electrical Manufacturers Association (NEMA) plan in the 1930s and are also included in the Equal Pay Act (1963) to define equal work (Gomberg 1947). The Hay System is perhaps the most widely used (Milkovich and Newman, 1993). The three Hay factors are know-how, problem solving, and accountability (note that Hay Associates does not define its guide chart–profile method as a variation of the point method) (Hay Associates 1981). Adapting factors from existing plans usually involves relying on the judgment of a task force or job evaluation committee. More often than not, the committee is made up of key decision makers (or their representatives) from various functions (or units, such as finance, operations, engineering, and marketing). Approaches vary, but typically it begins with a task force or committee representing key management players. To identify compensable factors involves getting answers to one central question: Based on our operating and strategic objectives, what should we value and pay for in our jobs? Obviously, custom designing factors is time consuming and expensive. The argument in favor of it rests on the premise that these factors are more likely to be work related, business related, and acceptable to the employees involved.

In terms of the optimal number of factors, it is generally recommended to stay below 10 in order to avoid dilution of effect, information overload, and factor redundancy. Five to 7 factors are usually a manageable number to capture the essence of jobs in an organization. With regard to the number of total points to be allocated across the factors, most firms choose either 500 or 1000 points.

2.4.3. Establish Factor Scales

Once the factors to be included in the plan are chosen, scales reflecting the different degrees within each factor are constructed. Each degree may also be anchored by the typical skills, tasks, and behaviors taken from benchmark jobs that illustrate each factor degree. Table 4 shows the National Metal Trade Association's scaling for the factor of knowledge.

Belcher (1974) suggests the following criteria for determining degrees:

1. Limit to the number necessary to distinguish among jobs.
2. Use understandable terminology.
3. Anchor degree definition with benchmark job titles.
4. Make it apparent how the degree applies to the job.

Using too many degrees makes it difficult for evaluators to accurately choose the appropriate degree and may result in a wide variance in total points assigned by different evaluators. The threat this poses to acceptance of the system is all too apparent.

Some plans employ 2D grids to define degrees. For example, in the Hay plan, degrees of the factor know-how are described by four levels of managerial know-how (limited, related, diverse, and comprehensive) and eight levels of technical know-how (ranging from professional mastery through elementary vocational). An evaluator may select among at least 32 (4×8) different combinations of managerial and technical know-how to evaluate a job.

2.4.4. Establish Factor Weights

Once the degrees have been assigned, the factor weights must be determined. Factor weights are important because different weights reflect differences in importance attached to each factor by the employer. There are two basic methods used to establish factor weights: committee judgment and statistical analysis. In the first, a standing compensation committee or a team of employees is asked to allocate 100% of value among the factors. Some structured decision process such as Delphi or other nominal group technique may be used to facilitate consensus (Elizur 1980). For the statistical method, which typically utilizes multiple regression analysis, the weights are empirically derived in

TABLE 4 Illustration of a Compensable Factor Scheme

I. Knowledge

This factor measures the knowledge or equivalent training required to perform the position duties.

First Degree

Use of reading and writing, adding and subtracting of whole numbers; following of instructions; use of fixed gauges, direct reading instruments and similar devices; where interpretation is not required.

Second Degree

Use of addition, subtraction, multiplication, and division of numbers including decimals and fractions; simple use of formulas, charts, tables, drawings, specifications, schedules, wiring diagrams; use of adjustable measuring instruments; checking of reports, forms, records and comparable data; where interpretation is required.

Third Degree

Use of mathematics together with the use of complicated drawings, specifications, charts, tables; various types of precision measuring instruments. Equivalent to 1 to 3 years applied trades training in a particular or specialized occupation.

Fourth Degree

Use of advanced trades mathematics, together with the use of complicated drawings, specifications, charts, tables, handbook formulas; all varieties of precision measuring instruments. Equivalent to complete accredited apprenticeship in a recognized trade, craft, or occupation; or equivalent to a 2-year technical college education.

Fifth Degree

Use of higher mathematics involved in the application of engineering principles and the performance of related practical operations, together with a comprehensive knowledge of the theories and practices of mechanical, electrical, chemical, civil or like engineering field. Equivalent to complete 4 years of technical college or university education.

Source: Milkovich and Newman 1993.

such a way as to correlate as closely as possible to a set of pay rates that is agreed upon by the parties involved (Delbecq et al. 1975). The criterion is usually the pay rate for benchmark jobs, and the predictors are the jobs' degree levels on each of the factors.

Initial results of either the committee judgment or statistical approach for deriving factor weights may not lead to completely satisfactory results. The correspondence between internal value (the job-evaluation results) and the external value (what the market says you should be paying) may not be sufficiently high. Several procedures are commonly used to strengthen this relationship. First, the sample of benchmark jobs may be changed through adding or deleting jobs. Second, the factor degree levels assigned to each benchmark job may be adjusted. Third, the pay structure serving as the criterion may be revised. And finally, the factor-weighting scheme may be modified. Thus, a task force beginning with exactly the same factors and degrees could end up with very different job-evaluation plans, depending on the benchmark jobs used, the pay rates chosen as the criterion, and the method employed to establish the weights.

2.4.5. Evaluate Jobs

To translate weights and factor scales into actual job points, the maximum number of points to be used in the system is first divided among the factors according to their weights. The points for each factor are then attached to that factor's scale. For example, if a factor is weighted 20% in a 500-point system, then a total of 100 points is assigned to this factor; and if there are five degrees on the factor, then each degree is worth 20 points.

In the point method, each job's relative value, and hence its location in the pay structure, is determined by the total points assigned to it. A job's total point value is the sum of the numerical values for each degree of compensable factor that the job possesses. In Table 5, the point plan has four factors: skills required, effort required, responsibility, and working conditions. There are five degrees for each factor.

In addition to factor definitions, the evaluator will be guided by benchmark jobs and written descriptions that illustrate each degree for each respective factor. Thus, the evaluator chooses a degree

TABLE 5 The Point Method of Job Evaluation: Factors, Weights, and Degrees

(3) Weights	(1) Factors	(2) Degrees				
40%	Skills required	1	2	3	4	5
30%	Effort required	1	2	3	4	5
20%	Responsibility	1	2	3	4	5
10%	Working conditions	1	2	3	4	5

Source: Milkovich and Newman 1993.

for each factor according to the correspondence between the job being evaluated and the benchmark jobs or descriptions for each factor scale. Then the ratings are multiplied by the factor weights and the products are summed. In the above example, skills required carries a greater weight (40% of the total points) for this employer than does working conditions (10% of the total points). Thus, a job's 240 total points may result from two degrees of skills required ($2 \times 40 = 80$), three each of effort required ($3 \times 30 = 90$) and responsibility ($3 \times 20 = 60$), and one of working conditions ($1 \times 10 = 10$); ($80 + 90 + 60 + 10 = 240$).

Once the total points for all jobs are computed and a hierarchy based on points established, then jobs are compared to each other to ensure that their relative locations in the hierarchy are acceptable. Almost without fail, certain naturally occurring clusters of jobs will emerge.

2.5. Single-Factor Systems

The premise underlying single-factor approaches is that the job content or value construct is unidimensional. In other words, proponents argue that internal value of jobs can be determined by evaluating them against each other on a single factor, instead of the more traditional 5- to 10-factor systems. The two most widely known single-factor plans are Jaques's time span of discretion (TSD) and Arthur Young's decision banding (Jaques 1970). In time span of discretion, each job is made up of tasks and each task is judged to have an implicit or explicit time before its consequences become evident. Jaques defines TSD as "the longest period of time in completing an assigned task that employees are expected to exercise discretion with regard to the pace and quality of the work without managerial review" (Jaques 1964). According to Jaques, TSD is distinct from job evaluation in that it represents measurement (of time units) rather than subjective judgement.

The single factor used in the decision banding method is the decision making required on the job (Patterson and Husband 1970). It identifies and describes six types of decisions that may be required on the job. In order from simplest to most complex, they are: defined, operational, process, interpretive, programming, and policy making. Under this approach, results of job analysis are examined to determine the highest level of decision-making required of the job. Each job is then placed in the corresponding decision band.

Over 50 years ago, Lawshe and others demonstrated that a few factors will yield practically the same results as many factors (Lawshe 1947). Some factors may have overlapping definitions and may fail to account for anything unique in the criterion chosen. In multifactor plans, 3 to 5 factors explained most of the variation in the job hierarchy. In a study conducted 30 years ago, a 21-factor plan produced the same job structure that could be generated using only 7 of the factors. Further, the jobs could be correctly slotted into classes using only 3 factors. Yet the company decided to keep the 21-factor plan because it was "accepted and doing the job."

3. OTHER METHODS OF VALUING JOBS

3.1. Market-Based Pay Systems

For every organization, prevailing wages in the labor market will affect compensation. For some jobs and some organizations, market wage levels and ability to pay are virtually the only determinants of compensation levels. An organization in a highly competitive industry may, by necessity, merely price jobs according to what the market dictates. For most companies, however, to take all their jobs (which may number in the hundreds or thousands) and compare them to the market is not realistic. One can only imagine the effort required for a company to conduct and/or participate in wage surveys for thousands of jobs every year. Alternatively, one computer company was able to slot thousands of jobs into 20 pay grades using a version of the point factor method.

Market pricing basically involves setting pay structures almost exclusively through reliance on rates paid in the external market. Employers following such an approach typically match a large percentage of their jobs with market data and collect as much summarized market data as possible. Opting for market pricing usually reflects more of an emphasis on external competitiveness and less of a focus on internal consistency (the relationships among jobs within the firm).

Market pricers often use the ranking method to determine the pay for jobs unique to their firms. Often called rank to market, it involves first determining the competitive rates for positions for which external market data is available and then slotting the remaining (nonbenchmark) jobs into the pay hierarchy. At Pfizer, for example, job analysis results in written job descriptions. This is immediately followed by labor market analysis and market pricing for as many jobs as possible. After that, the internal job relationships are reviewed to be sure they are ''reasonable in light of organization needs.'' The final step is pricing those jobs not included in the survey. These remaining jobs are compared to the survey positions ''in terms of their total value to Pfizer.'' This internal evaluation seeks to ensure consistency with promotion opportunities and to properly reflect ''cross-functional job values'' (e.g., production vs. clerical jobs).

3.2. Knowledge-Based Pay Systems

As we indicated earlier, some organizations consider individual employee characteristics, in accordance with internal organizational factors and external market conditions, in setting pay rates. Increasing foreign and domestic competition and rapid technological change have inspired innovative individual and team pay-for-performance and knowledge- and skill-based pay systems. Such systems are posited to engender (1) greater mutual commitment between individuals and organizations, and (2) stronger linkages between the rewards given to employees and the performance of the organization.

Technically, knowledge-based pay systems do not involve job evaluation. Instead, they are an alternative to systems that do involve job evaluation. Knowledge-based pay systems pay employees based on what they *know* rather than what particular job they are doing (Gupta., et al 1986). Generally, such systems base pay on the depth of knowledge in a particular field (e.g., scientists and teachers) (Luthans and Fox 1989). For instance, all else equal, a sixth-grade teacher with a Master's degree will be paid more than a sixth-grade teacher with a Bachelor's degree under this system.

3.3. Skill-Based Pay Systems

Similarly, skill-based pay systems reward employees for their breadth of knowledge pertaining to different jobs (e.g., proficiency in a number of various production jobs). For instance, if one person could operate machines A, B, and C, she may be paid $15 per hour (even if she only works on machine A all year). Her colleague may be qualified to work on machines A and C, and therefore he would only make $13 per hour (even if he worked on both machines over the course of the year). As can be seen, pay is driven by the quantity of tasks a person is qualified to perform.

The chief advantages of knowledge- and skill-based pay systems are leaner staffs and greater flexibility in scheduling. Advocates claim they benefit employees: job satisfaction increases because employees get a sense of having an impact on the organization, and motivation increases because pay and performance are closely linked. Potential disadvantages include higher pay rates and increased training costs, the administrative burden of maintaining records, the erosion of knowledge/skills if not used, and the challenge of managing an equitable job rotation. These disadvantages may or may not be offset by having a leaner workforce and greater productivity.

4. CREATING THE JOB-EVALUATION SYSTEM

4.1. Implementing Job Evaluation

The major decisions involved in the design and administration of job evaluation include:

1. What are the objectives of job evaluation?
2. Which job-evaluation method should be used?
3. Should a single plan or multiple plans be used?
4. Who should participate in designing the system?

4.2. Specifying the Macro Objectives of Job Evaluation

From a macro standpoint, job evaluation allows an organization to establish a pay structure that is internally equitable to employees and consistent with the goals of the organization. Once an organization decides on its strategic goals and its value system, job evaluation can help to reward jobs in a manner consistent with the strategic mission. For example, organizations in mature industries may decide that continued success depends on greater risk taking amongst employees, particularly in new product development. Compensable factors can be chosen to reinforce risk by valuing more highly those jobs with a strong risk-taking component. Once this emphasis on risk taking is communicated to employees, the first step in reshaping the value system has begun.

Since they guide the design and administration of job evaluation, strategic plans and business objectives need to be clearly and emphatically specified. Unfortunately, these initially established objectives too often get diluted, discarded, or muddled in the midst of all the statistical procedures performed. Another complication can be the bureaucracy that tends to accompany the administration of job evaluation. Job evaluation sometimes seems to exist for its own sake, rather than as an aid to achieving the organization's mission (Burns 1978). So an organization is best served by initially establishing its objectives for the process and using these objectives as a constant guide for its decisions.

4.3. Specifying the Micro Objectives of Job Evaluation

Some of the more micro objectives associated with job evaluation include:

Help foster equity by integrating pay with a job's contributions to the organization.
Assist employees to adapt to organization changes by improving their understanding of job content and what is valued in their work.
Establish a workable, agreed-upon pay structure.
Simplify and rationalize the pay relationships among jobs, and reduce the role that chance, favoritism, and bias may play.
Aid in setting pay for new, unique, or changing jobs.
Provide an agreed-upon device to reduce and resolve disputes and grievances.
Help ensure that the pay structure is consistent with the relationships among jobs, thereby supporting other human resource programs such as career planning, staffing, and training.

4.4. Choosing the Job-Evaluation Method

Obviously, the organization should adopt a job-evaluation method that is consistent with its job-evaluation objectives. More fundamentally, however, the organization should first decide whether job evaluation is necessary at all. In doing so, it should consider the following questions (Hills 1989):

Does management perceive meaningful differences between jobs?
Can legitimate criteria for distinguishing between jobs be articulated and operationalized?
Will job evaluation result in clear distinctions in employees' eyes?
Are jobs stable and will they remain stable in the future?
Is traditional job evaluation consistent with the organization's goals and strategies?
Do the benefits of job evaluation outweigh its costs?
Can job evaluation help the organization be more competitive?

Many employers design different job-evaluation plans for different job families. They do so because they believe that the work content of various job families is too diverse to be adequately evaluated using the same plan. For example, production jobs may vary in terms of working conditions and the physical skills required. But engineering and marketing jobs do not vary on these factors, nor are those factors particularly important in engineering or marketing work. Rather, other factors such as technical knowledge and skills and the degree of contact with external customers may be relevant. Another category of employees that might warrant special consideration, primarily due to supervisory responsibilities, is management.

The most common criteria for determining different job families include similar knowledge/skill/ability requirements, common licensing requirements, union jurisdiction, and career paths. Those who argue for multiple plans, each with unique compensable factors, claim that different job families have different and unique work characteristics. To design a single set of compensable factors capable of universal application, while technically feasible, risks emphasizing generalized commonalities among jobs and minimizing uniqueness and dissimilarities. Accurately gauging the similarities and dissimilarities in jobs is critical to establish and justify pay differentials. Therefore, more than one plan is often used for adequate evaluation.

Rather than using either a set of company-wide universal factors or entirely different sets of factors for each job family, some employers, such as Hewlett-Packard, start with a core set of common factors, then add other sets of specialized factors unique to particular occupational or functional areas (finance, manufacturing, software and systems, sales, management). These companies' experiences suggest that unique factors tailored to different job families are more likely to be both acceptable to employees and managers and easier to verify as work related than are generalized universal factors.

4.5. Deciding Who Will Participate in the Job Evaluation Process

Who should be involved in designing job evaluation? The choice is usually among compensation professionals, managers, and/or job incumbents. If job evaluation is to be an aid to managers and if maximizing employee understanding and acceptance is an important objective, then all these groups need to be included.

4.5.1. Compensation/Job-Evaluation Committees

A common approach to gaining acceptance and understanding of pay decisions is through use of a compensation (job-evaluation) committee. Membership on these committees seems to vary among firms. They all include representatives from key operating functions, and increasingly, with the emphasis on employee empowerment, they are including nonmanagerial employees. In some cases, the committee's role is only advisory; in others its approval may be required for all major decisions.

4.5.2. Employee–Manager Participation

Participation in the design and administration of compensation plans seems related to increased trust and commitment on the part of employees and managers. Lack of participation makes it easier for employees and managers to imagine ways the structure might have been rearranged to their personal liking. For example, an operating manager may wish to elevate the job title for a star performer in order to exceed the maximum pay permitted for the existing job title. This is less likely to occur if the people affected by job-evaluation outcomes are involved in the design and administration processes. Some evidence has been found, however, for the assertion that incumbents tend to rate their jobs higher than their supervisors. Hence, there is a need for a system of checks and balances in the process (Huber 1991).

4.5.3. Unions

Management probably will find it advantageous to include union representation as a source of ideas and to help promote acceptance of the results. For example, at both AT&T and Borg-Warner, union–management task forces have participated in the design of new job-evaluation systems. When this occurs, the union joins with management to identify, negotiate, and resolve problems related to the job-evaluation process. As noted below, not everyone buys into this notion of cooperation.

Other union leaders, however, feel that philosophical differences prevent their active participation in job evaluation (Burns 1978). They take the position that collective bargaining yields more equitable results than does job evaluation. In other cases, jobs are jointly evaluated by union and management representatives, and disagreements are submitted to an arbitrator.

5. MAINTAINING THE JOB-EVALUATION SYSTEM

As an administrative procedure, job evaluation invites give and take. Consensus building often requires active participation by all those involved. Employees, union representatives, and managers may be included in discussions about the pay differences across various jobs. Job evaluation even involves negotiations among executives or managers of different units or functions within the same organization. So viewed as an administrative procedure, job evaluation is used for resolving conflicts about pay differences that inevitably arise over time.

5.1. Handling Appeals and Reviews

Once the job structure or structures are established, compensation managers must ensure that they remain equitable. This requires seeing that jobs that employees feel have been incorrectly evaluated are reanalyzed and reevaluated. Likewise, new jobs or those that experience significant changes must be submitted to the evaluation process.

5.2. Training Job Evaluators

Once the job-evaluation system is complete, those who will be conducting job analyses and evaluations will require training, especially those evaluators who come from outside the human resource department. These employees may also need background information on the entire pay system and how it is related to the overall strategies for managing human resources and the organization's objectives.

5.3. Approving and Communicating the Results of the Job-Evaluation Process

When the job evaluations are completed, approval by higher levels in the organization (e.g., Vice President of Human Resources) is usually required. Essentially, the approval process serves as a control. It helps ensure that any changes that result from job evaluation are consistent with the organization's overall strategies and human resource practices. The particular approval process differs among organizations. Figure 4 is one example.

The emphasis on employee and manager understanding and acceptance of the job-evaluation system requires that communications occur during the entire process. Toward the end of the process,

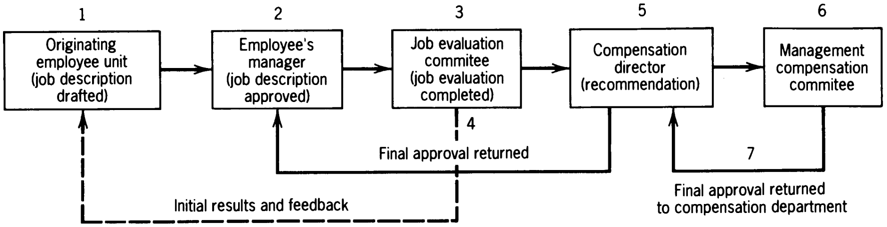

1
Originating employee unit (job description drafted)
2
Employee's manager (job description approved)
3
Job evaluation commitee (job evaluation completed)
4
5
Compensation director (recommendation)
6
Management compensation commitee
7
Final approval returned
Initial results and feedback
Final approval returned to compensation department

Figure 4 Job Evaluation Approval Process. (From Milkovich and Newman 1993)

the goals of the system, the parties' roles in it, and the final results will need to be thoroughly explained to all employees.

5.4. Using Information Technology in the Job-Evaluation Process

Almost every compensation consulting firm offers a computer-based job evaluation system. Their software does everything from analyzing the job-analysis questions to providing computer-generated job descriptions to predicting the pay classes for each job. Some caution is required, however, because "computer assisted" does not always mean a more efficient, more acceptable, or cheaper approach will evolve. The primary advantages for computer-aided job evaluation according to its advocates include:

Alleviation of the heavy paperwork and tremendous time saving
Marked increase in the accuracy of results
Creation of more detailed databases
Opportunity to conduct improved analyses (Rheaume and Jones 1988)

The most advanced use of computers for job evaluation is known as an expert system. Using the logic built by subject matter experts and coded into the computer, this software leads a job evaluator through a series of prompted questions as part of a decision tree to arrive at a job-evaluation decision (Mahmood et al. 1995).

But even with the assistance of computers, job evaluation remains a subjective process that involves substantial judgment. The computer is only a tool, and misused, it can generate impractical, illogical, and absurd results (Korukonda 1996).

5.5. Future Trends in the Job Evaluation Process

Job evaluation is not going to go away. It has emerged and evolved through the industrial, and now the informational, revolution. Unless everyone is paid the same, there will always be a need to establish and institutionalize a hierarchy of jobs in the organization. The process should, and will, continue to be improved upon. The use of computer software will dramatically simplify the administrative burdens of job evaluation. Furthermore, new technologies and processes will enable organizations to combine internal job-evaluation information with labor market data to strengthen the internal consistency–external competitiveness model discussed above.

6. EVALUATING THE JOB-EVALUATION SYSTEM

Job evaluation can take on the appearance of a bona fide measurement instrument (objective, numerical, generalizable, documented, and reliable). If it is viewed as such, then job evaluation can be judged according to precise technical standards. Just as with employment tests, the reliability and validity of job evaluation plans should be ascertained. In addition, the system should be evaluated for its utility, legality, and acceptability.

6.1. Reliability: Consistent Results

Job evaluation involves substantial judgment. Reliability refers to the consistency of results obtained from job evaluation conducted under different conditions. For example, to what extent do different job evaluators produce similar results for the same job? Few employers or consulting firms report the results of their studies. However, several research studies by academics have been reported (Arvey 1986; Schwab 1980; Snelgar 1983; Madigan 1985; Davis and Sauser 1993; Cunningham and Graham 1993; Supel 1990). These studies present a mixed picture; some report relatively high consistency (different evaluators assign the same jobs the same total point scores), while others report lower agreement on the values assigned to each specific compensable factor. Some evidence also reports that evaluators' background and training may affect the consistency of the evaluations. Interestingly, an evaluator's affiliation with union or management appears to have little effect on the consistency of the results (Lawshe and Farbo 1949; Harding et al. 1960; Moore 1946; Dertien 1981).

6.2. Validity: Legitimate Results

Validity is the degree to which a job-evaluation method yields the desired results. The desired results can be measured several ways: (1) the hit rate (percentage of correct decisions it makes), (2) convergence (agreement with results obtained from other job-evaluation plans), and (3) employee acceptance (employee and manager attitudes about the job-evaluation process and the results) (Fox 1962; Collins and Muchinsky 1993).

6.2.1. Hit Rates: Agreement with Predetermined Benchmark Structures

The hit rate approach focuses on the ability of the job-evaluation plan to replicate a predetermined, agreed-upon job structure. The agreed-upon structure, as discussed earlier, can be based on one of several criteria. The jobs' market rates, a structure negotiated with a union or a management committee, and rates for jobs held predominately by men are all examples.

Figure 5 shows the hit rates for a hypothetical job-evaluation plan. The agreed-upon structure has 49 benchmark jobs in it. This structure was derived through negotiation among managers serving on the job-evaluation committee along with market rates for these jobs. The new point factor job-evaluation system placed only 14, or 29%, of the jobs into their current (agreed-upon) pay classes. It came within ±1 pay class for 82% of the jobs in the agreed-upon structure. In a study conducted at Control Data Corporation, the reported hit rates for six different types of systems ranged from 49 to 73% of the jobs classified within ±1 class of their current agreed-upon classes (Gomez-Mejia et al. 1982). In another validation study, Madigan and Hoover applied two job-evaluation plans (a modification of the federal government's factor evaluation system and the position analysis questionnaire) to 206 job classes for the State of Michigan (Madigan and Hoover 1986). They reported hit rates ranging from 27 to 73%, depending on the scoring method used for the job-evaluation plans.

Is a job-evaluation plan valid (i.e., useful) if it can correctly slot only one-third of the jobs in the "right" level? As with so many questions in compensation, the answer is "it depends." It depends on the alternative approaches available, on the costs involved in designing and implementing these plans, on the magnitude of errors involved in designing and implementing these plans, and on the magnitude of errors involved in missing a "direct hit." If, for example, being within ±1 pay class translates into several hundred dollars in pay, then employees probably are not going to express much confidence in the "validity" of this plan. If, on the other hand, the pay difference between ±1 class is not great or the plan's results are treated only as an estimate to be adjusted by the job-evaluation committee, then its "validity" (usefulness) is more likely.

6.2.2. Convergence of Results

Job-evaluation plans can also be judged by the degree to which different plans yield similar results. The premise is that convergence of the results from independent methods increases the chances that the results, and hence the methods, are valid. Different results, on the other hand, point to lack of validity. For the best study to date on this issue, we again turn to Madigan's report on the results of three job-evaluation plans (guide chart, PAQ, and point plan) (Madigan 1985). He concludes that the three methods generate different and inconsistent job structures. Further, he states that the measurement adequacy of these three methods is open to serious question. An employee could have received up to $427 per month more (or less), depending on the job-evaluation method used.

These results are provocative. They are consistent with the proposition that job evaluation, as traditionally practiced and described in the literature, is not a measurement procedure. This is so because it fails to consistently exhibit properties of reliability and validity. However, it is important

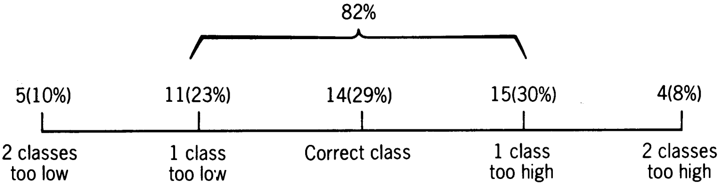


Figure 5 Illustration of Plan's Hit Rate as a Method to Judge the Validity of Job Evaluation Results. (From Milkovich and Newman 1993)

to maintain a proper perspective in interpreting these results. To date, the research has been limited to only a few employers. Further, few compensation professionals seem to consider job evaluation a measurement tool in the strict sense of that term. More often, it is viewed as a procedure to help rationalize an agreed-upon pay structure in terms of job- and business-related factors. As such, it becomes a process of give and take, not some immutable yardstick.

6.3. Utility: Cost-Efficient Results

The usefulness of any management system is a function of how well it accomplishes its objectives (Lawler 1986). Job evaluation is no different; it needs to be judged in terms of its objectives. Pay structures are intended to influence a wide variety of employee behaviors, ranging from staying with an employer to investing in additional training and willingness to take on new assignments. Consequently, the structures obtained through job evaluation should be evaluated in terms of their ability to affect such decisions. Unfortunately, little of this type of evaluation seems to be done.

The other side of utility concerns costs. How costly is job evaluation? Two types of costs associated with job evaluation can be identified: (1) design and administration costs and (2) labor costs that result from pay structure changes recommended by the job evaluation process. The labor cost effects will be unique for each application. Winstanley offers a rule of thumb of 1 to 3% of covered payroll (Winstanley). Experience suggests that costs can range from a few thousand dollars for a small organization to over $300,000 in consultant fees alone for major projects in firms like Digital Equipment, 3M, TRW, or Bank of America.

6.4. Nondiscriminatory: Legally Defensible Results

Much attention has been directed at job evaluation as both a potential source of bias against women and as a mechanism to reduce bias (Treiman and Hartmann 1981). We will discuss some of the studies of the effects of gender in job evaluation and then consider some recommendations offered to ensure bias-free job evaluation.

It has been widely speculated that job evaluation is susceptible to gender bias. To date, three ways that job evaluation can be biased against women have been studied (Schwab and Grams 1985).

First, direct bias occurs if jobs held predominantly by women are undervalued relative to jobs held predominantly by men, simply because of the *jobholder's gender*. The evidence to date is mixed regarding the proposition that the gender of the jobholder influences the evaluation of the job. For example, Arvey et al. found no effects on job evaluation results when they varied the gender of jobholders using photographs and recorded voices (Arvey et al. 1977). In this case, the evaluators rightfully focused on the work, not the worker. On the other hand, when two different job titles (special assistant—accounting and senior secretary—accounting) were studied, people assigned lower job-evaluation ratings to the female-stereotyped title "secretary" than to the more gender-neutral title, "assistant" (McShane 1990).

The second possible source of gender bias in job evaluation flows from the *gender of the individual evaluators*. Some argue that male evaluators may be less favorably disposed toward jobs held predominantly by women. To date, the research finds no evidence that the job evaluator's gender or the predominant gender of the job-evaluation committee biases job-evaluation results (Lewis and Stevens, 1990).

The third potential source of bias affects job evaluation indirectly through the *current wages paid for jobs*. In this case, job-evaluation results may be biased if the jobs held predominantly by women are incorrectly underpaid. Treiman and Hartmann argue that women's jobs are unfairly underpaid simply because women hold them (Trieman and Hartmann 1995). If this is the case and if job evaluation is based on the current wages paid in the market, then the job-evaluation results simply mirror any bias that exists for current pay rates. Considering that many job-evaluation plans are purposely structured to mirror the existing pay structure, it should not be surprising that the current wages for jobs influence the results of job evaluation, which accounts for this perpetual reinforcement. In one study, 400 experienced compensation administrators were sent information on current pay, market, and job-evaluation results (Rynes et al. 1989). They were asked to use this information to make pay decisions for a set of nine jobs. Half of the administrators received a job linked to men (i.e., over 70% of job holders were men, such as security guards) and the jobs given the other half were held predominately by women (e.g., over 70% of job holders were women, such as secretaries). The results revealed several things: (1) Market data had a substantially larger effect on pay decisions than did job evaluations or current pay data. (2) The jobs' gender had no effects. (3) There was a hint of possible bias against physical, nonoffice jobs over white-collar office jobs. This study is a unique look at several factors that may affect pay structures. Other factors, such as union pressures and turnover of high performers, that also affect job-evaluation decisions were not included.

The implications of this evidence are important. If, as some argue, market rates and current pay already reflect gender bias, then these biased pay rates could work indirectly through the job-evaluation process to deflate the evaluation of jobs held primarily by women (Grams and Schwab

1985). Clearly the criteria used in the design of job evaluation plans are crucial and need to be business and work related.

Several recommendations help to ensure that job evaluation plans are bias free (Remick 1984). Among them are:

1. Ensuring that the compensable factors and scales are defined to recognize the content of jobs held predominantly by women. For example, working conditions should include the noise and stress of office machines and the working conditions surrounding computers.
2. Ensuring that compensable factor weights are not consistently biased against jobs held predominantly by women. Are factors usually associated with these jobs always given less weight?
3. Ensuring that the plan is applied in as bias-free a manner as feasible. This includes ensuring that the job descriptions are bias free, that incumbent names are excluded from the job-evaluation process, and that women are part of the job-evaluation team and serve as evaluators.

Some writers see job evaluation as the best friend of those who wish to combat pay discrimination. Bates and Vail argue that without a properly designed and applied system, ‘‘employers will face an almost insurmountable task in persuading the government that ill-defined or whimsical methods of determining differences in job content and pay are a business necessity’’ (Bates and Vail 1984).

6.5. Acceptability: Sensible Results

Acceptance by the employees and managers is one key to a successful job-evaluation system. Any evaluation of the worthiness of pay practices must include assessing employee and manager acceptance. Several devices are used to assess and improve the acceptability of job evaluation. An obvious one is the inclusion of a formal appeals process, discussed earlier. Employees who feel their jobs are incorrectly evaluated should be able to request reanalysis and reevaluation. Most firms respond to such requests from managers, but few extend the process to all employees, unless those employees are represented by unions who have a negotiated grievance process. They often justify this differential treatment on the basis of a fear of being inundated with appeals. Employers who open the appeals process to all employees theorize that jobholders are the ones most familiar with the work performed and know the most about changes, misrepresentations, or oversights that pertain to their job. No matter what the outcome from the appeal, the results need to be explained in detail to any employee who requests that his or her job be reevaluated.

A second method of assessing acceptability is to include questions about the organization’s job structure in employee attitude surveys. Questions can assess perceptions of how useful job evaluation is as a management tool. Another method is to determine to what extent the system is actually being used. Evidence of acceptance and understanding can also be obtained by surveying employees to determine the percentage of employees who understand the reasons for job evaluation, the percentage of jobs with current descriptions, and the rate of requests for reevaluation.

As noted earlier, stakeholders of job evaluations extend beyond employees and managers to include unions and, some argue, comparable worth activists. The point is that acceptability is a somewhat vague test of the job evaluation—acceptable to whom is an open issue. Clearly managers and employees are important constituents because acceptance makes it a useful device; but others, inside and outside the organization, also have a stake in job evaluation and the pay structure.

7. SUMMARY

In exchange for services rendered, individuals receive compensation from organizations. This compensation is influenced by a wide variety of ever-changing dynamics, many of which are identified in this chapter. The central focus of this chapter, though, was on just one of these influences: the internal worth of jobs. We introduced the different systems for evaluating jobs, the procedures necessary to operationalize these systems, and the criteria for evaluating the effectiveness of these systems in an organization.

REFERENCES

Arvey, R. (1986), ‘‘Sex Bias in Job Evaluation Procedures,’’ *Personnel Psychology*, Summer, pp. 315–335.

Arvey, R., Passino, E. M., and Lounsbury, J. W. (1977), ‘‘Job Analysis Results as Influenced by Sex of Incumbent and Sex of Analyst,’’ *Journal of Applied Psychology*, Vol. 62, No. 4, pp. 411–416.

Bates, M. W., and Vail, R. G. (1984), ‘‘Job Evaluation and Equal Employment Opportunity: A Tool for Compliance—A Weapon for Defense,’’ *Employee Relations Law Journal*, Vol. 1, No. 4, pp. 535–546.

Belcher, D. W. (1974), *Compensation Administration*, Prentice Hall, Englewood Cliffs, NJ, pp. 151–152.

Benge, E. J., Burk, S. L. H., and Hay, E. N. (1941), *Manual of Job Evaluation*, Harper & Row, New York.

Bureau of National Affairs (1976), Personnel Policies Forum Survey No. 113, pp. 1–8.

Burns, M. (1978), *Understanding Job Evaluation*, Institute of Personnel Management, London.

Cartter, A. M. (1959), *Theories of Wages and Employment*, Irwin, New York.

Chesler, D. J. (1948), "Reliability and Comparability of Different Job Evaluation Systems," *Journal of Applied Psychology*, October, pp. 465–475.

Collins, J. M., and Muchinsky, P. M. (1993), "An Assessment of the Construct Validity of Three Job Evaluation Methods: A Field Experiment," *Academy of Management Journal*, Vol. 4, pp. 895–904.

Cunningham, J. B., and Graham, S. (1993), "Assessing the reliability of Four Job Evaluation Plans," *Canadian Journal of Administrative Sciences*, Vol. 1, pp. 31–47.

Davis, K. R., and Sauser, W. I. (1993), "A Comparison of Factor Weighting Methods in Job Evaluation: Implications for Compensation Systems," *Public Personnel Management,* Vol. 1, pp. 91–106.

Delbecq, A. L., Van De Ven, A. H., and Gustafson, D. H. (1975), *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*, Scott, Foresman & Co., Glenview, IL.

Dertien, M. G. (1981), "The Accuracy of Job Evaluation Plans," *Personnel Journal*, July, pp. 566–570.

Elizur, D. (1980), *Job Evaluation: A Systematic Approach*, Gower Press, London.

Emerson, S. M. (1991), "Job Evaluation: A Barrier to Excellence," *Compensation and Benefits Review*, Vol. 1, pp. 39–51.

Fowler, A. (1996), "How to Pick a Job Evaluation System," *People Management*, Vol. 2, pp. 42–43.

Fox, W. M. (1962), "Purpose and Validity in Job Evaluation," *Pesonnel Journal,* Vol. 41, pp. 432–437.

Gomberg, W. (1947), *A Labor Union Manual on Job Evaluation*, Roosevelt College, Labor Education Division, Chicago.

Gomez-Mejia, L. R., Page, R. C., and Tornow, W. W. (1982), "A Comparison of the Practical Utility of Traditional, Statistical, and Hybrid Job Evaluation Approaches," *Academy of Management Journal*, Vol. 25, pp. 790–809.

Grams, R. and Schwab, D. (1985), "Investigation of Systematic Gender-Related Error in Job Evaluation," *Academy of Management Journal,* Vol. 28, No. 2, pp. 279–290.

Gupta, N., Jenkins, G. D., and Curington, W. (1986), "Paying for Knowledge: Myths and Realities," *National Productivity Review*, Spring, pp. 107–123.

Harding, F. D., Madden, J. M., and Colson, K. (1960), "Analysis of a Job Evaluation System," *Journal of Applied Psychology*, Vol. 5, pp. 354–357.

Hay Associates (1981), *The Guide Chart-Profile Method of Job Evaluation*.

Hills, F. S. (1989), "Internal Pay Relationships," *Compensation and Benefits*, in L. Gomez-Mejia, Ed., Bureau of National Affairs, Washington, DC, pp. 3.29–3.69.

Huber, V. L. (1991), "Comparison of Supervisor–Incumbent and Female–Male Multidimensional Job Evaluation Ratings," *Journal of Applied Psychology*, Vol. 1, pp. 115–121.

Jaques, E. (1961), *Equitable Payment*, John Wiley & Sons, New York.

Jaques, E. (1964), *Time Span Handbook*, Heinemann, London.

Jaques, E. (1970), *Equitable Payment*, Heinemann, London.

Kalantari, B. (1995), "Dynamics of Job Evaluation and the Dilemma of Wage Disparity in the United States," *Journal of Business Ethics*, Vol. 5, pp. 397–403.

Kanin-Lovers, J., Richter, A. S., and Simon, R. (1995), "Competency-linked Role Evaluation-Management Owned and Operated," *Journal of Compensation and Benefits*, Vol. 10, pp. 53–58.

Korukonda, A. R. (1996), "Cognitive Processes and Computer Advances in Job Evaluation: Innovation in Reverse?" *Canadian Journal of Administrative Sciences*, Vol. 1, pp. 78–82.

Laabs, J. J. (1997), "Rating Jobs Against New Values," *Workforce*, Vol. 5, pp. 38–49.

Lawler, E. E., III (1986), "The New Pay," in *Current Issues in Human Resource Management,* S. Rynes and G. Milkovich, Eds., Richard D. Irwin, Homewood, IL.

Lawler, E. L. (1986), "What's Wrong With Point-Factor Job Evaluation?" *Compensation and Benefits Review*, Vol. 18, March–April, pp. 20–38.

Lawshe, C. H., Jr. (1947), "Studies in Job Evaluation 2: The Adequacy of Abbreviated Point Ratings for Hourly-Paid Jobs in Three Industrial Plants," *Journal of Applied Psychology,* Vol. 31, pp. 355–365.

Lawshe, C. H., Jr. and Farbo, P. C. (1949), "Studies in Job Evaluation 8: The Reliability of an Abbreviated Job Evaluation System," *Journal of Applied Psychology*, Vol. 33, pp. 158–166.

Lewis, C. T., and Stevens, C. K. (1990), "An Analysis of Job Evaluation Committee and Job Holder Gender Effects on Job Evaluation," *Public Personnel Management*, Vol. 3, pp. 271–278.

Livernash, E. R. (1957), "The Internal Wage Structure," in *New Concepts in Wage Determination*, G. Taylor and F. Pierson, Eds., McGraw-Hill, New York, pp. 143–172.

Luthans, F., and Fox, M. L. (1989), "Update on Skill-Based Pay," *Personnel*, March, pp. 26–32.

Madigan, R. M. (1985), "Comparable Worth Judgments: A Measurement Properties Analysis," *Journal of Applied Psychology*, Vol. 70, pp. 137–147.

Madigan, R. M., and Hoover, D. J. (1986), "Effects of Alternative Job Evaluation Methods on Decisions Involving Pay Equity," *Academy of Management Journal*, March, pp. 84–100.

Mahmood, M. A., Gowan, M. A., and Wang, S. P. (1995), "Developing a Prototype Job Evaluation Expert System: A Compensation Management Application," *Information Management*, Vol. 1, pp. 9–28.

McLagan, P. A. (1997), "Competencies, The Next Generation," *Training and Development*, Vol. 5, pp. 40–47.

McShane, S. L. (1990), "Two Tests of Direct Gender Bias in Job Evaluation Ratings," *Journal of Occupational Psychology*, Vol. 2, pp. 129–140.

Milkovich, G. T., and Newman, J. M. (1993), *Compensation*, 4th Ed., Richard D. Irwin, Homewood, IL.

Moore, F. G. (1946), "Statistical Problems in Job Evaluation," *Personnel,* September, pp. 125–136.

Nash, A. N., and Carroll, S. J., Jr. (1975) *The Management of Compensation*, Wadsworth, Belmont, CA.

Patterson, T. T., and Husband, T. M. (1970), "Decision-Making Responsibilities: Yardstick for Job Evaluation," *Compensation Review*, Second Quarter, pp. 21–31.

Pierson, D. (1983), "Labor Market Influences on Entry vs. Non-Entry Wages," *Nebraska Journal of Economics and Business*, Summer, pp. 7–18.

Quaid, M. (1993), "Job Evaluation as Institutional Myth," *Journal of Management Studies*, Vol. 2, pp. 239–260.

Remick, H. (1984), *Comparable Worth and Wage Discrimination.* Temple University Press, Philadelphia.

Rheaume, R. A., and Jones, W. W. (1988), "Automated Job Evaluations That Consolidate What Employees Do," *Computers in Personnel*, Summer, pp. 39–45.

Rynes, S., Weber, C., and Milkovich, G. (1989), "The Effects of Market Survey Rates, Job Evaluation, and Job Gender on Job Pay," *Journal of Applied Psychology*, Vol. 1, pp. 114–123.

Schwab, D. P. (1980), "Job Evaluation and Pay Setting: Concepts and Practices," in *Comparable Worth: Issues and Alternatives*, E. R. Livernash, Ed., Equal Employment Advisory Council, Washington, DC.

Schwab, D., and Grams, R. (1985), "Sex-Related Factors in Job Evaluation: A 'Real-World' Test," *Journal of Applied Psychology*, Vol. 70, No. 3, pp. 533–559.

Snelgar, R. J. (1983), "The Comparability of Job Evaluation Methods," *Personnel Psychology*, Vol. 36, pp. 371–380.

Supel, T. M. (1990), "Equivalence and Redundancy in the Point-Factor Job Evaluation System," *Compensation and Benefits Review*, Vol. 2, pp. 48–55.

Treiman, D. J., and Hartmann, H. I. (1981), *Women, Work, and Wages,* National Academy of Sciences, Washington, DC.

Winstanley, N., personal communication.

CHAPTER 35

Selection, Training, and Development of Personnel

ROBERT W. SWEZEY
InterScience America, Inc.

RICHARD B. PEARLSTEIN
American Red Cross

1. INTRODUCTION **920**

2. SELECTION **921**

- 2.1. Predictors 921
  - 2.1.1. Aptitude and Ability Tests 921
  - 2.1.2. Simulations 922
  - 2.1.3. Interviews and Biodata 922
  - 2.1.4. Work Samples 923
  - 2.1.5. References 923
  - 2.1.6. Others 923
- 2.2. Validity and Fairness 923
- 2.3. Performance Measurement and Criterion Assessment 924

3. TRAINING **924**

- 3.1. Differences between Training and Education 924
- 3.2. Training, Learning, Performance, and Outcomes 924
- 3.3. Analysis of Organizations, Jobs, Tasks, and People 925
- 3.4. Training Design and Development 926
- 3.5. Training Techniques, Strategies, and Technologies 928
  - 3.5.1. Simulation as a Training Technology 929
- 3.6. Acquisition, Retention, and Transfer of Learned Information 929
  - 3.6.1. Acquisition 929
  - 3.6.2. Retention 930
  - 3.6.3. Transfer 931
- 3.7. Training Teams and Groups 933
- 3.8. Training Evaluation 934

4. DEVELOPMENT **937**

- 4.1. Individual Performance Enhancement 937
- 4.2. Organizational Performance Enhancement 938
  - 4.2.1. Management and Leadership Development 938
  - 4.2.2. Organizational Development 939

5. THE FUTURE OF SELECTION, TRAINING, AND DEVELOPMENT **939**

- 5.1. Selection 939
- 5.2. Training 940
- 5.3. Development 940

REFERENCES **941**

1. INTRODUCTION

Selection, training, and development are the primary tools used to promote human performance in industrial engineering settings. In this chapter, we will review important research issues in these areas. We have elected to emphasize training over the equally important topics of selection and development, primarily because training can often have substantial impact in existing application areas. In many industrial engineering situations, staff members are already in place, and even if personnel selection is an option, it is performed by other agencies, such as the human resources

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

department. Second, "development" activities often consist of training—either the construction of new training programs or the use of existing ones, or both. Consequently, in our view, training (plus judicious use of other, ancillary, performance-improvement technologies) constitutes a serious potential impact area for industrial engineering situations.

Further, a major concern in the domain of selection research is the development of technologies designed to address the validity and reliability of instruments used to select personnel. ("Validity" deals with the question of determining the extent to which an assessment technique actually measures what it purports to measure, such as intelligence, clerical skill, and mathematical aptitude. "Reliability" addresses the extent to which such an instrument measures whatever it purports to measure consistently, that is, without major score differences upon repeated applications of the instrument to the same people.) Such topics as these, while very important, are beyond the scope of this chapter.

For the same reason, we have chosen not to deal with other widely researched topics in this domain, such as recruitment of personnel, employee retention, personnel turnover, metaanalysis (statistical technologies for collapsing data from multiple research programs in order to reach generalizable conclusions), and validity generalization (techniques designed to apply known measurement validities obtained in one situation to another). Readers interested in these topics may consult the many current textbooks available in these areas or an overall review article (Guion 1991). An authoritative reference for those involved in the methods used in selection and classification of personnel is the *Standards for Educational and Psychological Testing* (AERA et al. 1999).

2. SELECTION

In selection research, *predictor* refers to any technique or technology used to predict the success of individuals in the application context. In general, it is the goal of selection technologies to predict successfully how well people will perform if hired or selected for admission to a specific situation, such as a college. In the following pages, we briefly review several of the most popular predictor categories.

2.1. Predictors

2.1.1. Aptitude and Ability Tests

Very generally, *aptitude* refers to a capacity or suitability for *learning* in a particular domain or topic area (such as mechanical aptitude), whereas *ability* refers to a natural or inborn talent to *perform* in a given area (such as athletic ability). Development and application of tests of aptitude and ability have long been a major focus of activity in the area of selection research. In general, ability and aptitude tests consist of several subcategories. Cognitive ability tests address issues that involve remembering information, producing ideas, thinking, and comparing or evaluating data. Sensory motor tests include physical, perceptual, motor, and kinesetic measures. Personality, temperament, and motivational tests address those issues as predictors of job success. Additionally, so-called lie-detector or polygraph tests are increasingly used in some situations as a selection and/or screening device, despite the existence of a substantial amount of data that question their validity (see Guion 1991; Sackett and Decker 1979).

One series of projects that have recently addressed aptitude testing involves the development and validation of the (U.S.) Armed Services Vocational Aptitude Battery (ASVAB) (Welsh et al. 1990). This effort involved years of research and dozens of studies that applied and researched the contributions of aptitude-testing techniques to a huge variety of military job settings. To transfer these military technologies to civilian settings, researchers compared ASVAB counterparts to the civilian General Aptitude Test Battery (GATB). These comparisons were reviewed by Hunter (1986) and others, who have demonstrated that many ASVAB components are highly similar to their civilian counterparts, thus arguing that many of the military's scales and testing procedures are directly relevant for nonmilitary applications.

A second major research program in this area is known as Project A. That effort (see Campbell 1990 for a review) involved a number of complex research projects designed to validate predictors of hands-on job-performance measures and other criteria for U.S. Army jobs. Project A, conducted at a cost of approximately $25 million, is certainly one of the largest selection-research projects in history. It found that "core job performance" was predicted by general cognitive ability (assessed by the ASVAB), while ratings of effort and leadership, personal discipline, and physical fitness were best predicted by personality measures (McHenry et al. 1990). Other predictors produced only small increments in validity over general cognitive ability in predicting core job performance (Schmidt et al. 1992).

A common issue in selection research is the question of whether "general" ability is more important in determining job success than are various "special" abilities (which presumably are components of the more overall "general" factor). Various authors (Hunter 1986; Jensen 1986; Thorndike 1986; and many others) have argued in favor of the general factor, whereas others (cf. Prediger 1989) have challenged these conclusions, suggesting that specific aptitudes are more important in predicting job performance.

2.1.2. Simulations

Simulations are increasingly used as predictors in selection situations. Simulations are representations of job situations. Although presumably task relevant, they are nonetheless abstractions of real-world job tasks. Many simulations used in selection occur in so-called assessment centers, in which multiple techniques are used to judge candidates. In such situations, people are often appraised in groups so that both interpersonal variables and individual factors can be directly considered. Simulation exercises are often used to elicit behaviors considered relevant to particular jobs (or tasks that comprise the jobs) for which people are being selected. This includes so-called in-basket simulations for managerial positions, where hypothetical memos and corporate agenda matters are provided to job candidates in order to assess their decision-making (and other) skills.

Recently, computer-based simulations—ranging from low-fidelity ‘‘games’’ to highly sophisticated complex environments coupled with detailed performance-measurement systems—have been developed for use as predictors in selection and assessment situations. One example of this approach is team performance assessment technology (TPAT) (e.g., Swezey et al. 1999). This and similar methods present simulated task environments that assess how individuals and/or teams develop plans and strategies and adapt to changes in fluctuating task demands. This technology exposes respondents to complex environments that generate task-relevant challenges. At known times, participants have an opportunity to engage in strategic planning; at other times, emergencies may require decisive action. These technologies are designed to allow participants to function across a broad range of situations that include unique demands, incomplete information, and rapid change. They often employ a performance-measurement technology termed ‘‘quasi-experimental’’ (Streufert and Swezey 1985). Here, fixed events of importance require that each participant deal with precise issues under identical conditions. Other events, however, are directly influenced by actions of participants. In order to enable reliable measurement, fixed (preprogrammed) features are inserted to allow for comparisons among participants and for performance assessment against predetermined criteria. Other methodologies of this sort include the tactical naval decision making system (TANDEM), a low-fidelity simulation of a command, control, and communication environment (Dwyer et al. 1992; Weaver et al. 1993), and the team interactive decision exercise for teams incorporating distributed expertise (TIDE^2) (Hollenbeck et al. 1991).

2.1.3. Interviews and Biodata

Most reviews concerning the reliability and validity of interviews as a selection device ‘‘have ended with the depressing but persistent conclusion that they have neither’’ (Guion 1991, p. 347). Then why are they used at all? Guion provides four basic reasons for using interviews as selection devices in employment situations:

1. They serve a public relations role. Even if rejected, a potential employee who was interviewed in a competent, professional manner may leave with (and convey to others) the impression that he was treated fairly.
2. They can be useful for gathering ancillary information about a candidate. Although collecting some forms of ancillary information (such as personal, family, and social relationship data) is illegal, other (perfectly legitimate) information on such topics as work history and education that may have otherwise been unavailable can be collected in a well-structured interview protocol.
3. They can be used to measure applicant characteristics that may otherwise not be adequately measured, such as friendliness, ability to make a good first impression, and conversational etc.
4. Interviewers are themselves decision makers. In some circumstances, interviewers may be used to arrive at overall judgments about candidate suitability that can be used in the decision process.

Many alternative ways of conducting interviews have been studied. For instance, Martin and Nagao (1989) compared paper-and-pencil, face-to-face, and computerized versions of interviews. They found that socially desirable responses often occurred when impersonal modes of interviewing were used, but that applicants for high-status jobs often resented impersonal interviews. Tullar (1989) found that successful applicants are given longer interviews and dominate the conversation more than less-successful applicants.

As to biodata, information about an individual’s past, Hammer and Kleiman (1988) found that only 6.8% of 248 firms surveyed stated that they had ever used biodata in employment decisions, and only 0.4% indicated that they currently used biodata. Yet there is evidence that biodata has substantial validity as a selection device (Schmidt et al. 1992). The most frequent reasons for not using biodata were (a) lack of knowledge, (b) lack of methodological expertise, and (c) lack of personnel, funding, and time.

An important assumption in use of biodata as a selection tool is that past behaviors—often behaviors far in the past (e.g., high school achievements)—are good predictors of future behaviors. However Kleiman and Faley (1990) found that biodata items inquiring about *present* behaviors were just as valid as those focusing on past behaviors in predicting intention to reenlist in the Air National Guard; and Russell et al. (1990) found that retrospective life-history essays could serve as the source of valid biodata. These and other studies indicate that biodata may be a significantly underutilized source of information in personnel selection situations.

2.1.4. Work Samples

Unlike simulations, which are not the actual activities performed in the workplace but an abstraction or representation of aspects of those activities, work sample testing involves the actual activities performed on the job. These are measured or rated for use as a predictor. The logic applied to work sample testing is somewhat different from other selection technologies in that it employs a ''criterion-based'' perspective: the performance being tested *is* usually a subset of the actual task employed and is derived directly from the content domain of that task. This stands in contrast to other types of selection methodologies where a measure of some ''construct'' (such as personality, aptitude, ability, or intelligence) is measured and related statistically (usually correlated) with actual job performance (see Swezey 1981 for a detailed discussion of this issue).

In work sample testing, a portion of an actual clerical task that occurs everyday may be used to measure clerical performance (e.g., ''type this letter to Mr. Elliott''). Typically, such tests actually sample the work domain in some representative fashion; thus, typing a letter would be only one aspect of a clerical work sample test that may also include filing, telephone answering, and calculating tasks.

Work sample testing is less abstract than other predictive techniques in that actual performance measurement indices (e.g., counts, errors) or ratings (including checklists) may be used as indices of performance. Where possible, work sample testing (or its close cousin, simulation) is the preferred method for use in selection research and practice because one does not get mired down trying to decide how to apply abstract mental constructs to the prediction of job performance.

2.1.5. References

The use of references (statements from knowledgeable parties about the qualifications of a candidate for a particular position) as a predictor of job performance or success is the most underresearched—yet widely used—selection technique. The problem with using references as a selection technique is that, in most cases, the persons supplying the references have unspecified criteria. That is, unless the requesting parties provide explicit standards for use in providing the reference, referees may address whatever qualities, issues, or topics they wish in providing their comments. This wide-ranging, subjective, unstructured technique is therefore of highly questionable validity as a predictor of job performance. Nevertheless, the technique is widely used in employment and other application-oriented settings (e.g., college applications) because we typically value the unstructured opinions of knowledgeable people concerning assets and/or liabilities of an applicant or candidate. The key word here is ''knowledgeable.'' Unfortunately, we often have no way of determining a referee's level of knowledge about a candidate's merits.

2.1.6. Others

Among the many other predictors used as selection devices are such areas as graphology, genetic testing, investigative background checks, and preemployment drug testing. Of these, preemployment drug testing has undoubtedly received the most attention in the literature. In one noteworthy effort, Fraser and Kroeck (1989) suggested that the value of drug screening in selection situations may be very low. In partial reaction to this hypothesis, a number of investigations were reviewed by Schmidt et al. (1992). Those authors cite the work of Fraser and Kroeck and of Murphy et al. (1990), who, using college student samples, found that individual drug use and self-reported frequency of drug use were consistently correlated with disapproval of employee drug testing. However, no substantial associations between acceptability of drug testing and employment experience or qualifications were established.

2.2. Validity and Fairness

As mentioned above, test validity is a complex and ambiguous topic. Basically, one can break this domain into two primary categories: (1) psychometric validities, including criterion-based validities (which address—either in the future [predictive] or simultaneously [concurrent]—the relationship(s) among predictors and criteria), and construct validity (which relates predictors to psychological constructs, such as abilities, achievement, etc.); and (2) content (or job/task-related) validities. See Guion (1991) for a discussion of these topics.

Various U.S. court rulings have mandated that job-related validities be considered in establishing test fairness for most selection and classification situations. These issues are addressed in the latest *Standards for Educational and Psychological Testing* (1999) document.

2.3. Performance Measurement and Criterion Assessment

An oft-quoted principle is that the best predictor of future performance is past performance. However, current performance can be a good predictor as well. In selecting personnel to perform specific tasks, one may look at measurements of their performance on similar tasks in the past (e.g., performance appraisals, certification results, interviews, samples of past work). As noted earlier, one may also measure their current performance on the tasks for which one is selecting them, either by assessing them on a simulation of those tasks or, when feasible, by using work samples. A combination of both types of measures may be most effective. Measures of past performance show not only what personnel can do, but what they have done. In other words, past performance reflects motivation as well as skills. A disadvantage is that, because past performance occurred under different circumstances than those under which the candidates will perform in the future, ability to generalize from past measures is necessarily limited. Because measures of current performance show what candidates can do under the actual (or simulated) conditions in which they will be performing, using both types of measures provides a more thorough basis on which to predict actual behaviors.

Whether based on past performance or current performance, predictors must not only be accurate, they must also be relevant to the actual tasks to be performed. In other words, the criteria against which candidates' performance is assessed must match those of the tasks to be performed (Swezey 1981). One way to ensure that predictors match tasks is to use criterion-referenced assessment of predictor variables for employee selection. To do this, one must ensure that the predictors match the tasks for which one is selecting candidates as closely as possible on three relevant dimensions: the conditions under which the tasks will be performed, the actions required to perform the task, and the standards against which successful task performance will be measured. While it is better to have multiple predictors for each important job task (at least one measure of past performance and one of current performance), it may not always be possible to obtain multiple predictors. We do not recommend using a single predictor for multiple tasks—the predictor is likely to be neither reliable nor valid.

3. TRAINING

3.1. Differences between Training and Education

"Training" (the systematic, structured development of specific skills required to perform job tasks) differs from "education" (development of broad-based informational background and general skills). Most people are more familiar with education than training because (a) their experiences from elementary school through college usually fall within an educational model, and (b) much of what is called training in business and industry is actually more akin to education than training.

Education is important, even essential, for building broad skill areas, but training is the approach of preference for preparing people to perform specific tasks or jobs. Table 1 shows differences between education and training and between educators and trainers. Not all differences apply to any specific instance of training or education, but as a whole they illustrate the distinction between the two.

3.2. Training, Learning, Performance, and Outcomes

The *purpose* of training is to facilitate learning skills and knowledges required to perform specific tasks. The *outcome* of training is acceptable performance on those tasks. People often learn to perform tasks without training, but if training is effective, it enables people to learn faster and better. For example, a person might be able to learn to apply the principle of cantilevering to bridge construction through reading and trial and error, but the process would be lengthier and the outcomes would probably be less precise than if the person received systematic, well-designed training.

Because training is a subsystem of a larger organizational system, it should not be treated outside the context of the larger system. In organizations, training is usually requested in two types of circumstances: either current performance is inadequate or new performance is required. In other words, organizations train either to solve current problems or to take advantage of new opportunities. A second factor affecting training is whether it is intended for new or incumbent employees. Basically, training for new opportunities is essentially the same for new and incumbent employees. However, new employees may require prerequisite or remedial training prior to receiving training on new tasks to help take advantage of opportunities, whereas incumbents presumably would have already completed the prerequisite training (or would have otherwise mastered the necessary skills).

TABLE 1 Education and Educators Compared to Training and Trainers

Education	Training
Increases knowledge	Increases job performance
Broadens understanding	Focuses on job skills
Conceptualizes experience	Simulates experience
Asks "How much do you know?"	Asks "What can you do with what you know?"
Provides base of knowledge	Uses knowledge to build skills
Often requires broad generalization of knowledge	Requires knowledge to be applied in specific situations
Educator	**Trainer**
Dispenses information	Facilitates learning experience
Is primary learning resource	Allows participants to serve as learning resources
Stresses inquiry	Stresses practical application
Groups students according to ability (norm-referenced)	Tries to help all participants perform to specified level (criterion-referenced)
Focuses on information	Focuses on performance
Evaluates via classroom testing	Evaluates via on-the-job performance assessment

Just as learning can take place without training, organizational outcomes can be achieved without training. In fact, when acceptable performance on specific tasks is a desired organizational outcome, training is frequently *not* the appropriate way to achieve that outcome (see Gilbert 1978; Harless 1975; Mager and Pipe 1970). The type of analysis that determines factors contributing to performance of an organization's personnel subsystem has been called by a variety of names, including performance analysis (Mager and Pipe 1970), and front-end analysis (Harless 1975). Mager and Pipe (1970), noting that training is appropriate only when performance is unacceptable because performers lack necessary skills and knowledge, suggested this classic question for making the determination: "Could they do it if their lives depended on it?" If the answer is "no," then training is warranted. If the answer is "yes," the next question is, "What is preventing them from performing as required?" Mager and Pipe also suggested various ways of categorizing obstacles to desired performance outcomes that do not consist of skill/knowledge deficiencies (see also Harless 1975; Rummler and Brache 1990). These can be reduced to two broad categories:

1. Lack of environmental support (e.g., well-specified processes, job aids, tools and equipment, management expectations, adequate facilities)
2. Lack of appropriate motivational systems (e.g., clear and immediate feedback, appropriate consequences for action, well-designed reward systems)

For skill/knowledge deficiencies, training interventions are appropriate. Deficiencies involving environmental support and/or appropriate motivational systems should be remedied through other human performance interventions, such as providing clear instructions and developing reward systems. Regardless of the intervention, implementation should always include pilot testing and evaluation components. If desired outcomes are not achieved, the analysis, design, and/or implementation may have been flawed.

3.3. Analysis of Organizations, Jobs, Tasks, and People

Organizations consist of people performing tasks to achieve a particular mission. A job is defined as a collection of related tasks that helps accomplish significant elements of an organization's products and/or services. For training and other human performance-improvement interventions to be successful, they must address appropriate organizational levels and/or units. Organizational analysis seeks to determine relationships among an organization's inputs, functions, jobs, and products/services. This type of analysis builds upon front-end and performance analysis.

Rummler and Wilkins (1999) introduced a technique known as performance logic to diagnose an organization's problems and opportunities. The process highlights disconnects among procedures, skills, jobs, and desired product/service outcomes, pointing the way to appropriate training and other human performance-improvement interventions. Others (Mourier 1998) have also demonstrated successes using this approach.

Job analysis has several purposes. One deals with determining how to apportion an organization's tasks among jobs. This process identifies the tasks that belong with each job, the goal being to reduce overlap while providing clean demarcations between products and processes. From the perspective of an organization's human resources subsystem, job analysis can be used to organize tasks into clusters that take advantage of skill sets common to personnel classification categories (e.g., tool and die press operator, supervisory budget analyst) and consequently reduce the need for training. By distributing tasks appropriately among jobs, an organization may be able to select skilled job candidates, thereby reducing attendant training costs.

A second purpose of job analysis involves job redesign. During the 1970s, researchers addressed various ways to redesign jobs so that the nature of the jobs themselves would promote performance leading to desired organizational outcomes. For example, Lawler et al. (1973) found that several factors, such as variety, autonomy, and task identity, influenced employee satisfaction and improved job performance. Herzberg (1974) identified other factors, such as client relationship, new learning, and unique expertise, having similar effects on satisfaction and performance. Peterson and Duffany (1975) provide a good overview of the job redesign literature.

Needs analysis (sometimes called needs assessment) examines what should be done so that employees can better perform jobs. Needs analysis focuses on outcomes to determine optimal performance for jobs. Rossett (1987) has provided a detailed needs-analysis technique. Instead of needs analysis, many organizations unfortunately conduct a kind of wants analysis, a process that asks employees and/or supervisors to state what is needed to better perform jobs. Because employees and managers frequently cannot distinguish between their wants and their needs, wants analysis typically yields a laundry list of information that is not linked well to performance outcomes.

Finally, task analysis is a technique that determines the inputs, tools, and skills/knowledge necessary for successful task performance. In training situations, task analysis is used to determine the skill/knowledge requirements for personnel to accomplish necessary job outcomes. Skills gaps are defined as the difference between required skills/knowledge and those possessed by the individuals who are (or will be) performing jobs. Training typically focuses on eliminating the skills gaps. Helpful references for conducting task analyses include Carlisle (1986) and Zemke and Kramlinger (1982).

3.4. Training Design and Development

A comprehensive reference work with respect to training design and development (Goldstein 1993) refers to what is generally known as a systems approach to training. Among many other things, this approach includes four important components:

1. Feedback is continually used to modify the instructional process, thus training programs are never completely finished products but are always adaptive to information concerning the extent to which programs meet stated objectives.
2. Recognition exists that complicated interactions occur among components of training such as media, individual characteristics of the learner, and instructional strategies.
3. The systems approach to training provides a framework for reference to planning.
4. This view acknowledges that training programs are merely one component of a larger organizational system involving many variables, such as personnel issues, organization issues, and corporate policies.

Thus, in initiating training program development, it is necessary, according to Goldstein, to consider and analyze not only the tasks that comprise the training, but characteristics of the trainees and organizations involved.

This approach to instruction is essentially derived from procedures developed in behavioral psychology. Two generic principles underlie technological development in this area. First, the specific behaviors necessary to perform a task must be precisely described, and second, feedback (reinforcement for action) is utilized to encourage mastery.

This (and previous work) served as a partial basis for the military's instructional systems development (ISD) movement (Branson et al. 1975) and the subsequent refinement of procedures for criterion-referenced measurement technologies (Glaser 1963; Mager 1972; Swezey 1981).

For a discussion of the systems approach to training, we briefly review the Branson et al. (1975) ISD model. This model is possibly the most widely used and comprehensive method of training development. It has been in existence for over 20 years and has revolutionized the design of instruction in many military and civilian contexts. The evolutionary premises behind the model are that performance objectives are developed to address specific behavioral events (identified by task analyses), that criterion tests are developed to address the training objectives, and that training is essentially developed to teach students to pass the tests and thus achieve the requisite criteria.

The ISD method has five basic phases: analysis, design, development, implementation, and control (evaluation).

Phase one—analysis:

- **(a)** Perform a task analysis. Develop a detailed list of tasks, the conditions under which the tasks are performed, the skills and/or knowledge required for their performance, and the standards that indicate when successful performance has been achieved.
- **(b)** Select pertinent tasks/functions. Note, by observation of an actual operation, the extent to which each task is actually performed, the frequency with which it is performed, and the percentage of a worker's time devoted to it. Then make an assessment concerning the importance of each task to the overall job.
- **(c)** Construct job-performance measures from the viewpoint that job-performance requirements are the basis for making decisions about instructional methods.
- **(d)** Analyze training programs that already exist to determine their usefulness to the program being developed.
- **(e)** Identify the instructional setting/location for each task selected for instruction.

Phase two—design:

- **(a)** Describe student entry behavior. Classify, at the task level, categories into which each entry behavior falls (i.e., cognitive, procedural, decision making, problem solving, etc.). Then check whether each activity is within the existing performance repertoire of the trainees (something the potential students already know) or is a specialized behavior (something the students will have to learn).
- **(b)** Develop statements that translate the performance measures into terminal learning objectives. Consider the behavior to be exhibited, the conditions under which the learning will be demonstrated, and the standards of performance that are considered acceptable.
- **(c)** Develop criterion-referenced tests (CRTs). CRTs are tests that measure what a person needs to know or do in order to perform a job. Ensure that at least one test item measures each instructional objective. Determine whether the testing should be done by pencil and paper, by practical exercise, or by other methods. Criterion-referenced tests perform three essential functions:
    - They help to construct training by defining detailed performance goals.
    - They aid in diagnosis of whether a student has absorbed the training content because the tests are, in fact, operational definitions of performance objectives.
    - They provide a validation test for the training program.
- **(d)** Determine the sequence and structure of the training. Develop instructional strategies and the sequencing of instruction for the course.

Phase three—development:

- **(a)** Specify specific learning events and activities.
- **(b)** Specify an instructional management plan. Determine how instruction will be delivered and how the student will be guided through it: by group mode, self-paced mode, or combinations of the two. Determine how the curriculum will be presented (e.g., lectures, training aids, simulators, job-performance aids, television, demonstration, and computer based).
- **(c)** Review and select existing instructional materials for inclusion in the curriculum.
- **(d)** Develop new instructional materials.
- **(e)** Validate the instruction by having it reviewed by experts or job incumbents or by trying it out on typical students.

Phase four—implementation:

- **(a)** Implement the instructional management plan by ensuring that all materials, equipment, and facilities are available and ready for use. Before beginning instruction, schedule students and train instructors.
- **(b)** Conduct the instruction under the prescribed management plan in the designated setting under actual conditions using final course materials.

Phase five—control (evaluation):

- **(a)** Monitor course effectiveness, assess student progress, assess the quality of the course, and compare results with the original learning objectives.
- **(b)** Determine how graduates perform on the job.
- **(c)** Revise the course as appropriate.

3.5. Training Techniques, Strategies, and Technologies

Two factors known to affect quality of training are the techniques used to deliver the instruction and the methods or strategies used to convey the instructional content. One of the main tasks in the development of training, therefore, is the selection and implementation of appropriate instructional methods and techniques.

One function of technology-assessment research is to provide a basis for estimating the utility of particular technology-based approaches for training. Researchers have analyzed virtually every significant technology trend of the last two decades, including effects of programmed instruction (Kulik et al. 1980), computer-based instruction (Kulik and Kulik 1987), and visual based instruction (Cohen et al. 1981). The majority of these studies show no significant differences among technology groups and, when pooled, yield only slight advantages for innovative technologies. Kulik and associates report that the size of these statistically significant gains is in the order of 1.5 percentage points on a final exam. Cohen et al. (1981) conducted an analysis of 65 studies in which student achievement using traditional classroom-based instruction was compared with performance across a variety of video-based instructional media, including films, multimedia, and educational TV. They found that only one in four studies (approximately 26%) reported significant differences favoring visual-based instruction. The overall effect size reported by Cohen et al. was relatively small compared to studies of computer-based instruction or of interactive video.

Computer-based instruction and interactive video both represent areas where some data reporting advantages have emerged; specifically, reductions in the time required to reach threshold performance levels (Fletcher 1990; Kulik and Kulik 1987). Results of an analysis of 28 studies, conducted by Fletcher (1990), suggested that interactive video-based instruction increases achievement by an average of 0.50 standard deviations over conventional instruction (lecture, text, on-the-job training, videotape).

Results of decades of work in the area of media research led Clark (1983, 1994) to conclude that media have no significant influence on learning effectiveness but are mere vehicles for presenting instruction. According to Clark (1983, p. 445), "the best current evidence is that media are mere vehicles that deliver instruction but do not influence student achievement any more than the truck that delivers our groceries causes changes in our nutrition." Similarly, Schramm (1977) commented that "learning seems to be affected more by what is delivered than by the delivery system." Although students may prefer sophisticated and elegant media, learning appears largely unaffected by these features. Basically, all media can deliver either excellent or ineffective instruction. It appears that it is the instructional methods, strategies, and content that facilitate or hinder learning. Thus, many instructional technologies are considered to be equally effective provided that they are capable of dealing with the instructional methods required to achieve intended training objectives.

Meanwhile, teaching methods appear to be important variables influencing the effectiveness of instructional systems. Instructional methods define how the process of instruction occurs: what information is presented, in what level of detail, how the information is organized, how information is used by the learner, and how guidance and feedback are presented. The choice of a particular instructional method often limits the choice of presentation techniques. Technology-selection decisions, therefore, should be guided by: (1) capacity of the technology to accommodate the instructional method, (2) compatibility of the technology with the user environment, and (3) trade-offs that must be made between technology effectiveness and costs.

Conventional instructional strategies such as providing advance organizers (Allen 1973), identifying common errors (Hoban and Van Ormer 1950), and emphasizing critical elements in a demonstration (Travers 1967) work well with many technology forms, including video-based instructional technologies. Embedding practice questions within instructional sequences also appears to be a potent method supported by interactive video and computer-based instruction. Benefits of embedded questions apparently occur whether responses to the questions are overt or covert or whether the material is preceded or followed by the questions (Teather and Marchant 1974). Compared to research on instructional delivery systems, relatively little research has examined the application of instructional methods.

Morris and Rouse (1985) have suggested that strategies used to teach troubleshooting skills should combine both theoretical and procedural knowledge. This was confirmed by Swezey et al. (1991), who addressed these issues using a variety of instructional strategies. Three strategies, termed "procedural," "conceptual," and "integrated," were used. Results indicated that integrated training strategies that combine procedural and theoretical information facilitate retention of concepts and improve transfer to new settings. Swezey and Llaneras (1992) examined the utility of a technology that provides training program developers with guidance for use in selecting *both* instructional strategies and media. The technology distinguishes among 18 instructional strategies, as well as among 14 types of instructional media. Outputs are based upon decisions regarding the types of tasks for which training is to be developed, as well as consideration of various situational characteristics operating in the particular setting. Three general classes of situational characteristics are considered: (1) task attributes (motion, color, difficulty, etc.), (2) environmental attributes (hazardous, high workload, etc.), and (3) user attributes (level of experience and familiarity with the task). Results indicated that participants who had access to the selection technology made better presentation strategy, media, and study/practice strategy selections than those relying on their own knowledge.

3.5.1. Simulation as a Training Technology

As technology develops, progress is being made in linking various types of instructional devices. Fletcher (1990) has reviewed much of the literature in this area. In general, he has found that computer-based instructional techniques were often more effective than conventional instruction if they included interactive features such as tutorials. Fletcher also notes that the available studies provide little insight into the relative contributions of various features of interactive technology to learning.

Equipment simulators have been used extensively for training aircraft pilots. Jacobs et al. (1990), and Andrews et al. (1992) have reviewed this research. In general, it appears that simulator training combined with training in actual aircraft is more effective than training in the aircraft by itself. According to Tannenbaum and Yukl (1992), the realism of simulators is being enhanced greatly by continuing developments in video technology, speech simulation, and speech recognition; and advancements in networking technology have opened up new possibilities for large-scale simulator networking. Although simulators have been used extensively for training small teams in the military, networking allows larger groups of trainees to practice their skills in large-scale interactive simulations (Thorpe 1987; Alluisi 1991; Weaver et al. 1995). Often in these circumstances, the opponent is not merely a computer but other teams, and trainees are therefore often highly motivated by the competition. In such situations, as in the actual environment, simulations are designated to operate directly as well as to react to unpredictable developments. A thorough history of the development of one such complex, networked system, known as SIMNET, has been provided by Alluisi (1991). Comprehensive, system-based evaluations of these technologies, however, remain to be performed.

3.6. Acquisition, Retention, and Transfer of Learned Information

3.6.1. Acquisition

Much literature has indicated that it is necessary to separate the initial acquisition of a performance skill or concept from longer-term learning (Schmidt and Bjork 1992; Cormier 1984; Salmoni et al. 1984). Schmidt and Bjork (1992), for instance, have described a process by which characteristics of a task create an environment in which performance is depressed but learning is apparently enhanced. The implications of this phenomenon, known as contextual interference (Battig 1966; Shea and Morgan 1979), for training are that variations of task conditions, manipulation of task difficulty, or other means of inducing extra effort by trainees are likely to be beneficial for retention and transfer. Recent research by Schneider et al. (1995) investigated effects of contextual interference in rule-based learning, a skill underlying many applied tasks found in the aviation, military, and computer fields. Results mirrored those found in the verbal learning and motor skill-acquisition literatures, indicating that a random practice schedule leads to the best retention performance after initial acquisition but hinders initial learning. Thus, training conditions designed to achieve a particular training objective (long-term retention, transfer, and/or resistance to altered contexts) are not necessarily those that maximize performance during acquisition. This concept has significant implications for predicting trainee performance. Although a typical method employed to predict trainee performance involves monitoring acquisition data during training, using immediate past performance data as predictors to estimate future retention performance, this strategy may not appropriately index performance on the job or outside of the training environment. Further, initial performance of complex skills tends to be unstable and often is a poor indicator of final performance. Correlations between initial and final performance levels for a grammatical reasoning task, for example, reached only 0.31 (Kennedy et al. 1980), a moderate level at best. Using initial performance during learning as a predictor, therefore, may lead to inconsistent and/or erroneous training prescriptions.

Skill acquisition is believed by many to proceed in accordance with a number of stages or phases of improvement. Although the number of stages and their labels differ among researchers, the existence of such stages is supported by a large amount of research and theoretical development during the past 30 years. Traditional research in learning and memory posits a three-stage model for characterizing associative-type learning involving the process of differentiating between various stimuli or classes of stimuli to which responses are required (stimulus discrimination), learning the responses (response learning), and connecting the stimulus with a response (association). Empirical data suggest that this process is most efficient when materials are actively processed in a meaningful manner, rather than merely rehearsed via repetition (Craik and Lockhart 1972; Cofer et al. 1966).

Anderson (1982) has also proposed a three-stage model of skill acquisition, distinguishing among cognitive, associative, and autonomous stages. The cognitive stage corresponds to early practice in which a learner exerts effort to comprehend the nature of the task and how it should be performed.

In this stage, the learner often works from instructions or an example of how a task is to be performed (i.e., modeling or demonstration). Performance may be characterized by instability and slow growth, or by extremely rapid growth, depending upon task difficulty and degrees of prior experience of the learner. By the end of this stage, learners may have a basic understanding of task requirements, rules, and strategies for successful performance; however, these may not be fully elaborated. During the associative stage, declarative knowledge associated with a domain (e.g., facts, information, background knowledge, and general instruction about a skill acquired during the previous stage) is converted into procedural knowledge, which takes the form of what are called production rules (condition–action pairs). This process is known as knowledge compilation. It has been estimated that hundreds, or even thousands, of such production rules underlie complex skill development (Anderson 1990). Novice and expert performance is believed to be distinguished by the number and quality of production rules. Experts are believed to possess many more elaborate production rules than novices (Larkin 1981).

Similarly, Rumelhart and Norman (1978) recognized three kinds of learning processes: (1) the acquisition of facts in declarative memory (accretion), (2) the initial acquisition of procedures in procedural memory (restructuring), and (3) the modification of existing procedures to enhance reliability and efficiency (tuning). Kyllonen and Alluisi (1987) reviewed these concepts in relation to learning and forgetting facts and skills. Briefly, they suggest that new rules are added to established production systems through the process of accretion, fine-tuned during the process of tuning, and subsequently reorganized into more compact units during the restructuring process.

Rasmussen (1979) has also distinguished among three categories, or modes, of skilled behavior: skill based, rule based, and knowledge based. Skill-based tasks are composed of simple stimulus-response behaviors, which are learned by extensive rehearsal and are highly automated. Rule-based behavior is guided by conscious control and involves the application of appropriate procedures based on unambiguous decision rules. This process involves the ability to recognize specific well-defined situations that call for one rule rather than the other. Knowledge-based skills are used in situations in which familiar cues are absent and clear and definite rules do not always exist. Successful performance involves the discrimination and generalization of rule-based learning. Rasmussen proposes that, in contrast to Anderson's model, performers can move among these modes of performance as dictated by task demands. This general framework is useful in conceptualizing links between task content and the type of training required for proficiency in complex tasks.

3.6.2. Retention

Instructional designers must consider not only how to achieve more rapid, high-quality training, but also how well the skills taught during training will endure after acquisition. Further, what has been learned must be able to be successfully transferred or applied to a wide variety of tasks and job-specific settings. Swezey and Llaneras (1997) have recently reviewed this area.

Retention of learned material is often characterized as a monotonically decreasing function of the retention interval, falling sharply during the time immediately following acquisition and declining more slowly as additional time passes (Wixted and Ebbesen 1991; Ebbinghaus 1913; McGeoch and Irion 1952). There is general consistency in the loss of material over time. Subjects who have acquired a set of paired items, for example, consistently forget about 20% after a single day and approximately 50% after one week (Underwood 1966). Bahrick (1984), among others, demonstrated that although large parts of acquired knowledge may be lost rapidly, significant portions can also endure for extended intervals, even if not intentionally rehearsed.

Evidence suggests that the rate at which skills and knowledge decay in memory varies as a function of the degree of original learning; decay is slower if material has previously been mastered than if lower-level acquisition criteria were imposed (Loftus 1985). The slope and shape of retention functions also depend upon the specific type of material being tested as well as upon the methods used to assess retention. As meaningfulness of material to the student increases, for example, rate of forgetting appears to slow down. Further, recognition performance may be dramatically better than recall, or vice versa, depending simply upon how subjects are instructed (Tulving and Thomson 1973; Watkins and Tulving 1975). Also, various attributes of the learning environment, such as conditions of reinforcement, characteristics of the training apparatus, and habituation of responses, appear to be forgotten at different rates (Parsons et al. 1973).

Retention, like learning and motivation, cannot be observed directly; it must be inferred from performance following instruction or practice. To date, no uniform measurement system for indexing retention has been adopted. General indices of learning and retention used in the research literature over the past 100 years include a huge variety of methods for measuring recall, relearning, and recognition. Luh (1922), for instance, used five different measures to index retention: recognition, reconstruction, written reproduction, recall by anticipation, and relearning.

The list of variables known to reliably influence rate of forgetting of learned knowledge and skills is relatively short. In a recent review, Farr (1987) surveyed the literature and identified a list of variables known to influence long-term retention, including degree of original learning, characteristics of the learning task, and the instructional strategies used during initial acquisition. According to Farr, the single largest determinant of magnitude of retention appears to be the degree of original learning. In general, the greater the degree of learning, the slower will be the rate of forgetting (Underwood and Keppel 1963). This relationship is so strong that it has prompted some researchers to argue that any variable that leads to high initial levels of learning will facilitate retention (Prophet 1976; Hurlock and Montague 1982).

The organization and complexity of material to be learned also appear to have a powerful influence on retention. The efficiency of information acquisition, retrieval, and retention appears to depend to a large degree on how well the material has been organized (Cofer et al. 1966). Conceptually organized material appears to show considerably less memory loss and to be more generalizable than material that is not so organized.

Research has identified numerous variables that fall under the banner of strategies for skill and knowledge acquisition and retention, including spaced reviews, massed/distributed practice, part/whole learning, and feedback. An important variable with respect to forgetting is spaced practice. The "spacing effect" (the dependency of retention on the spacing of successive study sessions) suggests that material be studied at widely spaced intervals if retention is required. Similarly, research comparing distributed practice (involving multiple exposures of material over time) vs. massed practice (requiring concentrated exposure in a single session), has occurred in a wide variety of contexts, including the acquisition of skills associated with aircraft carrier landings (Wightman and Sistrunk 1987), word-processing skills (Bouzid and Crawshaw 1987), perceptual skills associated with military aircraft recognition (Jarrard and Wogalter 1992), and learning and retention of second-language vocabulary (Bloom and Shuell 1981). Most research in this area, however, has emphasized acquisition of verbal knowledge and motor skills. In general, research on the issue of distribution of practice has emphasized effects on acquisition and found that distribution of practice is superior to massed practice in most learning situations, long rest periods are superior to short rest periods, and short practice sessions between rests yield better performance than do long practice sessions (Rea and Modigliani 1988). However, no consistent relationships between these variables and long-term retention have emerged.

Techniques that help learners to build mental models that they can use to generate retrieval cues, recognize externally provided cues, and/or generate or reconstruct information have also generally facilitated retention (Kieras and Bovair 1984). Gagné (1978) identified three general strategies for enhancing long-term retention, including reminding learners of currently possessed knowledge that is related to the material to be learned, ensuring that the training makes repeated use of the information presented, and providing for and encouraging elaboration of the material during acquisition as well as during the retention interval.

Other factors that have been shown to induce forgetting include interference by competing material that has been previously or subsequently acquired and the events encountered by individuals between learning and the retention test. Information acquired during this interval may impair retention, while simple rehearsal or reexposure may facilitate retention. Additional factors influencing skill and knowledge retention include the length of the retention interval, the methods used to assess retention, and individual difference variables among trainees. The absolute amount of skill/knowledge decay, for example, tends to increase with time, while the rate of forgetting declines over time. Researchers have also postulated a critical period for skills loss during a nonutilization period at between six months and one year after training (O'Hara 1990). Psychomotor flight skills, for example, are retained many months longer than procedural flight skills (Prophet 1976), and decay of flight skills begins to accelerate after a six-month nonutilization period (Ruffner et al. 1984).

3.6.3. Transfer

The topic of transfer-of-training is integrally related to other training issues, such as learning, memory, retention, cognitive processing, and conditioning; these fields make up a large subset of the subject matter of applied psychology (Swezey and Llaneras 1997). In general, the term *transfer-of-training* concerns the way in which previous learning affects new learning or performance. The central question is how previous learning transfers to a new situation. The effect of previous learning may function either to improve or to retard new learning. The first of these is generally referred to as positive transfer, the second as negative transfer. (If new learning is unaffected by prior learning, zero transfer is said to have occurred.) Many training programs are based upon the assumption that what is learned during training will transfer to new situations and settings, most notably the operational environment. Although U.S. industries are estimated to spend billions annually on training and development, only a fraction of these expenditures (not more than 10%) are thought to result in performance transfer to the actual job situation (Georgenson 1982). Researchers, therefore, have sought to determine fundamental conditions or variables that influence transfer-of-training and to develop comprehensive theories and models that integrate and unify knowledge about these variables.

Two major historical viewpoints on transfer exist. The identical elements theory (first proposed by Thorndike and Woodworth 1901) suggests that transfer occurs in situations where identical elements exist in both original and transfer situations. Thus, in a new situation, a learner presumably takes advantage of what the new situation offers that is in common with the learner's earlier experiences. Alternatively, the transfer-through-principles perspective suggests that a learner need not necessarily be aware of similar elements in a situation in order for transfer to occur. This position suggests that previously used "principles" may be applied to occasion transfer. A simple example involves the principles of aerodynamics, learned from kite flying by the Wright brothers, and the application of these principles to airplane construction. Such was the position espoused by Judd (1908), who suggested that what makes transfer possible is not the objective identities between two learning tasks, but the appropriate generalization in the new situation of principles learned in the old. Hendriksen and Schroeder (1941) demonstrated this transfer-of-principles philosophy in a series of studies related to the refraction of light. Two groups were given practice shooting at an underwater target until each was able to hit the target consistently. The depth of the target was then changed, and one group was taught the principles of refraction of light through water while the second was not. In a subsequent session of target shooting, the trained group performed significantly better than did the untrained group. Thus, it was suggested that it may be possible to design effective training environments without a great deal of concern about similarity to the transfer situation, as long as relevant underlying principles are utilized.

A model developed by Miller (1954) attempted to describe relationships among simulation fidelity and training value in terms of cost. Miller hypothesized that as the degree of fidelity in a simulation increased, the costs of the associated training would increase as well. At low levels of fidelity, very little transfer value can be gained with incremental increases in fidelity. However, at greater levels of fidelity, larger transfer gains can be made from small increments in fidelity. Thus, Miller hypothesized a point of diminishing returns, where gains in transfer value are outweighed by higher costs. According to this view, changes in the requirements of training should be accompanied by corresponding changes in the degree of fidelity in simulation if adequate transfer is to be provided. Although Miller did not specify the appropriate degree of simulation for various tasks, subsequent work (Alessi 1988) suggests that the type of task and the trainee's level of learning are two parameters that interact with Miller's hypothesized relationships. To optimize the relationship among fidelity, transfer, and cost, therefore, one must first identify the amount of fidelity of simulation required to obtain large amount of transfer and the point where additional increments of transfer are not worth the added costs.

Another model, developed by Kinkade and Wheaton (1972), distinguishes among three components of simulation fidelity: equipment fidelity, environmental fidelity, and psychological fidelity. Equipment fidelity refers to the degree to which a training device duplicates the "appearance and feel" of the operational equipment. This characteristic of simulators has also been termed physical fidelity. Environmental, or functional, fidelity refers to the degree to which a training device duplicates the sensory stimulation received from the task situation. Psychological fidelity (a phenomenon that Parsons [1980], has termed "verisimilitude") addresses the degree to which a trainee perceives the training device as duplicating the operational equipment (equipment fidelity) and the task situation (environmental fidelity). Kinkade and Wheaton postulated that the optimal relationship among levels of equipment, environmental, and psychological fidelity varies as a function of the stage of learning. Thus, different degrees of fidelity may be appropriate at different stages of training.

The relationship between degree of fidelity and amount of transfer is complex. Fidelity and transfer relationships have been shown to vary as a function of many factors, including instructor ability, instructional techniques, types of simulators, student time on trainers, and measurement techniques (Hays and Singer 1989). Nevertheless, training designers often favor physical fidelity in a training device, rather than the system's functional characteristics, assuming that increasing levels of physical fidelity are associated with higher levels of transfer. The presumption that similarity facilitates transfer can be traced to the identical elements theory of transfer first proposed by Thorndike and Woodworth (1901). In fact, numerous studies show that similarity does not need to be especially high in order to generate positive transfer-of-training (see Hays 1980; Provenmire and Roscoe 1971; Valverde 1973). Hay's (1980) review of fidelity research in military training found no evidence of learning deficits due to lowered fidelity, and others have shown an advantage for lower-fidelity training devices, suggesting that the conditions of simulation that maximize transfer are not necessarily those of maximum fidelity. One possible reason for this may be that widespread use of such well-known techniques as corrective feedback and practice to mastery in simulators and training devices, which may act to increase transfer, may also decrease similarity. A second possibility is that introducing highly similar (but unimportant) components into a simulation diverts a student's attention from the main topics being trained, thereby causing reduced transfer, as is the case with manipulating the fidelity of motion systems in certain types of flight simulators (Lintern 1987; Swezey 1989).

A recent review by Alessi (1988) devotes considerable attention to debunking the myth that high fidelity necessarily facilitates transfer. After reviewing the literature in the area, Alessi concluded that transfer-of-training and fidelity are not linearly related but instead may follow an inverted-U-shaped function similar to that suggested by the Yerkes–Dodson law (see Swezey 1978; Welford 1968), which relates performance to stress. According to this view, increases in similarity first cause an increase and then a corresponding decrease in performance, presumably as a function of increasing cognitive load requirements associated with decreased stimulus similarity.

Current issues and problems in transfer span four general areas: (1) measurement of transfer, (2) variables influencing transfer, (3) models of transfer, and (4) application of our knowledge of transfer to applied settings. Research on transfer needs to be conducted with more relevant criterion measures for both generalization and skill maintenance. Hays and Singer (1989) provide a comprehensive review of this domain. A large proportion of the empirical research on transfer has concentrated on improving the design of training programs through the incorporation of learning principles, including identical elements, teaching of general principles, stimulus variability, and various conditions of practice. A critique of the transfer literature conducted by Baldwin and Ford (1988) also indicates that research in the area needs to take a more interactive perspective that attempts to develop and test frameworks to incorporate more complex interactions among training inputs. Many studies examining training design factors, for example, have relied primarily upon data collected using college students working on simple memory or motor tasks with immediate learning or retention as the criterion of interest. Generalizability, in many cases, is limited to short-term simple motor and memory tasks.

3.7. Training Teams and Groups

Because teams are so prevalent, it is important to understand how they function (Swezey and Salas 1992). Teamwork can take on various orientations, including behavioral and cognitive perspectives. Behaviorally oriented views of the teamwork process deal with aspects involved in coordinated events among members that lead to specific functions (e.g., tasks, missions, goals, or actions) that a team actually performs in order to accomplish its required output. Cognitive views of the teamwork domain deal with a team's shared processing characteristics, such as joint knowledge about a task, joint ability to perform aspects of the work, motivation levels, personalities, and thought processes (Salas et al. 1995). The extent to which mental models are shared among team members may help in understanding and training teamwork skills (Cannon-Bowers and Salas 1990). In order for members of a team to coordinate effectively with others, they must share knowledge of team requirements. This, in turn, can help to coordinate a team effectively. Since the range of each individual team member's abilities varies widely, the overall level of competency within a team is often dependent upon shared skills. An individual's ability to complete a team function competently and convey information to other members accurately is a strong influence on the overall team structure and performance.

Swezey et al. (1994) suggest that in order to complete a team task accurately, four things must happen: (1) there must be an exchange of competent information among team members; (2) there must be coordination within the team structure of the task requirements; (3) periodic adjustments must be made to respond to task demands; and (4) a known and agreed-upon organization must exist within the group.

It is important to consider such issues as team size, task characteristics, and feedback use when designing team training programs. It has been shown that team size has a direct influence on the performance and motivation of teams. As the size of the team increases, team resources also increase (Shaw 1976). Due to the increase in information available to larger teams, individual team members have a more readily accessible pool of information for use in addressing team tasks. This in turn can provide increases in creativity, the processing of information, and overall team effectiveness (Morgan and Lassiter 1992). However, the unique aspects associated with larger teams may do more harm than good due to the fact that the increase in size may also pose problems. Various studies (Indik 1965; Gerard et al. 1968) have found that problems in communication and level of participation may increase due to increases in team size. These two issues, when combined, may diminish team performance by placing increased stress on a team in the areas of coordination and communication workload (Shaw 1976).

Task goal characteristics are another important contributor to team performance. Given a task that varies in difficulty, it has been found that goals that appear to be challenging and that target a specific task tend to have a higher motivational impact than those that are more easily obtainable (Ilgen et al. 1987; Gladstein 1984; Goodman 1986; Steiner 1972). According to Ilgen et al. (1987), motivation directs how much time an individual is willing to put forth in accomplishing a team task.

The use of feedback is another factor that influences team motivation and performance. Feedback helps to motivate an individual concerning recently performed tasks (Salas et al. 1992). Researchers concur that the feedback should be given within a short period of time after the relevant performance (Dyer 1984; Nieva et al. 1978). The sequence and presentation of the feedback may also play significant roles in motivation. Salas et al. (1992) have noted that during the early stages of training, feedback should concentrate on *one* aspect of performance; however, later in time it should concentrate on *several* training components. It has been found that sequencing helps team training in that teams adjust to incorporate their feedback into the task(s) at hand (Briggs and Johnston 1967).

According to Salas and Cannon-Bowers (1995), a variety of methods and approaches are currently under development for use in building effective team training programs. One such technique is the developing technology of team task analysis (Levine and Baker 1991), a technique that, as it matures, may greatly facilitate the development of effective team training. The technology provides a means to distinguish team learning objectives for effective performance from individual objectives and is seen as a potential aid in identifying the teamwork-oriented behaviors (i.e., cues, events, actions) necessary for developing effective team training programs.

A second area is team-performance measurement. To be effective, team-performance measurement must assess the effectiveness of various teamwork components (such as team-member interactions) as well as intrateam cognitive and knowledge activities (such as decision making and communication) in the context of assessing overall performance of team tasks. Hall et al. (1994) have suggested the need for integrating team performance outcomes into any teamwork-measurement system. As team feedback is provided, it is also necessary to estimate the extent to which an individual team member is capable of performing his or her specific tasks within the team. Thus, any competently developed team performance-measurement system must be capable of addressing both team capabilities and individual capabilities separately and in combination.

Teamwork simulation exercises are a third developing technology cited by Salas and Cannon-Bowers (1995). The intent of such simulations is to provide trainees with direct behavioral cues designed to trigger competent teamwork behaviors. Essential components of such simulations include detailed scenarios or exercises where specific teamwork learning objectives are operationalized and incorporated into training.

Salas and Cannon-Bowers (1995) have also identified three generic training *methods* for use with teams; information based, demonstration based, and practice based. The information-based method involves the presentation of facts and knowledge via the use of such standard delivery techniques as lectures, discussions, and overheads. Using such group-based delivery methods, one can deliver relevant information simultaneously to large numbers of individuals. The methods are easy to use and costs are usually low. Information-based methods may be employed in many areas, such as helping teammates understand what is expected of them, what to look for in specific situations, and how and when to exchange knowledge.

Demonstration-based methods are performed, rather than presented, as are information-based methods. They offer students an opportunity to observe behaviors of experienced team members and thus of the behaviors expected of them. Such methods help to provide shared mental models among team members, as well as examples of how one is expected to handle oneself within complex, dynamic, and multifaceted situations (Salas and Cannon-Bowers 1995).

Practice-based methods are implemented via participatory activities such as role playing, modeling, and simulation techniques. These methods provide opportunities to practice specific learning objectives and receive feedback information. With such methods, trainees can build upon previous practice attempts until achieving the desired level(s) of success.

A final category of developmental teamwork technologies, according to Salas and Cannon-Bowers (1995), involves teamwork training implementation strategies. These authors discuss two such strategies: cross-training and team coordination training. In their view, cross-training, training all team members to understand and perform each other's tasks, is an important strategy for integrating inexperienced members into experienced groups. Team coordination training addresses the issue that each team member has specific duties and that those duties, when performed together, are designed to provide a coordinated output. Team coordination training involves the use of specific task-oriented strategies to implement coordination activities and has been successfully applied in the aviation and medical fields.

3.8. Training Evaluation

Training evaluation may serve a variety of purposes, including improving training and assessing the benefits of training. Formative evaluation consists of processes conducted while training is being developed. Summative evaluation consists of processes conducted after training has been implemented.

Because training typically does not remain static (a job's tasks are often modified, replaced, or added), the distinction between formative evaluation and summative evaluation is sometimes blurred (see Goldstein 1993). Formative evaluation focuses on improving training prior to implementation. It may also serve the purpose of assessing benefits of training in a preliminary way. Summative evaluation focuses on assessing the benefits of training to determine how training outcomes improve on-job performance. It may also serve the purpose of improving future training iterations. In terms of instructional systems design, summative evaluation completes the cycle that began with analysis. Figure 1 shows a simplified version of this cycle.

Both formative and summative aspects of evaluation feed back into the training analysis phase, which showed why training was necessary, the performances that training presumably would change, and the job outcomes that would be affected. Formative evaluation typically includes internal quality checks (such as review by subject matter experts, review against instructional standards, and editing for organization and clarity), as well as pilot testing. Formative evaluation may occur throughout the process shown in Figure 1. Summative evaluation includes assessments of after-training outcomes, both immediately following the conclusion of a training program and at later intervals.

Kirkpatrick (1994) has developed a widely used classification taxonomy of evaluation levels. Table 2 shows these levels and how they apply to both formative and summative evaluation.

Note that pilot testing conducted during formative evaluation should address, at a minimum, levels 1 and 2 of Kirkpatrick's taxonomy. Summative evaluation should address all four levels. Kirkpatrick (1994) has suggested that all four levels must be addressed for evaluation to be useful to organizations. Although using the level 1 evaluation standard as a measure of satisfaction can be useful, using it as an indicator of training effectiveness is misleading. Clark (1982), for instance, showed that about one-third of the time, a negative correlation exists between a learner's end-of-course self-rating and measures of learning.

Level 2 evaluation is obviously beneficial for determining whether training produces the learning it is intended to produce. As discussed above, we recommend criterion-referenced testing as the appropriate source of level 2 data for training (as opposed to education) evaluation. End-of-course criterion-referenced tests are designed to show whether or not trainees can perform adequately on measures of each course objective. In criterion-referenced measurement, the operative question is, "Can trainees perform the required task or not?" Contrast this with norm-referenced testing, which addresses the question, "How do learners' levels of knowledge compare?" In practice, a norm-referenced test could show a percentage of students excelling, while a criterion-referenced test on the same information could show all students as failing to meet certain training standards.

Conducted with a criterion-referenced approach, level 2 evaluation shows what trainees can do at the end of a course, but not what they will do on the job. There are several reasons for this, as noted above. First, there is the issue of retention over time. If learners do not have the opportunity to practice what they have learned in training, retention will decrease. Consequently, for training to be effective, it should be conducted close to the time when trainees need to use what they have learned. Transfer issues, as previously described, may also affect actual job performance and should be carefully considered in training design.

Another reason that level 2 evaluation may not predict how trainees will perform later in time is that outside factors may inhibit application of what is learned. As noted previously, organizations may lack appropriate motivational systems and environmental support for specific behaviors. Thus, if trainees are taught to perform behavior *y* in response to stimulus *x* in training but are required to perform behavior *z* in response to that stimulus on the job, the behavior learned in training will soon be extinguished. Similarly, if employees are taught to use certain tools during training but those tools are absent from the work environment, the tool-related behaviors learned during training will not be used.

Kirkpatrick's level 3 evaluation seeks to measure changes in behavior after training. Because factors other than training may influence later behaviors, measuring these changes successfully may

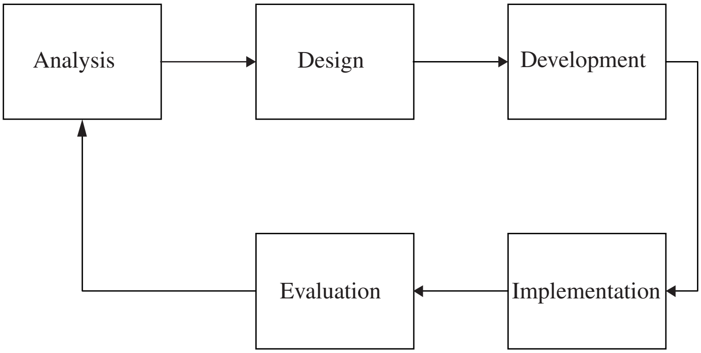


Figure 1 Evaluation's Role in an ISD Model.

TABLE 2 Kirkpatrick's Levels Applied to Formative and Summative Evaluation

	Type of Evaluation	
Level of Evaluation	Formative	Summative
1. **Reaction:** Participants' reaction to course (customer satisfaction).	"Dog food principle": Even if it's very nutritious, if the dogs don't like it, they won't eat it. Make sure participants are satisfied with the course.	Reactions can be used to distinguish between instructors, setting, audience, and other factors (e.g., are ratings for some instructors consistently higher than for others?).
2. **Learning:** Change measured at end of course.	Does the course meet its learning objectives? If it doesn't, revise and repilot test until it does.	Use end-of-course measures to determine needs for remediation and for improving future iterations of course.
3. **Behavior:** Change measured on the job.	Typically not included in formative evaluation.	If the course doesn't change behavior on the job, the investment in development has been wasted. Analyze causes for lack of change. Modify course, motivational factors, and/or environmental factors as necessary.
4. **Results:** Impact of change on job outcomes.	Typically not included in formative evaluation.	If the behavior change occurs but does not improve the job outcomes it targeted, the training may be flawed and/or external factors may be affecting the outcome. Reanalyze needs.

require an experiment. Table 3 shows an example of an experimental design permitting a level 3 evaluation.

Using the design shown in Table 3, training has been randomly assigned to group B. Baseline performance of both groups is then measured at the same time, before group B receives training. If the assignment is truly random, there should be no significant differences between the performance of trainees in groups A_1 and B_1. At the same point after training, job performance of both groups is again measured. If all other factors are equal (as they should be in random assignment), differences between groups A_2 and B_2 will show the effect of training on job behavior. Depending on the type of measurement, one may use various statistical indices, such as a chi squared test or phi coefficient as a test of statistical significance (see Swezey 1981 for a discussion of this issue.)

This design does not use level 2 evaluation data. However, by modifying the design so that measures of A_1 and B_1 behaviors occur at the end of training (rather than prior to training) and measures of A_2 and B_2 occur later (e.g., at some interval after training), one can use level 2 evaluation data to assess the relative effect of nontraining factors on job performance. In this case (level 2 evaluation), B_1 performance should be significantly higher than A_1 performance. But B_2 performance may not be significantly higher than A_2 performance, depending upon whether nontraining factors inhibit job performance. Again, statistical comparisons are used to determine significance of the effect(s).

TABLE 3 Experimental Design for Level 3 Evaluation

	Experimental Condition	
Points of Measurement	No Training	Training
Before Training	A_1	B_1
After Training	A_2	B_2

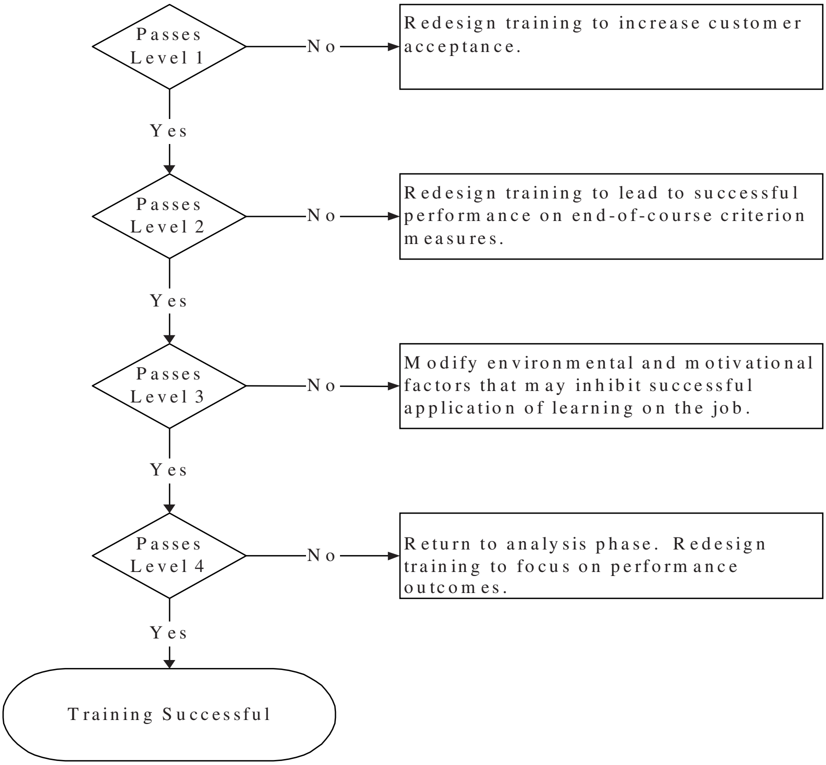

Passes Level 1
No
Redesign training to increase customer acceptance.
Yes
Passes Level 2
No
Redesign training to lead to successful performance on end-of-course criterion measures.
Yes
Passes Level 3
No
Modify environmental and motivational factors that may inhibit successful application of learning on the job.
Yes
Passes Level 4
No
Return to analysis phase. Redesign training to focus on performance outcomes.
Yes
Training Successful

Figure 2 Remedying Failures at Each Level of Evaluation.

Kirkpatrick's level 4 evaluation addresses the basic question, "So what?" Even if evaluation at levels 1 through 3 show positive results, if training does not lead to desired results on job performance, it may not be the proper solution or worth the expense. For example, if training is designed to lead to performance that reduces error rates, and these error rates do not decline adequately, training may not have been the correct solution to the problem situation. Figure 2 provides a brief overview of how to address failure at each of Kirkpatrick's four levels of evaluation.

4. DEVELOPMENT

The distinction between training and development is ill defined. For purposes of this chapter, we define development in organizational contexts as "Plans and activities intended to achieve long-term changes in employees and their organizations." As such, development may include education, training, organizational realignment, values clarification, modifications to feedback and reward systems, mentoring, and a host of other interventions. In this section, we briefly mention a few forms of development in terms of research results.

4.1. Individual Performance Enhancement

In the 1940s and 1950s, much research on individual performance enhancement focused on personality development. By the 1960s, however, it had become clear that a wide variety of personalities could function successfully in nearly all jobs. One of the most widely used psychometric tools in business and industry, the Myers–Briggs Type Indicator (MBTI), purports to measure personality types. However, both Druckman and Bjork (1991) and Pittenger (1993) showed that data obtained via the MBTI have very little predictive validity in terms of job performance. Although research into the predictive value of personality testing continues today, it remains questionable whether development that focuses on personality traits as predictors will lead to significant results with respect to job performance.

Recently, the issue of competencies has also become popular in training circles. Most research on competencies can be divided into two types:

1. Those that address specific job skills, e.g., "can calibrate a variety of instruments used to measure tensile strength"
2. Those that address traits, (sometimes disguised in behavioral terms), e.g., "demonstrates the ability to work well with others in a variety of situations"

Although training may be an appropriate development activity for the first type of competency, the second may not benefit from *any* development activity. For development activities to be useful, their outcomes must be clearly specified. If "works well with others" is defined in operational terms (e.g., others report that the individual contributes well in team situations), then development activities, such as training, coaching, and mentoring, may be appropriate.

A variety of experiential forms of training are in use as development activities. These range from wilderness survival courses to games and specially designed on-the-job training experiences that focus on an individual's performance. Spitzer (1986) reviews the preparation, conduct, and follow up of a variety of specially designed on-the-job training experiences. Most such experiential approaches have not been validated via careful research. Thiagarajan (1980) also summarizes advantages and disadvantages of games and other forms of experiential learning. He notes that, while experiential learning is often shown to produce the same results as more conventional forms of training, it occasionally leads to poorer skill acquisition. However, he argues that because such experiential techniques are often fun, they may gain trainee acceptance and therefore be a good vehicle for individual skill enhancement. We tend to disagree with this view.

Coaching and mentoring are other methods of performance enhancement used in development situations. In coaching, a coach works with a student during one or more sessions to help the student develop and apply skills on the job. Pearlstein and Pearlstein (1991) present an ISD-based approach to coaching, and Fournies (1987) describes a wide variety of contexts for coaching. Generally, mentoring is a long-term relationship between an experienced and a less-experienced individual that is designed to bring about changes in behavior via counseling, feedback, and a variety of other techniques. Murray (1991) provides an overview of factors and techniques contributing to effective mentoring.

Career planning as a development activity can also be an effective means of enhancing performance. In career planning, individuals and their supervisors and/or human resources departments develop and plan a series of steps to help align individual goals with organizational goals and career progression. These steps may include training and education (either on or off the job), coaching, mentoring, special assignments, and job rotations. Schein (1978) provide a detailed model for career planning. While most employee-assessment systems appear to be a sound basis for individual performance enhancement, their unsuccessful implementation may work against job performance. According to Scherkenbach (1988), Deming's total quality management (TQM) process argues against such systems precisely because poor implementation can cause major roadblocks to success of both individuals and organizations. Since most employee assessment systems are linked to pay decisions, employees who admit to development needs may, in effect, be arguing against raises. This link thus creates both unsuccessful implementations and possible conflicts of interest.

4.2. Organizational Performance Enhancement

Two commonly practiced forms of organizational performance enhancement are management and leadership development and organization development. Management and leadership development focuses on improving organizational outcomes by improving the performance of those responsible for the outcomes. Organization development focuses on developing a total organization, as opposed to some particular aspect(s) of the organization, such as its leadership, management, or employee skills (see Gallessich 1982; Vaill 1971). Pearlstein (1991) argues that leadership development cannot be separated from organization development because leaders cannot be developed outside of the context of organization-wide goals for development.

4.2.1. Management and Leadership Development

Stodgill (1974), in a summary of over 3000 books and articles, found hundreds of definitions of leaders and managers. Some of these define management as focusing on support of organizational activities in their current state, as contrasted with leadership, which focuses on changing an organization to meet new challenges. In this context, a distinction is often made based upon topic areas, with management focusing on basics such as making assignments, delegating, and feedback and leadership focusing on more advanced areas such as communicating vision, establishing shared values, and individual employee empowerment. However, in the context of current organizational challenges, arbitrary distinctions between management and leadership are increasingly less relevant. Vaill (1989) asserts that organizations often find themselves, like kayakers, in "permanent white water." This is even more the case now, with today's greater impacts of increasing globalization and rapidly accelerating technology.

Pearlstein (1992) notes that "Research (on leadership and management) . . . does not show conclusively that any leadership development training approach based on skill-building has led to long-term improvements in organizational functioning. Research does not show conclusively that any leadership development approach based on personal/awareness development or values clarification has led to long-term organizational improvements." Howard and Bray (1988) summarize the research of over 30 years of research on longitudinal management development studies. Complex and surprising results are the rule. For example, interpersonal skills, considered by many assessment centers to be one of the most critical predictors of management success, decline over time for many managers. Worse, managers selected in the 1980s had lower levels of interpersonal skills than those selected in the 1950s. Similarly, Kotter (1982) reports on intensive studies of 15 general managers and concludes that training does not seem as important to leaders' development as do childhood and early work experiences. McCall et al. (1988) summarize the responses of nearly 400 corporate leaders on factors important to their development and note that leaders cited a 12:1 ratio of nontraining to training factors. Finally, Pearlstein (1991) indicates that management and leadership development programs are usually generic, offered out of the context of specific management and leadership jobs, do not last long enough, and focus either on skills or on personal development but not on both.

4.2.2. Organization Development

Organization development (OD) is practiced in many large organizations today. OD practitioners use a form of experimentation called action research, which includes five major phases:

1. Reaching agreement on goals and procedures of the OD activity
2. Designing ways to diagnose needs for organizational change
3. Presenting diagnostic findings in a useful way
4. Using the diagnostic findings to plan an organizational change strategy.
5. Applying the change strategy and evaluating its outcomes

Both Lippitt et al. (1958) and Argyris (1970) provide examples of using this phased approach. The examples have in common behavioral approaches to collecting the data on which to base the change strategies. Pearlstein (1997) provides nine OD principles for those who wish to practice within a human performance technology context:

1. Develop work units (not necessarily the entire organization).
2. Make sure that you are interacting with the appropriate decision maker.
3. Start by reaching agreement on desired outcomes, measurements, roles, and activities.
4. Insist on behavioral measurement, but first listen carefully to the client's statements of values and visions.
5. Make sure that the client considers the impacts of change on individuals, work units, the entire organization, and the larger community of which the organization is a part.
6. Meet with all key players, both individually and in groups.
7. Seek consensus but don't withhold expertise.
8. Give interventions sufficient time to work.
9. Close well with the client by helping to: (a) collect and disseminate summative evaluation data, (b) celebrate organizational successes, and (c) plan next steps.

5. THE FUTURE OF SELECTION, TRAINING, AND DEVELOPMENT

We do not claim to hold a crystal ball; however, some aspects of selection, training, and development are likely to stay the same, while others are likely to change. In our view, those that will stay the same are based on well-established research results and apply independently of increasing globalization and accelerating technological development. Those that will change lack these attributes.

5.1. Selection

Selection tools will likely remain focused on measuring past and current behavior. Some tools that rely on assessing current performance may be slow to change because of difficulties associated with simulating tasks remotely. Many tools will become more widely used over the Internet. That is, organizations will increasingly use the Internet as a means for obtaining information previously gained via application forms, paper-based tests, interviews, and other selection devices. For example, many organizations now use Internet-based application forms that feed candidates' responses in standard formats directly into databases, which are then mined to match candidates' qualifications with job requirements. Similarly, because of globalization, it is likely that selection tools such as application forms, tests, and interviews will also become increasingly available in multilingual and multicultural versions. Further, simulations and work sample assessments will be increasingly used, also in different culturally oriented and language-based forms.

5.2. Training

The basic development process for training is likely to remain some form of systems-oriented instructional model, although cognitively oriented technologies will increasingly be developed. The well-researched ISD approach has been adapted for use in many cultures and leads to reliable outcomes regardless of the specific type of technology employed. As cognitive approaches become better able to predict factors that build expertise, they will be more frequently used in task analysis. Results may be used to design training that is more useful for "far transfer" situations—that is, to tasks relatively dissimilar to those used as examples during training (see Clark 1992). Still, results of cognitive analysis are likely to be used within the context of a systems-oriented instructional model.

Training delivery and administration systems are very likely candidates for change. Computer-based training (CBT), now widely available in CD-ROM format, is increasingly more available in other disk-based formats (e.g., DVD) and on the Internet as Web-based training (WBT). Limitations on processing power, memory, storage, and bandwidth are all decreasing rapidly. This is leading to an increased ability for CBT, regardless of its mode of delivery, to offer high levels of simulation. Streaming audio and streaming video, voice recognition, and other technological advances are rapidly improving simulation fidelity. Similarly, advances in manipulanda (e.g., tactile gloves, aircraft yokes) are also increasing simulation fidelity. The result is that training delivered to remote sites can use both simulation and work sample presentation more effectively.

Other distance learning approaches, such as virtual classrooms, are being used increasingly. Both synchronous (real-time classroom) and asynchronous (classes via forums and message boards) types of distance learning will become widely used. For example, it is likely that organizations will increasingly incorporate university courses directly into work environments. This can be done globally with, for example, Saudi managers taking economics courses offered by Japanese universities.

There are negative aspects to the rapid spread of distance learning techniques. Often media values are emphasized at the expense of sound instructional practice. For example, much CBT available today does not make use of the technology for branching to different points in instruction based upon learner responses. Although the technology for branching has existed since the 1960s (see, e.g., Markle 1969), much CBT available today is linear. Although some argue that learner control (the ability to move through lessons at will) is more important than programmed control, research findings suggest otherwise. Clark (1989), for instance, showed that given a choice of approach (structured or unstructured), CBT users were more likely to choose the alternative least useful for skill acquisition. Novice learners, who may have benefited from a structured approach, often chose an unstructured one, while experts, who could use their advanced ability to explore advantageously, were more likely to choose structured approaches. Similarly, the advance of virtual classrooms in corporate settings may lead to an emphasis on education, as opposed to training. Thus, employees may learn more generalities about topics, at the expense of learning how to perform tasks.

One recent development in training has been slow to evolve. Electronic performance support systems (EPSS), which combine online help, granular training, and expert advice functions, were popularized by Gery (1991). Although used in some corporations today, EPSS use does not seem to be accelerating, possibly because development costs are relatively high. In an EPSS, help is directly available on employee computers. In addition to highly contextual online help, brief, focused training modules are offered when users repeatedly request help on the same topics. Such systems use artificial intelligence techniques, such as algorithms based upon data derived from users' experiences, to offer expert advice. Thus, a user who wanted to advise a consumer on what lumber to purchase for a deck could select an expert advisor function to input the consumer's space requirements and house dimensions, and the computer could immediately provide appropriate sketches and lumber requirements. Such systems can dramatically reduce amounts of time spent in formal training.

5.3. Development

Like training, approaches for individual, management, and leadership development are likely to be offered increasingly on the Internet via virtual classrooms. Since many organizations continue to spend a large amount on development activities, there is little reason to suppose that they will discontinue such expenditures.

Acknowledgments

Portions of this chapter were adapted from Swezey and Llaneras (1997). The authors would like to thank Gloria Pearlstein and Lisa Dyson for their help with the preparation of this chapter.

REFERENCES

Alessi, S. M. (1988), "Fidelity in the Design of Instructional Simulations," *Journal of Computer-Based Instruction,* Vol. 15, No. 2, pp. 40–47.

Allen, W. H. (1973), "Research in Educational Media," in *Educational Media Yearbook,* J. W. Brown, Ed., R.R. Bowker, New York.

Alluisi, E. A. (1991), "The Development of Technology for Collective Training: SIMNET, a Case History," *Human Factors,* Vol. 33, No. 3, 343–362.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999), *Standards for Educational and Psychological Testing,* American Psychological Association, Washington, DC.

American Educational Research Association (1999), *Standards for Educational and Psychological Testing,* American Psychological Association, Washington, DC.

Anderson, J. R. (1982), "Acquisition of Cognitive Skill," *Psychological Review,* Vol. 89, pp. 369–406.

Anderson, J. R. (1990), *Cognitive Psychology and Its Implications,* 3rd Ed., W.H. Freeman, New York.

Andrews, D. H., Waag, W. L., and Bell, H. H. (1992), "Training Technologies Applied to Team Training: Military Examples," in *Teams: Their Training and Performance,* R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ.

Argyris, C. (1970), *Intervention Theory and Method: A Behavioral Science View,* Addison-Wesley, Reading, MA.

Bahrick, H. P. (1984), "Semantic Memory Content in Permastore: Fifty Years of Memory for Spanish Learned in School," *Journal of Experimental Psychology: General,* Vol. 113, pp. 1–29.

Baldwin, T. T., and Ford, J. K. (1988), "Transfer-of-Training: A Review and Directions for Future Research," *Personal Psychology,* Vol. 41, pp. 63–105.

Battig, W. F. (1966), "Facilitation and Interference," in *Acquisition of Skill,* E. A. Bilodeau, Ed., Academic Press, New York.

Bloom, K. C., and Shuell, T. J. (1981), "Effects of Massed and Distributed Practice on the Learning and Retention of Second-Language Vocabulary," *Journal of Educational Research,* Vol. 74, No. 4, pp. 245–248.

Bouzid, N., and Crawshaw, C. M. (1987), "Masses Versus Distributed Word Processor Training," *Applied Ergonomics,* Vol. 18, pp. 220–222.

Branson, R. K., Rayner, G. T., Coxx, J. L., Furman, J. P., King, F. J., and Hannum, W. J. (1975), "Interservice procedures for instructional systems development: Executive summary and model," ADA019486, U.S. Army Combat Arms Training Board, Fort Benning, GA.

Briggs, G. E., and Johnston, W. A. (1967), "Team Training," NAVTRADEVCEN-1327-1, AD-608 309, Naval Training Device Center, Port Washington, NY.

Campbell, J. (1990), "An Overview of the Army Selection and Classification Project (Project A)," *Personnel Psychology,* Vol. 43, pp. 231–239.

Cannon-Bowers, J. A., and Salas, E. (1990), "Cognitive Psychology and Team Training: Shared Mental Models in Complex Systems," Paper presented at the 5th Annual Conference of the Society for Industrial and Organizational Psychology, K. (Miami).

Carlisle, K. E. (1986), *Analyzing Jobs and Tasks*. Educational Technology Publications, Englewood Cliffs, NJ.

Clark, R. E. (1982), "Antagonism between Achievement and Enjoyment in ATI Studies," *Educational Psychology,* Vol. 17, No. 2, pp. 92–101.

Clark, R. E. (1983), "Reconsidering Research on Learning from Media," *Review of Educational Research,* Vol. 53, pp. 445–459.

Clark, R. E. (1989), "When Teaching Kills Learning: Research on Mathematics," in *Learning and Instruction: European Research in an International Context,* Mandl, H. N., Bennett, N., de Corte, E., and Freidrich, H. F., Eds., Vol. 2, Pergamon Press, London.

Clark, R. E. (1992), "How the Cognitive Sciences are Shaping the Profession," in *Handbook of Human Performance Technology,* Stolovitch, H. D., and Keeps, E. J., Eds., Jossey-Bass, San Francisco.

Clark, R. E. (1994), "Media Will Never Influence Learning," *Educational Technology Research and Development,* Vol. 42, No. 3, pp. 21–29.

Cofer, C. N., Bruce, D. R., and Reicher, G. M. (1966), "Clustering in Free Recall as a Function of Certain Methodological Variations," *Journal of Experimental Psychology,* Vol. 71, pp. 858–866.

Cohen, P. A., Ebeling, B. J., and Kulik, J. A. (1981), "A Meta-analysis of Outcome Studies of Visual-Based Instruction," *Educational Communication and Technology Journal,* Vol. 29, pp. 26–36.

Cormier, S. M. (1984), "Transfer-of-training: An interpretive review," Tech. Rep. No. 608, Army Research Institute, Alexandria, VA.

Craik, F. I. M., and Lockhart, R. S. (1972), "Levels of Processing: A Framework for Memory Research," *Journal of Verbal Learning and Verbal Behavior,* Vol. 11, pp. 671–684.

Druckman, D., and Bjork, R. A., Eds. (1991), *In the Mind's Eye: Enhancing Human Performance,* National Academy Press, Washington, DC.

Dwyer, D., Hall, J., Volpe, C., Cannon-Bowers, J. A., and Salas, E. (1992), "A Performance Assessment Task for Examining Tactical Decision Making under Stress," Tech. Rep. No. 92-002, U.S. Naval Training Systems, Center, Orlando, FL.

Dyer, J. L. (1984), "Review on Team Training and Performance: A State-of-the-Art Review," in *Human factors review,* F. A. Muckler, Ed., The Human Factors Society, Santa Monica, CA.

Ebbinghaus, H. (1913), *Memory: A Contribution to Experimental Psychology,* H. A. Ruger and C. E. Bussenius, Trans., Columbia University, New York (original work published 1885).

Farr, M. J. (1987), *The Long-Term Retention of Knowledge and Skills: A Cognitive and Instructional Perspective,* Springer-Verlag, New York.

Fletcher, J. D. (1990), "Effectiveness and Cost of Interactive Videodisc Instruction in Defense Training and Education," IDA Paper P-2372, Institute for Defense Analyses, Alexandria, VA.

Fournies, F. F. (1987), *Coaching for Improved Work Performance,* Liberty Hall Press, Blue Ridge Summit, PA.

Fraser, S. L., and Kroeck, K. G. (1989), "The Impact of Drug Screening on Selection Decisions," *Journal of Business Psychology,* Vol. 3, pp. 403–411.

Gagné, E. D. (1978), "Long-Term Retention of Information Following Learning from Prose, *Review of Educational Research,* Vol. 48, pp. 629–665.

Gallessich, J. (1982), *The Profession and Practice of Consultation,* Jossey-Bass, San Francisco.

Georgenson, D. L. (1982), "The Problem of Transfer Calls for Partnership," *Training and Development Journal,* Vol. 36, No. 10, pp. 75–78.

Gerard, H. B., Wilhelmy, R. A., and Conolley, E. S. (1968), "Conformity and Group Size," *Journal of Personality and Social Psychology,* Vol. 8, pp. 79–82.

Gery, G. J. (1991), *Electronic Performance Support Systems: How and Why to Remake the Workplace through the Strategic Application of Technology,* Weingarten, Boston.

Gilbert, T. F. (1978), *Human Competence: Engineering Worthy Performance,* McGraw-Hill, New York.

Gladstein, D. L. (1984), "Groups in Context: A Model of Task Group Effectiveness," *Administrative Science Quarterly,* Vol. 29, pp. 499–517.

Glaser, R. B. (1963), "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," *American Psychologist,* Vol. 18, pp. 519–521.

Goldstein, I. L. (1993), *Training in Organizations,* 3rd Ed., Brooks/Cole, Pacific Grove, CA.

Goodman, P. S., Ed. (1986), *Designing Effective Work Groups,* Jossey-Bass, San Francisco.

Guion, R. M. (1991), "Personnel Assessment, Selection, and Placement," in *Handbook of Industrial and Organizational Psychology,* M. D. Dunnette and L. M. Hough, Eds., Consulting Psychologists Press, Palo Alto, CA, pp. 327–397.

Hall, J. K., Dwyer, D. J., Cannon-Bowers, J. A., Salas, E., and Volpe, C. E. (1994). "Toward Assessing Team Tactical Decision Making under Stress: The Development of a Methodology for Structuring Team Training Scenarios," in *Proceedings of the 15th Annual Interservice/Industry Training Systems and Education Conference* (Washington, DC), pp. 87–98.

Hammer, E. G., and Kleiman, L. A. (1988), "Getting to Know You," *Personnel Administration,* Vol. 34, pp. 86–92.

Harless, J. H. (1975), *An Ounce of Analysis* (*Is Worth a Pound of Objectives*), Harless Performance Guild, Newman, GA.

Hays, R. T. (1980), "Simulator Fidelity: A Concept Paper," Tech. Rep. No. 490, Army Research Institute, Alexandria, VA.

Hays, R. T., and Singer, M. J. (1989), *Simulation Fidelity in Training System Design,* Springer, New York.

Hendriksen, G., and Schroeder, W. H. (1941), "Transfer-of-Training in Learning to Hit a Submerged Target," *Journal of Educational Psychology,* Vol. 32, pp. 205–213 (cited in Hilgard 1962).

Herzberg, F. (1974), "The Wise Old Turk," *Harvard Business Review,* Vol. 52, September–October, pp. 70–80.

Hilgard, E. R. (1962), *Introduction to Psychology,* 3d Ed., Harcourt Brace & World.

Hoban, C. F., and Van Ormer, E. B. (1950), "Instructional Film Research, 1918–1950," Tech. Rep. No. SDC 269-7-19, U.S. Naval Special Devices Center, Port Washington, NY (ERIC Document Reproduction Service No. ED 647 255).

Hollenbeck, J. R., Sego, D. J., Ilgen, D. R., and Major, D. A. (1991), "Team Interactive Decision Exercise for Teams Incorporating Distributed Expertise (TIDE^2): A Program and Paradigm for Team Research," Tech. Rep. No. 91-1, Michigan Sate University, East Lansing, MI.

Howard, A., and Bray, D. (1988), *Managerial Lives in Transition: Advancing Age and Changing Times,* Guilford Press, New York.

Hunter, J. E. (1986), "Cognitive Ability, Cognitive Aptitudes, Job Knowledge, and Job Performance," *Journal of Vocational Behavior,* Vol. 29, pp. 340–62.

Hurlock, R. E., and Montague, W. E. (1982), "Skill Retention and Its Primary Implications for Navy Tasks: An Analytical Review," NPRDC SR 82-21, Navy Personnel Research & Development Center, San Diego.

Ilgen, D. R., Shapiro, J., Salas, E., and Weiss, H. (1987), "Functions of Group Goals: Possible Generalizations from Individuals to Groups," Tech. Rep. No. 87-022, Naval Training Systems Center, Orlando, FL.

Indik, B. P. (1965), "Organizational Size and Member Participation: Some Empirical Test of Alternatives," *Human Relations,* Vol. 18, pp. 339–350.

Jacobs, J. W., Prince, C., Hays, R. T., and Salas, E. (1990), "A Meta-analysis of the Flight Simulator Training Research," NAVTRASYSCEN TR-89-006, Naval Training Systems Center, Orlando, FL.

Jarrard, S. W., and Wogalter, M. S. (1992), "Recognition of Non-studies Visual Depictions of Aircraft: Improvement by Distributed Presentation," in *Proceedings of the Human Factors Society 36th Annual Meeting* (Atlanta, October 12–16), pp. 1316–1320.

Jensen, A. R. (1986), "G: Artifact or Reality?" *Journal of Vocational Behavior,* Vol. 29, pp. 301–331.

Judd, C. H. (1908), "The Relation of Special Training and General Intelligence," *Educational Review,* Vol. 36, pp. 28–42.

Kennedy, R. S., Jones, M. B., and Harbeson, M. M. (1980), "Assessing Productivity and Well-Being in Navy Work-Places," in *Proceedings of the 13th Annual Meeting of the Human Factors Association of Canada* (Point Ideal, Lake of Bays, ON), Human Factors Association of Canada, Downsview, ON, pp. 8–13.

Kieras, D. E., and Bovair, S. (1984), "The Role of a Mental Model in Learning to Operate a Device," *Cognitive Science,* Vol. 8, pp. 255–273.

Kinkade, R. G., and Wheaton, G. R. (1972), "Training Device Design," in *Human Engineering Guide to Equipment Design,* H. P. Van Cott and R. G. Kinkade, Eds., U.S. Printing Office, Washington, DC.

Kirkpatrick, D. L. (1994), *Evaluating Training Programs: The Four Levels.* Barrett-Koehler, San Francisco.

Kleiman, L. S., and Faley, R. H. (1990), "A Comparative Analysis of the Empirical Validity of Past and Present-Oriented Biographical Items," *Journal of Business Psychology,* Vol. 4, pp. 431–437.

Kotter, J. (1982), *The General Managers,* Free Press, New York.

Kulik, J. A., and Kulik, C. (1987), "Review of Recent Research Literature on Computer-Based Instruction," *Contemporary Educational Psychology,* Vol. 12, pp. 222–230.

Kulik, J. A., Kulik, C., and Cohen, P. A. (1980), "Effectiveness of Computer-Based College Teaching: A Meta-analysis of Comparative Studies," *Review of Educational Research,* Vol. 50, pp. 525–544.

Kyllonen, P. C., and Alluisi, E. A. (1987), "Learning and Forgetting Facts and Skills," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York.

Larkin, J. (1981), "Enriching Formal Knowledge: A Model for Learning to Solve Textbook Physics Problems," in *Cognitive Skills and Their Acquisition,* J. R. Anderson, Ed., Erlbaum, Hillsdale, NJ.

Lawler, E. E., III, Hackman, J. R., and Kaufman, S. (1973), "Effects of Job Redesign: A Field Experiment," *Journal of Applied Social Psychology,* Vol. 3, No. 1, pp. 49–62.

Levine, E. L., and Baker, C. V. (1991), "Team Task Analysis: A Procedural Guide and Test of the Methodology," in *Methods and Tools for Understanding Teamwork: Research with Practical Implications,* E. Salas, Chair, Symposium presented at the 6th Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Lintern, G. (1987), "Flight Simulation Motion Systems Revisited," *Human Factors Society Bulletin,* Vol. 30, No. 12, pp. 1–3.

Lippitt, R., Watson, J., and Westley, B. (1958), *The Dynamics of Planned Change,* Harcourt, Brace & World, New York.

Loftus, G. R. (1985), "Evaluating Forgetting Curves," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* Vol. 11, pp. 397–406.

Luh, C. W. (1922), "The Conditions of Retention," *Psychological Monographs,* Whole No. 142, Vol. 31, No. 3.

Mager, R. F. (1972), *Goal Analysis,* Lear Siegler/Fearon, Belmont, CA.

Mager, R. F., and Pipe, P. (1970), *Analyzing Performance Problems or You Really Oughta Wanna,* Fearon, Belmont, CA.

Markle, S. M. (1969), *Good Frames and Bad: A Grammar of Frame Writing,* 2nd Ed., John Wiley & Sons, New York.

Martin, C. L., and Nagao, D. H. (1989), "Some Effects of Computerized Interviewing on Job Applicant Responses," *Journal of Applied Psychology,* Vol. 75, pp. 72–80.

McCall, M. W., Lombardo, M. M., and Morrison, A. M. (1988), *Lessons of Experience: How Successful Executives Develop on the Job,* Lexington Books, Lexington, MA.

McGeoch, J. A., and Irion, A. L. (1952), *The Psychology of Human Learning,* 2nd Ed., Longmans, Green & Co., New York.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. L., and Ashworth, S. (1990), "Project A Validity Results: The Relationship between Predictor and Criterion Domains," *Personnel Psychology,* Vol. 43, pp. 335–354.

Miller, R. B. (1954), *Psychological Considerations for the Design of Training Equipment,* American Institutes for Research, Pittsburgh.

Morgan, B. B., Jr., and Lassiter, D. L. (1992), "Team Composition and Staffing," in *Teams: Their Training and Performance,* R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ.

Morris, N. M., and Rouse, W. B. (1985), "Review and Evaluation of Empirical Research in Troubleshooting," *Human Factors,* Vol. 27, No. 5, pp. 503–530.

Mourier, P. (1998), "How to Implement Organizational Change That Produces Results," *Performance Improvement,* Vol. 37, No. 7, pp. 19–28.

Murphy, K. R., Thornton, G. C., and Reynolds, D. H. (1990), "College Students' Attitudes toward Employee Drug Testing Programs," *Personnel Psychology,* Vol. 43, pp. 615–31.

Murray, M. (1991), *Beyond the Myths and Magic of Mentoring: How to Facilitate an Effective Mentoring Program,* Jossey-Bass, San Francisco.

Nieva, V. F., Fleishman, E. A., and Rieck, A. (1978), "Team Dimensions: Their Identity, Their Measurement, and Their Relationships," Contract No. DAHC19-78-C-0001, Response Analysis Corporation, Washington, DC.

O'Hara, J. M. (1990), "The Retention of Skills Acquired through Simulator-Based Training," *Ergonomics,* Vol. 33, No. 9, pp. 1143–1153.

Parsons, H. M. (1980), *Aspects of a Research Program for Improving Training and Performance of Navy Teams,* Human Resources Research Organization, Alexandria, VA.

Parsons, P. J., Fogan, T., and Spear, N. E. (1973), "Short-Term Retention of Habituation in the Rat: A Developmental Study from Infancy to Old Age," *Journal of Comparative Psychological Psychology,* Vol. 84, pp. 545–553.

Pearlstein, G. B., and Pearlstein, R. B. (1991), "Helping Individuals Build Skills through ISD-Based Coaching," Paper presented at Association for Education and Communication Technology Annual Conference (Orlando, FL, March).

Pearlstein, R. B. (1991), "Who Empowers Leaders?" *Performance Improvement Quarterly,* Vol. 4, No. 4, pp. 12–20.

Pearlstein, R. B. (1992), "Leadership Basics: Behaviors and Beliefs," Paper presented at the 30th Annual Conference of the National Society for Performance and Instruction (Miami).

Pearlstein, R. B. (1997), "Organizational Development for Human Performance Technologists," in *The Guidebook for Performance Improvement: Working with Individuals and Organizations,* R. Kaufman, S. Thiagarajan, and P. MacGillis, Eds., Pfeiffer, San Francisco.

Peterson, R. O., and Duffany, B. H. (1975), "Job Enrichment and Redesign," in *The Training and Development Handbook by AT&T.* AT&T, New York.

Pittenger, D. J. (1993), "The Utility of the Myers-Briggs Type Indicator," *Review of Educational Research,* Vol. 63, No. 4, pp. 467–488.

Prediger, D. J. (1989), "Ability Differences across Occupations: More than g," *Journal of Vocational Behavior,* Vol. 34, pp. 1–27.

Prophet, W. W. (1976), "Long-Term Retention of Flying Skills: A Review of the Literature," HumRRO Final Report 76-35, ADA036977, Human Resources Research Organization, Alexandria, VA.

Provenmire, H. K., and Roscoe, S. N. (1971), "An Evaluation of Ground-Based Flight Trainers in Routine Primary Flight Training," *Human Factors,* Vol. 13, No. 2, pp. 109–116.

Rasmussen, J. (1979), "On the Structure of Knowledge: A Morphology of Mental Models in Man-Machine Systems Context," Report M-2192, Riso National Laboratory, Denmark.

Rea, C. P., and Modigliani, V. (1988), "Educational Implications of the Spacing Effect," in *Practical Aspects of Memory: Current Research and Issues,* Vol. 1, M. M. Gruenberg, P. E. Morris, and R. N. Sykes, Eds., John Wiley & Sons, Chichester, pp. 402–406.

Rossett, A. (1987), *Training Needs Assessment,* Educational Technology Publications, Englewood Cliffs, NJ.

Ruffner, J., Wick, W., and Bickley, W. (1984), "Retention of Helicopter Flight Skills: Is There a Critical Period for Proficiency Loss?" in *Proceedings of the Human Factors Society 28th Annual Meeting,* Human Factors Society, Santa Monica, CA.

Rumelhart, D. E., and Norman, D. A. (1978), "Accretion, Tuning and Restructuring: Three Modes of Learning," in *Semantic Factors in Cognition,* J. W. Cotton and R. L. Klatzky, Eds., Erlbaum, Hillsdale, NJ.

Rummler, G. A., and Brache, A. P. (1990), *Improving Performance: How to Manage the White Space on the Organization Chart,* Jossey-Bass, San Francisco.

Rummler, G. A., and Wilkins, C. L. (1999), "Performance Logic: Breaking the Code to Organization Performance," in *Proceedings of the Annual Conference of the International Association for Performance Improvement* (March), ISPI, Washington, DC.

Russell, C. J., Mattson, J., Devlin, S. E., and Atwater, D. (1990), "Predictive Validity of Biodata Items Generated from Retrospective Life Experience Essays," *Journal of Applied Psychology,* Vol. 75, pp. 569–580.

Sackett, P. R., and Decker, P. J. (1979), "Detection of Deception in the Employment Context: A Review and Critical Analysis," *Personnel Psychology,* Vol. 32, pp. 487–506.

Salas, E., and Cannon-Bowers, J. A. (1995), "Methods, Tools, and Strategies for Team Training," in *Training for 21st Century Technology: Applications of Psychological Research,* M. A. Quinones and A. Dutta, Eds., APA Press, Washington, DC.

Salas, E., Dickinson, T. L., Converse, S. A., and Tannenbaum, S. I. (1992), "Toward an Understanding of Team Performance and Training," in *Teams: Their Training and Performance,* R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ.

Salas, E., Cannon-Bowers, J. A., and Blickensderfer, E. L. (1995), "Team Performance and Training Research: Emerging Principles," *Journal of the Washington Academy of Sciences,* Vol. 83, No. 2, pp. 81–106.

Salmoni, A. W., Schmidt, R. A., and Walter, C. B. (1984), "Knowledge of Results and Motor Learning: A Review and Critical Appraisal," *Psychological Bulletin,* Vol. 95, pp. 355–386.

Schein, E. H. (1978), *Career Dynamics: Matching Individual and Organizational Needs,* Addison-Wesley, Reading, MA.

Scherkenbach, W. W. (1988), *The Deming Route to Quality and Productivity: Road Maps and Roadblocks,* Mercury Press, Rockville, MD.

Schmidt, R. A., and Bjork, R. A. (1992), "New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training," *Psychological Science,* Vol. 3, No. 4, pp. 207–217.

Schmidt, F., Ones, D., and Hunter, J. (1992), "Personnel Selection," *Annual Review of Psychology,* Vol. 43, pp. 627–670.

Schneider, V. I., Healy, A. F., Ericsson, K. A., and Bourne, L. E., Jr. (1995), "The Effects of Contextual Interference on the Acquisition and Retention of Logical Rules," in *Learning and Memory of Knowledge and Skills,* A. F. Healy and L. E. Bourne, Jr., Eds., Sage, Newbury Park, CA.

Schramm, W. (1977), *Big Media, Little Media,* Sage, Beverly Hills, CA.

Shaw, M. E. (1976), *Group Dynamics: The Psychology of Small Group Behavior,* McGraw-Hill, New York.

Shea, J. F., and Morgan, R. L. (1979), "Contextual Interference Effects on the Acquisition, Retention and Transfer of a Motor Skill," *Journal of Experimental Psychology: Human Learning and Memory,* Vol. 5, pp. 179–187.

Spitzer, D. R. (1986), *Improving Individual Performance,* Educational Technology Publications, Englewood Cliffs, NJ.

Steiner, I. D. (1972), *Group Process and Productivity,* Academic Press, New York.

Stodgill, R. M. (1974), *Handbook of Leadership,* Free Press, New York.

Streufert, S., and Swezey, R.W. (1985), "Simulation and Related Research Methods in Environmental Psychology," in *Advances in Environmental Psychology,* A. Baum and J. Singer, Eds., Erlbaum, Hillsdale, NJ, pp. 99–118.

Swezey, R. W. (1978), "Retention of Printed Materials and the Yerkes-Dodson Law," *Human Factors Society Bulletin,* Vol. 21, pp. 8–10.

Swezey, R. W. (1981), *Individual Performance Assessment: An Approach to Criterion-Referenced Test Development,* Reston, Reston, VA.

Swezey, R. W. (1989), "Generalization, Fidelity and Transfer-of-Training," *Human Factors Society Bulletin,* Vol. 32, No. 6, pp. 4–5.

Swezey, R. W., Hutcheson, T. D., Rohrer, M. W., Swezey, L. L., and Tirre, W. C. (1999), "Development of a Team Performance Assessment Device (TPAD): Final Report," InterScience America, Leesburg, VA (Report prepared under Contract No. F41624-97-C-5006 with the U.S. Air Force Research Laboratory, Brooks AFB, TX).

Swezey, R. W., and Llaneras, R. E. (1992), "Validation of an Aid for Selection of Instructional Media and Strategies," *Perceptual and Motor Skills,* Vol. 74, p. 35.

Swezey, R.W., and Llaneras, R.E. (1997), "Models in Training and Instruction," in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York. pp. 514–577.

Swezey, R. W., and Salas, E. (1992), "Guidelines for Use in Team Training Development," in *Teams, Their Training and Performance,* R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ.

Swezey, R. W., Perez, R. S., and Allen, J. A. (1991), "Effects of Instructional Strategy and Motion Presentation Conditions on the Acquisition and Transfer of Electromechanical Troubleshooting Skill," *Human Factors,* Vol. 33, No. 3, pp. 309–323.

Swezey, R. W., Meltzer, A. L., and Salas, E. (1994), "Issues Involved in Motivating Teams," in *Motivation: Theory and Research,* H. F. O'Neil, Jr. and M. Drillings, Eds., Erlbaum, Hillsdale, NJ. pp. 141–170.

Tannenbaum, S. I., and Yukl, G. (1992), "Training and Development in Work Organizations," *Annual Review of Psychology,* Vol. 43, pp. 399–441.

Teather, D. C. B., and Marchant, H. (1974), "Learning from Film with Particular Reference to the Effects of Cueing, Questioning, and Knowledge of Results," *Programmed Learning and Educational Technology,* Vol. 11, pp. 317–327.

Thiagarajan, S. (1980), *Experiential Learning Packages,* Educational Technology Publications, Englewood Cliffs, NJ.

Thorndike, R. L. (1986), "The Role of General Ability in Prediction," *Journal of Vocational Behavior,* Vol. 29, pp. 332–339.

Thorndike, E. L., and Woodworth, R. S. (1901), "The Influence of Improvement in One Mental Function upon the Efficiency of Other Functions," *Psychological Review,* Vol. 8, pp. 247–261.

Thorpe, J. A. (1987), "The New Technology of Large Scale Simulation Networking: Implications for Mastering the Art of Nonfighting," in *Proceedings of the 15th Annual Interservice/Industry Training Systems and Education Conference* (Washington, DC), pp. 87–98.

Travers, R. M. (1967), "Research and Theory Related to Audiovisual Information Transmission," U.S. Office of Education contract No. OES-16-006, Western Michigan University, Kalamazoo (ERIC Document Reproduction Service No. ED 081 245).

Tullar, W. L. (1989), "Relational Control in the Employment Interview," *Journal of Applied Psychology,* Vol. 74, pp. 971–977.

Tulving, E., and Thomson, D. M. (1973), "Encoding Specificity and Retrieval Processes in Episodic Memory," *Psychological Review,* Vol. 3, pp. 112–129.

Underwood, B. J. (1966), *Experimental Psychology,* 2nd Ed., Appleton-Century-Crofts, New York.

Underwood, B. J., and Keppel, G. (1963), "Retention as a Function of Degree of Learning and Letter-Sequencing Interference," *Psychological Monographs,* Vol. 77, Whole No. 567.

Vaill, P. B. (1971), "OD: A Grammatical Footnote," *Journal of Applied Behavioral Science,* Vol. 2, No. 2, p. 264.

Vaill, P. B. (1989), *Managing as a Performing Art,* Jossey-Bass, San Francisco.

Valverde, H. H. (1973), "A Review of Flight Simulator Transfer-of-Training Studies," *Human Factors,* Vol. 15, pp. 510–523.

Watkins, M. J., and Tulving, E. (1975), "Episodic Memory: When Recognition Fails," *Journal of Experimental Psychology, General,* Vol. 1, pp. 5–29.

Welford, A. T. (1968), *Fundamentals of Skill,* Methuen & Co, London.

Welsh, J. R., Watson, T. W., and Ree M. J. (1990), *Armed Services Vocational Aptitude Battery (ASVAB): Predicting Military Criteria from General and Specific Abilities,* Air Force Human Resources Laboratory, Brooks Air Force Base, TX.

Weaver, J. L., Morgan, B. B., Jr., Hall, J., and Compton, D. (1993), "Team Decision Making in the Command Information Center: Development of a Low-Fidelity Team Decision Making Task for Assessing the Effects of Teamwork Stressors," Technical Report submitted to the Naval Training Systems Center, Orlando, FL.

Weaver, J. L., Bowers, C. A., Salas, E., and Cannon-Bowers, J. A. (1995), "Networked Simulations: New Paradigms for Team Performance Research," *Behavior Research Methods, Instruments, and Computers,* Vol. 27, No. 1, pp. 12–24.

Wightman, D. C., and Sistrunk, F. (1987), "Part-Task Training Strategies in Simulated Carrier Landing Final-Approach Training," *Human Factors,* Vol. 29, pp. 245–254.

Wixted, J. T., and Ebbesen, E. B. (1991), "On the Form of Forgetting," *Psychological Science,* Vol. 2, pp. 409–415.

Zemke, R., and Kramlinger, T. (1982), *Figuring Things Out: A Trainer's Guide to Needs and Task Analysis,* Addison-Wesley, Reading, MA.

CHAPTER 36

Aligning Technological and Organizational Change

Ann Majchrzak
Najmedin Meshkati
University of Southern California

1. INTRODUCTION 949

2. WHY THE TOPIC IS CRITICAL TO INDUSTRIAL ENGINEERS 949
  2.1. Failures of Implementation of New Technology 949
  2.2. Why These High Failure Rates? 950

3. WHAT ARE THE STUMBLING BLOCKS TO ALIGNMENT? 952
  3.1. The Future of Technology Is Probabilistic 952
  3.2. Some Factors Are Less Malleable Than Others 952
  3.3. Alignment Requires a Cross-Functional Definition of the Problem 953
  3.4. Alignment Is Context Specific and Nonrepeatable 953
  3.5. Alignment Requires Comprehensive Solutions That Are Difficult to Identify and Realize 953
  3.6. Alignment Involves Long Planning Cycles, Where Observable Results and Knowing Whether You Made the Right Decisions Take Awhile 954
  3.7. Alignment Involves Compromises 954

4. HOW CAN TECHNOLOGY PLANNERS PURSUE ALIGNMENT DESPITE THESE DIFFICULTIES? 954
  4.1. Focus Alignment on Business Purpose, Driven by Competitive Need, Not as a Technology Fix to a Localized Problem 955
  4.2. Recognize the Breadth of Factors and Their Relationships That Must Be Designed to Achieve Alignment 955
  4.3. Understand the Role of Cultures in Alignment 955
  4.4. Make Organization and Technology Design Choices That Encourage Innovation 961
    4.4.1. Solutions That Enhance, Not Deskill Workers 962
    4.4.2. Solutions Should Be Human Centered 962
    4.4.3. Solutions That Integrate Across Processes, Not Bifurcate 963
    4.4.4. Solutions That Encourage Knowledge Recognition, Reuse, and Renewal 963
    4.4.5. Solutions Should Decentralize Continuous Improvement 963
  4.5. Agree on a Change Process for Achieving Alignment 963
  4.6. Use Decision Aids to Enhance Internal Understanding of Technology-Organizational Alignment 965
    4.6.1. TOP Modeler 965
    4.6.2. iCollaboration Tool for Technology-Organizational Realignment 966

5. CONCLUSIONS 968

REFERENCES 969

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

The installation of new technology, especially technology involving computers or microprocessors, virtually always involves some change to the organization and its members. Thus, the effective management of technological change must include the effective management of organizational change as well.

In this chapter, conclusions are presented from the literature and recent work by the authors concerning effective management of simultaneous change in technology and organizational design. The objective of this chapter is to impart to the practicing engineer the following four points:

1. There are clear relationships between technological and organizational changes.
2. Introduction of technological change is tantamount to introduction of a technological, organizational, and people (TOP) change.
3. In order to ensure that the full range of TOP options available to any organization is considered in the selection of any single set of TOP changes, the engineer as technology planner must strive to understand the entire set of anticipated TOP changes prior to implementing new technology.
4. Planned change strategies must be thoughtfully applied to facilitate successful progress through the TOP changes.

Technologies of primary interest here are computer-automated production and information technologies because these have received the most research attention in the last decade. Production technologies include computer-automated manufacturing (CAM) and computer-integrated manufacturing (CIM) and their component technologies such as flexible manufacturing cells (FMC), automated guided vehicles, and computer numerical control (CNC) machines. Information technologies include manufacturing resource planning (MRP), computer-aided design (CAD), computer-aided engineering analysis, electronic mail, collaborative technologies, transaction processing technologies such as enterprise resource planning (ERP) systems, supply chain management systems, and electronic commerce.

2. WHY THE TOPIC IS CRITICAL TO INDUSTRIAL ENGINEERS

2.1. Failures of Implementation of New Technology

Accumulated evidence indicates that the implementation of computer-automated technology has not achieved as much success as originally anticipated. The American Production and Inventory Control Society and the Organization for Industrial Research have estimated the failure rate of these technologies to be as high as 75% (Works 1987). In a study in which 55 managers in 41 organizations supplying or using CAM were interviewed, half of the CAM installations were reported as failures (Ettlie 1986). In a study of 95 flexible manufacturing systems in the United States and Japan, the FMSs in the United States were found to be so ineffectively used as to yield little of the flexibility had been achieved in Japan (Jaikumar 1986). Kalb (1987) reports a 30–70% failure rate of computerized manufacturing technologies. One new product-development manager of a large computer manufacturer reported that "Inadequately implemented new technologies cost our plants up to $1 million a day in unexpected losses." A major study of 2000 U.S. firms that had implemented new office systems revealed that at least 40% of these systems failed to achieve the intended results (Long 1989). Gibbs (1994) reports that for every six new large-scale software systems that are put into operation, two others are cancelled, with the average software development project overshooting its schedule by half. The Standish Group in 1995 reported that only 16% of information systems projects were judged to be successful, with 31% outright cancelled (*Wall Street Journal* 1998b). The Standish Group conducted another survey in 1997 of 360 information system professionals and found that 42% of corporate information technology projects were abandoned before completion and 33% were over budget or late (*Computerworld* 1997).

Examples of these failures abound. For example, after spending more than $17 million on a long-anticipated overhaul of Los Angeles's computerized payroll system, the city controller scrapped it (*Los Angeles Times* 1999). The London Ambulance Service computer-aided dispatch system deployed in 1992 was intended to provide an automatic vehicle-locating system, telephone call processing, and allocation buttons for crew to report on current status. The system was pulled because crews couldn't accurately indicate their status and dispatchers couldn't intervene to get the crews to the needed locations (Flowers 1997). The State of California cancelled deployment of an automated child-support system for automatically tracking parents across counties who do not have primary custody of their children after spending $100 million (*Los Angeles Times* 1997). Fox Meyer, once a $5 billion drug-distribution company, was unable to process the huge volume of orders from pharmacies after installing a $65 million ERP system. As a result, it filed for bankruptcy in 1996, was bought in 1997 for just $80 million, and filed a $500 million lawsuit against Andersen Consulting, the implementers of the ERP system (*Information Week* 1998; *Wall Street Journal* 1998b). Oxford Health Plans lost $363 million in 1997 when their new claims-processing system delayed claims processing and client billing (*Wall Street Journal* 1998a; Champy 1998). Computer systems were blamed for delaying the scheduled opening of the first deregulated electricity market in the United States (*Information Week* 1997). Hershey, the nation's largest candy maker, installed a $110 million ERP system in July 1999. Glitches in the system left many distributors and retailers with empty candy shelves in the season leading up to Halloween (*Wall Street Journal* 1999). Whirlpool reported that problems with a new ERP system and a high volume of orders combined to delay shipments of appliances to many distributors and retailers.

These failures are expensive. In an internal document of September 15, 1997, the information systems research firm MetaFax calculated an average yearly loss of $80 billion from a 30% cancellation rate and a $59 billion loss from a 50% over-budget rate. In 1997 alone (before the Y2K inflated IT expenditures), companies spent $250 billion on information technology; a 30–70% failure rate clearly means that billions of dollars are spent with disappointing results (*Wall Street Journal* 1998b). Aside from a disappointing return on investment, the impacts of failed technology investments include:

- Harm to the firm's reputation (where poor implementation gets blamed on the technology vendor or designer)
- Broken trust (where workers are unwilling to go the extra mile the next time)
- Reduced management credibility (because management can't deliver on promises)
- Slower learning curve (leading to crisis management as problems increase with implementation rather than decrease)
- Reduced improvement trajectory (since there is no time to explore opportunities for new technology or new business opportunities for existing technology)

2.2. Why These High Failure Rates?

In one of the first major studies on this problem of implementation, the Congressional Office of Technology Assessment concluded: "The main stumbling blocks in the near future for implementation of programmable automation technology are not technical, but rather are barriers of cost, organization of the factory, availability of appropriate skills, and social effects of the technologies" (OTA 1984, p. 94). A few years later, the Manufacturing Studies Board of the National Research Council conducted a study of 24 cases of the implementation of CAM and CIM technologies and concluded: "Realizing the full benefits of these technologies will require systematic change in the management of people and machines including planning, plant culture, plant organizations, job design, compensation, selection and training, and labor management relations" (MSB 1986). In a 1986 Yankee Consulting Group marketing survey of CAM and CIM users, the users reported that 75% of the difficulties they experienced with the technologies could be attributable to issues concerned with planning the use of the technology within the context of the organization (Criswell 1988).

Recent evidence continues to support the conclusion that a significant component of the complexity of technological change lies in the organizational changes often experienced. C. Jackson Grayson, Jr., then Chairman of the American Productivity and Quality Center in Houston, Texas, analyzed the 68 applications for the Malcolm Baldrige National Quality Award for 1988 and 1989 and found that a major reason for failing to meet the examination criteria was the neglect of and failure to integrate human and organizational aspects with technology investments (Grayson 1990). Peter Unterweger of the UAW Research Department, after extensive case study visits in the United States and abroad, concluded that the successes of technological implications can be attributable to: (a) hardware playing a subordinate role to organizational or human factors and (b) developing the technical and organizational systems in step with one another (Unterweger 1988). In a study of 2000 U.S. firms implementing new office systems, less than 10% of the failures were attributed to technical failures; the majority of the reasons given were human and organizational in nature (Long 1989). The MIT Commission on Industrial Productivity concluded from their extensive examination of the competitiveness of different American industries: "Reorganization and effective integration of human resources and changing technologies within companies is the principal driving force for future productivity growth" (Dertouzos et al. 1989). More recently, in a 1997 survey by the Standish Group of 365 IT executive managers, the top factors identified in application development project failures were poor management of requirements and user inputs (*Computerworld* 1998a). The 1997 MetaFax survey found the reasons for IS failures to include poor project planning and management. In a 1998 *Computerworld* survey of 365 IT executives, the top factors for software development project failures were the lack of user input and changing requirements (*Computerworld* 1998a). A careful study of six failures of information technology projects found that those projects that devoted more effort to the technology rather than to the organizational issues (such as awareness, training, and changes to organizational procedures) were more likely to fail (Flowers 1997). In short, these failures can be attributed to the inadequate integration of technical with social and organizational factors during the introduction of the technological change, called sociotechnical or TOP (for Technology, Organization, and People) integration. This recognition has led *The Wall Street Journal* to write: ‘‘What's emerging here is a search for a better balance between manpower and computer power’’ (1998b, p. 1).

Several cases of failures directly attributable to poor alignment of technology and organizational change can be cited. In one such example (Ciborra and Schneider 1990), a U.S. aircraft instruments plant implemented a computerized MRP system. Ten months into the implementation process, none of the expected gains in efficiency had materialized, despite clearly defined goals and plans, a sound economic evaluation, and a structured implementation plan. The major problem was that there was so much emphasis on following the rules created by the MRP system that clerks often hesitated to override the system's commands even when they knew that the commands did not make sense. Even useful localized innovations with the system, such as shortcuts and new rules of thumb, remained private know-how because localized practices were not sanctioned by management. Learning from mistakes was limited because effective job performance for the system designers was measured by adherence to best technical practice, not to shop-floor reality, and thus the system designers were not willing to have their competence questioned.

As another example, in 1997 Chrysler Financial tossed out a sophisticated financial package bought for the company's financial team. The problem was that the system was incompatible with the company's e-mail system. So the company adopted a less sophisticated approach that was more closely aligned with the way the financial staffers worked: instead of monitoring dealer activity with a 100% computerized system, the company instructed clerks to obtain information from dealers the old-fashioned way—over the phone—and enter the information quickly to make it available to financial staffers who wanted to know which dealers were moving a lot of cars or taking bad loans. According to the project director, the purely computerized solution would have cost many millions of dollars more and taken years to install, but ‘‘by adding some people into the equation, we could get 95% of what we needed’’ and take only 90 days to set it up (*Wall Street Journal* 1999, p. A26).

Another example of how advanced technology without correct organizational alignment in the automotive industry failed is presented by *The Economist:*

> [T]he giant Hamtramck plant in Detroit, which makes Cadillacs, is just five years old and heavily automated but ranks among the least competitive plants in the United States. Hamtramck is typical of GM's early efforts to beat the Japanese by throwing truckloads of cash into a new technology. Hamtramck had what is politely called a ‘‘very rough start-up’’. Its robots ran wild. Although the problems have now largely been tamed, GM learnt in a joint venture with Toyota that what really mattered in manufacturing was people. (*Economist* 1990).

As another example, British Airways put in a system at airport gates in which the screen was mounted horizontally, low on a desktop. Ticket agents looked down when checking in passengers; as a result, passengers saw only the top of the agent's head. The consultant on the project reported that they did this deliberately so there would be less eye contact and less schmoozing and the lines would be shorter. However, after installation, passengers complained; apparently fliers are naturally anxious and often need a little schmoozing, according to the consultant. The airline moved the screens to eye level (*Computerworld* 1998b).

Similarly, according to a survey of the artificial intelligence industry by *The Economist,* blind introduction of computers in the workplace by an American airline (which prefers to remain nameless) proved that people do not like taking orders from a machine when an expert system was installed to schedule the work of maintenance engineers (*Economist* 1992). The engineers simply rejected the system's plans and the computer system had to be withdrawn. But when, after suitable delay, the airline reintroduced more or less the same system for engineers to use when and if they wanted, it was much better received.

A final example of a project devoting too much attention to the technology side and too little to the organizational side is the London Ambulance system failure. In the formal inquiry on the failure, it was noted that the initial concept of the system was to fully automate ambulance dispatching; however, management clearly underestimated the difficulties involved in changing the deeply ingrained culture of London Ambulance and misjudged the industrial relations climate so that staff were alienated to the changes rather than brought on board. (Flowers 1997).

While much of this information supporting the important role of aligning technology and organizations is anecdotal, there have been several econometric studies of larger samples supporting this claim. A growing body of literature has established strong empirical links among such practices as high-involvement work practices, new technologies, and improved economic performance (MacDuffie 1995; Arthur 1992). Pil and MacDuffie (1996) examined the adoption of high-involvement work practices over a five-year period in 43 automobile assembly plants located around the world, their technologies (ranging from highly flexible to rigidly integrated), and their economic performance and found that the level of complementary human resource practices and technology was a key driver of successful introduction of high-involvement practices. Kelley (1996) conducted a survey of 973 plants manufacturing metal products and found that a participative bureaucracy (i.e., group-based employee participation that provides opportunities to reexamine old routines) is complementary to the productive use of information technology in the machining process. Osterman (1994) used data on 694 U.S. manufacturing establishments to examine the incidence of innovative work practices, defined as the use of teams, job rotation, quality circles, and total quality management. He found that having a technology that requires high levels of skills was one factor that led to the increased use of these innovative work practices.

To conclude, it should be clear that technological change often necessitates some organizational change. If both organizational and technological changes are not effectively integrated and managed to achieve alignment, the technological change will fail.

3. WHAT ARE THE STUMBLING BLOCKS TO ALIGNMENT?

If the benefits of aligning technology and organizational design are so clear, why isn't it done? We suggest that there are many reasons.

3.1. The Future of Technology Is Probabilistic

The technology S curve has been historically documented as describing technology change over the years (Twiss 1980; Martino 1983). The curve, plotted as the rate of change of a performance parameter (such as horsepower or lumens per watt) over time, has been found to consist of three periods: an early period of new invention, a middle period of technology improvement, and a late period of technology maturity. The technology S curve, however, is merely descriptive of past technology changes. While it can be used for an intelligent guess at the rate of technology change in the future, technology change is sufficiently unpredictable that it cannot be used to predict precisely when and how future change may occur. Fluctuating market demand and/or novelty in the technology base exacerbate the challenge. For example, at Intel, typically at least one third of new process equipment has never been previously used (Iansiti 1999). This probabilistic nature of the technology makes creating aligned technology and organizational solutions difficult because it cannot be known with any certainty what the future organizational-technology solution is likely to be over the long term.

In his study of six information technology project failures, Flowers (1997) concluded that the unpredictability of the technology is a primary complexity factor that contributes to project failure. The more that the technology is at the "bleeding" edge, the greater the complexity. Avoiding overcommitment to any one technology or organizational solution, avoiding escalationary behavior where more resources are thrown at the solution-generation process without adequate checks and balances, and maintaining project-reporting discipline in the face of uncertainty are suggested ways of managing the inherent probabilistic nature of technology.

3.2. Some Factors Are Less Malleable Than Others

A series of research studies on the process by which technologies and organizations are adapted when technologies are implemented into an organization have shown that adaptations of both technologies and the organization can occur (Barley 1986; Contractor and Eisenberg 1990; Orlikowski and Robey 1991; Orlikowski 1992; Giddens 1994; Rice 1994; Orlikowski et al. 1995; Rice and Gattiker 1999). However, in reality, some adaptations are less likely to occur because some of these factors tend to be less malleable (Barley 1986; Johnson and Rice 1987; Poole and DeSanctis 1990; Orlikowski 1992; DeSanctis and Poole 1994; Orlikowski and Yates 1994). One of these factors is the existing organizational structure. For example, Barley (1986) found evidence that one factor that tends to be less malleable is the existing power structure in the organization. Barley found that when a medical radiation device was installed into two separate hospitals, the work changed in accordance with the organizational structure, not vice versa. That is, in the hospital where the radiologists had more power in the organizational structure than the technicians, the rift between the two jobs became greater with the new technology. In contrast, in the hospital where technicians and radiologists were not separated hierarchically, the technology was used to share knowledge between the two. Another factor often found to be less malleable is what DeSanctis and Poole (1994) refer to as the "technology spirit," which they define as the intended uses of the technology by the developer or champion who influenced the developer. If the spirit is intended to displace workers, then this spirit is unlikely to be changed during implementation. Research contradicting this assertion has been conducted recently, however (Majchrzak et al. 2000). Moreover, Tyre and Orlikowski (1994) have found that malleability may be temporal, that is, that technologies and structures can be changed, but only during windows of opportunity that may periodically reopen as the technology is used. In their study, the authors found these windows to include new rethinking about the use of the technology or new needs for the technology that were not originally envisioned. These windows did not stay open for very long; thus, over the long term, some factors may have appeared to be less malleable than others.

In sum, then, a stumbling block to integrating TOP is determining which facets of TOP are malleable to facilitate the alignment; when one facet is not malleable, that puts additional pressure on the remaining facet to conform—a pressure that may not be achievable.

3.3. Alignment Requires a Cross-Functional Definition of the Problem

For a solution to be sociotechnically aligned, changes may be needed in all aspects of the organization, not just that which is under the purview of the industrial engineer or even the manufacturing manager. Changes may be required in the material-handling organization (which may not report to the manufacturing department), the purchasing department, or the human resources department. For example, Johnson and Kaplan (1987), in their study of just-in-time manufacturing, found that those departments that made changes in the incentive systems (a responsibility outside that of the manufacturing manager) were less likely to have implementation problems than companies that did not make such changes. This cross-functional nature of aligned solutions creates the problem that because the solution touches on everyone's responsibilities, it is essentially no one's responsibility (Whiston 1996). Thus, unless an organizational structure is created to explicitly recognize the cross-functional nature of the alignment, a single function—such as the industrial engineer—cannot create the alignment. As a result, resolving a cross-functional problem with a single function becomes difficult, if not impossible.

3.4. Alignment Is Context Specific and Nonrepeatable

A solution that achieves alignment between technology and organization is typically so context specific that it is not likely to be repeatable in its exact form for the next alignment problem that comes along. This is because of the many factors that must be considered in deriving a technology-organization solution. For example, the global introduction of a new technology product typically now requires some modification in each context in which it is introduced either because of the different needs of customers or different service or manufacturing environments (Iansiti 1999) As another example, altering even one technology factor, such as the degree to which the technology can diagnose its own failures, creates the need to change such organizational factors as the amount of skills that workers must have to operate the technology (Majchrzak 1988). As another example, human supervisory control of automated systems—such as is seen in an oil and gas pipeline control center—involves fault diagnosis, error detection and recovery, and safe handling of rare, critical, and nonroutine events and incidents; these activities require very specific system-dependent sets of skills and teamwork (Meshkati 1996).

This context-specific nature of technology-organization solutions contradicts the desire of many managers today to use ''cookie cutter'' or repeatable solutions, believing that such solutions will cost less than solutions tailored to each site (Jambekar and Nelson 1996; Kanz and Lam 1996). In addition, Kahneman et al. (1982) have found that the judgments of people in conditions of uncertainty are governed by the availability heuristic (or bias), whereby people judge the likelihood of something happening by how easily they can call other examples of the same thing to mind. If they have no other examples, they will create connections between examples, even though the connections are tenuous. As a result, they will believe that they have a repeatable solution even though one is not warranted.

For example, when globally implementing ERP systems, managers have a choice whether to roll out a single standardized ERP solution worldwide or to allow some issues (such as user interface screens or data structures) to have localized solutions. Forcing a single standardized implementation world-wide has been the preferred strategy in most implementations because it minimizes the complexity and resources required to accommodate to localized modifications (Cooke and Peterson 1998). However, implementers at Owens-Corning believe that part of their success in their global ERP implementation was attributable to allowing localized solutions, even though it was slightly more complicated in the beginning. They believe that allowing field locations to tailor some aspects of the ERP system not only ensured the buy-in and commitment of field personnel to the ERP project, but also ensured that the ERP system met each and every field location's particular needs.

Thus, another stumbling block to alignment is that alignment solutions are best construed as nonrepeatable and highly contextual—a concept that raises management concerns about the resources required to allow such contextualization.

3.5. Alignment Requires Comprehensive Solutions That Are Difficult to Identify and Realize

A solution aligned for technology and organization is a comprehensive one involving many factors. Today it is widely believed that in addition to strategy and structure, an organization's culture, technology, and people all have to be compatible. If you introduce change in technology, you should

expect to alter your corporate strategy to capitalize on the new capabilities, alter various departmental roles and relations, add personnel with new talents, and attempt to "manage" change in shared beliefs and values needed to facilitate use of the new technology. (Jambekar and Nelson 1996, p. 29.5)
Despite this need for integration, Iansiti (1999) charges that "technology choices are too often made in scattershot and reactive fashion, with technology possibilities chosen for their individual potential rather than from their system-level integration." Iansiti specifically suggests that only when there is a proactive process of technology integration—"one comprising a dedicated, authorized group of people armed with appropriate knowledge, experience, tools, and structure"—will results be delivered on time, lead times be shorter, resources be adequately utilized, and other performance measures be achieved. In a study of reengineering efforts, Hall et al. (1993) argue that many attempts at reengineering have failed because of a focus on too few of the factors needing to be changed. Instead, for reengineering to work, fundamental change is required in at least six elements: roles and responsibilities, measurements and incentives, organizational structure, information technology, shared values, and skills.

Thus, another stumbling block to alignment is the need to consider all these factors and their relationships. For many managers and industrial engineers, there are too many factors and relationships; as a result, it is far easier to focus mistakenly on only one or a few factors.

3.6. Alignment Involves Long Planning Cycles, Where Observable Results and Knowing Whether You Made the Right Decisions Take Awhile

Years ago, Lawrence and Lorsch (1967) helped us to recognize the importance of the time horizon of feedback from the environment in determining whether strategic and organizational decisions are the right decisions. In their research, they found that some departments had very quick time horizons, such as a manufacturing department that is structured and oriented to obtaining quick feedback from the environment. In contrast are departments with longer time horizons, such as a research and development department, in which the department is organized to expect feedback about their work over a much longer time period. Lawrence and Lorsch further found that these different time horizons of feedback created different needs for organizational structures, performance-monitoring systems, and personnel policies. The technology-development process can be characterized as one that has a long planning cycle so that the time horizon of feedback may be months or years. For example, the average CIM implementation may take up to 3 years to complete; while the implementation of a large ERP system takes at least 18 months. As a result, managers and engineers need to make decisions about the design of the technology-organization solution in the absence of any data from the field. While some of these decisions may be changed later if data from the field indicate a problem in the design, some of these decisions are changeable only at great cost. This creates a bias toward conservativeness, that is, making decisions that minimize risk. As a result, only those factors that decision makers have historical reason to believe should be changed are likely to be changed, increasing the probability of misalignment. Thus, another stumbling block to achieving alignment is the long planning cycle of technology-organizational change, which tends to create a bias against change because learning whether planning decisions are the right ones.

3.7. Alignment Involves Compromises

Given the many factors involved in deriving an aligned solution and the many functions affected by an aligned solution, the final aligned solution is unlikely to be an idealized solution. Rather, the final solution is likely to be the outcome of a series of negotiations among the relevant parties. For example, a labor union may not want to give up the career-progression ladder provided by specialized jobs and embrace cross-functional teamwork; management may not want to give up the decision-making control they enjoy and embrace autonomy among the teams. The process of negotiating these different positions to result in some amicable compromise may be difficult and frustrating, adding to the challenges imposed by alignment.

Information technology, because it tends to break down organizational barriers, turfs, and layers, could face opposition from individuals entrenched in the companies' hierarchy. For example, production planning, inventory control, and quality control will increasingly be under the control of front-line employees, and this will pose a major threat to low-level supervisors and middle managers (Osterman 1989) and may even lead to their extinction (Drucker 1988).

4. HOW CAN TECHNOLOGY PLANNERS PURSUE ALIGNMENT DESPITE THESE DIFFICULTIES?

The difficulties identified in Section 3 are real difficulties not likely to go away with new managers, new technologies, new industrial engineering skills, new organizational designs, or new motivations. Therefore, industrial engineers must identify ways to move past these difficulties. This means taking the difficulties into account when pursuing alignment, rather than ignoring them. Effort then is not spent on reducing the difficulties per se, but on managing them so that alignment can still be achieved. Below we propose several ways of pursuing alignment in ways that allow technology planners to move past the difficulties.

4.1. Focus Alignment on Business Purpose, Driven by Competitive Need, Not as a Technology Fix to a Localized Problem

The impact of the difficulties identified in Section 3 is often experienced as resistance to change. Managers argue against a technology; workers refuse to intervene to fix the technology; industrial engineers focus solely on the technology, refusing to consider work and job changes. This resistance to change is often a sign that the justification for the technology is weak. Weak justifications are those where the need for the technology is not driven by competitive advantage pursued by the firm. Porter (1985), Schlie and Goldhar (1995), Pine (1993), Goldman et al. (1995), and D'Aveni and Gunther (1994), among others, have emphasized the need for technology choices to be driven by the competitive advantage being pursued by the firm. Yet, as pointed out by Kanz and Lam (1996), traditional strategic management rarely adequately ties technology choices to strategic choices because of a lack of understanding of how technology choices are different from other types of strategic choices (such as new products or cost-cutting strategies). In a two-year study involving over 300 major firms, they found that while 50 executives believed their firms tied improvements in their IT infrastructure to a business strategy, only 10 firms were found to be doing so after a formal assessment. Moreover, while 190 executives believed their overall corporate strategies were driving the methodology for implementing their business plans, less than 20 strategies were actually doing so. The remainder were constrained by limitations in either organizational or IT culture and design (Sweat 1999).

Schlie (1996) offers specific suggestions for identifying how technology choices should be driven by competitive firm needs. He adopts Porter's (1985) strategic planning framework, which suggests that competitive advantage can be derived at any point along a firm's value chain (e.g., inbound logistics, outbound logistics, marketing/sales, procurement, R&D, human resource management, or firm infrastructure). For the point on the value chain that the firm decides to have a competitive advantage, that advantage can be achieved either through cost leadership (i.e., low cost, low price) or differentiation (i.e., uniqueness to the customer). Using this framework, Schlie (1996) suggests that firm management should first decide where in the value chain they will compete, and then how they will use technology to facilitate achieving their competitive advantage. In communicating this to plant personnel, then, justification of both the strategic choices as well as how technology helps the strategic choices is required. Schlie cautions, however, that some technologies can only be adequately justified for some of these strategic choices. He uses as an example the advanced manufacturing technologies CAM and CIM, pointing out that the contribution of these technologies to the competitive advantage of low cost is ambiguous and situation specific. Yet when firms justify their technology expenditures based on direct labor savings, that is precisely what they are suggesting. Thus, difficulties of alignment will not be overcome if the justification for the technology expenditure is suspect from the outset.

While there are many other strategic planning frameworks for integrating technology design choices with strategic choices (e.g., Burgelman and Rosenbloom 1999; Leonard-Barton 1995), the purpose here is not to elaborate the frameworks but to emphasize the need for the technology design choices to be driven by a business strategy—regardless of the framework used—and not by reactive problem solving.

4.2. Recognize the Breadth of Factors and Their Relationships That Must Be Designed to Achieve Alignment

It is apparent from Section 2.2 that the high failure rates of new technologies are due to the lack of alignment among technology and organizational factors. What are these factors? The U.S. industry's initiative on agile manufacturing (documented in Goldman et al. 1995) identified a range of factors, including the production hardware, the procurement process, and the skills of operators. The National Center for Manufacturing Sciences created a program to promote manufacturing firms to assess themselves on their excellence. The assessment contained 171 factors distributed across 14 areas ranging from supplier development to operations, from cost to flexibility, from health and safety to customer satisfaction. In a five-year industry–university collaborative effort funded by the National Center for Manufacturing Sciences, 16 sets of factors were identified that must be aligned (Majchrzak 1997; Majchrzak and Finley 1995), including:

- Business strategies
- Process variance-control strategies
- Norms of behavior

- Strategies for customer involvement
- Employee values
- Organizational values
- Reporting structure
- Performance measurement and reward systems
- Areas of decision-making authority
- Production process characteristics
- Task responsibilities and characteristics
- Tools, fixtures, and material characteristics
- Software characteristics
- Skill breadth and depth
- Information characteristics
- Equipment characteristics

Within each set, 5–100 specific features were identified, with a total of 300 specific design features needing to be designed to create an aligned organizational-technology solution for a new technology. In this five-year study, it was also found that achieving alignment meant that each of these factors needed to be supportive of each other factor. To determine whether a factor was supportive of another factor, each factor was assessed for the degree to which it supported different business strategies, such as minimizing throughput time or maximizing inventory turnover. Supportive factors were then those that together contributed to the same business strategy; inversely, misaligned solutions were those for which design features did not support similar business strategies.

Recognizing this range of factors and their relationships may seem overwhelming; but it can be done. The cross-functional teams and use of CAD technologies for developing the Boeing 777 aircraft present an excellent example of alignment. In designing the 777, Boeing created approximately 240 teams, which were labeled ‘‘design-build teams.’’ These teams included cross-functional representatives from engineering design, manufacturing, finance, operations, customer support, maintenance, tool designers, customers, and suppliers (Condit 1994). To communicate part designs, the teams used 100% digital design via the 3D CAD software and the networking of over 2000 workstations. This allowed the suppliers to have real-time interactive interface with the design data; tool designers too were able to get updated design data directly from the drawings to speed tool development. In addition, the CAD software's capability in performing preassembly checks and visualization of parts allowed sufficient interrogation to determine costly misalignments, interferences, gaps, confirmation of tolerances, and analysis of balances and stresses (Sherman and Souder 1996). In sum, the technology of CAD was aligned with the organizational structure of the cross-functional teams.

4.3. Understand the Role of Cultures in Alignment

Culture affects alignment by affecting the change process: changes that support the existing culture are easier to implement successfully than changes that cause the culture to change. At least two types of culture must be considered in designing a technology-organization solution: the national culture of the country and the culture of the organization.

According to Schein (1985), organizational culture is ‘‘a pattern of basic assumptions—invented, discovered, or developed by a given group as it learns to cope with its problems of external adaptation and internal integration—that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems.’’ Kotter and Heskett (1992, p. 4) contend that organizational culture has two levels that differ in terms of their visibility and their resistance to change:

> At the deeper and less visible level, culture refers to values that are shared by the people in a group and that tend to persist over time even when group membership changes. . . . At the more visible level, culture represents the behavior patterns or style of an organization that new employees are automatically encouraged to follow by their fellow employees. . . . Each level of culture has a natural tendency to influence the other.

Operationally, organizational culture is defined as a set of shared philosophies, ideologies, values, beliefs, expectations, attitudes, assumptions, and norms (Mitroff and Kilmann 1984). Cultural norms are the set of unwritten rules that guide behavior (Jackson 1960). Use of this concept allows the capturing of those dimensions of organizational life that may not be visible in the more rational and mechanical aspects of the organization.

Cultures can be characterized not only by their focus but also by their strength (O'Reilly 1989; Beyer 1992). Strong cultures exert greater conformity on organizational members than weak cultures. The stronger the culture, then, the more difficult it will be to implement a technology-organization alignment that contrasts with that culture. For example, if a knowledge-management repository is installed, workers are unlikely to contribute to the repository if there is a strong culture that encourages independence and heroism (Davenport 1994) Thus, in designing a technology-organization solution, the existing culture of the organization should be carefully considered and, if possible, used to foster the solution.

National culture, according to anthropologists, is the way of life of a people—the sum of their learned behavior patterns, attitudes, customs, and material goods. According to Azimi (1991), the culture of a society consists of a set of ideas and beliefs. These ideas and beliefs should have two principal characteristics or conditions: first, they should be accepted and admitted by the majority of the population; and second, the acceptance of these beliefs and ideas should not necessarily depend upon a scientific analysis, discussion, or convincing argument. Also, national culture, in the context of technology transfer and utilization, could operationally be defined as the ‘‘collective mental programming of peoples’ minds’’ (Hofstede 1980a).

National culture affects not only the safety but also the success and survival of any technology. National cultures differ on at least four basic dimensions: power distance, uncertainty avoidance, individualism-collectivism, and masculinity-femininity (Hofstede 1980b). Power distance is the extent to which a society accepts the fact that power in institutions and organizations is distributed unequally. It is an indication of the interpersonal power or influence between two entities, as perceived by the less powerful of the two (BCAG 1993). Uncertainty avoidance is the extent to which a society feels threatened by uncertain and ambiguous situations. It also refers to attempts to avoid these situations by providing greater career stability, establishing more formal rules, not tolerating deviant ideas and behaviors, and believing in absolute truths and the attainment of expertise. Individualism is characterized by a loosely knit social framework in which people are supposed to take care of themselves and their immediate families only, while collectivism is characterized by a tight social framework in which people distinguish between in-group and out-group; they expect their in-group members (e.g., relatives, clan, organization) to look after them, and in exchange they owe absolute loyalty to the group. The masculinity dimension expresses the extent to which the dominant values in a society are ‘‘masculine,’’ as evidenced by decisiveness, interpersonal directness, and machismo (Johnston 1993). Other characteristics of masculine cultures include assertiveness, the acquisition of money and material goods, and a relative lack of empathy and reduced perceived importance for quality-of-life issues. This dimension can also be described as a measure of the need for ostentatious manliness in the society (BCAG 1993). Femininity, the opposite pole of this continuum, represents relatively lower assertiveness and greater empathy and concern for issues regarding the quality of life.

The four cultural dimensions discussed above also have significant implications for most complex technological systems’ performance, reliability, and safety. For instance, according to Helmreich (1994) and Helmreich and Sherman (1994), there is evidence that operators with high power distance and high uncertainty avoidance prefer and place a ‘‘very high importance’’ on automation. Furthermore, it is known that the primary purpose of regulations is to standardize, systematize, and impersonalize operations. This is done, to a large extent, by ensuring adherence to (standard and emergency) operating procedures. On many occasions it requires replacing operators’ habits with desirable intentions that are prescribed in procedures or enforced by regulations. However, according to several studies, an operator’s culturally driven habit is a more potent predictor of behavior than his or her intentions, and there could be occasions on which intentions cease to have an effect on operators’ behavior (Landis et al. 1978). This fact places in question the effectiveness of those regulations and procedures that are incompatible with operators’ culturally driven habits.

A major, though subtle, factor affecting the safety and performance of a technological system is the degree of compatibility between its organizational culture and the national culture of the host country. It is an inevitable reality that groups and organizations within a society also develop cultures that significantly affect how the members think and perform (Schein 1985).

Demel (1991) and Demel and Meshkati (1989) conducted an extensive field study to explore how the performance of U.S.-owned manufacturing plants in other countries is affected by both the national culture of the host country and the organizational culture of the subsidiary plant. A manufacturing plant division of a large American multinational corporation was examined in three countries: Puerto Rico, the United States, and Mexico. Hofstede’s (1980a) Values Survey Module for national culture and Reynolds’s (1986) Survey of Organizational Culture were administered. Performance measures (i.e., production, safety, and quality) were collected through the use of secondary research.

The purpose of this investigation was threefold:

1. To determine whether there were any differences among the national cultures of Puerto Rico, the United States, and Mexico
2. To find out whether there were any differences between the organizational cultures of the three manufacturing plants

3. To establish whether there was any compatibility between the organizational culture of the plants and the national culture of the three countries, and examine whether the compatibility (or incompatibility) affected their performance in terms of production yields, quality, safety, and cycle time

Although the results of this study indicate that there are differences among the national culture dimensions of Puerto Rico, the United States, and Mexico, no significant differences were found between the organizational cultures of the three plants. This may be due to selection criteria, by which candidates, by assessment of their behavioral styles, beliefs, and values, may have been carefully screened to fit in with the existing organizational culture, Additionally, socialization may have been another factor. This means that the company may have had in-house programs and intense interaction during training, which can create a shared experience, an informal network, and a company language. These training events often include songs, picnics, and sporting events that build a sense of community and feeling of togetherness. Also, the company may have had artifacts, the first level of organizational culture, such as posters, cards, and pens that remind the employees of the organization's visions, values, and corporate goals and promote the organization's culture.

Therefore, it seems that a ''total transfer'' has been realized by this multinational corporation. Because these manufacturing plants produce similar products, they must obtain uniform quality in their production centers. To gain this uniformity, this company has transferred its technical installations, machines, and organization. Moreover, to fulfill this purpose, the company chooses its employees according to highly selective criteria. Notwithstanding, Hofstede's research demonstrates that even within a large multinational corporation known for its strong culture and socialization efforts, national culture continues to play a major role in differentiating work values (Hofstede 1980a).

There are concepts in the dimensions of organizational culture that may correspond to the same concepts of the dimensions of national culture:

The power distance dimension of national culture addresses the same issues as the perceived oligarchy dimension of organizational culture. They both refer to the nature of decision making; in countries where power distance is large, only a few individuals from the top make the decisions. Uncertainty avoidance and perceived change address the concepts of stability, change, and risk taking. One extreme is the tendency to be cautious and conservative, such as in avoiding risk and change when possible in adopting different programs or procedures. The other is the predisposition to change products or procedures, especially when confronted with new challenges and opportunities—in other words, taking risks and making decisions. Uncertainty avoidance may also be related to perceived tradition in the sense that if the employees have a clear perception of ''how things are to be done'' in the organization, their fear of uncertainties and ambiguities will be reduced. An agreement to a perceived tradition in the organization complements well a country with high uncertainty avoidance. Individualism–collectivism and perceived cooperation address the concepts of cooperation between employees and trust and assistance among colleagues at work. In a collectivist country, cooperation and trust among employees are perceived more favorably than in an individualist country.

The perceived tradition of the organizational culture may also be related to individualism-collectivism in the sense that if members of an organization have shared values and know what their company stands for and what standards they are to uphold, they are more likely to feel as if they are an important part of the organization. They are motivated because life in the organization has meaning for them. Ceremonies of the organizational culture and rewards given to honor top performance are very important to employees in any organization. However, the types of ceremonies or rewards that will motivate employees may vary across cultures, depending on whether the country has a masculine orientation, where money and promotion are important, or a feminine orientation, where relationships and working conditions are important. If given properly, these may keep the values, beliefs, and goals uppermost in the employees' minds and hearts.

Cultural differences may play significant roles in achieving the success of the corporations' performance. The findings of this study could have important managerial implications. First, an organizational culture that fits one society might not be readily transferable to other societies. The organizational culture of the company should be compatible with the culture of the society the company is transferring to. There needs to be a good match between the internal variety of the organization and the external variety from the host country. When the cultural differences are understood, the law of requisite variety can then be applied as a concept to investigate systematically the influence of culture on the performance of the multinational corporations' manufacturing plants. This law may be useful for examining environmental variety in the new cultural settings. Second, the findings have confirmed that cultural compatibility between the multinational corporations' organizational culture and the culture of the countries they are operating in plays a significant role in the performance of the corporations' manufacturing plants.

Therefore, it can be suggested that the decision concerning which management system or method to promote should be based on specific human, cultural, social, and deeply rooted local behavior patterns. It is critical for multinational corporations operating in different cultures from their own to ensure and enhance cultural compatibility for the success of their operations. As a consequence, it can be recommended that no organizational culture should be transferred without prior analysis and recommendations for adjustment and adaptation to the foreign countries' cultures and conditions. This research has given a clear view of the potential that currently exists for supervising and evaluating cultural and behavioral aspects of organizations as affected by their external environment and their relationship to the performance of the organizations. Culture, both national and organizational, will become an increasingly important concept for technology transfer.

Results showed that while there were differences between the national cultures of the three countries, there were no significant differences between the organizational cultures of the three manufacturing plants. It is noteworthy that the rank order of the performance indicators for these plants was in exact concordance with the rank order of the compatibility between the organizational culture and the national culture of the host country: Mexico had the highest overall cultural compatibility and the highest performance; Puerto Rico had high overall compatibility and the next-highest overall performance; and the United States had the lowest cultural compatibility and the lowest overall performance.

Meshkati has recently studied the concept of a "safety culture." Nuclear reactor operators' responses to nuclear power plant disturbances is shown in Figure 1 (Meshkati et al. 1994, adapted from Rasmussen 1992). The operators are constantly receiving data from the displays in the control

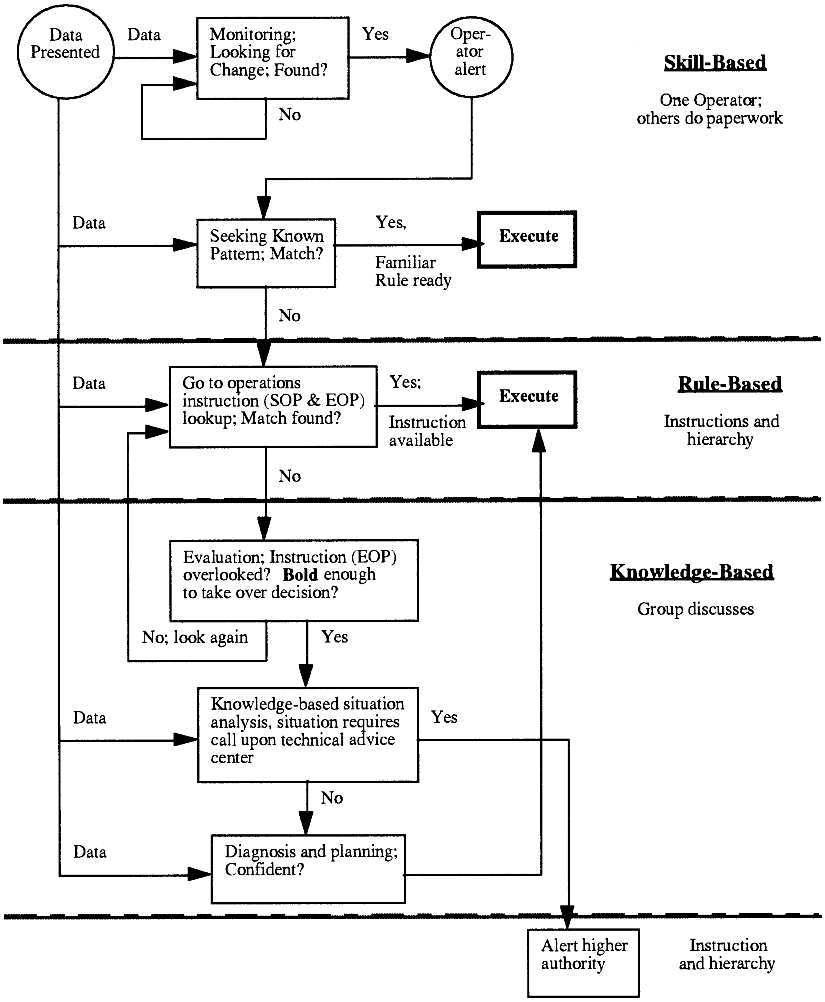

Data Presented
Data
Monitoring; Looking for Change; Found?
Yes
Oper-ator alert
No
Skill-Based
One Operator; others do paperwork
Data
Seeking Known Pattern; Match?
Yes,
Familiar Rule ready
Execute
No
Data
Go to operations instruction (SOP & EOP) lookup; Match found?
Yes;
Instruction available
Execute
Rule-Based
Instructions and hierarchy
No
Evaluation; Instruction (EOP) overlooked? Bold enough to take over decision?
Knowledge-Based
Group discusses
No; look again
Yes
Data
Knowledge-based situation analysis, situation requires call upon technical advice center
Yes
No
Data
Diagnosis and planning; Confident?
Alert higher authority
Instruction and hierarchy

Figure 1 Model for Nuclear Power Plant Operators' Responses to Disturbances. (Adapted from Rasmussen 1992)

room and looking for change or deviation from standards or routines in the plant. It is contended that their responses during transition from the rule-based to the knowledge-based level of cognitive control, especially in the knowledge-based level, are affected by the safety culture of the plant and are also moderated or influenced by their cultural background. Their responses could start a vicious cycle, which in turn could lead to inaction, which wastes valuable time and control room resources. Breaking this vicious cycle requires boldness to make or take over decisions so that the search for possible answers to the unfamiliar situation does not continue unnecessarily and indefinitely. It is contended that the boldness is strongly culturally driven and is a function of the plant's organizational culture and reward system and the regulatory environment. Boldness, of course, is also influenced by operators' personality traits, risk taking, and perception (as mentioned before), which are also strongly cultural. Other important aspects of the national culture include hierarchical power distance and rule orientation (Lammers and Hickson 1979) which govern the acceptable behavior and could determine the upper bound of operators' boldness.

According to the International Atomic Energy Agency, two general components of the safety culture are the necessary framework within an organization whose development and maintenance is the responsibility of management hierarchy and the attitude of staff at all different levels in responding to and benefiting from the framework (IAEA 1991). Also, the requirements of individual employees for achieving safety culture at the installation are a questioning attitude, a rigorous and prudent approach, and necessary communication. However, it should be noted that other dimensions of national culture—uncertainty avoidance, individualism–collectivism, and masculinity–femininity—while interacting with these general components and requirements, could either resonate with and strengthen or attenuate safety culture. For instance, the questioning attitude of operators is greatly influenced by the power distance, rule orientation, and uncertainty avoidance of the societal environment and the openness in the organizational culture of the plant. A rigorous and prudent approach that involves understanding the work procedures, complying with procedure, being alert for the unexpected, and so on is moderated by power distance and uncertainty avoidance in the culture and by the sacredness of procedures, the criticality of step-by-step compliance, and a definite organizational system at the plant. Communication which involves obtaining information from others, transmitting information to others, and so on, is a function of all the dimensions of national culture as well as the steepness and rigidity of the hierarchical organizational structure of the plant.

The nuclear industry shares many safety-related issues and concerns with the aviation industry, and there is a continuous transfer of information between them (e.g., EPRI 1984). Cultural and other human factors considerations affecting the performance of a cockpit crew are, to a large extent, similar to those affecting nuclear plant control room operators. Therefore, it is worth referring briefly to a fatal accident involving a passenger airplane in which, according to an investigation by the U.S. National Transportation Safety Board (NTSB 1991), national cultural factors within the cockpit and between it and the air traffic control tower contributed significantly to the crash. Avianca flight 052 (AV052) (Avianca is the airline of Colombia), a Boeing 707, crashed in Cove Neck, New York, on January 25, 1990, and 73 of the 158 persons aboard were killed. According to the NTSB:

> The NTSB determines that the probable cause of this accident was the failure of the flight crew to adequately manage the airplane's fuel load, *and their failure to communicate an emergency fuel situation to air traffic control* before fuel exhaustion occurred. (NTSB 1991, p. 76, emphasis added)
>
> The word "*priority*" was used in procedures' manuals provided by the Boeing Company to the airlines. A captain from Avianca Airlines testified that the use by the first officer of the word "priority," rather than "*emergency,*" may have resulted from training at Boeing. . . . He stated that these personnel received the *impression* from the training that the words priority and emergency *conveyed the same meaning* to air traffic control. . . . The controllers stated that, although they would do their utmost to assist a flight that requested "priority," *the word would not require a specific response* and that *if a pilot is in a low fuel emergency and needs emergency handling, he should use the word "emergency."* (NTSB 1991, p. 63; emphasis added)

The NTSB concluded:

> The first officer, who made all recorded radio transmissions in English, *never used the word "Emergency,"* even when he radioed that two engines had flamed out, and he did not use the appropriate phraseology published in United States aeronautical publications to communicate to air traffic control the flight's minimum fuel status. (NTSB 1991, p. 75, emphasis added)

Helmreich's (1994) comprehensive analysis of the AV052 accident thoroughly addresses the role of cultural factors. His contention is that

> had air traffic controllers been aware of cultural norms that may influence crews from other cultures, they might have communicated more options and queried the crew more fully regarding the flight status. . . . The possibility that behavior on this [flight] was dictated in part by norms of national culture cannot be dismissed. It seems likely that national culture may have contributed to [the crew's behavior and decision

> making]. . . . Finally, mistaken cultural assumptions arising from the interaction of two vastly different national cultures [i.e., crew and ATC] may have prevented effective use of the air traffic control system. (Helmreich 1994, p. 282)

These conclusions have been corroborated in principle by several other studies: an operator's culturally driven habit is a more potent predictor of behavior than his or her intentions, and there could be occasions on which intentions cease to have an effect on operators' behavior (Landis et al. 1978). This fact brings to question the effectiveness of those (safety-related) regulations and procedures that are incompatible with operators' culturally driven habits.

According to Helmreich (1994):

> In a culture where group harmony is valued above individual needs, there was probably a tendency to remain silent while hoping that the captain would "save the day." There have been reported instances in other collectivist, high power distance cultures where *crews have chosen to die in a crash rather than disrupt group harmony and authority* and bring accompanying shame upon their family and in-group. (Emphasis added)
>
> High Uncertainty Avoidance may have played a role [in this accident] by locking the crew into a course of action and preventing discussion of alternatives and review of the implications of the current course of action. High Uncertainty Avoidance is associated with a tendency to be inflexible once a decision has been made as a means of avoiding the discomfort associated with uncertainty.

Moreover, the importance of the cultural factors vis-à-vis automation in the aviation industry is further highlighted by two recently published studies. Helmreich and Merritt (1998), in their study of national culture and flightdeck automation, surveyed 5705 pilots across 11 nations and report that "the lack of consensus in automation attitudes, both within and between nations, is disturbing." They conclude that there is a need for clear explication of the philosophy governing the design of automation. Most recently, the U.S. Federal Aviation Administration Human Factors Study Team issued a report (FAA 1996). The team identified several "vulnerabilities" in flight crew management of automation and situation awareness that are caused by a number of interrelated deficiencies in the current aviation system, such as "insufficient understanding and consideration of cultural differences in design, training, operations, and evaluation." They recommend a host of further studies, under the title of "Cultural and Language Differences." Moreover, they include pilots' understanding of automation capabilities and limitations, differences in pilot decision regarding when and whether to use different automation capabilities, the effects of training, and the influence of organizational and national cultural background on decisions to use automation.

4.4. Make Organization and Technology Design Choices That Encourage Innovation

The difficulties discussed in Section 3 suggest that even when a comprehensive technology-organization solution is devised, the unpredictability of the process by which technologies and organizational change unfolds will inevitably lead to unplanned events. Simply creating a portfolio of contingency plans is likely to be insufficient because contingencies to cover all unplanned events cannot be identified in advance. Thus, technology-organization solutions are more likely to be successful when they allow for innovation at the individual and group level. That is, even if careful plans have been made for everything from critical technical features for maintainability to redesigned job descriptions and performance-incentive systems, changes to these features, descriptions, and systems should be not only permitted but encouraged as personnel struggle to make the technology suit their work process.

In a careful analysis of six failed information systems developments, Flowers (1997) found that one of the main reasons for failure was an attitude in which failure, or association with failure, was likely to result in scapegoating or possible loss of employment or else have a severe effect upon the careers of the individual or individuals involved. For example, in the report of the inquiry on the failure of the London Ambulance system, the negative effect on the implementation of a senior manager was noted: the senior manager instilled a fear of failure by being very powerful, with a determination not to be deflected off course. Kelley (1996), in a survey of almost 1000 manufacturing plants, found that group-based employee participation mechanisms that supported the reexamination of old routines and taking advantage of informal shortcuts that employees had worked out on their own were complementary—especially in higher technology firms—to the productive use of information technology in the machining process.

Another example of the need for individual and group-level innovation is a recent study of the implementation of a collaborative technology to allow an interorganizational virtual (i.e., distributed across time and location) team to conceptualize and develop a new product. (Majchrzak et al. 2000). The eight-person team was encouraged to indicate the features they wanted in a collaborative technology. They asked for a central repository on a central server that could capture all types of knowledge (from text to drawings), mechanisms for cataloguing the knowledge for easy retrieval later (such as keywords, dates, author identification, and reference links to previous related entries), mechanisms for being informed when new knowledge relevant to their area of expertise was entered into the knowledge base (e.g., profiling their interests coupled with e-mail notification when an entry fit that profile), ability to link desktop applications interactively to the knowledge base (called hot links), templates for commonly captured knowledge (such as for meeting agendas, meeting minutes, action items, decision rationale), and access anywhere by anyone anytime (24 × 7 access by team members and managers). A system was developed to these specifications. Then the team was encouraged to develop a set of coordination norms for how to conduct their creative engineering design work virtually using the collaborative technology. They created a new work process that would encourage all members of the team (including suppliers and specialists) and external managers to enter all knowledge asynchronously into the knowledge base and for each member then to comment on the entries as need be. The team worked for 10 months and successfully developed a breakthrough product. What is relevant for this discussion is that while the team had the opportunity to create its own technology and work process at the outset, in the end it changed every single one of its norms and most of the ways in which it used the technology. Thus, while there was careful planning prior to the beginning of the team's work—far more planning than would normally be permitted in many organizations today—the team still found it necessary to make changes. The team was fortunate because it were encouraged and able to make those changes as they became necessary. The technology was designed sufficiently flexibly so that entries could be identified using simple searches rather than the complex navigation tools that they thought they might need. Management was sufficiently flexible that when the team asked them to stop using the technology, they obliged. The team's work process was sufficiently flexible that when asynchronous communication proved insufficient, they were able to add a ‘‘meet-me'' teleconference line so that all future encounters could be synchronously conducted using both the collaborative technology and the teleconference capability. Thus, the team succeeded not only because there had been careful planning, but because they could also innovate their work process and the technology as problems arose.

Thus, critical to the success of technology-organization alignment is that the technology-organization solution be designed to encourage localized innovation (Johnson and Rice 1987; Rogers 1995), that is, innovation required to make a particular technology-organization solution work in a particular context with a particular set of people. Characteristics of solutions that allow localized innovation include:

4.4.1. Solutions That Enhance, Not Deskill Workers

When workers are deskilled from a technology-organization solution, they do not have the knowledge to be able to intervene when necessary, identify problems, formulate solutions, and then implement the solutions. Thus, solutions must not permit deskilling. Technologies that avoid deskilling are those that allow workers to understand what the technology is doing and how it is doing it and provide workers with ways to intervene in the process to perform the planning, thinking, and evaluation work, leaving the routine work to the technology (Majchrzak 1988). The collaborative technology used by the virtual team members described in Majchrzak et al. (2000) was entirely open, with no hidden formula, hidden menus, or hidden processing; thus, the team was able to evolve the technology to the point where they could it make useful to them. CNC machines that hide processing logic from the operators are examples of technologies that violate this principle and thus inhibit innovation.

4.4.2. Solutions Should Be Human Centered

A broader proposition than that solutions should not deskill workers is that solutions should be human centered, that is, solutions should focus on how people use information, not simply on how to design a better, faster, cheaper machine. Davenport (1994) lists guidelines for designing human-centered information systems:

- Focus on broad information types, rather than on specific computerized data.
- Emphasize information use and sharing rather than information provision.
- Assume transience of solutions rather than permanence.
- Assume multiple rather than single meanings of terms.
- Continue design and reinvention until desired behavior is achieved enterprise wide rather than stopping the design process when it is done or system is built.
- Build point-specific structures rather than enterprise-wide structures.
- Assume compliance is gained over time through influence rather than dictated policy.
- Let individuals design their own information environments rather than attempt to control those environments.

Human-centered automation initiative is a good example of the technologies that attempt to avoid deskilling of human operators (Billings 1996). Loss of situation awareness, which could have been caused by ''glass cockpit'' and overautomation, have been cited as a major cause of many aviation mishaps (Jentsch et al. 1999; Sarter and Woods 1997). Also, in many cases, because of the aforementioned issues, automation only aggravates the situation and becomes part of the problem rather than the solution. For example, in the context of aviation, automation is even more problematic because it ''amplifies [crew] individual difference'' (Graeber 1994) and ''it amplifies what is good and it amplifies what is bad'' (Wiener 1994). Furthermore, the automated devices themselves still need to be operated and monitored by the very human whose caprice they were designed to avoid. Thus, the error is not eliminated, only relocated. The automation system itself, as a technological entity, has a failure potential that could result in accidents. The problem arises when an automated system fails; it inevitably requires human intervention to fix it in a relatively short time. The same operators who have been out of the loop, may have ''lost the bubble'' (Weick 1990) with respect to cause and effect of the system failure and been deskilled, must now skillfully engage in those very activities that require their contributions to save the day (Meshkati 1996; Roberts and Grabowski 1996).

Deskilling is not necessarily limited to technical skills; blind automation tends to undermine interpersonal skills as well as encourage performance in isolated workstations and ingrains an individualistic culture in the organization. According to an analysis of high-reliability systems such as flight operations on aircraft carriers by Weick and Roberts (1993), a culture that encourages individualism, survival of the fittest, macho heroics, and can-do reactions is often counterproductive and accident prone. Furthermore, interpersonal skills are not a luxury but a necessity in high-reliability organizations.

4.4.3. Solutions That Integrate Across Processes, Not Bifurcate

Technology-organization solutions that create more differentiation between jobs hurt innovation because the problems that arise during implementation are rarely limited to an action that a single person holding a single job can solve. For example, an engineer may not have realized that by speeding up a processing step, she has added a greater queue for inspection, which, if left unresolved, will lead to quicker but more faulty inspections. To solve this problem requires that both quality control and manufacturing work together. For this reason, solutions that are focused on improvements to entire processes—that is, a process-based view of the organization—tend to be more successfully implemented than solutions that are focused on individual functions (Majchrzak and Wang 1996).

4.4.4. Solutions That Encourage Knowledge Recognition, Reuse, and Renewal

Localized innovation can be costly if the solutions themselves are not captured for later potential reuse. That is, if every single context is allowed to experiment through trial and error and generate different ways to handle the problems that arise during solution implementation, and this experimentation is done without the benefit of a knowledge base or technical staff to support knowledge transfer across contexts, the end cost of all the localized solutions can be very high. Thus, each site should have available, and be encouraged to use, a knowledge repository that describes various ways to resolve the different difficulties it is likely to encounter. Moreover, to make such a knowledge repository effective, each context should be encouraged to contribute to the knowledge repository so that future implementations can benefit from their learning (McDermott 1999). For example, a percentage of consultants' pay at Ernst & Young is determined by their contribution to the central knowledge repository (called Ernie) and the uses by other consultants made of their entries.

4.4.5. Solutions Should Decentralize Continuous Improvement

For people to engage in localized innovation, they must both be motivated to do so and have the ability to do it. Motivation can be encouraged by the provision of incentives through reward-and-recognition programs as well as by management offering a consistent message and modeling behavior that everyone should continuously improve what they do. Ability to innovate can be provided through such classic continuous improvement skills as ''five whys,'' Pareto analysis, graphing actual vs. expected outcomes over time, and group problem-solving techniques. Finally, a cycle of continuously evolving on-site experimentation that enables technology and context eventually to ''fit'' should be encouraged (Leonard-Barton 1988). ''Such experimentation can range from scientific investigations of new materials to beta testing a product prototype with potential customers, and from mathematically simulating product performance to studying product aesthetics via physical models'' (Iansiti 1999, p. 3–55). Expecting everyone to become knowledgeable about continuous improvement techniques and then motivating everyone to use those techniques can help to encourage localized innovation.

4.5. Agree on a Change Process for Achieving Alignment

The difficulties identified in Section 3 are better managed when the process by which the technology-organization solution is designed and implemented is an orderly, known, and repeatable process.

People become anxious when they are thrust into chaotic situations over which they have no control and which affect their jobs and possibly their job security and careers (Driver et al. 1993). People are given some sense of control when they know what the process is: how decisions will be made, by whom, and when, and what their role is in the decision-making and implementation process. Sociotechnical systems design suggests the following nine steps in the design and implementation of a technology-organization solution (Emery 1993; Taylor and Felten 1993):

1. Initial scanning of the production (or transformation) system and its environment to identify the main inputs, transforming process, outputs, and types of variances the system will encounter
2. Identification of the main phases of the transformation process
3. Identification of the key process variances and their interrelationships
4. Analysis of the social system, including the organizational structure, responsibility chart for controlling variances, ancillary activities, physical and temporal relationships, extent to which workers share knowledge of each others' roles, payment system, how roles fill psychological needs of employees, and possible areas of maloperation
5. Interviews to learn about people's perceptions of their roles
6. Analysis of relationship between maintenance activities and the transformation system
7. Relationship of transformation system with suppliers, users, and other functional organizations
8. Identification of impact on system of strategic or development plans and general policies
9. Preparation of proposals for change

These nine steps have been created to optimize stakeholder participation in the process, where stakeholders include everyone from managers to engineers, from suppliers to users, from maintainers to operators. A similar process is participative design (Eason 1988; Beyer and Holtzblatt 1998), tenets of which include:

- No one who hasn't managed a database should be allowed to program one.
- People who use the system help design the infrastructure.
- The best information about how a system will be used comes from in-context dialogue and role playing.
- Prototyping is only valuable when it is done cooperatively between users and developers.
- Users are experts about their work and thus are experts about the system; developers are technical consultants.
- Employees must have access to relevant information, must be able to take independent positions on problems, must be able to participate in all decision making, must be able to facilitate rapid prototyping, must have room to make alternative technical and/or organizational arrangements, must have management support but not control, must not have fear of layoffs, must be given adequate time to participate, and must be able to conduct all work in public.

The participative design process first involves establishing a steering committee of managers who will ultimately be responsible for ensuring that adequate resources are allocated to the project. The steering committee is charged with chartering a design team and specifying the boundaries of the redesign effort being considered and the resources management is willing to allocate. The design team then proceeds to work closely with the technical staff first to create a set of alternative organizational and technical solutions and then to assess each one against a set of criteria developed with the steering committee. The selected solutions are then developed by the design and technical personnel, with ever-increasing depth. The concept is that stakeholders are involved before the technology or organizational solutions are derived and then continue to be involved as the design evolves and eventually makes the transition to implementation (Bodker and Gronbaek 1991; Clement and Van den Besselaar 1993; Kensing and Munk-Madsen 1993; Damodaran 1996; Leonard and Rayport 1997).

Both the participative design and STS processes also focus on starting the change process early. Typically, managers wait to worry about alignment after the technology has been designed, and possibly purchased. Because too many organizational and other technology choices have now been constrained, this is too late (Majchrzak 1988). For example, if the data entry screens for an enterprise resource-planning system are designed not to allow clerks to see the next steps in the data flow, and the organizational implications of this design choice have not been considered in advance, clerks may well misunderstand the system and input the wrong data, leading to too many orders being sent to the wrong locations. This is what happened at Yamaha. If the enterprise resource-planning system had been designed simultaneously with the jobs of clerks, then the need to present data flow information would have been more apparent and the costly redesign of the user interface would not have been required. Thus, starting the change process before the technology has been designed is critical to achieve alignment.

Finally, a change process must include all best practices of any project management structure, from metrics and milestones to skilled project managers and contract administration. Too often, the implementation of a technology-organization solution is not given the organizational sanction of a project and instead is decomposed into the various functional responsibilities, with the integration being assigned to somebody's already full plate of responsibilities. A project manager is needed who is responsible for the entire life cycle, from design to implementation to use, and whose performance is based on both outcome metrics (e.g., the extent to which the solution contributed to the business objectives) and process metrics (e.g., did people involved in the design find that their time was well spent?). The project manager needs to report to a steering committee of representatives from each stakeholder community, and the steering committee should hold the program manager accountable to following best-practice project-management principles such as

1. Clear descriptions of specifications and milestones that the solution must meet
2. Risk identification, tracking, and mitigation
3. Early and iterative prototyping with clear testing plans including all stakeholders
4. A formal process for tracking requests for changes in specifications

Finally, given the key role played by the project manager, the job should not be given to just anyone. While small technology-organizational solutions might be handled by an inexperienced project manager, provided there is some formal mentoring, the larger the project, the greater the need for experience. With larger projects, for example, experience is required in contract administration (e.g., devising contracts with service providers that offer them the type of incentives that align their interests with yours), coordination of distributed teams, managing scope creep, and balancing conflicts of interests. These skills are not specific to a technology; but they are specific to project management expertise. The Conference Board, for example, found that problematic enterprise resource planning installations were attributable to the use of poor project-management principles that were specific not to the technology but rather to the scale of the change required (Cooke and Peterson 1998). Thus, a planned change process is critical to the ability to overcome difficulties of alignment.

4.6. Use Decision Aids to Enhance Internal Understanding of Technology-Organizational Alignment

A final recommendation for managing the difficulties of alignment is to use computerized decision aids that have a sufficient knowledge base to offer guidance on how to align technology and organizational options under various contexts. In this way, a company can reduce its reliance on outside consultants while it iteratively strives for better and better alignment. Pacific Gas & Electric seemed to appreciate the need to take technology and organizational development in-house, according to the *Wall Street Journal* (1998b), when it decided to use a team of 300 company employees in its $200 million, four-year effort to rebuild the company's aging computer system. Big-name consulting firms were eschewed, in favor of small consulting firms, but only in supporting roles. As with any simulation package used in industrial engineering, such a decision aid should allow what-if modeling, that is, the ability to try out different technology-organizational solutions and see which are more likely to achieve the desired outcomes at the least cost. Such a decision aid should also incorporate the latest best practices on what other firms have been able to achieve when aligning their organizations and technologies. Finally, such a decision aid should help the firm to conduct a cross-functional comprehensive assessment of the current alignment state of the firm and compare it to the to-be state to identify the high-priority gaps that any new solution should resolve. In this section, we describe two decision aids, TOP Modeler (www.topintegration.com) and iCollaboration software (www.adexa.com).

4.6.1. TOP Modeler

TOP Modeler is a dynamic organization analysis, design, and reengineering tool (Majchrzak and Gasser 2000). TOP Modeler uses a flexible, dynamic modeling framework to deliver a large, well-validated base of scientific and best-practice knowledge on integrating the technology, organizational, and people (TOP) aspects of advanced business enterprises. The current focus of TOP Modeler's knowledge base is advanced manufacturing enterprises, although it can be expanded to other types of enterprises. TOP Modeler's knowledge base was developed with a $10 million, five-year investment of the U.S. Air Force ManTech program, the National Center for Manufacturing Sciences, Digital Equipment Corporation, Texas Instruments, Hewlett-Packard, Hughes, General Motors, and the University of Southern California.

Users have the choice of using TOP Modeler to evaluate their current organization or evaluate their alternative "to-be" future states. Users do this by describing their business strategies and being informed by TOP Modeler of an ideal organizational profile customized to their business strategies. Then users can describe features of their current or proposed future organization and be informed by TOP Modeler of prioritized gaps that need to be closed if business strategies are to be achieved. There are three sets of business strategies contained in TOP Modeler: business objectives, process variance control strategies, and organizational values. TOP Modeler also contains knowledge about the relationships among 11 sets of enterprise features, including information resources, production process characteristics, empowerment characteristics, employee values, customer involvement strategies, skills, reporting structure characteristics, norms, activities, general technology characteristics, and performance measures and rewards.

The TOP Modeler system has a graphical, interactive interface; a large, thorough, state-of-the-art knowledge representation; and a flexible system architecture. TOP Modeler contains a tremendous depth of scientific and best-practice knowledge—including principles of ISO-9000, NCMS's Manufacturing 2000, etc.—on more than 30,000 relationships among strategic and business attributes of the enterprise. It allows users to align, analyze, and prioritize these attributes, working from business strategies to implementation and back. The user of TOP Modeler interacts primarily with a screen that we call the ferris wheel. This screen provides an immediate, intuitive understanding of what it means to have TOP integration in the workplace: TOP integration requires that numerous different aspects of the workplace (e.g., employee values, information, and responsibilities for activities) must all be aligned around core organizational factors (e.g., business objectives) if optimum organizational performance is to be achieved.

TOP Modeler has been used in over 50 applications of organizational redesign, business process redesign, or implementation of new manufacturing technology. The companies that have used it have ranged from very small companies to very large companies, located in the United States, Brazil, and Switzerland. Some of the uses we have been informed about include:

- Use by government-sponsored manufacturing consultants (e.g., Switzerland's CCSO) to help small companies develop strategic plans for restructuring (in one case, the tool helped the consultant understand that the company's initial strategic plan was unlikely to succeed until management agreed to reduce the amount of variation that it allowed in its process).
- Use by large software vendors (e.g., EDS) to help a company decide to not relocate its plant from one foreign country to another (because the expense of closing the "gaps" created by the move was likely to be too high).
- Use by a large manufacturing company (General Motors) to decide whether a joint venture plant was ready to be opened (they decided on delaying the opening because the tool helped to surface differences of opinion in how to manage the workforce).
- Use by a small manufacturing company (Scantron) to decide whether its best practices needed improving (the tool helped the company to discover that while it did indeed have many best practices, it needed to involve the workforce more closer with the supplier and customer base, an action the company subsequently took).
- Use in a large technology change effort at a large manufacturing company (Hewlett-Packard) to help identify the workforce and organizational changes needed for the new production technology to operate correctly (resulting in a substantial improvement in ramp-up time when the new product and production process was introduced).
- Use by a redesign effort of a maintenance crew (at Texas Instruments) to determine that the team-based approach they had envisioned needed several important improvements prior to start-up.
- Use by a strategic planning committee at a large manufacturing company to identify areas of misalignment among elements of a new strategic plan (in this case between quality and throughput time).
- Use by a manufacturing division manager to verify his current business strategy, which had been given to him by his group manager. As a consequence of using TOP Modeler, he discovered that he had agreed to a business objective of new product development without having the authority over the necessary people, skills, and other resources to deliver on that objective. He went back to his group manager to renegotiate these resources.

These are just a few examples of the uses made of TOP Modeler. We believe that with a decision aid, the difficulties of achieving alignment are substantially reduced.

4.6.2. iCollaboration Tool for Technology-Organizational Realignment

Another powerful decision aid for the (intra- and interenterprise) technology-organization alignment that is currently being used by many companies in different industries is the state-of-the-art, Internet-enabled Adexa company's iCollaboration software (www.adexa.com). These manufacturing and service industries include electronics, semiconductor, textile and apparel, and automotive. Adexa tools provide a continuous and dynamic picture of a manufacturing enterprise supply chain and operational status at all times.

iCollaboration software suite enables users to dynamically address and monitor operational planning, materials management, order management and request for quotes (RFQs), strategic and operational change propagation, new product management, collaborative customer and demand planning, and customer satisfaction.

The main feature of the iCollaboration suite include:

1. An integrated/synchronized set of supply chain and operations planning tools that cover the strategic planning (facilities, products, supplies
2. Supply chain planning (sourcing, making, storing)
3. Material and capacity planner; detailed make and deliver, factory planning
4. Reactive dynamic scheduler shop-floor scheduling, dispatch lists

Figure 2 shows a conceptual architecture of Adexa iCollaboration suite and how it interfaces both intra- and interenterprise. The following is a review of some specific areas, examples, and improvements where the iCollaboration tool could help the enterprise in systematic integration of technology into an organizational setting.

4.6.2.1. Operational Planning Operational planning in this context encompasses all the activities: forecasting, demand planning, sourcing, production planning, and shipping. The strategic plan should be aligned with the tactical and operational so that every decision at all levels is consistent; any deviations should be immediately relayed and approved or rejected. The supply chain planner (SCP) and the global strategic planner (GSP) tools create strategic plans that are feasible at lower levels, thus eliminating/minimizing unnecessary monitoring. Each planner (demand, materials, production, etc.) is working from the most current plans and has visibility into what is possible, which leads to linear-empowered organizations. Every change in market demand or supply chain problem is immediately reflected and visible, allowing an optimal replan based on prevalent situations. Materials, manufacturing resources, skills, and shippers all operate in synch, thus eliminating expeditors and facilitators. The GSP, SCP, material and capacity planner (MCP), collaborative demand planner (CDP), collaborative supply planner (CSP), and collaborative enterprise Planner (CEP) tools enable

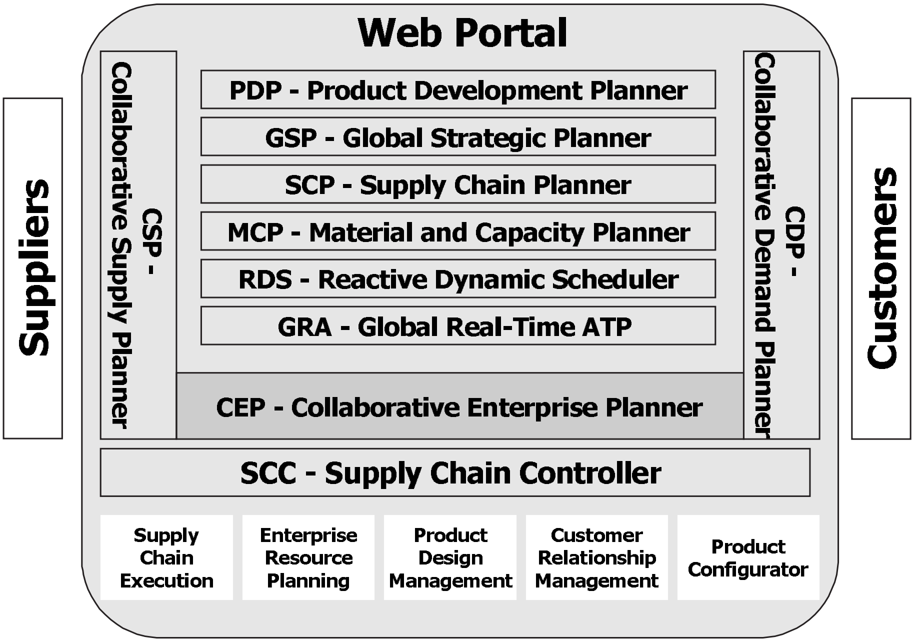

Web Portal
Suppliers
CSP - Collaborative Supply Planner
PDP - Product Development Planner
GSP - Global Strategic Planner
SCP - Supply Chain Planner
MCP - Material and Capacity Planner
RDS - Reactive Dynamic Scheduler
GRA - Global Real-Time ATP
CEP - Collaborative Enterprise Planner
CDP - Collaborative Demand Planner
Customers
SCC - Supply Chain Controller
Supply Chain Execution
Enterprise Resource Planning
Product Design Management
Customer Relationship Management
Product Configurator

Figure 2 The Architecture of the iCollaboration Tool.

the management to align its organizations and coordinate functions (e.g., centralized vs. decentralized planning, ability to manage customer or product lines by one person or department) effectively to meet its business goals.

4.6.2.2. Materials Management The materials organization or individual responsible for raw materials, subassembly suppliers, feeder plants, and finished goods management needs full visibility into all requirements and changes as they happen. Based on the enterprise, these functions may be aligned by product family, facility, or material type. The material managers have access to the latest long-term forecasts and plans, any market changes, order status and changes, effectivities, part substitutions, and all specific rules as they apply to the vendor or supplier, and the latest company rules regards products, materials, customers, or priorities. The enterprise is thus free to align the organization to achieve lowest inventories, lowest material-acquisition costs, best vendor contracts (reduced set of reliable suppliers, quality, etc.), effective end-of-life planning, and reduced obsolescence.

4.6.2.3. Order Management and Request for Quotes (*RFQs*) An organization, which is responsible for the first line of attack on responding to RFQs, order changes, new orders, new customers, should be able to respond rapidly and accurately to delivery capability and costs. More importantly, the response should be based on the current plant loads and reflects the true deliverable lead times and capabilities.

4.6.2.4. Strategic and Operational Change Propagation As is the norm, strategies and operational activities change for various internal or external reasons. Most organizations without access to the right technology manage this change by incurring high costs in terms of additional people both to convey the message of change and to manage and monitor. Visibility and instant change propagation in either direction allow enterprises to respond only when necessary, and they are guided by a system-oriented decision so that their responses are optimal and effective immediately.

4.6.2.5. New Product Management New product development, engineering, materials, sales, and production functions require seamless collaboration. Business processes that take advantage of these functionalities can be implemented so that new product introduction is as much a part of day-to-day operations as the making and delivery of current products. There may not necessarily be a need for any special organizations or staffing to meet new product introductions. These products become akin to new demands on resources; and in fact, with the added visibility and speed of change propagation, the enterprise can develop better-quality products and more of them. This can be done because an enterprise utilizing a tool such as iCollaboration can easily try out more ideas and functions simultaneously, which increases the ability of the enterprise to ramp up production faster

4.6.2.6. Collaborative Customer/Demand Planning The CDP tool allows the customer-facing individuals to evaluate and analyze the demands by sales organizations, geography, product managers, and manufacturing and product planners to interact and control all activities seamlessly and consistent with enterprise goals of maximizing profitability and related corporate strategies. The application of this tool may result in the synchronization among the entire sales and customer relationship teams, in conjunction with customer relationship management (CRM) integration, which would produce customer satisfaction.

4.6.2.7. Customer Satisfaction Customer satisfaction, as measured by product delivery due date performance, accurate order fill rate, response to quotes, and response to changes in orders, can be significantly enhanced by creating an empowered customer facing organization that is enabled and empowered. It should be noted that this issue is one of the most critical determinants of success for today's e-commerce businesses. With the iCollaboration tools, an organization can create customer-facing organizations that may be aligned with full customer responsibility, product responsibility, order responsibility, or any combination of those. These organizations or individuals are independent, do not have to call someone, and yet are in synch with all other supporting organizations.

5. CONCLUSIONS

A review of possible decisions leaves a long list of do's and don'ts for implementing new technology. Some of the more important ones are:

- Don't regard new technology and automation as a quick fix for basic manufacturing or human resource problems; look to the firm's entire human–organization–technology infrastructure as the fix.
- Don't assume that human resource problems can be resolved after the equipment is installed; some of the problems may have to do with the specific equipment selected.

- Do expect that multiple different configurations of matches of human–organization–technology infrastructure elements are equally effective as long as the organization can undergo all the needed changes.
- Do expect to redesign jobs of operators, technical support staff, and supervisors.
- Do involve marketing staff in resources planning.
- Don't look for broad-brush deskilling for skill upgrading of the workforce with new technology; rather, some new skills will be required and others will no longer be needed.
- Don't make direct labor the prime economic target of new technology; the displacement of direct labor is only a small part of the economic benefit of the new technology.
- Do perform a training-needs analysis prior to any employee training relating to the implementation of the new technology; do expect a substantial increase in training cost.
- Do expect that the union–management relationship will change dramatically.
- Do begin facing the dilemma of changing the organizational structure to meet both coordination and differentiation needs.
- Do expect resistance; begin convincing managers and the workforce of the need for change before installing the new technology.
- Do use a multidisciplinary project team to implement any new technology in the workplace.
- Do assess and incorporate all aspects of new technology in the implementation decision making, such as social and environmental impacts.
- Do ensure a thorough understanding of the dimensions of local national culture.
- Do ascertain a determination of the extent of national culture match with those of organizational culture of the technological system (to be implemented).
- Do ensure that the effects of cultural variables on the interactions between human operators and automation in control centers of technological systems are fully considered.

The foregoing ideas concerning technology alignment with organization are of paramount importance for the companies in this emerging era of e-commerce and e-business, which is the ideal test bed for the full implementations of these ideas. The new industrial revolution being precipitated by the advent of the e-commerce phenomenon is probably the most transformative event in human history, with the far-reaching capability to change everything from the way we work to the way we learn and play. The e-commerce industry has empowered customers more than ever and has further required seamless coordination and information exchange among inter- and intraorganizational units responsible for dealing with customers and suppliers. For instance, Dell Computer, a pioneer and a harbinger of the e-commerce industry, has created a fully integrated value chain that allows for a three-way information partnership with its suppliers and customers, thereby improving efficiency and sharing the benefits across the entire chain. (*Economist* 1999). Product managers, manufacturing, and product planners need to interact and control all activities seamlessly with full transparency to the customer. This holistic approach necessitates the utmost efficiency of the entire production life cycle and eventually requires addressing all environmental impacts of e-business.

The new world order for business and industry mandates that technology implementation be comprehensive and must encourage continuous evolution and that it needs tools to help the process of implementation.

Acknowledgments

Najmedin Meshkati would like to acknowledge the contributions of his former USC graduate student and research assistant, Ms. Gail Demel, for the cross-cultural study, undergraduate student Mr. Joseph Deato for his review and analysis of the material, and Dr. Cyrus K. Hadavi of Adexa for providing insightful information on enterprise resource planning and iCollaboration.

REFERENCES

Arthur, J. (1992) "The Link Between Business Strategy and Industrial Relations Systems in American Steel Minimills," *Industrial and Labor Relations Review,* Vol. 45, No. 3, pp. 588–506.

Azimi, H. (1991), *Circles of Underdevelopment in Iranian Economy,* Naey, Teheran (in Farsi).

Barley, S. R. (1986), "Technology as an Occasion for Structuring: Evidence from Observations of CT Scanners and the Social Order of Radiology Departments," *Administrative Science Quarterly,* Vol. 31, pp. 78–108.

Beyer, J. M. (1992), "Metaphors, Misunderstandings and Mischief: A Commentary," *Organization Science,* Vol. 3, No. 4, pp. 467–474.

Beyer, H., and Holtzblatt, K. (1998), *Contextual Design,* Morgan Kaufmann, San Francisco.

Billing, C. E. (1996), *Aviation Automation: The Search for a Human-Centered Approach,* Lawrence Erlbaum Associates, Mahwah, NJ.

Bodker, S., and Gronbaek, K. (1991), ‘‘Cooperative Prototyping: Users and Designers in Mutual Activity,’’ *International Journal of Man–Machine Studies,* Vol. 34, pp. 453–478.

Boeing Commercial Aircraft Group (BCAG) (1993), ‘‘Crew Factor Accidents: Regional Perspective,’’ in *Proceedings of the 22nd Technical Conference of the International Air Transport Association* (*IATA*) *on Human Factors in Aviation* (Montreal, October 4–8), IATA, Montreal, pp. 45–61.

Burgelman, R. A., and Rosenbloom, R. S. (1999), ‘‘Design and Implementation of Technology Strategy: An Evolutionary Perspective,’’ in *The Technology Management Handbook,* R. C. Dorf, Ed., CRC Press, Boca Raton, FL.

Champy, J. (1998), ‘‘What Went Wrong at Oxford Health,’’ *Computerworld,* January 26, p. 78.

Ciborra, C., and Schneider, L. (1990), ‘‘Transforming the Practives and Contexts of Management, Work, and Technology,’’ Paper presented at the Technology and the Future of Work Conference (Stanford, CA, March 28–30).

Clement, A., and Van den Besselaar, P. A. (1993), ‘‘Retrospective Look at PD Projects,’’ *Communications of the ACM,* Vol. 36, No. 4, pp. 29–39.

Computerworld (1997), ‘‘Workflow Software Aids App Development,’’ November 3, p. 57.

Computerworld (1998a), Briefs, June 22, p. 55.

Computerworld (1998b), ‘‘The Bad News,’’ August 10, p. 53.

Condit, P. M. (1994), ‘‘Focusing on the Customer: How Boeing Does It,’’ *Research-Technology Management,* Vol. 37, No. 1, pp. 33–37.

Contractor, N., and Eisenberg, E. (1990), ‘‘Communication Networks and New Media in Organizations,’’ in *Organizational and Communication Technology,* J. Fulk and C. Steinfield, Eds., Sage, Newbury Park, CA.

Cooke, D. P., and Peterson, W. J. (1998), *SAP Implementation: Strategies and Results,* Conference Board, New York.

Criswell, H. (1998), ‘‘Human System: The People and Politics of CIM,’’ Paper presented at UTOFACT Conference (Chicago).

Damodaran, L. (1996), ‘‘User Involvement in the System Design Process,’’ *Behavior and Information Technology,* Vol. 15, No. 6, pp. 363–377.

D’Aveni, R. A., and Gunther, R. (1994), *Hypercompetition: Managing the Dynamics of Strategic Maneuvering,* Free Press, New York.

Davenport, T. (1994), ‘‘Saving IT’s Soul: Human-Centered Information Management,’’ *Harvard Business Review,* March–April, Vol. 72, pp. 119–131.

Demel, G. (1991), ‘‘Influences of Culture on the Performance of Manufacturing Plants of American Multinational Corporations in Other Countries: A Macroergonomics Analysis,’’ Master’s Thesis, Institute of Safety and Systems Management, University of Southern California, Los Angeles.

Demel, G., and Meshkati, N. (1989), ‘‘Requisite variety: A concept to analyze the effects of cultural context for technology transfer,’’ in *Proceedings of the 33rd Annual Meeting of the Human Factors Society.* Santa Monica, CA: Human Factors Society, pp. 765–769.

Dertouzos, M., Lester, R., Salon, R., and The MIT Commission on Industrial Productivity (1988), *Made in America: Regaining the Productive Edge,* MIT Press, Cambridge, MA.

DeSanctis, G., and Poole, M. (1994), ‘‘Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory,’’ *Organization Science,* Vol. 5, No. 2, pp. 121–147.

Driver, M. J., Brousseau, K. R., and Hunsaker, P. L. (1993), *The Dynamic Decision Maker: Five Decision Styles for Executive and Business Success,* Jossey-Bass, San Francisco.

Drucker, P. F. (1988), ‘‘The Coming of the New Organization,’’ *Harvard Business Review,* Vol. 66, January–February, 45–53.

Eason, K. (1988), Information Technology and Organizational Change, Taylor & Francis, London.

Economist, The (1990), ‘‘Detroit’s Big Three: Are America’s Carmakers Headed for the Junkyard?’’ April 14.

Economist, The (1992), ‘‘Cogito, Ergo Something (Computers and human intelligence: A survey of artificial intelligence),’’ March 14.

Economist, The (1999), ‘‘Business and the Internet: The Net Imperative,’’ June 26.

Electric Power Research Institute (EPRI) (1984), *Commercial Aviation Experience of Value to the Nuclear Industry,* Prepared by Los Alamos Technical Associates, Inc. EPRI NP.3364, January, EPRI, Palo Alto, CA.

Emery, F. (1993), "The Nine-Step Model," in *The Social Engagement of Social Science: A Tavistock Anthology,* E. Trist and H. Murray, Eds., University of Pennsylvania Press, Philadelphia.

Ettlie, J. (1986),"Implementing Manufacturing Technologies: Lessons from Experience," in *Managing Technological Innovation: Organizational Strategies for Implementing Advanced Manufacturing Technologies,* D. D. Davis, Ed., Jossey-Bass, San Francisco.

Federal Aviation Administration (FAA) (1996), "The Interfaces Between Flightcrews and Modern Flight Deck Systems," FAA Washington, DC.

Flowers, S. (1997), "Information Systems Failure: Identifying the Critical Failure Factors," *Failure and Lessons Learned in Information Technology Management,* Vol. 1, pp. 19–29.

Gibbs, W. (1994), "Software's Chronic Crisis," *Scientific American,* March, pp. 72–81.

Giddens, A. (1994), *The Constitution of Society: Outline of the Theory of Structuration,* University of California Press, Berkeley.

Goldman, S. L., Nagel, R. N., and Preiss, K. (1995), *Agile Competitors and Virtual Organizations,* Van Nostrand Reinhold, New York.

Graeber, R. C. (1994), "Integrating Human Factors Knowledge into Automated Flight Deck Design," Invited presentation at the International Civil Aviation Organization (ICAO) Flight Safety and Human Factors Seminar (Amsterdam, May 18).

Grayson, C. (1990), "Strategic Leadership," Paper presented at the Conference on Technology and the Future of Work (Stanford, CA, March 28–30).

Hall, G., Rosenthal, J., and Wade, J. (1993), "How to Make Reengineering Really Work," *Harvard Business Review,* Vol. 71, November–December, pp. 119–133.

Helmreich, R. L. (1994), "Anatomy of a System Accident: Avianca Flight 052," *International Journal of Aviation Psychology,* Vol. 4, No. 3, pp. 265–284.

Helmreich, R. L., and Sherman, P. (1994), "Flightcrew Perspective on Automation: A Cross-Cultural Perspective." *Report of the Seventh ICAO Flight Safety and Human Factors Regional Seminar,* Montreal, Canada: International Civil Aviation Organization (ICAO), pp. 442–453.

Helmreich, R. L., and Merritt, A. (1998), *Culture at Work in Aviation and Medicine: National, Organizational, and Professional Influences,* Ashgate, Brookfield, VT.

Hofstede, G. (1980a), *Culture's Consequences,* Sage, Beverly Hills, CA.

Hofstede, G. (1980b), "Motivation, Leadership, and Organization: Do American Theories Apply Abroad?" *Organizational Dynamics,* Vol. 9, Summer, pp. 42–63.

Iansiti, M. (1999), "Technology Integration: Matching Technology and Context," in *The Technology Management Handbook,* R. C. Dorf, Ed., CRC Press, Boca Raton, FL.

Information Week (1997), "California Targets March for Electric Utility System," December 30.

Information Week (1998), "Andersen Sued on R/3," July 6, p. 136.

International Atomic Energy Agency (IAEA) (1991), *Safety Culture,* Safety Series No. 75-INSAG-4, IAEA, Vienna.

Jackson, J. M. (1960), "Structural Characteristics of Norms," in *The Dynamics of Instructional Groups: Socio-Psychological Aspects of Teaching and Learning,* M. B. Henry, Ed., University of Chicago Press, Chicago.

Jaikumar, R. (1986), "Post-Industrial Manufacturing," *Harvard Business Review,* November–December.

Jambekar, A. B., and Nelson, P. A. (1996), "Barriers to Implementation of a Structure for Managing Technology," in *Handbook of Technology Management,* G. H. Gaynor, Ed., McGraw-Hill, New York.

Jentsch, F., Barnett, J, Bowers, C. A., and Salas, E. (1999), "Who Is Flying This Plane Anyway? What Mishaps Tell Us about Crew Member Role Assignment and Air Crew Situation Awareness," *Human Factors,* Vol. 41, No. 1, pp. 1–14.

Johnson, B., and Rice, R. E. (1987), *Managing Organizational Innovation,* Columbia University Press, New York.

Johnson, H. T., and Kaplan, R. S. (1987), *Relevance Lost: The Rise and Fall of Management Accounting,* Harvard Business School Press, Boston.

Johnston, A. N. (1993), "CRM: Cross-cultural Perspectives," in *Cockpit Resource Management,* E. L. Wiener, B. G. Kanki, and R. L. Helmreich, Eds., Academic Press, San Diego, pp. 367–397.

Kalb, B. (1987), "Automation's Myth: Is CIM Threatening Today's Management?" *Automotive Industries,* December.

Kahneman, D., Slovic, P., and Tversky, A. (1982), *Judgment under Uncertainty: Heuristic and Biases,* Cambridge University Press, New York.

Kanz, J., and Lam, D. (1996), ‘‘Technology, Strategy, and Competitiveness,’’ in *Handbook of Technology Management,* G. H. Gaynor, Ed., McGraw-Hill, New York.

Kelley, M. R. (1996), ‘‘Participative Bureaucracy and Productivity in the Machined Products Sector,’’ *Industrial Relations,* Vol. 35, No. 3, pp. 374–398.

Kensing, F., and Munk-Madsen, A. (1993), ‘‘PD: Structure in the Toolbox,’’ *Communications of the ACM,* Vol. 36, No. 4, pp. 78–85.

Kotter, J. P., and Heskett, J. L. (1992), *Corporate Culture and Performance,* Free Press, New York.

Lammers, C. J., and Hickson, D. J. (1979), ‘‘A Cross-national and Cross-institutional Typology of Organizations,’’ In *Organizations,* C. J. Lammers and D. J. Hickson, Eds., Routledge & Kegan Paul, London, pp. 420–434.

Landis, D., Triandis, H. C., and Adamopoulos, J. (1978), ‘‘Habit and Behavioral Intentions as Predictors of Social Behavior,’’ *Journal of Social Psychology,* Vol. 106, pp. 227–237.

Lawrence, P. R., and Lorsch, J. W. (1967), *Organization and Environment: Managing Differentiation and Integration,* Graduate School of Business Administration, Harvard University, Boston.

Leonard-Barton, D. (1995), *Wellsprings of Knowledge: Building and Sustaining the Sources of Innovation,* Harvard Business School Press, Boston.

Leonard-Barton, D. (1998), ‘‘Implementation as Mutual Adaptation of Technology and Organization,’’ *Research Policy,* Vol. 17, No. 5, pp. 251–267.

Leonard, D., and Rayport, J. (1997), ‘‘Spark Innovation through Empathic Design,’’ *Harvard Business Review,* Vol. 75, November–December, pp. 103–113.

Long, R. J. (1989), ‘‘Human Issues in New Office Technology,’’ in *Computers in the Human Context: Information Technology, Productivity, and People,* T. Forester, Ed., MIT Press, Cambridge, MA.

Los Angeles Times (1997), ‘‘Snarled Child Support Computer Project Dies,’’ November 21, p. A1.

Los Angeles Times (1999), ‘‘$17 Million Later, Tuttle Scraps Computer Overhaul,’’ p. B-1.

MacDuffie, J. P. (1995), ‘‘Human Resource Bundles and Manufacturing Performance: Organizational Logic and Flexible Production Systems in the World Auto Industry,’’ *Industrial and Labor Relations Review,* Vol. 48, No. 2, pp. 199–221.

Majchrzak, A. (1988), *The Human Side of Factory Automation,* Jossey-Bass, San Francisco.

Majchrzak, A. (1997), ‘‘What to Do When You Can’t Have It All: Toward a Theory of Sociotechnical Dependencies,’’ *Human Relations,* Vol. 50, No. 5, pp. 535–565.

Majchrzak, A., and Gasser, L. (2000), ‘‘TOP-MODELER: Supporting Complex Strategic and Operational Decisionmaking,’’ *Information, Knowledge, Systems Management,* Vol. 2, No. 1.

Majchrzak, A., and Finley, L. (1995), ‘‘A Practical Theory and Tool for Specifying Sociotechnical Requirements to Achieve Organizational Effectiveness,’’ In *The Symbiosis of Work and Technology,* J. Benders, J. deHaan and D. Bennett, Eds., Taylor & Francis, London.

Majchrzak, A., and Wang, Q. (1996), ‘‘Breaking the Functional Mindset in Process Organizations,’’ *Harvard Business Review,* Vol. 74, No. 5, September–October, pp. 92–99.

Majchrzak, A., Rice, R. E., Malhotra, A., King, N., and Ba, S. (2000), ‘‘Technology Adaptation: The Case of a Computer-Supported Inter-Organizational Virtual Team,’’ *Management Information Sciences Quarterly,* Vol. 24, No. 4, pp. 569–600.

Manufacturing Studies Board (MSB) (1988), Committee on the Effective Implementation of Advanced Manufacturing Technology, National Research Council, National Academy of Sciences, *Human Resource Practice for Implementing Advanced Manufacturing Technology,* National Academy Press, Washington, DC.

Martino, J. P. (1983), *Technological Forecasting for Decision Making,* 2nd ed., North Holland, New York.

McDermott, R. (1999). ‘‘Why Information Technology Inspired but Cannot Deliver Knowledge Management,’’ *California Management Review,* Vol. 41, No. 4, pp. 103–117.

Meshkati, N. (1996), ‘‘Organizational and Safety Factors in Automated Oil and Gas Pipeline Systems,’’ in *Automation and Human Performance: Theory and Applications,* R. Parasuraman and M. Mouloua, Eds., Erlbaum, Hillsdale, NJ, pp. 427–446.

Meshkati, N., Buller, B. J., and Azadeh, M. A. (1994), *Integration of Workstation, Job, and Team Structure Design in the Control Rooms of Nuclear Power Plants: Experimental and simultation Studies of Operators’ Decision Styles and Crew Composition While Using Ecological and Traditional User Interfaces,* Vol. 1, Grant report prepared for the U.S. Nuclear Regulatory Commission, (Grant # NRC-04-91-102), University of Southern California, Los Angeles.

Mitroff, I. I., and Kilmann, R. H. (1984), *Corporate Tragedies: Product Tampering, Sabotage, and Other Catastrophes,* Praeger, New York.

National Transportation Safety Board (NTSB) (1991), *Aircraft Accident Report: Avianca, the Airline of Columbia, Boeing 707-321B,* HK 2016 Fuel Exhaustion, Cove Neck, New York, January 25, 1990, Report No. NTSB-AAR-91-04, NTSB, Washington, DC.

Office of Technology Assessment (OTA) (1984), *Computerized Manufacturing Automation,* Library of Congress #84-601053, OTA. Washington, DC.

O'Reilly, C. A. (1989), "Corporations, Culture, and Commitment: Motivation and Social Control in Organizations," *California Management Review,* Vol. 31, No. 4, pp. 9–25.

Orlikowski, W. J. (1992), "The Duality of Technology: Rethinking the Concept of Technology in Organizations," *Organization Science,* Vol. 3, No. 3, pp. 398–427.

Orlikowski, W. J., and Robey, D. (1991), "Information Technology and the Structuring of Organizations," *Information Systems Research,* Vol. 2, pp. 143–169.

Orlikowski, W. J., Yates, J., Okamura, K., and Fujimoto, M. (1995), "Shaping Electronic Communication: The Metastructuring of Technology in the Context of Use," *Organization Science,* Vol. 6, No. 4, pp. 423–443.

Orlikowski, W. J., and Yates, J. (1994), "Genre Repertoire: The Structuring of Communication in Organizations," *Administrative Science Quarterly,* Vol. 39, No. 4, pp. 541–574.

Osterman, P. (1989), in *The Challenge of New Technology to Labor–Management Relations* BLMR 135. Summary of a conference sponsored by the Bureau of Labor–Management Relations and Cooperative Programs, U.S. Department of Labor, Washington, DC.

Osterman, P. (1994), "How Common is Workplace Transformation and Who Adopts It?" *Industrial and Labor Relations Review,* Vol. 47, No. 2, pp. 173–188.

Pil, F., and MacDuffie, J. P. (1996), "The Adoption of High-Involvement Work Practices," *Industrial Relations,* Vol. 35, No. 3, pp. 423–455.

Pine, B. J. (1993), *Mass Customization: The New Frontier in Business Competition,* Harvard Business School Press, Boston.

Poole, M. S., and DeSanctis, G. (1990), "Understanding the Use of Group Decision Support Systems: The Theory of Adaptive Structuration," in *Organizational and Communication Technology,* J. Fulk and C. Steinfeld, Eds., Sage, Newbury Park, CA, pp. 173–193.

Porter, M. E. (1985), *Competitive Advantage,* Free Press, New York.

Rasmussen, J. (1992), Personal communication.

Reynolds, P. D. (1986), "Organizational Culture as Related to Industry, Position and Performance," *Journal of Management Studies,* Vol. 23, pp. 333–345.

Rice, R. (1994), "Network Analysis and Computer-Mediated Communication Systems," in *Advances in Social and Behavioral Science from Social Network Analysis,* L. Wasserman and J. Galaskiewicz, Eds., Sage, Newbury Park, CA, pp. 167–203.

Rice, R., and Gattiker, U. (1999), "New Media and Organizational Structuring of Meaning and Relations," in *New Handbook of Organizational Communication,* F. Jablin and L. Putnam, Eds., Sage, Newbury Park, CA.

Roberts, K. H., and Grabowski, M. (1996), "Organization, Technology, and Structuring," in *Handbook of Organization Studies,* S. R. Clegg, C. Hardy, and W. R. Nord, Eds., Sage, Thousand Oaks, CA, pp. 409–423.

Rogers, E. (1995), *Diffusion of Innovations,* 4th Ed., Free Press, New York.

Sarter, N., and Woods, D. D. (1997), "Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-320," *Human Factors,* Vol. 39, No. 4, pp. 553–569.

Schein, E. H. (1985), *Organizational Culture and Leadership,* Jossey-Bass, San Francisco.

Schlie, T. W. (1996), "The Contribution of Technology to Competitive Advantage," in *Handbook of Technology Management,* G. H. Gaynor, Ed., McGraw-Hill, New York.

Schlie, T. W., and Goldhar, J. D. (1995), "Advanced Manufacturing and New Direction for Competitive Strategy," *Journal of Business Research,* Vol. 33, No. 2, pp. 13–114.

Sherman, J. D., and Sounder, W. E. (1996), "Factors Influencing Effective Integration in Technical Organizations," in *Handbook of Technology Management,* G. H. Gaynor, Ed., McGraw-Hill, New York.

Sweat, J. (1999), "New Products Help Companies Link Apps Across and Beyond Their Organization," *Information Week,* June 14, http://www.informationweek.com/738/integrate.htm.

Taylor, J. C., and Felten, D. F. (1993), *Performance by Design,* Prentice Hall, Englewood Cliffs, NJ.

Twiss, B. (1980), *Managing Technological Innovation,* Longman Group, Harlow, Essex.

Tyre, M. J., and Orlikowski, W. J. (1994), ‘‘Windows of Opportunity—Temporal Patterns of Technological Adaptation in Organizations,’’ *Organization Science,* Vol. 5, No. 1, pp. 98–118.

Unterweger, P. (1988), ‘‘The Human Factor in the Factory of the Future,’’ in *Success Factors for Implementing Change: A Manufacturing Viewpoint,* K. Balache, Ed., Society of Manufacturing Engineers, Dearborn, MI.

Wall Street Journal (1998a), ‘‘At Oxford Health Financial ‘Controls’ Were out of Control,’’ April 29, p. 1.

Wall Street Journal, The (1998b), ‘‘Some Firms, Let Down by Costly Computers, Opt to ‘De-engineer,’ ’’ April 30, p. 1.

Wall Street Journal (1999), ‘‘Europe’s SAP Scrambles to Stem Big Glitches,’’ November, p. A26.

Weick, K. E. (1990), ‘‘Technology as Equivoque: Sensemaking in New Technologies,’’ in *Technology and Organizations,* P. S. Goodman and L. Sproull, Eds., Jossey-Bass, San Francisco.

Weick, K. E., and Roberts, H. K. (1993), ‘‘Collective Mind in Organizations: Heedful Interrelating on Flight Decks,’’ *Administrative Science Quarterly,* Vol. 38, pp. 357–381.

Whiston, T. G. (1996), ‘‘The Need for Interdisciplinary Endeavor and Improved Interfunctional Relationships,’’ in *Handbook of Technology Management,* G. H. Gaynor, Ed., McGraw-Hill, New York.

Wiener, E. (1994), ‘‘Integrating Practices and Procedures into Organizational Policies and Philosophies,’’ Invited presentation at the International Civil Aviation Organization (ICAO) Flight Safety and Human Factors Seminar (Amsterdam, May 18).

Works, M. (1987), ‘‘Cost Justification and New Technology Addressing Management’s No. 1 to the Funding of CIM,’’ in *A Program Guide for CIM Implementation,* 2nd Ed., L. Bertain and L. Hales, Eds., SME, Dearborn, MI.

CHAPTER 37

Teams and Team Management and Leadership

François Sainfort
Georgia Institute of Technology

Alvaro D. Taveira
University of Wisconsin-Whitewater

Neeraj K. Arora
National Cancer Institute

Michael J. Smith
University of Wisconsin-Madison

1. INTRODUCTION **975**
2. TEAMS: TYPES, CHARACTERISTICS, AND DESIGN **976**
  2.1. Types of Teams 976
  2.2. Characteristics of Teams 977
  2.3. Team Design 977
3. QUALITY IMPROVEMENT AND PARTICIPATORY ERGONOMICS TEAMS **978**
  3.1. Teams in the Context of Quality Improvement 978
  3.2. Participatory Ergonomics 980
4. KEY SUCCESS FACTORS FOR EFFECTIVE TEAMS **981**
5. TEAM EFFECTIVENESS **983**
  5.1. Structure Variables 983
    5.1.1. Organizational Characteristics 983
    5.1.2. Team Characteristics 984
    5.1.3. Task Characteristics 985
  5.2. Process Variables 985
    5.2.1. Task Issues 986
    5.2.2. Relationship Issues 986
    5.2.3. Leadership 987
  5.3. Outcome Variables 987
6. IMPACT OF TEAMS **987**
  6.1. Impact on Management 988
  6.2. Impact on Employees 988
REFERENCES **989**

This chapter examines the state of the research and practice on teamwork. We review and present main guidelines for the choice of teams, the effective utilization of teams in organizations, as well as available approaches to assess team outcomes. The impacts of teams on different segments of the organizations and future challenges for research and practice are discussed.

1. INTRODUCTION

Teamwork has been recommended as an organizational design feature for many years, as a way to improve productivity, quality, and employee satisfaction (Lawler and Mohrman 1987). This is especially true in today's organizational environment, with increased global competition and a more demanding workforce (Katzenbach and Smith 1993). The increased attention to teams has become widespread particularly in the context of total quality management (TQM) or quality improvement (QI), which relies heavily on teamwork (Dean and Bowen 1994; Deming 1986).

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

Teamwork represents one form of work organization that can have large positive and/or negative effects on the different elements of the work system and on human outcomes, such as performance, attitudes, well being, and health. Conceiving work as a social system and advocating the necessity of both technical and social performance optimizations as necessary for organizational effectiveness, the sociotechnical theory provides several arguments and examples supporting teamwork. The germinal study, which gave the essential evidence for the development of the sociotechnical field, was a team experience observed in the English mining industry during the 1950s. In this new form of work organization, a set of relatively autonomous work groups performed a complete collection of tasks interchanging roles and shifts and regulating their affairs with a minimum of supervision. This experience was considered a way of recovering the group cohesion and self-regulation concomitantly with higher level of mechanization (Trist 1981). The group had the power to participate in decisions concerning work arrangements, and these changes resulted in increased cooperation between task groups, personal commitment from the participants, and reduction in absenteeism and the number of accidents.

A GM assembly plant located in Fremont, California was considered until 1982, the year it ended operations, the worst plant in the GM system and in the auto industry as whole. For years, the plant presented dismay levels of quality, low productivity, and prevalent problems of absenteeism and turnover. The plant was reopened two years later under a new joint venture with Toyota. Changes focusing primarily on the relationship between workers and management, the organizational structure, and the widespread use of teamwork transformed the plant in one of the most productive ones of the GM system, with high levels of employee satisfaction and very low levels of absenteeism (Levine, 1995).

Examples like those above illustrate the advantages of using teams to perform a variety of organizational assignments and tasks. Supporters of teamwork have promoted teamwork on many grounds, highlighting its potential for increased productivity, quality, job satisfaction, organizational commitment, and increased acceptance of change, among others. Teamwork is the preferred strategy to increase employee involvement in the workplace. Indeed, terms such as *participatory management, employee involvement,* and participation are frequently equated with teamwork. According to Lawler (1986), employee involvement affects five major determinants of organizational effectiveness: motivation, satisfaction, acceptance of change, problem solving, and communication. Lawler states that employee involvement can create a connection between a particular level of performance and the perception of a favorable consequence. Involvement in teams can provide rewards beyond those allocated by the organization, such as money and promotion: it can supply intrinsic rewards, that is, accomplishment, personal growth, and so on. Similarly, Lawler argues that allowing people to participate in the definition of the procedures and methods utilized in their daily activities is an effective way to improve those methods and can motivate employees to produce a better-quality job. Teams also can ease the process of implementation of organizational changes and avoid the ''not-invented-here'' perception and create commitment with the implementation of change.

It has been argued that in many circumstances the effort of a group of people generates effective solutions that would not be produced by the same individuals working independently. This superior outcome would result not only from the greater pooled knowledge available to the group members, but also from the interaction process among them, from the mutual influence on each other's thinking. This process has been termed *collective intelligence* (Wechsler 1971) and refers to a mode of cooperative thinking that goes beyond simple collective behavior. Finally, system reliability is also assumed to improve through employee participation since it increases the human knowledge variety and enables the workers to understand their role in making the system more efficient and safer.

While there are a number of positive elements associated with teamwork in general, there are also some potential negative elements. Furthermore, these elements are not constant but rather depend on the type of team as well as the organizational environment in which teams operate. In the next section, we propose a typology of teams, review key characteristics of teams, and address the issue of designing teams.

2. TEAMS: TYPES, CHARACTERISTICS, AND DESIGN

2.1. Types of Teams

Sundstrom et al. (1990, p. 120) have defined work teams as ''interdependent collections of individuals who share responsibility for specific outcomes for their organizations.'' Teams can vary a great deal in the way they are designed, managed, and implemented. Various forms of teamwork have been proposed and applied, from temporary teams (e.g., ad hoc committees, quality improvement teams, project teams) to permanent teams (e.g., manufacturing crews, maintenance crews). Temporary teams are usually set up when some problem occurs or some change needs to be implemented and/or to better manage the change process. Teams also vary greatly in terms of the amount of autonomy and authority they have. For example, manager-led teams have responsibility only for the execution of work (Medsker and Campion 1997). On the other hand, self-managed teams can have a large amount of autonomy and decide on issues such as work organization and performance monitoring. Somewhere in between, semiautonomous work groups will experience limited degrees of autonomy and decision making over such issues. Finally, teams vary significantly in terms of the task or the nature of work to be performed. Sundstrom et al. (1990) propose four broad categories of work team applications: (1) advice and involvement, (2) production and service, (3) projects and development, and (4) action and negotiation.

2.2. Characteristics of Teams

Lawler (1986) lists the following characteristics of work teams: membership, work area coverage, training, meetings, supervision, reward systems, decision-making responsibility, installation process, and size. Sundstrom et al. (1990) have proposed that work team effectiveness is dynamically interrelated with organizational context, boundaries and team development. Hackman (1987) has proposed a normative model of group effectiveness. The model identifies three process criteria: effort, knowledge, and the appropriateness of task performance strategies. Increases in these three criteria, given task configurations, should improve the overall effectiveness of the group. According to Hackman (1987), the basic levers to change the process criteria are group design, organizational context, and synergy.

2.3. Team Design

For the design of work groups, three different levels of criteria need to be considered. First, global issues on strategy selection need to be defined—that is, decisions regarding the appropriateness of teamwork for the situation at hand, what type of team would be most adequate, and the amount of authority/autonomy granted to the team need to be made. Second, the specifics of the group design and mechanics need to be decided upon, including matters of size and composition/membership, work area coverage or tasks, and coordination. Finally, in agreement with team members, issues related to the team performance and duration need to be defined. This includes reward systems, duration, and performance/effectiveness assessment, all issues that are determinant in making continuation, change, or termination decisions throughout the life of the team.

Decisions at the strategic level are critical and difficult to make. The adequacy of teamwork for a given situation can be assessed through criteria depicted in Tables 3 and 4 in Medsker and Campion (2000) (Chapter 33 of this Handbook). As proposed by Medsker and Campion, Table 3 summarizes advantages and disadvantages of team design as compared to individual job design. Table 4 in Medsker and Campion proposes a list of systematic questions for which affirmative answers support the use of teamwork for the situation at hand. The choice of the appropriate type of team depends on the application but also on the risks and opportunities offered by the different team configurations, as shown in Table 1. The next decision is on amount of authority/autonomy provided to the group. This decision is difficult to make and depends on other characteristics such as the organizational

TABLE 1 Risks and Opportunities for Work Teams

Teams	Risks	Opportunities
Top management teams	Underbounded; absence of organizational context	Self-designing; influence over key organizational conditions
Task Forces	Team and work both new	Clear purpose and deadline
Professional support groups	Dependency on others for work	Using and honing professional expertise
Performing groups	Skimpy organizational supports	Play that is fueled by competition and/or audiences
Human service teams	Emotional drain; struggle for control	Inherent significance of helping people
Customer service teams	Loss of involvement with parent organization	Bridging between parent organization and its customers
Production teams	Retreat into technology; insulation from end users	Continuity of work; ability to home both the team design and the product

Adapted from J. R. Hackman, “Creating More Effective Work Groups in Organizations,” in *Groups That Work* (*And Those That Don’t*), J. R. Hackman, Ed., copyright © 1990 Jossey-Bass, Inc., a subsidiary of John Wiley & Sons, Inc. Reprinted by permission.

culture, the nature of the team's tasks, the skills and knowledge of the team members, and the training received and to be received. Such a decision has important implications for management and employee involvement, which will be addressed in Section 6.

Decisions at the tactical level, that is, the specifics of the group design and mechanics are usually easier to make and are negotiated with team members. This includes matters of size and composition/membership, work area coverage or tasks, and coordination mechanisms. For many teams, the optimal size is difficult to determine. In fact, a variety of factors may affect team size. Obviously the primary factor is the size and scope of a required project or set of tasks. However, several other factors can influence team size (it should be noted that all factors are not necessarily applicable to all types of teams). Factors affecting team size include:

- Amount of work to be done
- Amount of time available to do the work
- Amount of work any one person can do in the available time
- Differentiation of tasks to be performed in sequence
- Number of integrated tasks required
- Balancing of tasks assignments
- Cycle time required
- Variety of skills, competences, knowledge bases required
- Need for reserve team members
- Technological capabilities

Finally, at the third level, decisions regarding team performance and duration should be negotiated and made prior to engaging teamwork. Section 5 provides a comprehensive, structured list of variables affecting and/or defining team performance. Such characteristics can be used to develop a system to measure and monitor team performance over time.

As mentioned above, teams are widely used in today's organizational environment, with increased global competition and a more demanding workforce (Katzenbach and Smith 1993). The next section describes two important current applications of teamwork.

3. QUALITY IMPROVEMENT AND PARTICIPATORY ERGONOMICS TEAMS

Teamwork has been the backbone of quality improvement. More recently, teamwork has been used in the context of participatory ergonomics (PE). However, while QI teams primarily focus on activities related to identifying, designing, and implementing improvements in both work processes and products/services, PE teams primarily focus on improvement of working conditions. The following sections review the state of the art for both applications of teamwork.

3.1. Teams in the Context of Quality Improvement

Employee participation, particularly through teamwork, is one of the essential foundations of QI. Different from many other management approaches that present teamwork effectiveness as contingent to several aspects, QI supports the use of teams without any specific provisions (Dean and Bowen 1994).

A variety of quality-related teams can exist within organizations. Kano (1993) categorizes teams into three types: (1) functional teams, which are ongoing voluntary problem-solving groups made up of workers in the same workplace; (2) quality building-in-process teams, in which a set of related jobs are shared, with the goal of building quality into a product during the assembling process; and (3) task/project teams, which are ad hoc groups comprised of staff or line managers, who disband once the task is completed.

Quality circle (QC) is one of the most widely discussed and adopted forms of teamwork (Cotton 1993). QCs are project/problem-solving teams that have been defined as small groups of volunteers from the same work area who meet regularly to identify, analyze, and solve quality-related problems in their area of responsibility (Wayne et al. 1986). These groups usually consist of 8 to 10 employees who meet on a regular basis, such as one hour per week. In many QCs, the supervisor is designated as the circle leader. Formal training in problem-solving techniques is often a part of circle meetings.

The claimed benefits of QCs include quality and cost awareness; reduction in conflict and improved communications; higher morale, motivation, and productivity; and cost savings (Head et al. 1986). The effect of this type of teamwork on employee attitudes is assumed to be the primary reason for their success (Head et al. 1986). Marks et al. (1986) propose that QC participation will lead to enriched jobs, with employees experiencing work as more meaningful, obtaining greater knowledge of the results of their work, and gaining a greater sense of responsibility. Enriched jobs are the result of increased skill variety, task identity, task significance, autonomy, and feedback (Hackman and Oldham 1980). These job characteristics may then be related to outcomes that include higher job satisfaction and motivation.

Teamwork can enrich jobs through different mechanisms. Skill variety can be increased from both project activity and one's role within the team. Activities such as data collection and analysis, problem solving, presenting information to groups, and group decision making are key elements in quality-related teamwork, which may not be part of workers' daily routines. Team projects can typically be expected to increase the variety of stimuli to which employees are exposed (Rafaeli 1985).

An essential element of quality-related teamwork is providing feedback to participants (Head et al. 1987). Feedback may be provided through data collection conducted by the team. Interaction within the group and with outside groups (e.g., management, customers) is another potential source of feedback. Team activity increases the frequency of communication among coworkers and supervisors, and may include those outside of the team as well (Buch and Raban 1990; Rafaeli 1985; Marks et al. 1986).

At the team level, the worker may experience a degree of control that is higher than one would expect at the individual level. Autonomy is actually expected to increase among quality team members (Head et al. 1987; Rafaeli 1985). Teams may have control over the content and sequence of activities. In addition, team members may be given control over specific tasks within the group, such as data collection and analysis.

However, the data regarding these hypothesized relationships are somewhat inconsistent. Marks et al. (1986) found that QC participation in a manufacturing firm influenced work attitudes that were directly related to QC involvement: participation, decision making, group communication, worthwhile accomplishments, and enhancing the opportunities and skills needed for advancement. There was no improvement found in job challenge, personal responsibility, and overall job satisfaction. Rafaeli (1985), in a study of QCs in a manufacturing firm, did find QC involvement to be related to the job dimension of task variety, but not to autonomy. In addition, Rafaeli found no relationship between QC involvement and job satisfaction. Head et al. (1986) also studied QC participation in a manufacturing setting, and found no significant differences on any of the core job dimensions, nor in satisfaction or motivation measures. While Mohrman and Novelli (1985) did find improvements in job satisfaction for warehouse employees, there were then decreases in satisfaction after one year. Nonparticipants were significantly lower in feedback and involvement in decision making. Buch and Raban (1990) reported improvements in QC members' perceptions of certain job dimensions, such as autonomy and communication. However, they did not find any difference between members and nonmembers in overall job satisfaction. Finally, Jennings (1988) found QC participation to be related to negative outcomes, namely role conflict and stress.

These conflicting results may be due to the time period in which the different studies were conducted. Longitudinal studies of QCs have shown a consistent pattern of diminishing effects over time. Griffin (1988) found increases in both job satisfaction and organizational commitment for the first one and a half years of a QC program in a manufacturing plant, which were followed by decreases in these measures over a three-year period. Mohrman and Novelli (1985) found a similar pattern of results for job satisfaction. In a qualitative study of manufacturing and banking organizations, Doherty et al. (1989) found an increase, followed by a decrease, in perceived communication, participation, and employee/management relations for a team suggestion program.

The patterns from longitudinal studies indicate that QCs might not have lasting effects. Perhaps long-lasting attitudinal changes should not be expected from a program that accounts for such a small proportion of employees' total work time (Wood et al. 1983). Overall, studies of QCs seem to show a slight impact on satisfaction and commitment. Cotton (1993) argues that direct, teamwork-related attitudes, such as perceptions of influence, are more affected by QCs, while general attitudes, such as job satisfaction and organizational commitment, are less affected.

It has been suggested that, in order to obtain greater benefits, teamwork should be part of a more comprehensive program (Head et al. 1986). Quality circles are a parallel structure in the organization and may not have the authority or resources to affect change (Lawler and Mohrman 1987). QCs may not be related to the day-to-day work done in organizations, and nonparticipants may feel left out, resulting in a negative backlash (Lawler 1986). Demands may be increased for both participants and nonparticipants. For participants, there are the additional duties of going to team meetings and training sessions, while nonparticipants may occasionally have to fill in for participants who are away from their jobs. The main drawback with QCs, according to Lawler and Mohrman (1987), is that they are not well integrated into the organization, in terms of management philosophy, technical and organizational redesign, personnel policies, and training.

Quality improvement (QI) uses quality teams that are similar to QCs in that they address specific problem areas, employ statistical tools, provide group process and problem-solving training to team members, and use team facilitators. Both QCs and QI teams use the PDCA cycle (Plan, Do, Check, Act) and the QC Story (i.e., seven-step problem-solving method) as their primary problem-solving methodologies. However, there are differences in the context of quality teams under QI that may result in better integration within the organization.

Carr and Littman (1993) identify several differences between QI teams and QCs. QCs are often limited to employees and front-line supervisors, while QI teams include members from management as well. Involving management in quality teams can reduce management resistance and fear. QCs in general have a more limited focus than QI teams in both issues addressed and composition of teams. While QCs generally include only the members of a specific work area, QI teams may be cross-functional, including members from different units within the organization. Teams such as this can deal with broader organizational issues and can implement solutions that are more likely to be accepted and effective since more stakeholders are involved. Teams under QI have the potential for a high degree of integration into the organization through greater involvement of management and the existence of more broadly based teams.

3.2. Participatory Ergonomics

Perhaps one the fastest-growing applications of teamwork has been in the field of ergonomics. The use of teams to evaluate, design, and implement jobs and workstations is relatively recent but has met widespread acceptance. A clear indication of this trend is the growing number of submissions on the topic in most national and international ergonomics conferences and the inclusion of employee participation as one of the basic requirements in the proposed OSHA Ergonomics Standard (OSHA 1999).

Participatory ergonomics can be understood as a spinoff of the activity of quality-related teams focusing on working conditions. Noro and Imada created the term *participatory ergonomics* (PE) in 1984 with the main assumption that ergonomics is bounded by the degree to which people are involved in conducting this technology. According to Imada (1991), PE requires users (the real beneficiaries of ergonomics) to be directly involved in developing and implementing ergonomics. Wilson (1995) more recently proposed a more comprehensive definition of PE as ''the involvement of people in planning and controlling a significant amount of their own work activities, with sufficient knowledge and power to influence both processes and outcomes in order to achieve desirable goals.''

Imada (1991) points out three major arguments in support of worker involvement in ergonomics. First, ergonomics being an intuitive science, which in many cases simply organizes the knowledge the workers are already using, it can validate the workers' accumulated experience. Second, people are more likely to support and adopt solutions they feel responsible for. Involving users/workers in the ergonomic process has the potential to transform them into makers and supporters of the process rather than passive recipients. Finally, developing and implementing technology enables the workers to modify and correct occurring problems continuously.

Participatory ergonomics can be an effective tool for disseminating ergonomic information allowing for the utilization of this knowledge in a company-wide basis. It is evident that professional ergonomists will not be available to deal with all the situations existent in an entire organization and that there is a need to motivate, train, and provide resources to workers to analyze and intervene in their work settings.

Participatory ergonomics sees end users' contributions as indispensable elements of its scientific methodology. It stresses the validity of simple tools and workers' experience in problem solution and denies that these characteristics result in nonscientific outcomes. Employees or end users are in most situations in the best position to identify the strengths and weaknesses of the work situations. Their involvement in the analysis and redesign of their workplace can lead to better designs as well as increase their and the company's knowledge on the process.

This approach stresses the relevance of ''small wins'' (Weick 1984), a series of concrete, complete, implemented contributions that can construct a pattern of progress. The nature of these small victories allows the workers to see the next step, the next improvement, and it constitutes a gradual, involving movement towards organizational change. Participatory ergonomics is seen occasionally either as method to design and implement specific workplace changes or a work organization method in place regardless of the presence of change. Wilson and Haines (1997) argue that the participatory process is in itself more important than the focus of that participation since a ''flexible and robust'' process may support the implementation of any change.

Participatory ergonomics emphasizes self-control and self-determination and provides workers more control over their working conditions. This approach also offers potential for reduced job strain through increased social interaction and support. In fact, worker involvement has been shown to be the most common feature among effective stress-management programs (Karasek 1992).

Participatory ergonomics has been implemented through a variety of different organizational approaches and team designs, and no clear unifying model has been proposed or seems likely to be achieved (Vink et al. 1992). Liker et al. (1989) describe six different models of participation based on either direct or representative participation and on different levels of worker input. Wilson and

Haines (1997) describe seven dimensions of participatory ergonomics, characterizing the level of participation (e.g., workstation, organization), focus (e.g., product, workstation, job), purpose (e.g., design, implementation), timeline (e.g., continuous, discrete), involvement (e.g., direct, representative), coupling (e.g., direct, remote), and requirement (e.g., voluntary, necessary).

The tools employed in participatory ergonomics clearly reflect the importance given to the simplicity and meaningfulness of the methods. Group processes follow procedures that mirror those used in quality-related teams. The most common techniques utilized are also derived from QI applications, including Pareto analysis, brainstorming, cause-and-effect diagram, flowcharts, and several forms of displaying quantitative information. Other tools for observation, time measurement, workstation analysis, and problem solving have been developed to address the needs of teams working on ergonomic issues. See Noro and Imada (1991) and Wilson and Haines (1997) for a more complete account of available methods.

Participatory ergonomics is a main feature of most successful ergonomic programs, as emphasized in the National Institute for Occupational Safety and Health's (NIOSH) elements of ergonomics programs (Cohen et al. 1997). This approach has been increasingly common among interventions aimed at improving productivity and reducing both physical and mental workloads. Successful applications of PE have been reported in many industries, including meatpacking (Garg and Moore 1997), health care (Evanoff et al. 1999), automotive (Joseph 1986; Orta-Anes 1991; Keyserling and Hankins 1994), office work (Vink et al. 1995; Haims and Carayon 1998), and agriculture (Sutjana et al. 1999).

Even though the potential benefits are significant, PE faces the same difficulties as do other teamwork initiatives discussed elsewhere in this chapter. Difficulties for the successful implementation of PE include lack of commitment from management and its use as decoy for other purposes such as undermining the union or reducing management influence (Day 1998). The skeptical stance of management regarding the need for ergonomic improvement, the lack of worker awareness of ergonomic deficiencies, and labor management conflict are also highlighted as possible hurdles for PE development (Vink et al. 1992).

4. KEY SUCCESS FACTORS FOR EFFECTIVE TEAMS

Teamwork is not a simplistic, mechanistic work organization technique that can be applied easily with immediate results. To the contrary, it can be a complex management approach that demands well-planned support in all its phases to be effective. Teamwork is not a panacea and is not suitable to all organizational contexts.

Some basic insights into the suitability of teams can be derived from Kanter's (1983) thoughts on participatory management. *Mutatis mutandis,* the use of teamwork (participation) is appropriate for situations related to staying ahead of change, gaining new sources of expertise, involving all who know something about the subject, achieving consensus in controversial matters, building commitment, dealing with problems belonging to no one by organizational assignment, balancing vested interests, avoiding hasty decisions, handling conflicting views, and developing and educating people through their participation. On the other hand, teamwork can be inadequate when a solution for the problem is already available, when nobody really cares about the issue, and when there is no time for discussion.

Kanter (1983) suggests that for participation to be effective, the following elements are required: leadership (particularly for the initiation of the process), a clearly designed management structure, assignment of meaningful and feasible tasks with clear boundaries and parameters, a time frame, scheduling of reports and accountability, information and training for participants, recognition and rewards for teams' efforts (extrinsic rewards), delegation of control but no abdication of responsibility, and a clear process of formation of the participatory groups as well as their evolution and ending and the transfer of learning from them.

Management support has been widely recognized as the fundamental condition for the implementation of teamwork initiatives (Carr and Litman 1993; Hyman and Mason 1995; Kocham et al. 1984). Without continued support and commitment from management, team efforts are doomed to failure. The role of top management is to initiate the teamwork process, setting up policy and guidelines, establishing the infrastructure for team functioning, providing resources, promoting participation, guidance, and cooperation, and assigning a meaningful and feasible task. Tang et al. (1991) report a relationship between upper-management attendance and team members' participation and between middle-management attendance and teams' problem-solving activities. In a study of 154 QC applications from 1978 to 1988, Park and Golembiewski (1991) found middle-management attitude to be the strongest predictor of team success. Employees who are involved in team projects that receive low levels of management support may become frustrated due to lack of resources and cooperation. This may in turn result in negative attitudes, not only towards the project itself, but also towards the job and organization.

Hackman (1990) points out that unclear or insufficient authority or mandate, which relate to the lack of support from top management, are critical impediments to team achievement. Hackman indicates some consequential issues for teams with regard to group authority. First, the team needs to have authority to manage its own affairs. Second, teams need a stable authority structure. Finally, the timing and substance of interventions by authoritative figures. Interventions by authoritative figures can be most effective as the beginning of team development and can be particularly harmful if done on concerns that the group sees as theirs.

For ad hoc teams in particular, the clarity and importance of the team charge also play an important role in the definition and achievement of success. The team charge should be specific and relevant from the participants' and organization's perspectives.

Time limits are powerful organizing factors that shape team performance, particularly for ad hoc teams (Gersick and Davis-Sacks 1990). The available time guides the pace of work and the selection of strategies employed by teams. The lack of clear timelines can cause problems for teams making adopted strategies inadequate and impacting negatively the members' motivation. Time landmarks can in some situations be provided to the team through other avenues, such as a training delivery schedule (Taveira 1996).

The careful definition of team composition is emphasized as an essential aspect of team success in the literature (Carr and Littman 1993; Larson and LaFasto 1989; Scholtes 1988; Kanter 1983). These authors indicate that the absence of members with key expertise or critical organizational linkages can be a sticking point for teams. Both technical and organizational aspects need to be observed in team composition.

The team leader role is essential as an external linkage between the group and upper management, as a promoter of involvement, and as a coordinator and facilitator of communication inside the group (Taveira 1996; Scholtes 1988). Another facet of the team leader's position, serving as a role model, is highlighted by Bolman and Deal's (1992) symbolic tenet: "example rather than command holds a team together." The diligence of the leader in his effort of coordinating and supporting the team can motivate members to volunteer for work assignments and ease the distribution of assignments.

Training is considered to be the one of the most essential resource for team success. It can provide fundamental principles and procedures for its functioning. Training can impart ground rules, guidelines for internal and external communication, and favored ways to make decisions. Training sessions can provide opportunities to discuss and learn with other teams and be conducive to a perception of support and predictability about oncoming tasks and group development. It can introduce the team to a number of procedures and behaviors that enhance communication and involvement (Taveira 1996). Training in problem solving, data collection and analysis, and group decision making is necessary for employees to fully contribute to the group process.

Training is seen as fundamental for giving the team structure for a sound internal process (Hackman 1990). In the specific case of "one-project" teams, where a nonroutine task is undertaken by a new mix of people, training may be critical. Since such groups are unlikely to have established routines for coordinating members' efforts or for determining how work and influence will be distributed among them, training may provide vital guidelines (Gersick and Davis-Sacks 1990).

Moreland and Levine (1992) define commitment as an emotional bond between a person and a group. These authors point out two prevalent theories on commitment: (1) people are committed to a group insofar as it generates more rewards and fewer costs than do other groups to which they already belong or that they could join; (2) commitment depends primarily on how important a group is to someone's social identity. This second theory implies that a need for self-enhancement leads people to feel more committed to groups that seem more successful. A logical extension could be that early success increases the member's commitment to the group.

Correspondingly, Hackman (1990) asserts that groups that begin well and achieve some early wins often trigger a self-sustained upward trend in performance. Hackman delineates a two-factor hypothesis in this regard. The first factor is the quality of the group's initial design, and the second is the occurrence of positive or negative events that trigger the spiral.

Consensus is frequently referred to as the preferred decision-making strategy for teams. Shared definitions of consensus and clear procedures to put this mode of decision making in place are needed. Consensus is defined by Scholtes (1988) as a process of finding a proposal that is acceptable enough that all members can support it and no member opposes to it. Consensus requires time and active participation from team members (Carr and Littman 1993). It demands mutual respect (listening), open-mindedness, and effort at conflict resolution.

Amason et al. (1995) characterize the management of conflicts as "the crux of team effectiveness." They assert that effective teams manage conflict in a way that contributes to its objective. Less-effective teams either avoid conflict, which leads to compromised decisions, or let it seriously disrupt the group process. The authors divide conflict into two types of cognitive and affective. Cognitive conflict focuses on the substance of the issues under discussion. Examination, comparison, and conciliation of opinions characterize it. Cognitive conflict is useful because it invites team members to consider their perspectives from different angles and question underlying assumptions. It can improve members' understanding and commitment to the team's objectives. Affective conflict focuses on disagreements on personal matters and contributes to distrust, bitterness, cynicism, and indifference among team members. Amason et al. believe that task orientation, an inclusive strategy, and open communications are the key elements in fostering cognitive conflict, while avoiding affective conflict.

Nemeth (1992), analyzing the role of dissent on groups, highlights the importance of a vocal minority. According to Nemeth, the expression of disagreeing views, even when they are wrong, encourages attention and thought processes that enable the identification of new facts and solutions and promotes the quality of group decision making and performance. Disagreement may preclude precipitated action (Kanter 1983) and the occurrence of ‘‘groupthink’’ (Janis 1982), in which alternative courses of action are not considered. Minority views stimulate divergent thought processes, adoption of multiple perspectives, and the use of multiple strategies for problem solution.

Gersick and Davis-Sacks (1990) postulate that the challenge of the group is to find the correct equilibrium of independence from and sensitivity to outsiders. The authors add that balancing the relationship between the team and the outside context becomes more complicated when outsiders have dissimilar expectations. The team's composition also influences this balancing since it represents a specific aggregate of dispositions toward, and information about, those stakeholders.

References can be found in the literature to the idea that styles of participatory management should match the organizational culture. Locke et al. (1986) state that the ‘‘manager's natural style of leadership’’ (including decision making, autonomy, and employee control) must also be considered. Similarly, the goals of participatory management should be matched with employee knowledge, skills, and ability. Assessment of the organization's characteristics and needs concerning the implementation of participatory management is fundamental. Particularly, a careful training plan must be developed aimed at motivating and preparing people for teamwork.

Additionally, that an organization that has historically promoted from within and has good employee relations may benefit from more from teamwork than an organization that has high employee turnover rates and does not invest in long-term development of human resources.

There may be other drawbacks to participation in teams. Divided loyalties between the group and the organizational segments to which the members belong may result in peer pressure against participation. Coercion and retaliation against team members by management is a possibility. Members may also be frustrated if team recommendations are not acted upon or if their efforts are not recognized by top management.

Duffy and Salvendy (1997, 1999) using a survey of 103 electronic component manufacturers using concurrent engineering, reached conclusions extremely consistent with the experience of groups working in other organizational contexts. Their findings confirmed the importance of team members' proximity for success in product development and concurrent engineering efforts. Physical proximity increases communication frequency, and teams are likely to be more successful if they communicate more frequently. Successful group work was found to be dependent on a reward structure that reflects group achievement as well as individual achievement. Team size, problem-solving effectiveness, and technical training were also found to contribute to success. The perceived value of communication between different concurrent engineering functions/roles was found to be significantly related to quality of communication and concurrent engineering success.

The implementation of teamwork demands organizational resources, not only in financial terms, but also in time and organizational effort required for planning, training, meetings, and other activities. Therefore, resources must be allocated, and contrary to what some believe, effective teamwork is not free, and spontaneity is not the only driving force behind it.

5. TEAM EFFECTIVENESS

Team performance can be approached in many ways. The following model (developed specifically for QI teams) adopts a systems perspective to conceptualizing team performance by classifying the various factors affecting or related to performance into three broad categories derived from the structure–process–outcome paradigm espoused by Donabedian (1992). The model is displayed in Figure 1 and discussed in detail below. While it was developed in the context of quality improvement, many or all of its elements apply to other teams as well.

5.1. Structure Variables

Structure variables are the contextual parameters that may impact team processes and outcomes. We identified the following three dimensions within which the different structure variables could be classified: organizational characteristics, team characteristics, and task characteristics.

5.1.1. Organizational Characteristics

Researchers have discussed the impact of several organizational variables on project outcomes. Three factors stand out: top-management support, middle-management support, and sufficiency of resources.

- *Top-management support* of project teams has been stressed in terms of the extent to which the management encourages the team, provides constructive feedback, actively champions the pro-

STRUCTURE

I ORGANIZATION CHARACTERISTICS

(1) Top Management Support
(2) Middle Management Support
(3) Sufficiency of Resources

II TEAM CHARACTERISTICS

(1) Team Heterogeneity
(2) Team Expertise
(3) Team Authority
(4) Preference for Team Work
(5) Familiarity
(6) Team Size
(7) Training and Experience

III TASK CHARACTERISTICS

(1) Task Complexity
(2) Tension for Change
(3) Clear Direction And Boundaries

PROCESS

I TASK ISSUES

(1) Meetings Management
(2) Use of Tools
(3) Involvement of Key Affected Personnel
(4) External Expertise
(5) Solution Implementation

II RELATIONSHIP ISSUES

(1) Harmony
(2) Potency
(3) Participatory Decision Making
(4) Workload Sharing
(5) Commitment
(6) Communication
(7) Rewards and Recognition

III LEADERSHIP

(1) Team Leader Characteristics
(2) Team Leadership: Consideration
(3) Team Leadership: Initiating Structure

OUTCOME

(1) Benefits to Indivdual Team Members
(2) Cross-Functional Cooperation
(3) Team Efficiency
(4) Team Effectiveness: Qualitative Measures
(5) Team Effectiveness: Quantitative Measures

Figure 1 Model of Team Effectiveness.

ject, regularly reviews the team's progress, and rewards the team for its performance (Waguespack 1994; Mosel and Shamp 1993; McGrath 1964; Smith and Hukill 1994; Van de Ven 1980; CHSRA 1995; Rollins 1994).

- *Middle-management support* is deemed important for successful QI team performance because team members need to be given time off by their supervisors from their routine jobs to work on team issues (Gladstein 1984; Davis 1993). Team members have often reported difficulty in obtaining permission from their supervisors to attend team meetings (CHSRA 1995). Indeed, lack of encouragement and recognition and inadequate freedom provided by supervisors has been shown to contribute to delays in successful completion of projects (Early and Godfrey 1995).
- *Sufficiency of resources,* although an aspect of top-management support, has been so consistently linked with project outcome that it warrants separate discussion. Availability of adequate training, access to data resources, ongoing consulting on issues related to data collection, analysis, and presentation, adequate allocation of finances, and availability of administrative support are some of the resources cited as important in studies on team effectiveness (Mosel and Shamp 1993; Levin 1992; Smith and Hukill 1994; Early and Godfrey 1995; Gustafson et al. 1992; Nieva et al. 1978; CHSRA 1995).

5.1.2. Team Characteristics

Team characteristics, or group composition, has received significant attention in studies on team effectiveness. We identified the following seven distinct aspects of team composition that are likely to impact QI team performance: team heterogeneity, team expertise, team authority, preference for teamwork, familiarity, team size, and quality-improvement training and expertise.

- *Team heterogeneity* refers to the mix of personalities, attitudes, skills, background, abilities, rank, and experience among team members. Several studies have discussed the significant impact of team heterogeneity on team effectiveness (Guzzo and Dickson 1996; Campion et al. 1993; Morgan and Lassiter 1992; Salas et al. 1992).
- *Team expertise* assesses the team's ability to solve the assigned problem. A QI team would possess the expertise required to complete a project successfully if it included members who had expert knowledge of the functional area under study as well as adequate training and experience in the methods of quality improvement. In particular, studies show that successful teams have adequate representation of all departments affected by the process under study,

especially the process owners who have intimate knowledge of the process. In addition, successful teams also include members with prior QI teamwork experience (Rollins et al. 1994; Flowers et al. 1992; Johnson and Nash 1993; CHSRA 1995).

- *Team authority* assesses the relative power of the team within the organization that would facilitate the completion of the project efficiently and successfully. For instance, Gustafson et al. (1992) suggest that the reputation of the team leader and other team members among the affected parties significantly impacts the implementation success of solutions. Involvement of people in a position of authority on the team, such as department heads, and inclusion of opinion leaders, i.e., people whose opinions are well respected by the affected parties, helps in overcoming resistance to change and eases the process of solution implementation (CHSRA 1995; Davis 1993; Nieva et al. 1978; Rollins et al. 1994).
- *Preference for teamwork* is another important element of team composition. Campion et al. (1993) cite research that shows that employees who prefer to work in teams are likely to be more satisfied and effective as members of a team.
- *Familiarity* among team members may lead to improved group dynamics and hence better team effectiveness. Team members who are more familiar with each other may be more likely to work together better and exhibit higher levels of team performance (Misterek 1995).
- *Team size* has been found to have an important effect on team performance (Morgan and Lassiter 1992). Although larger teams result in increased resources, which may lead to improved team effectiveness, they may also lead to difficulties in coordination and reduced involvement of team members (Campion et al. 1993; Morgan and Lassiter 1992). Campion et al. (1993) suggest that teams should be staffed to the smallest number needed to carry out the project.
- *Quality-improvement training and experience* of team members has also been shown to affect the outcome of QI projects significantly. Although it features as an aspect of overall team expertise, we consider it important enough in the context of QI project teams to discuss separately. Case studies of successful QI teams show that most of the members of the teams had either participated on other QI projects or at least received some form of prior QI training, such as familiarity with the common QI tools and group processes (CHSRA 1995).

5.1.3. Task Characteristics

Task characteristics are factors that are specific to the problem assigned to the project team. We classify the various task characteristics deemed important in previous research studies into three categories:

1. *Task complexity* has been studied at two levels: (a) as a measure of the complexity of the process being studied, e.g., number of departments affected by the process (CHSRA 1995; Misterek 1995), the difficulty of measuring the process quantitatively (Davis 1993; Juran 1994); and (b) as a measure of the complexity of the goals assigned to the team, e.g., scope of the project, number of goals to be accomplished (Juran 1994).
2. *Tension for change* assesses the importance, severity, and significance of the project. Greater tension for change leads to higher motivation for solving the problem (Juran 1994; Gustafson et al. 1992; Van de Ven 1980). In order to be successful, projects should (a) be selected based on data-driven evidence of the existence of the problem (Mosel and Shamp 1993; Rollins et al. 1994; CHSRA 1995); (b) focus on processes that are a cause of dissatisfaction among the process owners (Gustafson et al. 1992); and (c) be considered important areas for improvement by management (Mosel and Shamp 1993).
3. *Clear directions and boundaries* refer to the extent to which management provides the team with a clear mandate. The clarity with which management describes the problem, its requirements, and project goals and explains the available team resources and constraints has been discussed as directly affecting team processes and outcomes (Misterek 1995; Levin 1992; Gustafson et al. 1992; Fleishman and Zaccaro 1992).

5.2. Process Variables

Mosel and Shamp (1993) and Levin (1992) classify process variables into two core dimensions: (1) task or project dimension, which consists of processes that are directly related to solving the assigned problem, such as use of QI tools, efficient planning of meetings, and solution generation and implementation, and (2) relationship or socioemotional dimension, which deals with the dynamics and relationships among team members, such as communication and harmony. Since team leadership impacts both task and relationship issues, we consider it separately as a third dimension of process variables. We discuss these three dimensions of process variables next.

5.2.1. *Task Issues*

The following five task variables have been shown to impact team outcomes and other team processes:

1. Efficient *meetings management* has been shown to result in sustained member involvement and improved overall efficiency with which the team solves the problem (CHSRA 1995). In particular, the advantages of mutually establishing team norms up front (such as meeting times, frequency, and length), and advanced planning of meeting agenda and assignment of responsibility to members for specific agenda items have been highlighted (Davis 1993; Juran 1994).
2. *Quality-improvement tools* aid the team at various stages of the project. Effective use of tools has been shown to help teams keep track of their activities, clarify understanding of the system, help identify problems and solution, help maintain focus, and aid in decision making and data collection and analyses (Plsek 1995; Scholtes 1988; CHSRA 1995; Levin 1992).
3. *Involvement of key personnel,* especially those who are directly affected by the process being studied, significantly improves the chances of success of a QI project (Gustafson et al. 1992). For instance, involvement of process owners during the various stages of problem exploration, solution design, and implementation results in a better understanding of the problem by the team and leads to solutions that are more likely to be accepted and implemented smoothly (Rollins et al. 1994; CHSRA 1995; Van de Ven 1980).
4. *External expertise* refers to sources outside the organization that may be helpful to the team during the various stages of the project. For instance, comparison of current levels of performance with industry standards often helps in providing data-based evidence of the severity of the problem, thereby resulting in increased management support (CHSRA 1995). Networking with other organizations that have successfully solved similar problems and identifying benchmarks helps teams develop successful solutions (Gustafson et al. 1992). Examples of other sources of external expertise that can help teams better understand the problem, and design effective solutions include literature, consultants, and clearinghouses (Rollins et al. 1994; CHSRA 1995).
5. Poor *solution implementation* not only may lead to significant delays in completion of a project (Early and Godfrey 1995) but may also result in the failure of the team's solutions in resulting in any substantial improvement (Gustafson et al. 1992). In order to implement its solutions successfully, the team needs to get buy-in from the process owners (Juran 1994; Rollins et al. 1994; Gustafson et al. 1992; Johnson and Nash 1993; CHSRA 1995). In order to evaluate and demonstrate the advantages of their solutions, the team needs to develop easy to measure process and outcome variables and must have in place a well-designed data-collection strategy (Juran 1994; CHSRA 1995). In addition, feedback from the process owners should be obtained to facilitate further improvement of the process (Gustafson et al. 1992).

5.2.2. *Relationship Issues*

The relationship-based variables that have been shown to impact a team's performance are as follows:

- Team *harmony* refers to the ability of team members to manage conflict and work together as a cohesive unit. The extent to which team members cooperate with one and other and work well together has been shown to affect team outcomes positively (Misterek 1995; Guzzo and Dickson 1996; Mosel and Shamp 1993).
- Team *potency* has been defined as a team's collective belief that it can be effective (Guzzo and Dickson 1996). It is similar to Bandura's (1982) notion of self-efficacy. Campion et al. (1993) have demonstrated a positive relationship between team potency and outcomes such as team productivity, effectiveness, and team member satisfaction.
- *Participatory decision making (PDM)* refers to involvement of all team members on important team decisions. Participation in decisions results in an increase in members' sense of responsibility and involvement in the team's task (Campion et al. 1993). PDM style has been shown to be a common characteristic of successful QI teams (CHSRA 1995; Scholtes 1988).
- *Workload sharing,* similar to PDM, ascertains the extent of balanced participation among members of a team. Teams where most of the members contribute equally to the work have been shown to be more productive and successful (CHSRA 1995; Campion et al. 1993; Scholtes 1988).
- *Commitment* of team members to the team's goals is one of the driving forces behind effective teams (Waguespack 1994). Successful QI teams have reported member motivation and commitment to improve the process as being a critical factor in their success (CHSRA 1995).
- *Communication* is a critical component of teamwork because it serves as the linking mechanism among the various processes of team functioning (Rosenstein 1994). For instance, Davis (1993)

discusses the impact of open communications among members of a QI team on team member commitment. Studies have also demonstrated a positive association between open communication and team performance and team member satisfaction (Gladstein 1984; Campion et al. 1993).

- *Rewards and recognition* help motivate team members and enhance their commitment to team goals (CHSRA 1995). Levin (1992) suggests that formally celebrating the achievement of various project milestones helps a QI team maintain its momentum, motivation, and enthusiasm for accomplishing the project successfully.

5.2.3. Leadership

The impact of team leadership on team performance has been extensively researched. The role of a team leader is to provide direction, structure, and support to other team members (Dickinson et al. 1992). The behavior and competence of the team leader has been shown to affect significantly both team processes and outcomes (CHSRA 1995; Mosel and Shamp 1993). Rosenstein (1994) divides leadership into two distinct but correlated behaviors:

1. *Consideration,* which focuses more on the relationship-based team processes, e.g., the extent to which the team leader facilitates open communication among team members
2. *Initiating structure,* which focuses more on the task-oriented team processes, e.g., the ability of the team leader to plan and coordinate the team's activities capably.
3. In addition to these two behavioral factors, researchers have also emphasized the team leader's *overall characteristics,* such as skills and commitment (Smith and Hukill 1994; Juran 1994). We therefore evaluate the dimension of leadership on all three factors.

5.3. Outcome Variables

Outcome variables constitute the results of the team's performance. We identified four different outcome variables that have been the focus of existing research studies on team performance: benefits to individual team members, cross-functional cooperation, team efficiency, and team effectiveness.

1. *Benefits to individual team members* assesses the influence of the team experience on individual team members (Hackman 1987). Increased job satisfaction, a feeling of accomplishment, and a more problem-focused approach to the daily work are some of the benefits that members derive as a result of participating on teams (Juran 1994; Campion et al. 1993).
2. Improvement in *cross-functional cooperation* is a very common positive outcome of successful QI team efforts (CHSRA 1995). Studies have shown that participation in QI teams by members of different departments often results in the development of mutual trust and respect across departments and a greater understanding of the system, which leads to improved interdepartmental cooperation and communication (Rollins et al. 1994; Juran 1994).
3. The output of the team's effort is measured both in terms of efficiency and effectiveness. *Team efficiency* assumes importance because organizations implementing TQM often complain about the time required to experience significant improvement (Early and Godfrey 1995). In a study of causes of delays in completing QI projects, Early and Godfrey (1995) report that up to 62.8% of the total time taken by the teams could have been avoided. Team productivity has also been a key outcome measure in studies of various other work groups (e.g., Campion et al. 1993).
4. *Team effectiveness* can be assessed by both qualitative and quantitative measures (Landy and Farr 1983; Guzzo and Dickson 1996). Qualitative measures are more subjective and judgmental, such as ratings that require individuals to evaluate the performance of the team (e.g., Campion et al. 1993). Quantitative measures, on the other hand, are objective and nonjudgmental, such as reduction in length of stay, dollars saved, and reduction in error rate (e.g., CHSRA 1995).

6. IMPACT OF TEAMS

Teamwork represents one form of work organization that can have large positive and/or negative effects on the different elements of the work system and on human outcomes, such as performance, attitudes, well being, and health. Given the variety of team characteristics and organizational settings, it is likely that the impact of teamwork on the work system will be highly variable. Some teams may provide for positive characteristics, such as increased autonomy and more interesting tasks, whereas other teams may produce production pressures and tightened management control (Lawler 1986). One important issue in team design is the degree of authority and autonomy (Medsker and Campion 1997; Goodman et al. 1988). It is, therefore, important to examine the impact of teamwork on the task and organizational elements of the work system. Tasks performed by teams are typically of different nature of tasks performed by individual employees. Understanding the physical and psychosocial characteristics of the tasks performed by the team and the members of the team is highly significant for ergonomists. Teams can provide opportunities for reducing the physical and psychosocial repetitiveness of tasks performed by individual employees. This is true only if employees have sufficient training on the different tasks and if rotation among tasks occurs. In some instances, the increased authority and autonomy provided to teams may allow employees to influence their work rhythms and production schedules. This may have beneficial physical impact if adequate work–rest schedules are used. On the other hand, members of the team may work very hard at the beginning of the shift in order to rest at the end of the day. This overload at the beginning of the shift may have some physical health consequences, such as cumulative trauma disorders. A more balanced workload over the entire shift is preferred. In other instances, teamwork has been accompanied by tightened management control (Barker 1993) and electronic and peer surveillance (Sewell 1998). In conclusion, the impact of teamwork on work organization and ergonomics is largely undetermined and depends on a range of factors. However, teamwork can provide many opportunities to improve elements of the work system.

6.1. Impact on Management

The upper managerial levels of organizations have been traditionally targeted in the efforts to sell teamwork. For these management segments, the benefits would come in improvements to the whole organization success and the possibility of spending more time working at the strategic level once the daily decisions can be undertaken by the employee teams. However, one group whose needs are frequently overlooked when implementing employee involvement programs is the middle managers or supervisors. Because the supervisors are a part of management, it is often assumed that they will buy into the philosophies adopted by upper management. Otherwise, according to studies by Klein (1984), even though 72% of supervisors view participation programs as being good for the company and 60% see them as good for employees, less than 31% view them as beneficial to themselves. This perspective is clearly portrayed by Kanter (1983): "participation is something the top orders the middle to do for the bottom." Concerns among supervisors relate to job security, job definition, and additional work created to implement these programs (Klein 1984). A common fear is that employee participation would take supervisors out of the chain of command. Supervisors typically have attained their positions via promotions intended to reward them for outstanding performance as a worker. Sharing their supervisory tasks can be seen as a loss of status to less-deserving workers. Support from first-line supervisors is essential for success of overall participation programs. Some successful experiences in obtaining this support have included the introduction of presentations to upper management, by supervisors, about teamwork activities and creation of teams for forepersons themselves (Harrison 1992).

6.2. Impact on Employees

It is considered that today's better trained and educated workers have expectations greater than basic pay, benefits, and a safe place to work. According to Lawler (1986), these enlarged expectations include participating in meaningful decisions. On the other side, potential problems from the employee perspective need to be addressed. The literature on the subject of employee involvement has dedicated much less emphasis on the problems than on the benefits. Indeed, when these problems are discussed, they are almost always seen from the perspective of the organization and its management. Very little has been written about the problems from the workers' standpoint.

Regarding the negative consequences of teamwork experienced by workers, Baloff and Doherty (1988) state that it can be very disruptive, especially during the crucial start-up period of employee involvement. These authors classify the negative consequences into three categories. First, participants may be subjected to peer-group pressure against what is perceived as collaboration with management in ways that endanger employees' interests. Second, the participants' manager may attempt to coerce them during the group activity, or they may retaliate against the participants if the results of their involvement displease them. Third, participants may have difficulty adapting psychologically at the end of a highly motivating participation effort if they are thrust back into narrow, rigidly defined tasks. Lawler (1986) expresses similar concern about some types of participation that do not match the overall structure of organization and inevitably will produce frustrated expectations among the workers.

On the more negative side of the spectrum of the assessments on teams, Parker and Slaughter (1994, 1988) see them as a way of undermining the union and exploiting workers. Team concept is seen as part of "management by stress," whereby production is speeded up and management actually exerts *more* control on employees. According to these authors the "work rationalization" that used to be done by management is being made now by the employees themselves. The authors point out that peer pressure related to this kind of involvement is even more restrictive than the hierarchy itself. They state that there are several myths about teamwork, that the team concept involves: job security, increased productivity, more control by workers, working smarter not harder, workers with more skills, stronger unions, and feeling of teamwork in the shop floor. The authors conclude that teams themselves are not harmful, but rather the way management has put them into practice and the underlying motivations.

All in all, however, if teamwork is properly chosen as a form of work design and if teams are well designed and managed, teamwork can effectively improve productivity, quality, and employee satisfaction.

REFERENCES

Amason, A. C., Thompson, K. R., Hochwarter, W. A., and Harrison, A. W. (1995), "Conflict: An Important Dimension in Successful Teams," *Organizational Dynamics,* Vol. 24, No. 2, pp. 20–35.

Aram, J. D., Morgan, C. P., and Esbeck, E. S. (1971), "Relation of Collaborative Interpersonal Relationships to Individual Satisfaction and Organizational Performance," *Administrative Science Quarterly,* Vol. 16, No. 3, pp. 289–296.

Baloff, N., and Doherty, E. M. (1985), "Potential Pitfalls in Employee Participation," *Organizational Dynamics,* Vol. 17, No. 3, pp. 51–62.

Bandura, A. (1982), "Self-Efficacy Mechanism in Human Agency," *American Psychologist,* Vol. 37, pp. 122–147.

Barker, J. R. (1993), "Tightening the Iron Cage: Concertive Control in Self-Managing Teams," *Administrative Science Quarterly,* Vol. 38, No. 2, pp. 408–437.

Barness, Z. I., Shortell, S. M., Gillies, R. R., Hughes, E. F., O'Brien, J. L., Bohr, D., Izui, C., and Kralovec, P. (1993), "The Quality March," *Hospitals and Health Networks,* Vol. 67, No. 23, pp. 52–55.

Berwick, D. M. (1994), "Managing Quality: The Next Five Years," *Quality Letter for Healthcare Leaders,* Vol. 6, No. 6, pp. 1–7.

Bolman, L. G., and Deal, T. E. (1992), "What Makes a Team Work?," *Organizational Dynamics,* Vol. 21, Autumn, pp. 35–44.

Buch, K., and Raban, A. (1990), "Quality Circles: How Effective Are They in Improving Employee Performance and Attitudes?," *Psychology—A Journal of Human Behavior,* Vol. 27, pp. 11–17.

Byham, W. C., and Nelson, G. D. (1994), "Using Empowerment to Make Quality Work in Health Care," *Quality Management in Health Care,* Vol. 2, No. 3, pp. 5–14.

Caldwell, C. (1995), *Mentoring Strategic Change in Health Care: An Action Guide,* ASQC Quality Press, Madison, WI.

Campion, M. A., Medsker, G. J., and Higgs, A. C. (1993), "Relations Between Work Group Characteristics and Effectiveness: Implications for Designing Effective Work Groups," *Personnel Psychology,* Vol. 46, pp. 823–850.

Caplan, R. D. (1972), "Organizational Stress and Individual Strain: A Social-Psychology Study of Risk Factors in Coronary Heart Disease Among Administrators, Engineers, and Scientists," Ph.D. dissertation, University of Michigan, *Dissertation Abstracts International,* Vol. 32, pp. 6706B–6707B.

Caplan, R. D., Cobb, S., French, J. R. P., Jr., Van Harrison, R., and Pinneau, S. R., Jr. (1975), *Job Demands and Worker Health,* HEW Publication No. (NIOSH) 75-160, U.S. Department of Health, Education, and Welfare, NIOSH, Washington, DC.

Carr, D. K., and Littman, I. D. (1993), *Excellence in Government: Total Quality Management for the 1990's,* Coopers & Lybrand, Arlington, VA.

Center for Health Systems Research and Analysis (CHSRA) (1995), *Quality Improvement Support System: Clinical Quality Improvement Case Studies,* Madison, WI.

Cohen, A. L., Gjessing, C. C., Fine, L. J., Bernard, B. P., and McGlothlin, J. D. (1997), *Elements of Ergonomics Programs: A Primer Based on Workplace Evaluations of Musculoskeletal Disorders,* National Institute for Occupational Safety and Health, Cincinnati.

Cohen, J., and Cohen, P. (1983), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* 2nd Ed., Erlbaum, Hillsdale, NJ.

Cook, J., and Wall, T. D. (1980), "New Work Attitude Measures of Trust, Organizational Commitment, and Personal Need Non-fulfillment," *Journal of Organizational Psychology,* Vol. 53, pp. 39–52.

Cooper, R. G. (1980), *Project Newprod: What Makes a New Product a Winner?,* Quebec Industrial Innovation Center, Montreal.

Corbett, C., and Pennypacker, B. (1992), "Using a Quality Improvement Team to Reduce Patient Falls," *Journal of Healthcare Quality,* Vol. 14, No. 5, pp. 38–54.

Cotton, J. L. (1993), *Employee Involvement: Methods for Improving Performance and Work Attitudes,* Sage, Newbury Park, CA.

Cronbach, L. J. (1951), ''Coefficient Alpha and the Internal Structure of Tests,'' *Psychometrika,* Vol. 16, pp. 97–334.

Davis, R. N. (1993), ''Cross-functional Clinical Teams: Significant Improvement in Operating Room Quality and Productivity,'' *Journal of the Society for Health Systems,* Vol. 4, No. 1, pp. 34–47.

Day, D. (1998), ''Participatory Ergonomics—A Practical Guide for the Plant Manager,'' in *Ergonomics in Manufacturing: Raising Productivity Through Workplace Improvement,* W. Karwowski and G. Salvendy, Eds., Engineering & Management Press, Norcross, GA.

Dean, J. W., and Bowen, D. E. (1994), ''Management Theory and Total Quality: Improving Research and Practice Through Theory Development,'' *Academy of Management Review,* Vol. 19, pp. 392–418.

Deming, W. E. (1986), *Out of the Crisis,* Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, MA.

Dickinson, T. L., McIntyre, R. M., Ruggeberg, B. J., Yanushefski, A., Hamill, L. S., and Vick, A. L. (1992), ''A Conceptual Framework for Developing Team Process Measures of Decision-Making Performances,'' Final Report, Naval Training Systems Center, Human Factors Division, Orlando, FL.

Doherty, E. M., Nord, W. R., and McAdams, J. L. (1989), ''Gainsharing and Organization Development: A Productive Synergy,'' *Journal of Applied Behavioral Science,* Vol. 25, pp. 209–229.

Donabedian, A. (1992), ''The Role of Outcomes in Quality Assessment and Assurance,'' *Quality Review Bulletin,* pp. 356–360.

Duffy, V. G., and Salvendy, G. (1997), ''Prediction of Effectiveness of Concurrent Engineering in Electronics Manufacturing in the U.S,'' *Human Factors and Ergonomics in Manufacturing,* Vol. 7, No. 4, pp. 351–373.

Duffy, V. G., and Salvendy, G. (1999), ''The Impact of Organizational Ergonomics on Work Effectiveness with Special Reference to Concurrent Engineering in Manufacturing Industries,'' *Ergonomics,* Vol. 42, No. 4, pp. 614–637.

Early, J. F., and Godfrey, A. B. (1995), ''But It Takes Too Long,'' *Quality Progress,* Vol. 28, No. 7, pp. 51–55.

Evanoff, B. A., Bohr, P. C., and Wolf, L. D. (1999), ''Effects of a Participatory Ergonomics Team Among Hospital Orderlies,'' *American Journal of Industrial Medicine,* Vol. 35, No. 4, pp. 358–365.

Fleishman, E. A., and Zaccaro, S. J. (1992), ''Toward a Taxonomy of Team Performance Functions, in *Teams: Their Training and Performance,* R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ, pp. 31–56.

Flowers, P., Dzierba, S., and Baker, O. (1992), ''A Continuous Quality Improvement Team Approach to Adverse Drug Reaction Reporting,'' *Topics in Hospital Pharmacy Management,* Vol. 12, No. 2, pp. 60–67.

Garg, A., and Moore, J. S. (1997), ''Participatory Ergonomics in a Red Meat Packing Plant, Part 1: Evidence of Long Term Effectiveness,'' *American Industrial Hygiene Association Journal,* Vol. 58, No. 2, pp. 127–131.

Gersick, C. J. G., and Davis-Sacks, M. L. (1990), ''Summary; Task Forces,'' in *Groups That Work* (*And Those That Don't*), J. R. Hackman, Ed., Jossey-Bass, San Francisco, pp. 147–153.

Gladstein, D. L. (1984), ''Groups in Context: A Model for Task Group Effectiveness,'' *Administrative Science Quarterly,* Vol. 29, pp. 499–517.

Goodman, P. S., Devadas, S., and Hughson, T. L. (1988), ''Groups and Productivity: Analyzing the Effectiveness of Self-Managing Teams,'' in *Productivity in Organizations,* J. P. Campbell, R. J. Campbell, and Associates, Eds., Jossey-Bass, San Francisco, pp. 295–327.

Griffin, R. W. (1988), ''Consequences of Quality Circles in an Industrial Setting: A Longitudinal Assessment,'' *Academy of Management Journal,* Vol. 31, pp. 338–358.

Gustafson, D. H., and Hundt, A. S. (1995), ''Findings of Innovation Research Applied to Quality Management Principles for Health Care,'' *Health Care Management Review,* Vol. 20, No. 2, pp. 16–33.

Gustafson, D. H., Cats-Baril, W. L., and Alemi, F. (1992), *Systems to Support Health Policy Analysis: Theory, Models, and Uses,* Health Administration Press, Ann Arbor, MI, pp. 339–357.

Guzzo, R. A., and Dickson, M. W. (1996), ''Teams in Organizations: Recent Research on Performance and Effectiveness,'' *Annual Review of Psychology,* Vol. 47, pp. 307–338.

Hackman, J. R. (1987), "The Design of Work Teams, in *Handbook of Organizational Behavior,* J. Lorsch, Ed., Prentice Hall, Englewood Cliffs, NJ, pp. 315–342.

Hackman, J. R. (1990), "Creating More Effective Work Groups in Organizations," in *Groups That Work* (*And Those That Don't*), J. R. Hackman, Ed., Jossey-Bass, San Francisco, pp. 479–508.

Hackman, J. R., and Oldham, G. (1980), *Work Redesign,* Addison-Wesley, Reading, MA.

Hackman, J. R., and Wageman, R. (1995), "Total Quality Management: Empirical, Conceptual, and Practical Issues," *Administrative Science Quarterly,* Vol. 40, pp. 309–342.

Haims, M. C., and Carayon, P. (1998), "Theory and Practice for the Implementation of 'In-house,' Continuous Improvement Participatory Ergonomic Programs," *Applied Ergonomics,* Vol. 29, No. 6, pp. 461–472.

Harrison, E. L. (1992), "The Impact of Employee Involvement on Supervisors," *National Productivity Review,* Vol. 11, Autumn, pp. 447–452.

Head, T. C., Molleston, J. L., Sorenson, P. F., and Gargano, J. (1986), "The Impact of Implementing a Quality Circles Intervention on Employee Task Perceptions," *Group and Organization Studies,* Vol. 11, pp. 360–373.

Hurrell, J. J., Jr., and McLaney, M. A. (1989) "Control, Job Demands, and Job Satisfaction," in *Job Control and Worker Health,* S. L. Sauter, J. J. Hurrell, and C. L. Cooper, Eds., John Wiley & Sons, New York.

Hyman, J., and Mason, R. (1995), *Managing Employee Involvement and Participation,* Sage, London.

Imada, A. (1994), "Participatory Ergonomics: Definition and Recent Developments," in *Proceedings of the 12th Triennial Congress of the International Ergonomics Association,* International Ergonomics Society, Toronto, Canada, pp. 30–33.

Imada, A. (1991), "The Rationale for Participatory Ergonomics," in *Participatory Ergonomics,* K. Noro and A. Imada, Eds., Taylor & Francis, London.

Ingle, S. (1982), "How to Avoid Quality Circle Failure in Your Company," *Training and Development Journal,* Vol. 36, No. 6, pp. 54–59.

James, C., Taveira, A., Sainfort, F., Carayon, P., and Smith, M. J. (1996), "Developing a Comprehensive Quality Management Assessment Instrument for the Public Sector: Results of a Pilot Study," in *Human Factors in Organizational Design Management—V,* O. Brown and H. W. Hendrick, Eds., North-Holland, Amsterdam, pp. 517–522.

James, L. R., Demaree, R. G., and Wolf, G. (1984), "Estimating Within-group Interrater Reliability with and without Response Bias," *Journal of Applied Psychology,* Vol. 69, No. 1, pp. 85–98.

Janis, I. L. (1982), *Victims of Groupthink; A Psychological Study of Foreign-Policy Decisions and Fiascoes,* Houghton Mifflin, Boston.

Jennings, K. R. (1988), "Testing a Model of Quality Circle Processes: Implications for Practice and Consultation," *Consultation—An International Journal,* Vol. 7, No. 1, pp. 19–28.

Johnson, N. E., and Nash, D. B. (1993), "Key Factors in the Implementation of a Clinical Quality Improvement Project: Successes and Challenges," *American Journal of Medical Quality,* Vol. 8, No. 3, pp. 118–122.

Johnson, W. R. (1992), "The Impact of Quality Circle Participation on Job Satisfaction and Organizational Commitment," *Psychology—A Journal of Human Behavior,* Vol. 29, No. 1, pp. 1–11.

Joint Commission Accreditation for Healthcare Organizations (JCAHO) (1994), *1995 Accreditation Manual for Hospitals,* JCAHO, Chicago.

Jones, A. P., James, L. R., Bruni, J. R., Hornick, C. W., and Sells, S. B. (1979), "Psychological Climate: Dimensions and Relationships of Individual and Aggregated Work Environment Perceptions," *Organizational Behavior and Human Performance,* Vol. 23, pp. 201–250.

Joseph, B. S. (1986), "A Participative Ergonomic Control Program in a U.S. Automotive Plant: Evaluation and Implications," Ph.D. dissertation, University of Michigan.

Juran, D. (1994), "Achieving Sustained Quantifiable Results in an Interdepartmental Quality Improvement Project," *Joint Commission Journal on Quality Improvement,* Vol. 20, No. 3, pp. 105–119.

Kano, N. (1994), "The House of Quality," *Academy of Management Review,* Vol. 19, pp. 325–337.

Kano, N. (1993), "A Perspective on Quality Activities in American Firms," *California Management Review,* Vol. 35, No. 3, pp. 12–31.

Kanter, R. M. (1983), *The Change Masters,* Simon & Schuster, New York.

Karasek, R. (1992), "Stress Prevention through Work Reorganization: A Summary of 19 International Case Studies," *Conditions of Work Digest,* Vol. 11, No. 2, pp. 23–41.

Katzenbach, J. R., and Smith, D. K. (1993), *The Wisdom of Teams: Creating the High Performance Organization,* HarperCollins, New York.

Keyserling, W. M., and Hankins, S. E. (1994), ''Effectiveness of Plant-Based Committees in Recognizing and Controlling Ergonomic Risk Factors Associated with Musculoskeletal Problems in the Automotive Industry,'' in *Proceedings of the XII Congress of the International Ergonomics Association,* Toronto, Vol. 3, pp. 346–348.

Kinlaw, D. C. (1990), *Developing Superior Work Teams,* Lexington Books, Lexington, MA, pp. 162–187.

Klein, J. (1984), ''Why Supervisors Resist Employee Involvement,'' *Harvard Business Review,* Vol. 62, September–October, pp. 87–95.

Kochan, T. A., Katz, H. C., and Mower, N. R. (1984), *Worker Participation and American Unions: Threat or Opportunity?,* Upjohn Institute, Kalamazoo, MI.

Kossek, E. E. (1989), ''The Acceptance of Human Resource Innovation by Multiple Constituencies,'' *Personnel Psychology,* Vol. 42, No. 2, pp. 263–281.

Landy, F. J., and Farr, J. L. (1983), *The Measurement of Work Performance: Methods, Theory, and Application,* Academic Press, New York.

Larson, C. E., and LaFasto, F. M. J. (1989), *Teamwork: What Must Go Right, What Can Go Wrong,* Sage, Newbury Park, CA.

Lawler, E. E., III (1986), *High-Involvement Management,* Jossey-Bass, San Francisco.

Lawler, E. E., III, Morhman, S. A., and Ledford, G. E., Jr. (1992), *Employee Participation and Total Quality Management,* Jossey-Bass, San Francisco.

Lawler, E. E., and Mohrman, S. A. (1987), ''Quality Circles: After the Honeymoon,'' *Organizational Dynamics,* Vol. 15, No. 4, pp. 42–54.

Ledford, G. E., Lawler, E. E., and Mohrman, S. A. (1988), ''The Quality Circle and Its Variations,'' in *Productivity in Organizations: New Perspectives from Industrial and Organizational Psychology,* J. P. Campbell and R. J. Campbell, Eds., Jossey-Bass, San Francisco.

Levine, D. I. (1995), ''Reinventing the Workplace: How Business and Employees Can Both Win,'' The Brookings Institution, Washington, DC.

Levin, I. M. (1992), ''Facilitating Quality Improvement Team Performance: A Developmental Perspective,'' *Total Quality Management,* Vol. 3, No. 3, pp. 307–332.

Liker, J. K., Nagamachi, M., and Lifshitz, Y. R. (1989), ''A Comparative Analysis of Participatory Programs in the U.S. and Japan Manufacturing Plants,'' *International Journal of Industrial Ergonomics,* Vol. 3, pp. 185–189.

Locke, E. A., Schweiger, D. M., and Latham, G. P. (1986), ''Participation in Decision Making: When Should It Be Used? *Organizational Dynamics,* Winter, pp. 65–79.

Marks, M. L., Mirvis, P. H., Hackett, E. J., and Grady, J. F. (1986), ''Employee Participation in a Quality Circle Program: Impact on Quality of Work Life, Productivity, and Absenteeism,'' *Journal of Applied Psychology,* Vol. 71, pp. 61–69.

McGrath, J. E. (1964), *Social Psychology: A Brief Introduction,* Holt, New York.

McLaney, M. A., and Hurrell, J. J., Jr. (1988), ''Control, Stress, and Job Satisfaction in Canadian Nurses,'' *Work and Stress,* Vol. 2, No. 3, pp. 217–224.

Medsker, G. J., and Campion, M. A. (2000), ''Job and Team Design,'' in *Handbook of Industrial Engineering,* 3rd Ed., G. Salvendy, Ed., John Wiley & Sons, New York.

Melum, M. M., and Sinioris, M. E. (1993), ''Total Quality Management in Health Care: Taking Stock,'' *Quality Management in Health Care,* Vol. 1, No. 4, pp. 59–63.

Misterek, S. D. A (1995), ''The Performance of Cross-Functional Quality Improvement Project Teams,'' Ph.D. dissertation, University of Minnesota.

Mohrman, S. A., and Novelli, L. (1985), ''Beyond Testimonials: Learning from a Quality Circles Programme,'' *Journal of Occupational Behavior,* Vol. 6, No. 2, pp. 93–110.

Moreland, R. L., and Levine, J. M. (1992), ''Problem Identification by Groups,'' in *Group Process and Productivity,* S. Worchel, W. Wood, and J. A. Simpson, Eds., Sage, Newberry Park, CA, pp. 17–47.

Morgan, B. B., Jr., and Lassister, D. L. (1992), ''Team Composition and Staffing,'' in *Teams: Their Training and Performance,* R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ, pp. 75–100.

Mosel, D., and Shamp, M. J. (1993), ''Enhancing Quality Improvement Team Effectiveness,'' *Quality Management in Health Care,* Vol. 1, No. 2, pp. 47–57.

Nemeth, C. J. (1992), ''Minority Dissent as a Stimulant to Group Performance,'' in *Group Process and Productivity,* S. Worchel, W. Wood, and J. A. Simpson, Eds., Sage, Newbury Park, CA, pp. 95–111.

Nieva, V. F., Fleishman, E. A., and Reick, A. M. (1978), *Team Dimensions: Their Identity, Their Measurement, and Their Relationships,* ARRO, Washington, DC.

Noro, K., and Imada, A. (1991), *Participatory Ergonomics,* Taylor & Francis, London.

Nunnally, J. C. (1978), *Psychometric Theory,* McGraw-Hill, New York.

Orta-Anes, L. (1991), "Employee Participation in Reduction of Ergonomic Risk Factors: Attitudes and Perception of Effects," Ph.D. dissertation, University of Michigan.

Osborne D., and Gaebler, T. (1992), *Reinventing Government,* Addison-Wesley, Reading, MA.

Occupational Safety and Health Administration (OSHA) (1999), Title 29, Code of Federal Regulations, Part 1910 Subpart Y, 1910.913.

Park, S. J., and Golembiewski, R. T. (1991), "An Examination of the Determinants of Successful QC Programs: Testing the Influence of Eleven Situational Features," *Organization Development Journal,* Vol. 9, No. 4, pp. 38–49.

Parker, M., and Slaughter, J. (1994), *Working Smart: A Union Guide to Participation Programs and Reengineering,* Labor Notes, Detroit, MI.

Plsek, P. E. (1995), "Techniques for Managing Quality," *Hospital and Health Services Administration Special CQI Issue,* Vol. 40, No. 1, pp. 50–79.

Plsek, P. E., and Laffel, G. (1993), "From the Editor," *Quality Management in Health Care,* Vol. 1, No. 2.

Quinn, R. and Staines, G. L. (1979), *The 1977 Quality of Employment Survey,* Institute for Social Research, University of Michigan, Ann Arbor, MI.

Rafaeli, A. (1985), "Quality Circles and Employee Attitudes," *Personnel Psychology,* Vol. 38, pp. 603–615.

Rollins, D., Thomasson, C., and Sperry, B. (1994), "Improving Antibiotic Delivery Time to Pneumonia Patients: Continuous Quality Improvement in Action," *Journal of Nursing Care Quality,* Vol. 8, No. 2, pp. 22–31.

Rosenstein, R. (1994), "The Teamwork Components Model: An Analysis Using Structural Equation Modeling," Ph.D. dissertation, Old Dominion University.

Sainfort, F., Mosgaller, T., Van Rensselaer, G., and Smith, M. J. (1996), "The Baldrige Award Criteria for Evaluating TQM Institutionalization," in *Human Factors in Organizational Design and Management—V,* O. Brown and H. Hendrick, Eds., North-Holland, Amsterdam, pp. 517–522.

Salas, E., Dickinson, T. L., Converse, S. A., and Tannenbaum, S. I. (1992), "Toward an Understanding of Team Performance and Training," in *Teams: Their Training and Performance,* R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ, pp. 3–29.

Scholtes, P. R. (1988), *The Team Handbook,* Joiner Associates, Madison, WI.

Seashore, S. E., Lawler, E. E., Mirvis, P., and Cammann, C., Eds. (1982), *Observing and Measuring Organizational Change: A Guide to Field Practice,* John Wiley & Sons, New York.

Sewell, G. (1998), "The Discipline of Teams: The Control of Team-Based Industrial Work through Electronic and Peer Surveillance," *Administrative Science Quarterly,* Vol. 43, No. 2, p. 397.

Shortell, S. M., O'Brien, J. L., Carman, J. M., Foster, R. W., Hughes, E. F. X., Boerstler, H., and O'Connor, E. J. (1995), "Assessing the Impact of Continuous Quality Improvement/Total Quality Management: Concept Versus Implementation, *Health Services Research,* Vol. 30, No. 2, pp. 377–401.

Shortell, S. M., O'Brien, J. L., Hughes, E. F. X., Carman, J. M., Foster, R. W., Boerstler, H., and O'Connor, E. J. (1994), "Assessing the Progress of TQM in US Hospitals: Findings from Two Studies," *The Quality Letter for Healthcare Leaders,* Vol. 6, No. 3, pp. 14–17.

Sims, H. P., Szilagyi, A. D., and Keller, R. T. (1976), "The Measurement of Job Characteristics," *Academy of Management Journal,* Vol. 19, pp. 195–212.

Smith, G. B., and Hukill, E. (1994), "Quality Work Improvement Groups: From Paper to Reality," *Journal of Nursing Care Quality,* Vol. 8, No. 4, pp. 1–12.

Steel, R. P., Mento, A. J., Dilla, B. L., Ovalle, N. K., and Lloyd, R. F. (1985), "Factors Influencing the Success and Failure of Two Quality Circle Programs," *Journal of Management,* Vol. 11, No. 1, pp. 99–119.

Stravinskas, J. (1991), "Analysis of the Factors Impacting Quality Teams," *45th Annual Quality Congress,* May 1991, Milwaukee, WI, pp. 159–162.

Sundstrom, E., De Meuse, K. P., and Futrell, D. (1990), "Work Teams: Applications and Effectiveness," *American Psychologist,* Vol. 45, No. 2, pp. 120–133.

Sutjana, D. P., Adiputra, N., Manuaba, A., and O'Neill, D. (1999), "Improvement of Sickle Quality Through Ergonomi Participatory Approach at Batunya Village Tabana Regency," *Journal of Occupational Health,* Vol. 41, No. 2, pp. 131–135.

Tang, T. L., Tollison, P. S., and Whiteside, H. D. (1991), "Managers' Attendance and the Effectiveness of Small Work Groups: The Case of Quality Circles," *Journal of Social Psychology,* Vol. 131, pp. 335–344.

Taveira, A. D. (1996), "A Successful Quality Improvement Team Project in the Public Sector: A Retrospective Investigation," Ph.D. dissertation, University of Wisconsin-Madison.

Trist, E. (1981), *The Evolution of Socio-technical Systems,* Ontario Ministry of Labor—Quality of Working Life Centre, Toronto.

Van de Ven, A. H. (1980), "Problem Solving, Planning, and Innovation. Part I: Test of the Program Planning Model," *Human Relations,* Vol. 33, No. 10, pp. 711–740.

Varney, G. H. (1989), *Building Productive Teams: An Action Guide and Resource Book,* Jossey-Bass, San Francisco.

Vink, P., Peeters, M., Grundeman, R. W., Smulders, P. G., Kompier, M. A., and Dul, J. (1995), "A Participatory Ergonomics Approach to Reduce Mental and Physical Workload," *International Journal of Industrial Ergonomics,* Vol. 15, pp. 389–396.

Vink, P., Lorussen, E., Wortel, E., and Dul, J. (1992), "Experiences in Participatory Ergonomics: Results of Roundtable Session During the 11th IEA Congress, Paris, July 1991," *Ergonomics,* Vol. 35, No. 2, pp. 123–127.

Waguespack, B. G. (1994), "Development of a Team Culture Indicator," Ph.D. dissertation, Louisiana State University.

Wayne, S. J., Griffin, R. W., and Bateman, T. S. (1986), "Improving the Effectiveness of Quality Circles," *Personnel Administrator,* Vol. 31, No. 3, pp. 79–88.

Wechsler, D. (1971), "Concept of Collective Intelligence," *American Psychologist,* Vol. 26, No. 10, pp. 904–907.

Weick, K. E. (1984), "Small Wins: Redefining the Scale of Social Problems," *American Psychologist,* Vol. 39, pp. 40–49.

Wilson, J. R. (1995), "Ergonomics and Participation," in *Evaluation of Human Work: A Practical Ergonomics Methodology,* J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London, pp. 1071–1096.

Wilson, J. R., and Haines, H. M. (1997), "Participatory Ergonomics," in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York.

Wood, R., Hull, F., and Azumi, K. (1983), "Evaluating Quality Circles: The American Application," *California Management Review,* Vol. 26, pp. 37–49.

Wycoff, M. A., and Skogan, W. K. (1993), "Community Policing in Madison: Quality from the Inside Out," National Institute of Justice Research Report.

Ziegenfuss, J. T. (1994), "Toward a General Procedure for Quality Improvement: The Double Track Process," *American Journal of Medical Quality,* Vol. 9, No. 2, pp. 90–97.

CHAPTER 38

Performance Management

Martin P. Finegan
KPMG

Douglas K. Smith
Author and Consultant

1. INTRODUCTION 995
2. THE CHANGE IMPERATIVE 996
  2.1. Forces of Change 996
  2.2. A Changing View of Performance Itself: The Balanced Scorecard 997
  2.3. New Mental Assumptions for Mastering Performance and Change 998
  2.4. Disciplines of the Performing and Learning Organization 999
  2.5. Work as Process and Competition as Time Based 1000
  2.6. High-Performance Organizations 1000
  2.7. The Impermanent Organization 1001
3. PERFORMANCE SUCCESS: GOAL SETTING AND METRICS 1001
  3.1. Performance Fundamentals and Obstacles 1002
    3.1.1. Natural Human Anxieties 1002
    3.1.2. Difficulty Expressing Nonfinancial Outcomes 1002
    3.1.3. Flawed Assumptions 1002
    3.1.4. The Legacy of Financial Management 1002
    3.1.5. The Intrusive Complexity of the Megaproject or Megaprogram 1002
  3.2. Overcoming the Obstacles: Making Performance Measurable 1003
    3.2.1. Picking Relevant and Specific Metrics 1003
    3.2.2. Using the Four Yardsticks 1004
    3.2.3. Articulating SMART Performance Goals 1005
4. COMBINING THE FORMAL AND INFORMAL ORGANIZATIONS: THE CHALLENGE OF ALIGNMENT IN A FAST-CHANGING WORLD 1005
  4.1. Identifying the Working Arenas That Matter to the Challenge at Hand 1006
  4.2. Using Logic to Achieve Alignment across the Relevant Working Arenas 1007
  4.3. Leadership of Both the Formal and Informal Organization During Periods of Change 1008
  4.4. Bringing it All Together 1009
5. CONCLUDING THOUGHTS 1010
REFERENCES 1010

1. INTRODUCTION

This chapter is about performance management. Performance relates to the measurable outcomes or results achieved by an organization. Management relates to the actions or activities an organization deploys to improve its desired outcomes. This chapter focuses on three major themes:

1. Key ideas organizations have found helpful in dealing with powerful and fundamental forces of change at work in the world

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2. How goal setting and metrics can improve performance at the individual, team, working-group and organizational-unit levels
3. How the formal and informal aspects of an organization each contribute to managing performance in a world of change

We have grounded this chapter in our own extensive experience as well as the writings of leading commentators. Our objective is to provide readers with a blend of best practices. In particular, we seek to avoid—and advise readers to avoid—selecting any single approach to performance management as ‘‘the best and only.’’ Rather, we promote a ‘‘both/and’’ view of the world, in which readers carefully craft that blend of approaches that best fits the needs and challenges ahead of them. This contrasts with the far too dominant ‘‘either/or’’ mindset that maintains, incorrectly, that performance is best managed by selecting a single comprehensive approach through some debate grounded in the proposition that either approach A or approach B is best. In our experience, it usually turns out that both approach A and approach B are relevant to the challenge of managing performance in a changing world.

This chapter concentrates on providing guidance that can be applied, practiced, and perfected by any individuals and teams in any organization. The concepts and techniques we put forward do not depend on senior-management support. We do not attempt to put forward a comprehensive performance-management system model. Our focus is on performance in an environment of change and on how individuals, teams, and groups can significantly improve their mindsets and approaches to improving performance. You and your colleagues can begin to make a performance difference tomorrow for yourself and your organizations. We will provide some ideas and perspectives on formal aspects to integrating performance management and to formal top-management-driven approaches to change. But by and large, our ideas are for you, and you can use them wherever you sit in an organization and on whatever performance challenge arises.

2. THE CHANGE IMPERATIVE

After summarizing the fundamental forces of change that so often determine the nature of today’s performance challenges, this section reviews a series of key concepts and ideas useful in managing performance itself. These include:

- The balanced scorecard: a change in what performance means and how its is measured
- New mental assumptions for managing performance and change
- Disciplines of learning and performing organizations
- Work as process and competition as time based
- Characteristics of high performance organizations
- The trend toward impermanent organizations and alliances

2.1. Forces of Change

If you are young enough, the world of change is all you have. For others, managing performance is very different today than in years past. Regardless of your age and experience, you should have a picture of the fundamental forces at work that shape and determine the challenges ahead of yourselves and your organizations. One of our favorite frameworks comes from John Kotter, a professor at Harvard Business School. Figure 1 shows Professor Kotter’s summary of the economic and social forces driving the need for major change.

These forces are fundamental because they will not be going away any time soon and many market responses to them are irreversible. Globalization and internet technologies are but two examples of changes with permanent and lasting impact.

New market and competitive pressures represent both danger and opportunity. Organizations everywhere are attempting to capitalize on the opportunities and to mitigate their risks from the dangers. The economic headline stories we read about and listen to every day are all in some way responses or reactions (or the lack thereof) to the fundamental forces.

Frameworks like Kotter’s help and encourage readers to look at the external environment for the drivers of change. They give clues as to what is important, how organizations might adapt and lead and, if probed, possible clues about what’s around the corner. Naturally, organizations must continue to look at themselves as well. But far too many organizations under perform as a result of only looking inward for both the causes and solutions to better performance. The ability to understand change and adapt to it more quickly than others is among the most important dimensions of managing performance for organizations to master as we enter the 21st century.

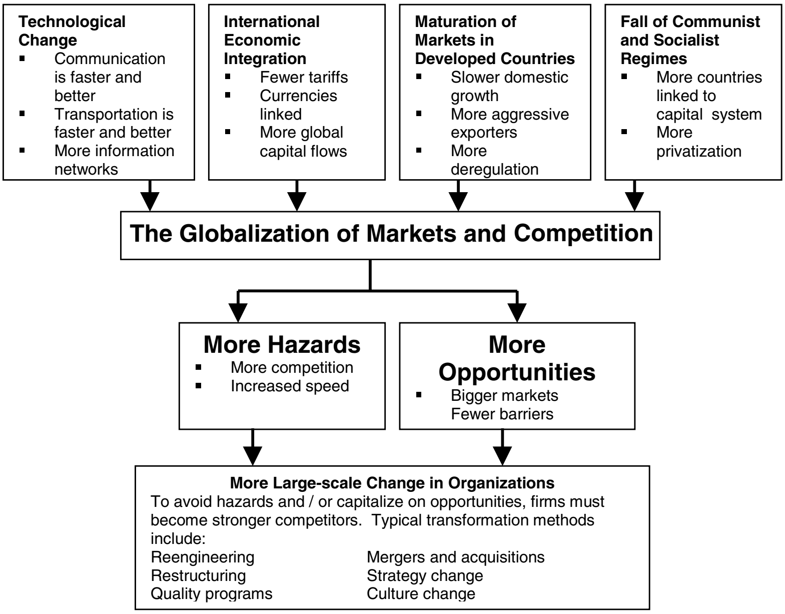

Technological Change
Communication is faster and better
Transportation is faster and better
More information networks
International Economic Integration
Fewer tariffs
Currencies linked
More global capital flows
Maturation of Markets in Developed Countries
Slower domestic growth
More aggressive exporters
More deregulation
Fall of Communist and Socialist Regimes
More countries linked to capital system
More privatization
The Globalization of Markets and Competition
More Hazards
More competition
Increased speed
More Opportunities
Bigger markets
Fewer barriers
More Large-scale Change in Organizations
To avoid hazards and / or capitalize on opportunities, firms must become stronger competitors. Typical transformation methods include:
Reengineering
Restructuring
Quality programs
Mergers and acquisitions
Strategy change
Culture change

Figure 1 Economic and Social Forces Driving the Need for Major Change in Organizations. (Adapted from Kotter 1996)

2.2. A Changing View of Performance Itself: The Balanced Scorecard

The forces of change have altered what performance itself means. Gone are the days when financial results were all that mattered. In today's world, organizations must deliver a combination of financial and nonfinancial performance outcomes. Readers who do not understand the new scorecard of performance will fall into the three main traps of a financial-only approach to performance management:

1. *Unsustainability:* Achieving sustained organizational performance demands outcomes and results that benefit all of the constituencies that matter. Shareholders provide opportunities and rewards to people of the enterprise to deliver value to customers, who generate returns to shareholders, who in turn provide opportunities to the people of the organization, and so on. If you substitute citizens or beneficiaries for customers of the enterprise, you can apply this concept to profit and not-for-profit organizations. Focusing solely on financial measures will create shortfalls in customer service, employee satisfaction, or product/process quality. The wise executive looks at financial measures as lagging measures and seeks other more direct measures of contributing performance to serve as leading measures.
2. *Demotivation:* Executives at the top are motivated by performance measures because they also receive big paydays from achieving them. But today's competitive environment requires tremendous energy and contribution from people throughout the organization. For most people in the organization, financial measures are a too-distant indicator of success or failure, to which many have contributed. Concentrating solely on the financial dimension of measurement is not motivating. It can go further toward being demotivating if people suspect that leaders are concentrating on financial measures out of self-interest.
3. *Confusion:* People need to see how and why their contributions make a difference. Financial goals alone will not reach or connect with very many individuals or groups. These people can become confused and resort to activity-based goals to fill their void.

This new scorecard was first popularized by Kaplan and Norton (1995). All organizations have multiple constituencies such as shareholders, customers, employees, and strategic partners. Each of these constituencies has performance needs and concerns that must be met if the organization hopes to survive and thrive. Kaplan and Norton's scorecard emphasizes the need to convert an organization's strategy into a series of linked performance metrics and outcomes across a 4D logic that suggests that a firm's financial results directly arise from results that matter to customers, which in turn arise from results of internal processes, which in turn arise from results that matter to the people of the organization in terms of learning and growth.

Smith (1999) fundamentally improved on Kaplan and Norton's thinking by suggesting that the balanced scorecard can be both more balanced and more integrated by replacing the linear logic of people-to-process-to-customer-to-shareholder with a reinforcing, integrated logic wherein results for each constituency both leads and lags results for others. Accordingly, managing performance in a sustainable way looks more like Figure 2, which depicts a philosophy for never-ending success.

When viewed in this way, financial and nonfinancial goals all reinforce and link to one another. Moreover, the goals support a narrative or story of success that will not fall victim to unsustainability, demotivation, and confusion.

2.3. New Mental Assumptions for Mastering Performance and Change

Managing both financial and nonfinancial performance in a world of change demands that readers know how to manage change. Having said that, the first and foremost principle of managing change (see Smith 1996) is to keep *performance* the primary objective of managing change, not *change*. Far too many people and organizations do the opposite. If readers are to avoid this trap, they must work hard to connect real and sustainable performance achievements to the changes underway in their organizations.

With clear and compelling performance objectives in mind, readers must avoid a variety of world-views that do not respond to the challenges at hand in today's fast moving world. Kanter (1983) foresaw many of the new and different ways for people and organizations to picture and respond to change. She called for a ''necessary shift from segmentalist to integrative assumptions.'' Today, we might paraphrase her thoughts as shifting from ''stovepipe'' to ''horizontal'' views of work and organization (see Smith 1996). Here are what Kanter described as ''old'' vs. ''new'' assumptions:

- *Old assumption #1:* Organizations and their subunits can operate as closed systems, controlling whatever is needed for their operation. They can be understood on their own terms, according to their internal dynamics, without much reference to their environment, location in a larger social structure, or links to other organizations or individuals.
- *Old assumption #2:* Social entities, whether collective or individual, have relatively free choice, limited only by their own abilities. But since there is also consensus about the means as well as the ends of these entities, there is clarity and singularity of purpose. Thus, organizations can have a clear goal; for the corporation, this is profit maximization.

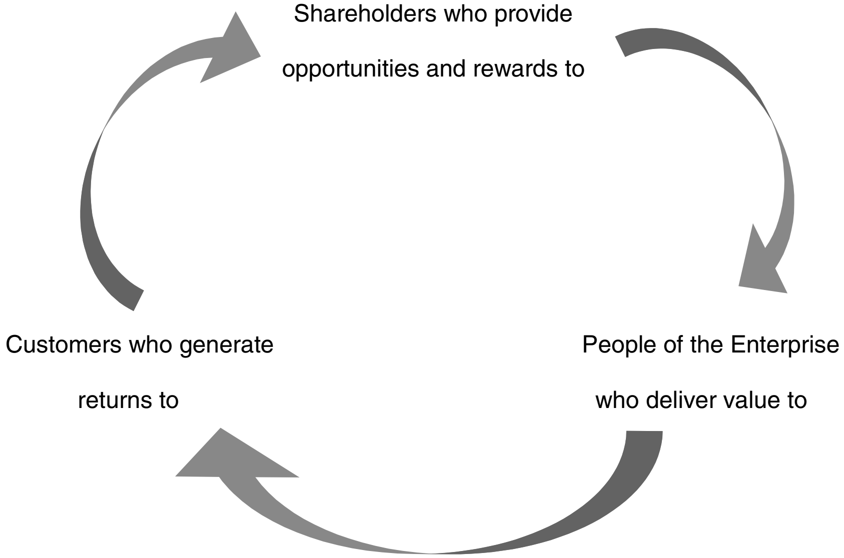


Figure 2 The Performance Cycle. (Adapted from Smith 1999)

- *Old assumption #3:* The individual, taken alone, is the critical unit as well as the ultimate actor. Problems in social life therefore stem from three individual characteristics: *failure of will* (or inadequate motivation), *incompetence* (or differences in talent), and *greed* (or the single-minded pursuit of self interest). There is therefore little need to look beyond these individual characteristics, abilities, or motives to understand why the coordinated social activities we call institutional patterns do not always produce the desired social goods.
- *Old assumption #4:* Differentiation of organizations and their units is not only possible but necessary. Specialization is desirable for both individuals and organizations; neither should be asked to go beyond their primary purposes. The ideal organization is divided into functional specialties clearly bounded from one another, and managers develop by moving up within a functional area.

Here are the new assumptions that Kanter puts forward as alternatives that are more responsive to the external pressures in our changing world:

- *New assumption #1:* Organizations and their parts are in fact open systems, necessarily depending on others to supply much of what is needed for their operations. Their behavior can best be understood in terms of their relationships to their context, their connections—or nonconnections—with other organizations or other units.
- *New assumption #2:* The choices of social entities, whether collective or individual, are constrained by the decisions of others. Consensus about both means and ends is unlikely; there will be multiple views reflecting the many others trying to shape organizational purposes. Thus, singular and clear goals are impossible; goals are themselves the result of bargaining processes.
- *New assumption #3:* The individual may still be the ultimate—or really the only—actor, but the actions often stem from the context in which the individual operates. Leadership therefore consists increasingly of the design of settings that provide tools for and stimulate constructive, productive individual actions.
- *New assumption #4:* Differentiation of activities and their assignment to specialists is important, but coordination is perhaps even more critical a problem, and thus it is important to avoid overspecialization and to find ways to connect specialists and help them to communicate.

The contrast between old and new is sharp. In the old-assumption world, the manager was in control of both the external and the internal. In the new assumption-based world, uncertainty dominates and the need to be fluid and lead by influence rather than control has become the norm. An organization cannot become a strong 21st-century performer if it remains dominated by the old assumptions. It will not be able to change to adapt to new market needs, new technologies, or new employee mindsets. The old model is too slow and costly because it produces unnecessary hierarchy and unneeded specialization.

It takes a very different mindset and perspective to thrive in a world dominated by the new assumptions. One cannot be successful at both in the same ways. Success in each requires a different mental model.

2.4. Disciplines of the Performing and Learning Organization

A variety of new mental models and disciplines have arisen in response to the shifting assumptions so well described by Kanter. Peter Senge is perhaps best known for triggering the search for new disciplines. He suggests five disciplines that distinguish the learning organization from old-world organizations that do not learn (or perform) (Senge et al. 1992):

- *Personal mastery:* learning to expand personal capacity to create the results we most desire and creating an organizational environment that encourages all of its members to develop themselves toward the goals and purposes they choose.
- *Mental models:* reflecting upon, continually clarifying, and improving our internal pictures of the world and seeing how they shape our actions and decisions.
- *Shared vision:* building a sense of commitment in a group by developing shared images of the future we seek to create and the principles and guiding practices by which we hope to get there.
- *Team learning:* transforming conversational and collective thinking skills so that groups of people can reliably develop intelligence and ability greater than the sum of individual members' talents.
- *Systems thinking:* a way of thinking about, and a language for describing and understanding, the forces and interrelationships that shape the behavior of systems. This discipline helps us see how to change systems more effectively and to act more in tune with the larger processes of the natural and economic world.

Each of these disciplines requires a commitment to practice to improve our views and skills in each area. Critically, readers who seek to master these disciplines must do so with an eye on performance itself. Readers and their organizations gain nothing when efforts seek to make people and organizations become learning organizations in the absence of a strong link to performance. Organizations must be both learning and performing organizations.

2.5. Work as Process and Competition as Time Based

Innovation, quality, and continuous improvement have emerged as primary challenges for organizations to apply learning and performance disciplines in the face of change. Innovation draws on and responds to the technological drivers so present in today's world. But innovation also applies to nontechnological challenges. Indeed, for at least the past two decades, competitive success has gone to those who add value to products and services through information, technology, process improvement, and customer service. They must continually ask themselves how to do better and how to do it faster. The old-world adage of "if it ain't broke, don't fix it" has been replaced by "if it ain't broke, fix it!"

At the heart of this reality is quality and continuous improvement. Quality measures itself from the eyes of the customer and views all work as process. Performance of processes is measured by defects as defined by customers, whether internal or external. Defects themselves have many dimensions. But, with the publication of *Competing against Time* (Stalk and Hout 1990), organizations throughout the world were put on notice that speed was now a reality of success. Whether internal or external, customers want error- and defect-free products and services and they want them *fast*. Or, we should say, *faster*. Organizations who master innovation, quality, and continuous improvement never settle for today's results. They continually seek to improve and do so by setting and meeting goals for reducing defects and errors and increasing speed of processes.

2.6. High-Performance Organizations

Many commentators have concluded that organizations cannot succeed in this new world without a fundamentally different set of characteristics. Figure 3 shows contrasting lists of characteristics from a survey of writers, thinkers, and executives.

Traditional	*High Performance*
• Internally driven design	• Customer and environmentally focused design
• Highly controlled fractionated units	• Empowered and autonomous units
• Ambiguous requirements	• Clear direction and goals
• Inspection of errors	• Control of variance at the source
• Technical system dominance	• Socio-technical integration
• Limited information flow	• Accessible information flow
• Fractionated, narrow jobs	• Enriched and shared jobs
• Controlling and restrictive human resources practices	• Empowering human resources practices
• Controlling management structure, process, and culture	• Empowering management structure, process, and culture
• Static designs dependent on senior management redesign	• Capacity to reconfigure

Figure 3 Contrasting Traditional and High-Performance Organizations. (*Source:* Nadler et al. 1992)

High-Performing Organizations	*Low-Performing Organizations*
• Informal	• Rank is right
• Experimental	• Little risk taking
• Action-oriented	• Formal relationships
• High cooperation	• Privileges and perks
• Low defensiveness	• Many status symbols
• High levels of trust	• Rules rigidly enforced
• Little second-guessing	• Slow action / great care
• Few trappings of power	• Much protective paperwork
• High respect for learning	• Decision making at the top
• Few rules and high flexibility	• High levels of fear and anxiety
• Low levels of anxiety and fear	• Your problem is yours, not ours
• Empowered team members	• Well-defined chain of command
• Failures seen as problems to solve	• Learning limited to formal training
• Decisions made at the action point	• Many information-giving meetings
• People easily cross organizational lines	• Trouble puts people on defensive
	• Little problem solving below top
• Much informal problem solving	• Crossing organizational lines forbidden
• Willingness to take risks	

Figure 4 High-Performing vs. Low-Performing Team Characteristics. (*Source:* Synectics, Inc.)

Synectics, an innovation firm, has captured the above list in a slightly different way. Figure 4 contrasts the spirit of innovation in high-performing vs. low performing organizations.

2.7. The Impermanent Organization

Finally, we wish to comment on the trend toward impermanent organizations and alliances. More and more often, organizations respond to performance and change challenges by setting up temporary alliances and networks, both within and beyond the boundaries of the formal organization. It could turn out that this model of the temporary organization and alliance formed to bring a new innovation to market will in fact become the norm. Some have suggested this as one very viable scenario, which Malone and Laubacher (1998) dub the ‘‘e-Lance economy.’’ Malone and Laubacher put forward the idea of many small temporary organizations forming, reforming, and recombining as a way of delivering on customer needs in the future. While they concede that this may be an extreme scenario, it is not an impossible one. Their research is part of an ongoing and significant series of projects at MIT around the 21st-century organization. Another research theme is the continued importance of process management in the future, putting processes alongside products in terms of performance-management importance. One view is certain: 20 years from now, very different business models than the ones we know today will have become the norm.

3. PERFORMANCE SUCCESS: GOAL SETTING AND METRICS

Let’s first look at a few fundamental flaws in how many organizations approach performance. As stressed by Smith (1999), ‘‘Performance begins with focusing on outcomes instead of activities.’’ Yet most people in most organizations do the reverse. With the exception of financial results, most goals are activity based instead of outcome based. Such goals read like ‘‘develop plans to reduce errors’’ or ‘‘research what customers want.’’ These are activities, not outcomes. They do not let the people involved know when they have succeeded, or even how their efforts matter to their own success and that of their organizations.

3.1. Performance Fundamentals and Obstacles

A variety of obstacles and bad habits explain this misplaced emphasis on activities instead of outcomes. At their root lie the old assumptions, financial-focus, internal orientation, and silo organization models we reviewed above. These obstacles and bad habits include:

3.1.1. Natural Human Anxieties

Most people get nervous about the specificity with which their personal success or failure will be measured. We like some flexibility to say we did the right things and that any lack of desired outcome is due to extenuating circumstances. A common tactic is to declare any outcome outside our complete control as unachievable. The problem with this is that for most people in an organization this leaves a narrow set of activities. The further you are from the front line to the customer, the more tempting and common this tactic becomes.

3.1.2. Difficulty Expressing Nonfinancial Outcomes

It is not easy to state nonfinancial goals in an outcome-based fashion. Yet so many performance challenges are first and best measured in nonfinancial ways. It is hard work and personally risky to move beyond the goal of completing the activities and expose your performance to a measure of how effective that activity is where it counts, in the eyes of customers, employees, and strategic partners. The basic anxiety and aversion to setting real outcomes as goals will always be around, particularly when new and different challenges confront us. A key to success is to control the anxiety rather than letting it control you.

3.1.3. Flawed Assumptions

In many instances, people falsely assume performance outcome-based goals exist when they don’t. People in organizations, especially the ones who have achieved a degree of success, often claim they already know what the critical outcomes are and how to articulate them, when in reality they don’t. Or people will elude the responsibility to state outcomes by claiming the outcomes themselves are implied in the activities or plans afoot. Or they will refer to the boss, expecting he or she has it all under control. All of these excuses are mere ruses to avoid the responsibility to specifically and expressly articulate the outcomes by which any effort can be monitored for success.

3.1.4. The Legacy of Financial Management

The financial scorecard has dominated performance business measurement in the modern corporation. As reviewed in the Section 1, the financial-only approach to performance management fails to account for performance outcomes that matter to customers, employees, and strategic partners. It produces organizational cultures that are short-term focused and have difficulty breaking out of the silo approach to work. Why? Because functional organizations are uniquely suited to cost accounting. With the introduction of activity-based accounting by Cooper and Kaplan, organizations were given the chance to move toward a process view of work and still keep their numbers straight. That can help, but it is not enough. Until organizations seriously set and achieve outcome-based goals that are both financial and nonfinancial and link to one another, those organizations will continue to manage performance suboptimally.

3.1.5. The Intrusive Complexity of the Megaproject or Megaprogram

Also standing in the way of outcome-based performance management is the grand illusion of a complete solution to a firm’s information systems. The promise of information technology systems that provide organizations with an integrated approach to transaction management and performance reporting has been a major preoccupation of management teams ever since computers, and personal computers in particular, have become both accessible and affordable (most recently in the form of enterprise resource planning [ERP] systems).

SAP, Oracle, and PeopleSoft are a few of the more popular ERP software providers who have experienced phenomenal success in the 1990s. While the drive to implement new systems was accelerated by the now-infamous Y2K problem, the promise of integrated and flexible information flow throughout an organization had great appeal. These systems were also very much a part of the broader ‘‘transformation’’ programs that many organizations were pursuing at the same time. Many comprehensive transformation frameworks and methodologies that have emerged over the past decade were built on the success that business process reengineering had in the early 1990s. These programs redesigned the people, process and technology of an organization to bring about the performance promise of transformation. Reengineering programs require at least two years to complete and are delivered through massive teams following very detailed methodology scripts. Completing the activities alone is often exhausting and risky. But their promised paybacks are huge, ranging from industry leadership to a chance to survive (and hopefully thrive once again).

Because of the long timeframes associated with the effort and with being able to see the reward, the value received from the effort is difficult to measure. There is a time lag between the team's implementation activities and the outcome. This is true of many things strategic. Consulting teams (and executive sponsors) are often onto their next assignments long before outcomes can be realized as they were defined in the business case.

A focus on performance outcomes for strategic initiatives most often gets lost or mired in the operational systems that are used in most companies. These systems are designed to support the tactics of an organization, which are very often bounded by the time cycles inherent in the formal budgeting and planning systems. All of these realities overwhelm the manager trying to create performance change. The bigger and more complex the organization, the more complicated the improvement of formal performance-management systems.

Many of the large consulting firms (certainly the ones showing annual growth rates in the 30–35% range during the past decade) play to the formal side of organization performance, bringing frameworks and methodologies that require large consulting teams that provide comprehensive solutions to performance management. At the same time, many corporate executives and managers are in need of "having it all integrated" for the promise of accelerated decision making and improved information flow.

Each of these goals has merit and the results can provide large payback. The problem is that in far too many situations, the payoff does not come because of the sheer complexity of the solutions. Much of the implementation cost and business case payback for these endeavors deals with taking activities out of the process. With the advent of the Internet, completely new business models are being pursued for connecting products or services with customers. The sheer size and cost of these approaches require a focus on the formal systems.

So to the list of obstacles to making performance measurable we add this significant pressure to focus on large, complex projects. As consulting firms and their clients have gained experience with "transformation" over the past decade, they have added more emphasis on the informal systems and the people aspect of change. However, their business models still require large teams that will continue to have a bias toward changing the formal systems rather than working at the informal.

3.2. Overcoming the Obstacles: Making Performance Measurable

So what can you do to overcome this formidable list of obstacles? Getting focused on performance outcomes rather than activities is the place to begin. But it is not enough on its own. You will need more to sustain your focus. There are three additional aspects to performance management:

1. Picking relevant and specific metrics
2. Using the "four yardsticks"
3. Articulating SMART goals

Let's take a brief look at the most important aspects behind each of these attributes.

3.2.1. Picking Relevant and Specific Metrics

Sometimes metrics are obvious; other times, the best measures seem elusive. Revenues, profits, and market share are universally recognized as effective metrics of competitive superiority and financial performance. But no universally recognized measures have emerged for such challenges as customer satisfaction, quality, partnering with others, being the preferred provider, innovation, and being the best place to work. Management teams must learn to avoid getting stuck because of the absence of an already accepted standard of measure. They must be willing to work together to pick measures that make sense for their particular challenges.

A good set of measures will have a proper blend of qualitative and quantitative metrics. Consider a company's aspirations to build partnering relationships with key suppliers or customers. Certain threshold goals for the amount of business conducted with each potential partner can provide quantitative and objective performance outcomes. However, it is easy to imagine an outcome where these measures are met but a true partnering relationship is not achieved. Partnership implies a variety of subjective and qualitative characteristics such as trust, consulting each other on critical matters, sharing knowledge, and relying on each other for difficult challenges. Using metrics around these important outcomes will require special candor and honesty from both partners to track behaviors, learn from results, and motivate improvement for both parties in the relationship. Using these types of subjective measures is acceptable and may in fact provide superior results. But you must understand the difficulties associated with such use and overcome them.

Here are several additional pieces of guidance about selecting measures that, while obvious, are often sources of frustration.

- Many metrics require extra work and effort. This will be true for almost all measures that do not already exist in an organization. So if the challenge is new, the best measures will most likely also be new.
- If the measure is new, you will not have a baseline. Organizations and managers must be willing to use their gut feeling as to their baseline performance level. Researching ranges of normal or best-in-class measures can give clues. It is not important to be exact, only to have a sufficient measure to allow the group to move.
- Some measurement criteria will demand contributions from people or groups who are not under your control or authority. In fact, most serious challenges in an organization will require all or many departments to contribute. It will take extra work to get all groups or departments aligned with both the goals and the measures.
- Some metrics are leading indicators of success, while others are lagging indicators. Revenues, profits, and market share are common examples of lagging indicators and therefore are routinely overused. Leading indicators must also be developed to get at the drivers behind the financial success.

The key is to work hard enough at it to make good measurement choices and then stick with them long enough to learn from them. Organizations and managers must overcome the anxieties and frustrations that come with outcome-based performance measures and learn how to select and use the best measures. The following section on the four yardsticks can help you become increasingly comfortable in choosing and sticking with the best measures of progress.

3.2.2. Using the Four Yardsticks

All performance challenges are measurable by some combination of the following:

- Speed/time
- Cost
- On-spec/expected quality
- Positive yields

The first two are quantitative and objective and the second two a blend of objective/subjective and quantitative/qualitative. Becoming adept at the use of these yardsticks will take you a long way toward overcoming the anxieties and obstacles inherent in performance outcome-based goals.

3.2.2.1. Speed/Time Process management is the most common application of this metric. We use it anytime we need to measure how long it takes to complete some activity or process. It is one of the measures that usually requires extra work. Most processes in an organization cross multiple department boundaries, but not neatly. The extra work comes in the need to be specific about beginning and ending points. The scope you place on the process end points will depend on your goal and level of ambition behind the process. For example, an order-generation and fulfillment process that is designed to be both the fastest and totally customer-driven will need to go beyond receipt of delivery and include process steps that measure your customers' use and satisfaction levels. If the process you want to measure is complex, you also must define the specific steps to the process so that you will understand where to concentrate efforts and where your efforts are paying off.

There are six choices you need to make when applying a speed/time metric:

1. What is the process or series of work steps you wish to measure?
2. What step starts the clock?
3. What step stops the clock?
4. What unit of time makes the most sense?
5. What number and frequency of items going through the process must meet your speed requirements?
6. What adjustments to roles and resources (e.g., systems) are needed to do the work of measurement and achieve the goals?

Some of the more fundamental processes in organizations in addition to order fulfillment include new product/service development and introduction, customer service, integrated supply chain, and the hiring/development/retention of people.

3.2.2.2. Cost Cost is clearly the most familiar of the four yardsticks. But here too we can point to nuances. Historically, organizations have focused mostly on the unit costs of materials or activities. These units paralleled organization silo structures. This approach to costing still makes sense for certain functionally sensitive performance challenges. On the other hand, many of today's process-based performance challenges demand the use of activity-based costing instead of unit costing.

3.2.2.3. On-Spec/Expected Quality Product and service specifications generally derive from production, operational, and service-level standards, legal and regulatory requirements, and customer and competitive demands. Another way of viewing specifications is through a company's value proposition, which includes the functions, features, and attributes invested in a product or service in order to win customer loyalty. Some dimensions of quality are highly engineered and easily defined and measured. Others can become abstract and very hard to measure unless the specifications are stated and well defined.

This is key when using this family of metrics. When customer expectations are unknown or poorly defined, they cannot be intentionally achieved. You cannot set or achieve goals related to aspects of performance that you cannot even define. The approach to quality and continuous improvement reviewed above therefore, emphasizes the need to let the customer define quality, to consider any deviation from that a defect, and to set specific goals about reducing defects on a continual basis.

3.2.2.4. Positive Yields This final yardstick category is designed to deal with more abstract or unknown dimensions to customer expectations. It is also a catch-all category whose measures reflect positive and constructive output or yield of organizational effort. Yields are often prone to subjective or qualitative measures, and their purpose is to get at the measurement of newer performance challenges such as alliances or strategic partnering, "delighting customers" or core competencies. While it is hard to reduce these aspirations to specific or quantifiable measurement, the subjective or qualitative measures in this area can be very effective as long as they can be assessed and tracked with effective candor and honesty. (Note how this brings us back to Kanter's "new assumptions" and Synectics' high-performance organization attributes.)

Good performance goals nearly always reflect a combination of two or more of the four yardsticks. Moreover, the first two yardsticks (speed/time and cost) measure the effort or investment *put into* organizational action, while the second two (on-spec/expected quality and positive yields) measure benefits you *get out of* that effort or investment. The best goals typically have at least one performance outcome related to effort put in and at least one outcome related to the benefits produced by that effort.

3.2.3. Articulating SMART Performance Goals

People setting outcome-based goals can benefit from using the SMART acronym as a checklist of items that characterize goals that are specific, relevant, aggressive yet achievable, relevant to the challenge at hand, and time bound. Thus, goals are SMART when they are:

- *Specific:* Answers questions such as "at what?" "for whom?" and "by how much?"
- *Measurable:* Learning requires feedback, which requires measurement. Metrics might be objective or subjective, as long as they are assessable.
- *Aggressive* (yet *Achievable*): Each "A" is significant. *Aggressiveness* suggests stretch, which provides inspiration. *Achievable* allows for a more sustained pursuit because most people will not stay the course for long if goals are not credible. Setting goals that are both aggressive and achievable allows people and organizations to gain all the advantages of stretch goals without creating illusions about what is possible.
- *Relevant:* Goals must relate to the performance challenge at hand. This includes a focus on leading indicators, which are harder to define and riskier to achieve than the more commonly relied-on lagging indicators around financial performance (i.e. revenue, profits).
- *Time bound:* The final specific measure relates to time and answering the question "by when?" You cannot define success without knowing when time is up.

4. COMBINING THE FORMAL AND INFORMAL ORGANIZATIONS: THE CHALLENGE OF ALIGNMENT IN A FAST-CHANGING WORLD

Organizations can hardly be called organized if people throughout them are pursuing goals that are random, unaligned, and conflicting. Managing performance, then, must include an effort to align goals, both against performance challenges and across the various parts of the organization (and increasingly, effort beyond the organization itself, e.g., by strategic partners). Until a decade ago, alignment was considered a simple task, one conducted mostly at budgeting and planning time to be certain all the numbers added up. Today, alignment is far more subtle and challenging. To meet that challenge, readers must be certain they are paying attention to only the relevant challenges, the relevant metrics, and the relevant parts of both the formal and the informal organization.

The formal organization equates to formal hierarchy. It reflects the official directions, activities, and behavior that leaders want to see. The informal organization relates to the actual organizational behavior. It reflects the behaviors that individuals and teams exhibit regardless of official leadership. Both are necessary for truly successful performance management and both are real in contributing to outcomes. Readers will fall into a trap if they worry only about alignment among the official, formal organization.

4.1. Identifying the Working Arenas That Matter to the Challenge at Hand

To avoid that trap, readers need to learn about “working arenas,” which consist of any part of an organization, whether formal or informal, and whether inside the organization or beyond it (e.g., strategic partner), where work happens that matters to performance. Working arenas are where people make performance happen. They include and go well beyond any single individual’s job. Today’s organizations exhibit much more variety in how work gets structured, as Figure 5 shows.

A real shift has occurred from hierarchy-based work structures to more horizontal and open work structures. This reinforces our message earlier about the pace of change and how it forces us to consider these newer forms of structuring work to create connection and speed. That is exactly what you and your colleagues need to do with your performance goals—create connections so that you can increase your speed during implementation. The absolute is that you must set outcome-based goals that fit all the relevant working arenas, not just the jobs, of the people who must achieve those goals.

Fit is not a new concept. It has been attached for a long time to the time-honored managerial maxim that formal accountability matches formal responsibility. However, today this maxim is impossible to apply. Performance challenges have too much overlap across formal structures. And they change far too often to live within any set of formal control processes around all we do. Instead, you should apply a new version of fit—where accountability for outcome-based goals must fit the working arenas of those involved in achieving the goals. The right column in Figure 5 implies both formal (departments, businesses, jobs) and informal (teams, initiatives, processes) working arenas. Fit still makes sense; it just needs to cover a broader spectrum of how work actually gets done. Our argument is that more and more high-performing organizations are learning to apply fit to a broader array of

1950s ➔ 1980s	1990s ➔ 21st century working arenas
• Jobs	• Processes
• Departments	• Teams
• Functions	• Joint ventures
• Projects	• Projects
• Business	• Initiatives
• Corporate headquarters	• Task forces
	• Businesses
	• Alliances
	• Programs
	• Communities of practice
	• Centers of excellence
	• Etc.

Figure 5 Working Arenas: Past and Future. (*Source:* Smith 1999)

informal approaches. It's still all about being effective and efficient—but with greater speed and nimbleness.

The concept of working arenas can help you divide up work in ways that include but go beyond the formal job–department–function–business model. It frees you to ask a critical series of six questions that make it easier to apply performance-based outcome goals, including:

1. What is the performance challenge at hand?
2. What outcomes would indicate success at this challenge?
3. What are the working arenas relevant to this challenge?
4. To which of those working arenas do I (or we) contribute?
5. What metrics make the most sense for these working arenas?
6. What SMART outcome-based goals should we set and pursue for each of these working arenas?

If you were to think through any specific performance challenge in your organization, several patterns would emerge. First, most people contribute to only two or three working arenas at any one time. Second, most contributions come first and foremost in the context of the individual or the team. Third, without identifying the working arena that makes the most sense for any given performance challenge, it is hard for people to confidently believe they are organized for success.

4.2. Using Logic to Achieve Alignment across the Relevant Working Arenas

In the silo or pyramid model of organization that dominated for most of the 20th century, alignment was a matter of adding up costs and revenues to be certain that budgets and plans made sense. Given the entirely formal character of organizations and the relatively small number of working arenas (see Figure 5), this made sense. However, in the fast-moving, flexible, "real" world of today, there are far too many different kind of working arenas, far too many different kind of performance challenges, and (as we discussed in Section 3.2), far more kinds of metrics for success. Budgeting- and planning-driven approaches to alignment that focus solely on being sure the numbers add up do not work in this world.

Instead, readers must get comfortable with two approaches to alignment, quantitative and qualitative. Both of these are logical. If a team is working to increase the speed of a critical process such as new product development, they might measure success by speed and quality. Those metrics will not add up or roll up to quantitatively support the entire business's revenue and profit goals. But those metrics do logically reinforce the entire business's strategy to grow through innovation. In the new world of alignment, then, it is most critical to ask whether the outcome-based goals set across a relevant series of working arenas are logically aligned, not just arithmetically aligned.

Table 1 lists many of today's most compelling performance challenges and suggests whether readers will find it easier to discover logical alignment quantitatively, qualitatively, or through some combination of both.

TABLE 1 Aligning Performance Challenges

Today's Performance Challenges	Quantitative Alignment	Qualitative Alignment
Core competencies		X
Customer service	X	X
Diversity	X	X
Electronic commerce	X	
Growth	X	
Innovation	X	X
Mergers/acquisitions	X	X
Profitability	X	
Reengineering	X	X
Relationship-based marketing		X
Speed	X	X
Strategy	X	
Teams		X
Technology	X	
Total quality	X	X
Values/behaviors/best place to work		X

Source: Smith 1999.

4.3. Leadership of Both the Formal and Informal Organization During Periods of Change

Leaders who must manage performance in the face of change will be more likely to succeed if they attend to both the formal and informal aspects of the organization. Those who focus solely on the formal, official organization will fail. Not only will they fall into the trap of believing that formal goal alignment is in place so long as the financial numbers add up, but they will also fail to address the most critical performance-management challenge of all: how to get existing employees and managers to take the risk to do things differently.

When performance itself depends on lots and lots of existing people learning new skills, behaviors and working relationships, leaders are faced with behavior-driven change. By contrast, if a new strategy or direction can be accomplished based on existing skills, then leaders are faced with decision-driven change. Many of the old assumptions reviewed in Section 2 work better for decision-driven change than for behavior-driven change. However, an ever-increasing number of performance challenges now require leaders to master the disciplines for managing behavior-driven change.

Here are four questions that will help you tell the difference between decision-driven and behavioral-driven change:

1. Does all or any significant part of your organization have to get very good at one or more things that it is not good at today?
2. Do lots of already employed people have to change specific skills, behaviors, and/or working relationships?
3. Does your organization have a positive record of success with changes of the type you are considering?
4. Do those people who must implement the new decisions and directions understand what they need to do and urgently believe the time to act is now?

If the answer is no to the first two questions and yes to the second two, you can employ a decision-driven change approach. If the answer is yes to the first two questions and no to the second two, you are facing behavior-driven change.

If you do face decision-driven change, we suggest following the following best practices (Kotter 1996):

1. Establishing a sense of urgency
    - Examining the market and competitive realities
    - Identifying and discussing crises, potential crises, or major opportunities
2. Creating the guiding coalition
    - Putting together a group with enough power to lead the change
    - Getting the group to work together like a team
3. Developing a vision and strategy
    - Creating vision to help direct the change effort
    - Developing strategies for achieving that vision
4. Communicating the change vision
    - Using all available avenues to communicate vision constantly
    - Having the guiding coalition role model the behavior expected of employees
5. Empowering broad-based action
    - Getting rid of obstacles
    - Changing systems or structures that undermine the change vision
    - Encouraging risk taking and nontraditional ideas, activities, and actions
6. Generating short-term wins
    - Planning for visible improvements in performance, or ‘‘wins’’
    - Creating those wins
    - Visibly recognizing and rewarding people who made the wins possible
7. Consolidating gains and producing more change
    - Using increased credibility to change all systems, structures, and policies that don’t fit together and don’t fit the transformation vision
    - Hiring, promoting, and developing people who implement the change vision
    - Reinvigorating the process with new projects, themes, and change agents

8. Anchoring new approaches in the culture
    - Creating better performance through customer- and productivity-oriented behavior, more and better leadership, and more effective management
    - Articulating the connections between new behaviors and organizational success
    - Developing means to ensure leadership development and succession.

If, however, you are confronted with behavior-driven change, Kotter's transformational leadership approaches will not necessarily work. Indeed, as extensively discussed in Smith (1996), study after study shows that up to four out of five change efforts either fail or seriously suboptimize. And the root cause of these failures lie in leaders who follow decision-driven approaches like those suggested by Kotter when, in fact, they face behavior-driven change.

Managing performance through a period of behavior-driven change demands a different approach. Here is a series of best practices:

1. Keep performance, not change, as the primary objective of behavior and skill change.
2. Focus on continually increasing the number of people taking responsibility for their own performance and change.
3. Ensure that each person always knows why his or her performance and change matter to the purpose and results of the whole organization.
4. Put people in a position to learn by doing and provide them with the information and support needed just in time to perform.
5. Embrace improvisation; experiment and learn; be willing to fail.
6. Use performance to drive change whenever demanded.
7. Concentrate organization designs on the work people do, not on the decision-making authority they have.
8. Create and focus your two scarcest resources during behavioral-driven change—energy and meaningful language.
9. Harmonize and integrate the change initiatives in your organization, including those that are decision driven as well as behavior driven.
10. Practice leadership based on the courage to live the change you wish to bring about—walk the talk.

Please take these and practice. Get others to try. They will make a huge and lasting difference for you personally and for your organizations.

4.4. Bringing It All Together

We will close with a few suggestions on how your organization can put it all together into an integrated performance outcomes-management system. The concepts, frameworks, and techniques presented in this chapter can be deployed to establish an outcomes-management system in your organization. The objective is performance. You should seek to establish a system and set of practices to help the people of your enterprise routinely set and update the SMART outcome-based goals that matter most to success as well as to choose which management disciplines to use to achieve their goals. The outcomes-management system will enable everyone to see how the goals in your organization fit together and make sense from a variety of critical perspectives, from top management (the whole) to each small group, working arena, and performance challenge and each individual throughout the organization.

Figure 6 presents an overview of the design of what a business outcomes-management system might look like.

Figure 6 summarizes the skeletal design of an outcomes-management system for any single business. It brings visibility to the outcomes that matter most to customers, shareholders, and people of the business. It then links these outcomes to the most critical functions, processes, and initiatives that contribute to overall success. This design can be extended to the multibusiness corporation and can be driven down through the organization to every critical shared service or working arena that is critical to success. See Smith (1999) for more illustrations of how this cascading performance outcomes-management system model can work.

An integrated performance model requires that you create the critical linkages, bring visibility to the interdependencies among working arenas, and drive performance through aggressive planning and execution. But avoid the trap of spending all the organization energy around the activity of creating the integrated plan. Make sure you remember the paramount rule: focus on performance outcomes, not activity.

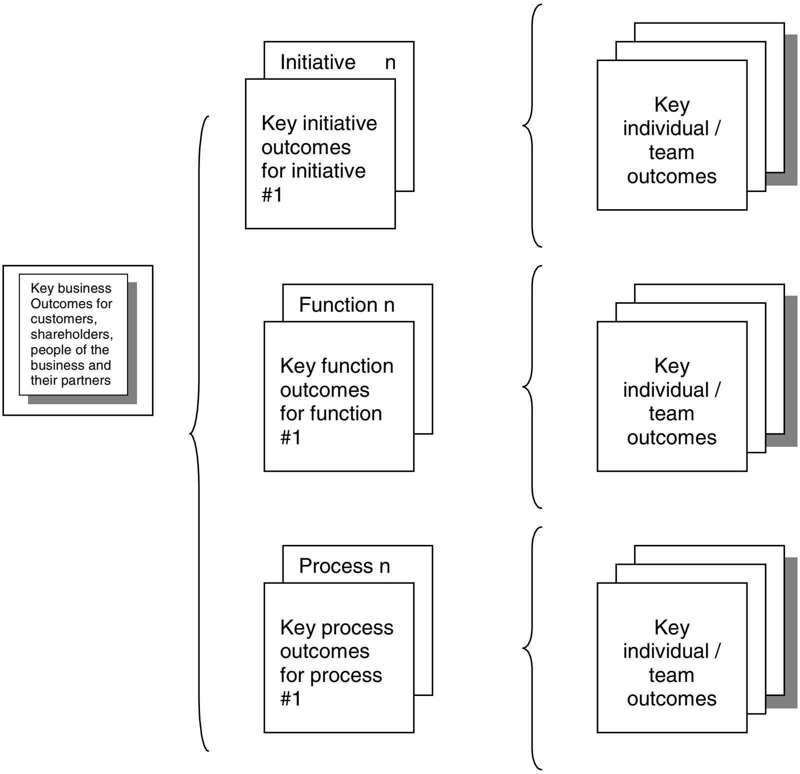

Key business Outcomes for customers, shareholders, people of the business and their partners
Initiative n
Key initiative outcomes for initiative #1
Function n
Key function outcomes for function #1
Process n
Key process outcomes for process #1
Key individual / team outcomes
Key individual / team outcomes
Key individual / team outcomes

Figure 6 Business Outcomes Management System: Design. (Adapted from Smith 1999)

5. CONCLUDING THOUGHTS

We have presented a view of how organizations should view performance in order to improve it. We recommend that you combine the approaches that effect the formal and informal systems in an organization. Success comes not from a single program, slogan, or initiative, but from working the total picture.

We have covered a lot of ground around our original three themes: change, measurement, and combining informal and formal organizational approaches to achieve performance success. We have tried to share with you our experiences and to give you ideas and techniques that will enhance your perspectives on performance management and will also allow you try new things today. We hope we have succeeded.

REFERENCES

Kanter, R. M. (1983), *The Change Masters,* Simon & Schuster, New York.

Kaplan, R., and Norton, D. (1995), *The Balanced Scorecard,* Harvard Business School Press, Boston.

Kotter J. P. (1996), *Leading Change,* Harvard Business School Press, Boston.

Malone, T. W., and Laubacher, R. J. (1998), "The Dawn of the E-Lance Economy," *Harvard Business Review,* Vol. 76, September–October, pp. 145–152.

Nadler, D. A., Gerstein, M. S., Shaw, R. B. and Associates (1992), *Organizational Architecture,* Jossey-Bass, San Francisco.

Senge, P., Ross, R., Smith, B., Roberts, C., and Kleiner, A. (1992), *The Fifth Discipline Fieldbook,* Doubleday, New York.

Smith, D. K. (1996), *Taking Charge of Change,* Addison-Wesley, Reading, MA.

Smith, D. K. (1999), *Make Success Measurable,* John Wiley & Sons, New York.

Stalk, G., and Hout, T. M. (1990), *Competing against Time: Time-Based Competition is Reshaping Global Markets,* Free Press, New York.

III.B

Human Factors and Ergonomics

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 39

Cognitive Tasks

NICOLAS MARMARAS
National Technical University of Athens

TOM KONTOGIANNIS
Technical University of Crete

1. INTRODUCTION 1013

2. MODELS OF HUMAN COGNITION AND DESIGN PRINCIPLES 1014
  2.1. The Human Information-Processing Model 1014
    2.1.1. The Model 1014
    2.1.2. Practical Implications 1016
  2.2. The Action-Cycle Model 1017
    2.2.1. The Model 1017
    2.2.2. The Gulfs of Execution and Evaluation 1018
    2.2.3. Using the Action-Cycle Model in Design 1018
  2.3. The Skill-, Rule-, and Knowledge-Based Model 1019
    2.3.1. The Model 1019
    2.3.2. Using the SRK Model 1020

3. DIAGNOSIS, DECISION MAKING, AND ERGONOMICS 1021
  3.1. Diagnosis 1022
  3.2. Decision Making 1023
  3.3. Supporting Diagnosis and Decision Making 1024

4. COGNITIVE TASK ANALYSIS 1024
  4.1. A Framework for Cognitive Task Analysis 1025
  4.2. Techniques for Cognitive Task Analysis 1028
    4.2.1. Hierarchical Task Analysis 1028
    4.2.2. Critical Decision Method 1028

5. DESIGNING COGNITIVE AIDS FOR COMPLEX COGNITIVE TASKS 1032
  5.1. The Case of a Cognitive Aid for CNC Lathe Programming 1032
    5.1.1. Cognitive Task Analysis 1032
    5.1.2. User Requirements for a Cognitive Aid Supporting CNC Lathe Programming 1034
  5.2. The Case of a Cognitive Aid for Managerial Planning 1034

6. CONCLUDING REMARKS 1037

REFERENCES 1037

1. INTRODUCTION

In any purposeful human activity there is a blend of physical components or manipulations and cognitive components, such as information processing, situation assessment, decision making, and planning. In many tasks, however, the cognitive components are more demanding and crucial for the task performance than the physical components. We call these tasks *cognitive*. Design, managerial and production planning, computer programming, medical diagnosis, process control, air traffic control, and fault diagnosis in technological systems are typical examples of cognitive tasks.

Traditionally, cognitive tasks have been carried out by white-collar workers, middle and upper cadres of enterprises, as well as several freelance professionals. With the advent of information technology and automation in modern industrial settings, however, the role of blue-collar workers has changed from manual controllers to supervisors and diagnosticians. In other words, developments in technology are likely to affect the requirements about knowledge and skills and the way operators interact with systems. Consequently, cognitive tasks abound in modern industries.

Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

In view of these changes in cognitive and collaborative demands, ergonomics must play a crucial role in matching technological options and user requirements. *Cognitive ergonomics* or *cognitive engineering* is an emerging branch of ergonomics that places particular emphasis on the analysis of cognitive processes—for example, diagnosis, decision making, and planning—required of operators in modern industries. Cognitive ergonomics aims to enhance performance of cognitive tasks by means of several interventions, including:

- User-centered design of human–machine interaction
- Design of information technology systems that support cognitive tasks (e.g., cognitive aids)
- Development of training programs
- Work redesign to manage cognitive workload and increase human reliability

Successful ergonomic interventions in the area of cognitive tasks require a thorough understanding not only of the demands of the work situation, but also of user strategies in performing cognitive tasks and of limitations in human cognition. In some cases, the artifacts or tools used to carried out a task may impose their own constraints and limitations (e.g., navigating through a large number of VDU pages), which add up to the total work demands. In this sense, the analysis of cognitive tasks should examine the interaction of users both with their work environment and with artifacts or tools; the latter is very important as modern artifacts (e.g., control panels, electronic procedures, expert systems) become increasingly sophisticated.

As a result, this chapter puts particular emphasis on how to design man–machine interfaces and cognitive aids so that human performance is sustained in work environments where information may be unreliable, events be difficult to predict, goals have conflicting effects, and performance be time constrained. Other types of situations are also considered, as they entail performance of cognitive tasks. Variations from everyday situations are often associated with an increase in human errors as operators are required to perform several cognitive tasks, such as detecting variations, tailoring old methods or devising new ones, and monitoring performance for errors. Sections 2 and 3 introduce human performance models that provide the basic framework for a cognitive analysis of the work environment, artifact constraints, and user strategies. Design principles and guidelines derived from these models are also presented in these sections. Section 4 describes the methodology of cognitive task analysis. Finally, Section 5 discusses two case studies presenting the cognitive analysis carried out for the design of cognitive aids for complex tasks.

2. MODELS OF HUMAN COGNITION AND DESIGN PRINCIPLES

Models of human cognition that have been extensively used in ergonomics to develop guidelines and design principles fall into two broad categories. Models in the first category have been based on the classical paradigm of experimental psychology—also called the behavioral approach—focusing mainly on information-processing stages. Behavioral models view humans as "fallible machines" and try to determine the limitations of human cognition in a neutral fashion independent from the context of performance, the goals and intentions of the users, and the background or history of previous actions. On the other hand, more recent models of human cognition have been developed mainly through field studies and the analysis of real-world situations. Quite a few of these cognitive models have been inspired by Gibson's (1979) work on ecological psychology, emphasizing the role of a person's intentions, goals, and history as central determinants of human behavior.

The human factors literature is rich in behavioral and cognitive models of human performance. Because of space limitations, however, only three generic models of human performance will be presented here. They have found extensive applications. Section 2.1 presents a behavioral model developed by Wickens (1992), the human information-processing model. Sections 2.2 and 2.3 present two cognitive models, the action-cycle model of Norman (1988) and the skill-, rule-, and knowledge-based model of Rasmussen (1986).

2.1. The Human Information-Processing Model

2.1.1. The Model

The experimental psychology literature is rich in human performance models that focus on how humans perceive and process information. Wickens (1992) has summarized this literature into a generic model of information processing (Figure 1). This model draws upon a computer metaphor whereby information is perceived by appropriate sensors, help-up and processed in a temporal memory (i.e., the working memory or RAM memory), and finally acted upon through dedicated actuators. Long-term memory (corresponding to a permanent form of memory, e.g., the hard disk) can be used to store well-practiced work methods or algorithms for future use. A brief description of the human information-processing model is given on page 1015.

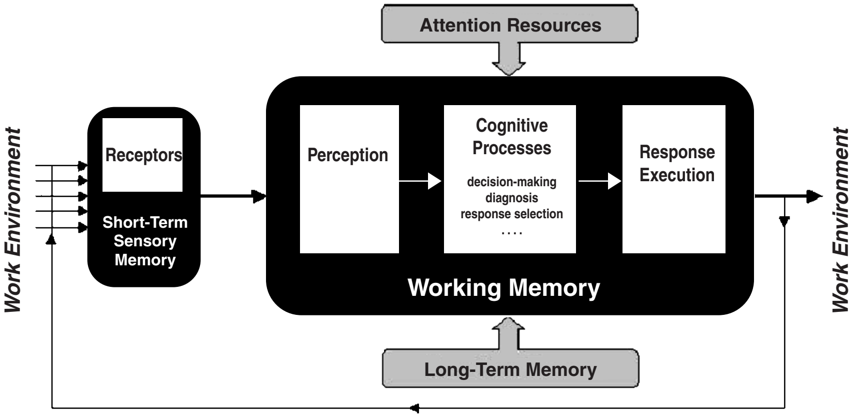


Figure 1 The Human Information-Processing Model. (Adapted from Wickens 1992)

Information is captured by sensory systems or receptors, such as the visual, auditory, vestibular, gustatory, olfactive, tactile, and kinesthetic systems. Each sensory system is equipped with a central storage mechanism called short-term sensory store (STSS) or simply *short-term memory*. STSS prolongs a representation of the physical stimulus for a short period after the stimulus has terminated. It appears that environmental information is stored in the STSS even if attention is diverted elsewhere. STSS is characterized by two types of limitations: (1) the storage capacity, which is the amount of information that can be stored in the STSS, and (2) the decay, which is how long information stays in the STSS. Although there is strong experimental evidence about these limitations, there is some controversy regarding the numerical values of their limits. For example, experiments have shown that the short-term visual memory capacity varies from 7 to 17 letters and the auditory capacity from 4.4 to 6.2 letters (Card et al. 1986). With regard to memory decay rate, experiments have shown that it varies from 70 to 1000 msec for visual short-term memory whereas for the auditory short-term memory the decay varies from 900 to 3500 msec (Card et al. 1986).

In the next stage, perception, stimuli stored in STSS are processed further, that is, the perceiver becomes conscious of their presence, he/she recognizes, identifies, or classifies them. For example, the driver first sees a “red” traffic light, then detects it, recognizes that it is a signal related to the driving task, and identifies it as a stop sign. A large number of different physical stimuli may be assigned to a single perceptual category. For example, a, A, **a**, *a* and the sound “a” all generate the same categorical perception: the letter “a.” At the same time, other characteristics or dimensions of the stimulus are also processed, such as whether the letter is spoken by a male or female voice, whether it is written in upper- or lowercase, and so on.

This example shows that stimulus perception and encoding is dependent upon available attention resources and personal goals and knowledge as stored in the long-term memory (LTM). This is the memory where perceived information and knowledge acquired through learning and training are stored permanently. As in working memory, the information in long-term memory can have any combination of auditory, spatial, and semantic characteristics. Knowledge can be either procedural (i.e., how to do things) or declarative (i.e., knowledge of facts). The goals of the person provide a workspace within which perceived stimuli and past experiences or methods retrieved from LTM are combined and processed further. This workspace is often called working memory because the person has to be conscious of and remember all presented or retrieved information. However, the capacity of working memory also seems to be limited. Miller (1956) was the first to define the capacity of working memory. In a series of experiments where participants carried out absolute judgment tasks, Miller found that the capacity of working memory varied between five and nine items or “chunks”* of information when full attention was deployed. Cognitive processes, such as diagnosis, decision making, and planning, can operate within the same workspace or working memory space. Attention is the main regulatory mechanism for determining when control should pass from perception to cognitive processes or retrieval processes from LTM.

---

*The term *chunk* is used to define a set of adjacent stimulus units that are closely tied together by associations in a person’s long-term memory. A chunk may be a letter, a digit, a word, a phrase, a shape, or any other unit.

At the final stage of information processing (response execution), the responses chosen in the previous stages are executed. A response can be any kind of action, such as eye or head movements, hand or leg movement, and verbal responses. Attention resources are also required at this stage because intrinsic (e.g., kinesthetic) and/or extrinsic feedback (e.g., visual, auditory) can be used to monitor the consequences of the actions performed.

A straightforward implication of the information-processing model shown in Figure 1 is that performance can become faster and more accurate when certain mental functions become ''automated'' through increased practice. For instance, familiarity with machine drawings may enable maintenance technicians to focus on the cognitive aspects of the task, such as troubleshooting; less-experienced technicians would spend much more time in identifying technical components from the drawings and carrying out appropriate tests. As people acquire more experience, they are better able to time-share tasks because well-practiced aspects of the job become automated (that is, they require less attention and effort).

2.1.2. Practical Implications

The information-processing model has been very useful in examining the mechanisms underlying several mental functions (e.g., perception, judgment, memory, attention, and response selection) and the work factors that affect them. For instance, signal detection theory (Green and Swets 1988) describes a human detection mechanism based on the properties of response criterion and sensitivity. Work factors, such as knowledge of action results, introduction of false signals, and access to images of defectives, can increase human detection and provide a basis for ergonomic interventions. Other mental functions (e.g., perception) have also been considered in terms of several processing modes, such as serial and parallel processing, and have contributed to principles for designing man–machine interfaces (e.g., head-up displays that facilitate parallel processing of data superimposed on one another). In fact, there are several information-processing models looking at specific mental functions, all of which subscribe to the same behavioral approach.

Because information-processing models have looked at the mechanisms of mental functions and the work conditions that degrade or improve them, ergonomists have often turned to them to generate guidelines for interface design, training development, and job aids design. To facilitate information processing, for instance, control panel information could be cast into meaningful chunks, thus increasing the amount of information to be processed as a single perceptual unit. For example, dials and controls can be designed and laid out on a control panel according to the following ergonomic principles (Helander 1987; McCormick and Sanders 1987):

- *Frequency of use* and *criticality:* Dials and controls that are frequently used, or are of special importance, should be placed in prominent positions, for example in the center of the control panel.
- *Sequential consistency:* When a particular procedure is always executed in a sequential order, controls and dials should be arranged according this order.
- *Topological consistency:* Where the physical location of the controlled items is important, the layout of dials should reflect their geographical arrangement.
- *Functional grouping:* Dials and controls that are related to a particular function should be placed together.

Many systems, however, fail to adhere to these principles—for example, the layout of stove burner controls fails to conform to the topological consistency principle (see Figure 2). Controls located beside their respective burners (Figure 2[*b*] and 2[*d*]) are compatible and will eliminate confusions caused by arrangements shown in Figure 2(*a*) and 2(*c*).

Information-processing models have also found many applications in the measurement of mental workload. The limited-resource model of human attention (Allport 1980) has been extensively used to examine the degree to which different tasks may rely upon similar psychological resources or mental functions. Presumably, the higher the reliance upon similar mental functions, the higher the mental workload experienced by the human operator. Measures of workload, hence, have tended to rely upon the degree of sharing similar mental functions and their limited capacities.

Information-processing models, however, seem insufficient to account for human behavior in many cognitive tasks where knowledge and strategy play an important role. Recent studies in tactical decision making (Serfaty et al. 1998), for instance, have shown that experienced operators are able to maintain their mental workload at low levels, even when work demands increase, because they can change their strategies and their use of mental functions. Under time pressure, for instance, crews may change from an explicit mode of communication to an implicit mode whereby information is made available without previous requests; the crew leader may keep all members informed of the ''big picture'' so that they can volunteer information when necessary without excessive communications. Performance of experienced personnel may become adapted to the demands of the situation,

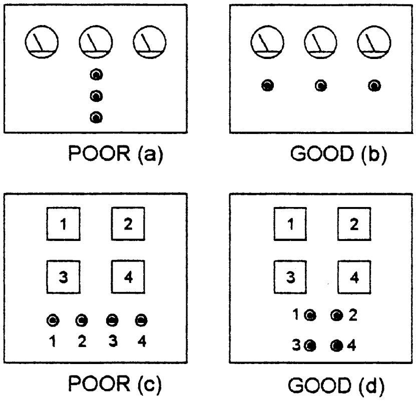

1
2
3
4
1 2 3 4
1 2
3 4
POOR (a)
GOOD (b)
POOR (c)
GOOD (d)

Figure 2 Alternative Layouts of Stove Burner Controls.

thus overcoming high workload imposed by previous methods of work and communication. This interaction of expert knowledge and strategy in using mental functions has been better explored by two other models of human cognition that are described below.

2.2. The Action-Cycle Model

2.2.1. The Model

Many artifacts seem rather difficult to use, often leading to frustrations and human errors. Norman (1988) was particularly interested in how equipment design could be benefit from models of human performance. He developed the action-cycle model (see Figure 3), which examines how people set themselves and achieve goals by acting upon the external world. This action–perception cycle entails two main cognitive processes by which people implement goals (i.e., execution process) and make further adjustments on the basis of perceived changes and evaluations of goals and intentions (i.e., evaluation process).

The starting point of any action is some notion of what is wanted, that is, the goal to be achieved. In many real tasks, this goal may be imprecisely specified with regard to actions that would lead to the desired goal, such as, “I want to write a letter.” To lead to actions, human goals must be transformed into specific statements of what is to be done. These statements are called intentions. For example, I may decide to write the letter by using a pencil or a computer or by dictating it to my secretary. To satisfy intentions, a detailed sequence of actions must be thought of (i.e., planning) and executed by manipulating several objects in the world. On the other hand, the evaluation side of

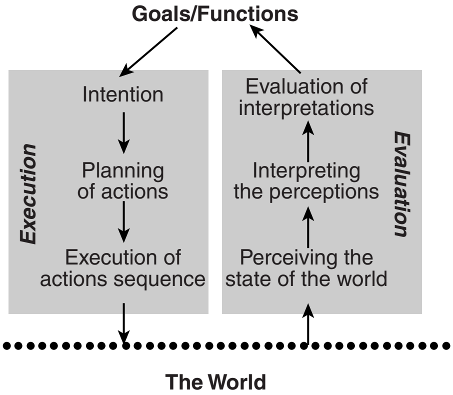

Goals/Functions
Intention
Planning of actions
Execution of actions sequence
Execution
Evaluation of interpretations
Interpreting the perceptions
Perceiving the state of the world
Evaluation
The World

Figure 3 The Action-Cycle Model. (Adapted from Norman 1988)

things has three stages: first, perceiving what happened in the world; second, making sense of it in terms of needs and intentions (i.e., interpretation); and finally, comparing what happened with what was wanted (i.e., evaluation). The action-cycle model consists of seven action stages: one for goals (forming the goal), three for execution (forming the intention, specifying actions, executing actions), and three for evaluation (perceiving the state of the world, interpreting the state of the world, evaluating the outcome).

As Norman (1988, p. 48) points out, the action-cycle model is an approximate model rather than a complete psychological theory. The seven stages are almost certainly not discrete entities. Most behavior does not require going through all stages in sequence, and most tasks are not carried out by single actions. There may be numerous sequences, and the whole task may last hours or even days. There is a continual feedback loop, in which the results of one action cycle are used to direct further ones, goals lead to subgoals, and intentions lead to subintentions. There are action cycles in which goals are forgotten, discarded, or reformulated.

2.2.2. The Gulfs of Execution and Evaluation

The action-cycle model can help us understand many difficulties and errors in using artifacts. Difficulties in use are related to the distance (or amount of mental work) between intentions and possible physical actions, or between observed states of the artifact and interpretations. In other words, problems arise either because the mappings between intended actions and equipment mechanisms are insufficiently understood, or because action feedback is rather poor. There are several *gulfs* that separate the mental representations of the person from the physical states of the environment (Norman 1988).

The *gulf of execution* reflects the difference between intentions and allowable actions. The more the system allows a person to do the intended actions directly, without any extra mental effort, the smaller the gulf of execution is. A small gulf of execution ensures high usability of equipment. Consider, for example, faucets for cold and hot water. The user intention is to control two things or parameters: the water temperature and the volume. Consequently, users should be able to do that with two controls, one for each parameter. This would ensure a good mapping between intentions and allowable actions. In conventional settings, however, one faucet controls the volume of cold water and the other the volume of hot water. To obtain water of a desired volume and temperature, users must try several combinations of faucet adjustments, hence losing valuable time and water. This is an example of bad mapping between intentions and allowable actions (a large gulf of execution).

The *gulf of evaluation* reflects the amount of effort that users must exert to interpret the physical state of the artifact and determine how well their intentions have been met. The gulf of evaluation is small when the artifact provides information about its state that is easy to get and easy to interpret and matches the way the user thinks of the artifact.

2.2.3. Using the Action-Cycle Model in Design

According to the action-cycle model, the usability of an artifact can be increased when its design bridges the gulfs of execution and evaluation. The seven-stage structure of the model can be cast as a list of questions to consider when designing artifacts (Norman 1988). Specifically, the designer may ask how easily the user of the artifact can:

1. Determine the function of the artifact (i.e., setting a goal)
2. Perceive what actions are possible (i.e., forming intentions)
3. Determine what physical actions would satisfy intentions (i.e., planning)
4. Perform the physical actions by manipulating the controls (i.e., executing)
5. Perceive what state the artifact is in (i.e., perceiving state)?
6. Achieve his or her intentions and expectations (i.e., interpreting states)
7. Tell whether the artifact is in a desired state or intentions should be changed (i.e., evaluating intentions and goals)

A successful application of the action-cycle model in the domain of human–computer interaction regards direct manipulation interfaces (Shneiderman 1983; Hutchins et al. 1986). These interfaces bridge the gulfs of execution and evaluation by incorporating the following properties (Shneiderman 1982, p. 251):

- Visual representation of the objects of interest
- Direct manipulation of objects, instead of commands with complex syntax
- Incremental operations that their effects are immediately visible and, on most occasions, reversible

The action-cycle model has been translated into a new design philosophy that views design as knowledge in the world. Norman (1988) argues that the knowledge required to do a job can be distributed partly in the head and partly in the world. For instance, the physical properties of objects may constrain the order in which parts can be put together, moved, picked up, or otherwise manipulated. Apart from these physical constraints, there are also cultural constraints, which rely upon accepted cultural conventions. For instance, turning a part clockwise is the culturally defined standard for attaching a part to another while counterclockwise movement usually results in dismantling a part. Because of these natural and artificial or cultural constraints, the number of alternatives for any particular situation is reduced, as is the amount of knowledge required within human memory. Knowledge in the world, in terms of constraints and labels, explains why many assembly tasks can be performed very precisely even when technicians cannot recall the sequence they followed in dismantling equipment; physical and cultural constraints reduce the alternative ways in which parts can be assembled. Therefore, cognitive tasks are more easily done when part of the required knowledge is available externally—either explicit in the world (i.e., labels) or readily derived through constraints.

2.3. The Skill-, Rule-, and Knowledge-Based Model

2.3.1. The Model

In a study of troubleshooting tasks in real work situations, Rasmussen (1983) observed that people control their interaction with the environment in different modes. The interaction depends on a proper match between the features of the work domain and the requirements of control modes. According to the skill-, rule-, and knowledge-based model (SRK model), control and performance of human activities seem to be a function of a hierarchically organized control system (Rasmussen et al. 1994). Cognitive control operates at three levels: skill-based, or automatic control; rule-based, or conditional control; and knowledge-based, or compensatory control (see Figure 4).

At the lowest level, skill-based behavior, human performance is governed by patterns of preprogrammed behaviors represented as analog structures in a time–space domain in human memory. This mode of behavior is characteristic of well-practiced and routine situations whereby open-loop or feedforward control makes performance faster. Skill-based behavior is the result of extensive practice where people develop a repertoire of cue–response patterns suited to specific situations. When a familiar situation is recognized, a response is activated, tailored, and applied to the situation. Neither any conscious analysis of the situation nor any sort of deliberation of alternative solutions is required.

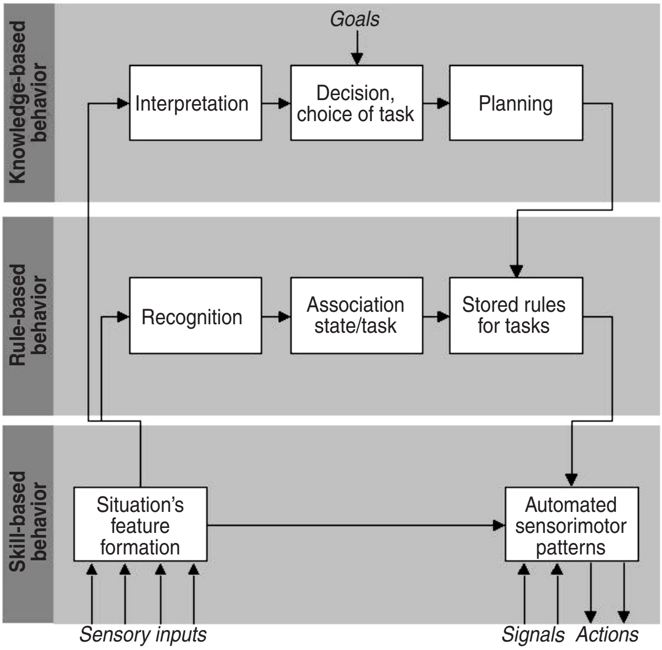


Figure 4 The Skill-, Rule-, and Knowledge-Based Model. (Adapted from Rasmussen 1986)

At the middle level, rule-based behavior, human performance is governed by conditional rules of the type:

$$\textit{if} \quad \langle \text{state } X \rangle \quad \textit{then} \quad \langle \text{action } Y \rangle$$

Initially, stored rules are formulated at a general level. They subsequently are supplemented by further details from the work environment. Behavior at this level requires conscious preparation: first recognition of the need for action, followed by retrieval of past rules or methods and finally, composition of new rules through either self-motivation or through instruction. Rule-based behavior is slower and more cognitively demanding than skill-based behavior. Rule-based behavior can be compared to the function of an expert system where situations are matched to a database of production rules and responses are produced by retrieving or combining different rules. People may combine rules into macrorules by collapsing conditions and responses into a single unit. With increasing practice, these macrorules may become temporal–spatial patterns requiring less conscious attention; this illustrates the transition from rule-based to skill-based behavior.

At the highest level, knowledge-based behavior, performance is governed by a thorough analysis of the situation and a systematic comparison of alternative means for action. Goals are explicitly formulated and alternative plans are compared rationally to maximize efficiency and minimize risks. Alternatives are considered and tested either physically, by trial and error, or conceptually, by means of thought experiments. The way that the internal structure of the work system is represented by the user is extremely important for performance. Knowledge-based behavior is slower and more cognitively demanding than rule-based behavior because it requires access to an internal or mental model of the system as well as laborious comparisons of work methods to find the most optimal one.

Knowledge-based behavior is characteristic of unfamiliar situations. As expertise evolves, a shift to lower levels of behavior occurs. This does not mean, however, that experienced operators always work at the skill-based level. Depending on the novelty of the situation, experts may move at higher levels of behavior when uncertain about a decision. The shift to the appropriate level of behavior is another characteristic of expertise and graceful performance in complex work environments.

2.3.2. Using the SRK Model

The SRK model has been used by Reason (1990) to provide a framework for assigning errors to several categories related to the three levels of behavior. Deviations from current intentions due to execution failures and/or storage failures, for instance, are errors related to skill-based behavior (slips and lapses). Misclassifications of the situation, leading to the application of the wrong rule or incorrect procedure, are errors occurring at the level rule-based behavior. Finally, errors due to limitations in cognitive resources—''bounded rationality''—and incomplete or incorrect knowledge are characteristic of knowledge-based behavior. More subtle forms of errors may occur when experienced operators fail to shift to higher levels of behavior. In these cases, operators continue skill- or rule-based behavior although the situation calls for analytical comparison of options and reassessment of the situation (e.g., knowledge-based processing).

The SRK model also provides a useful framework for designing man–machine interfaces for complex systems. Vicente and Rasmussen (1992) advance the concept of ecological interfaces that exploit the powerful human capabilities of perception and action, at the same time providing appropriate support for more effortful and error-prone cognitive processes. Ecological interfaces aim at presenting information ''in such a way as not to force cognitive control to a higher level than the demands of the task require, while at the same time providing the appropriate support for all three levels'' (Vicente and Rasmussen 1992, p. 598). In order to achieve these goals, interfaces should obey the following three principles:

1. *Support skill-based behavior:* The operator should be able act directly on the display while the structure of presented information should be isomorphic to the part–whole structure of eye and hand movements.
2. *Support rule-based behavior:* The interface should provide a consistent one-to-one mapping between constraints of the work domain and cues or signs presented on the interface.
3. *Support knowledge-based behavior:* The interface should represent the work domain in the form of an abstraction hierarchy* to serve as an externalized mental model that supports problem solving.

---

* Abstraction hierarchy is a framework proposed by Rasmussen (1985; Rasmussen et al. 1994) that is useful for representing the cognitive constraints of a complex work environment. For example, the constraints related to process control have been found to belong to five hierarchically ordered levels (Vicente and Rasmussen 1992, p. 592): the purpose for which the system was designed (functional purpose); the intended causal structure of the process in terms of mass, energy, information, or value flows (abstract purpose); the basic functions that the plant is designed to achieve (generalized function); the characteristics of the components and the connections between them (physical function); and the appearance and special location of these components (physical form).

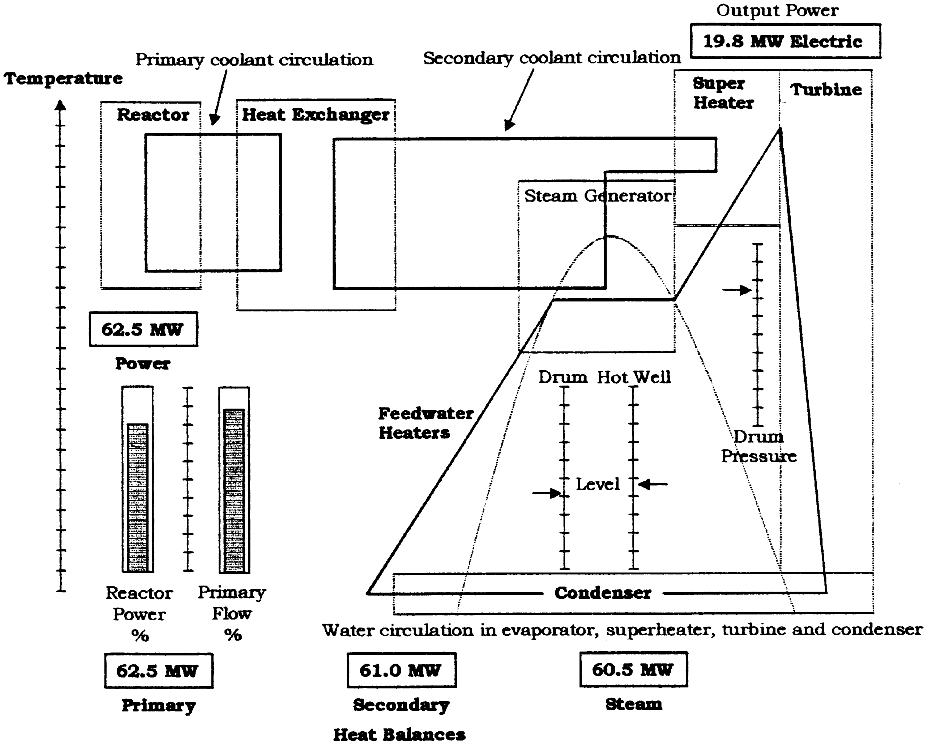

Output Power
19.8 MW Electric
Primary coolant circulation
Secondary coolant circulation
Temperature
Reactor
Heat Exchanger
Super Heater
Turbine
Steam Generator
62.5 MW
Power
Drum Hot Well
Feedwater Heaters
Level
Drum Pressure
Condenser
Reactor Power %
Primary Flow %
Water circulation in evaporator, superheater, turbine and condenser
62.5 MW
Primary
61.0 MW
Secondary
60.5 MW
Steam
Heat Balances

Figure 5 An Ecological Interface for Monitoring the Operation of a Nuclear Power Plant (From Lindsay and Staffon 1988, by permission. Work performed for U.S. Department of Energy at Argonne National Laboratory, Contract No. W-31,109-ENG-38)

An example of an ecological interface that supports direct perception of higher-level functional properties is shown in Figure 5. This interface, designed by Lindsay and Staffon (1988), is used for monitoring the operation of a nuclear power plant. Graphical patterns show the temperature profiles of the coolant in the primary and secondary circulation systems as well as the profiles of the feedwater in the steam generator, superheater, turbine, and condenser. Note that while the display is based on primary sensor data, the functional status of the system can be directly read from the interrelationships among data. The Rankine cycle display presents the constraints of the feedwater system in a graphic visualization format so that workers can use it in guiding their actions; in doing this, workers are able to rely on their perceptual processes rather than analytical processes (e.g., by having to solve a differential equation or engage in any other abstract manipulation). Plant transients and sensor failures can be shown as distortions of the rectangles representing the primary and secondary circulation, or as temperature points of the feedwater system outside the Rankine cycle. The display can be perceived, at the discretion of the observer, at the level of the physical implications of the temperature readings, of the state of the coolant circuits, and of the flow of the energy.

3. DIAGNOSIS, DECISION MAKING AND ERGONOMICS

After this brief review of behavioral and cognitive models, we can now focus on two complex cognitive processes: diagnosis and decision making. These cognitive processes are brought into play in many problem-solving situations where task goals may be insufficiently specified and responses may not benefit from past knowledge. These characteristics are common to many problem-solving scenarios and affect how people shift to different modes of cognitive control. Problem solvers may use experiences from similar cases in the past, apply generic rules related to a whole category of problems, or try alternative courses of actions and assess their results. In other words, optimizing problem solving on the basis of knowledge-based behavior may be time consuming and laborious. People tend to use several heuristics to regulate their performance between rule-based and knowledge-based processing. This increases task speed but may result in errors that are difficult to recover from.

In following, we present a set of common heuristics followed by experienced diagnosticians and decision makers, potential biases and errors that may arise, and finally ergonomic interventions how to support human performance.

3.1. Diagnosis

Diagnosis is a cognitive process whereby a person tries to identify the causes of an undesirable event or situation. Technical failures and medical problems are two well-known application areas of human diagnosis. Figure 6 presents some typical stages of the diagnosis process. Diagnosis starts with the perception of signals alerting one to a system's failure or malfunction. Following this, diagnosticians may choose whether to search for more information to develop a mental representation of the current system state. At the same time, knowledge about system structure and functioning can be retrieved from the long-term memory. On the basis of this evidence, diagnosticians may generate hypotheses about possible causes of the failure. Pending upon further tests, hypotheses may be confirmed, completing the diagnosis process, or rejected, leading to selection of new hypotheses. Compensation for failures may start when the diagnostician feels confident that a correct interpretation has been made of the situation.

Fault diagnosis is a demanding cognitive activity whereby the particular diagnostic strategy will be influenced by the amount of information to be processed in developing a mental model of the situation, the number of hypotheses consistent with available evidence, and the activities required for testing hypotheses. In turn, these factors are influenced by several characteristics of the system or the work environment, including:

- The number of interacting components of the system
- The degrees of freedom in the operation of the system
- The number of system components that can fail simultaneously
- The transparency of the mechanisms of the system

Time constraints and high risk may also add up, thus increasing the difficulty of fault diagnosis. A particularly demanding situation is dynamic fault management (Woods 1994), whereby operators have to maintain system functions despite technical failures or disturbances. Typical fields of practice

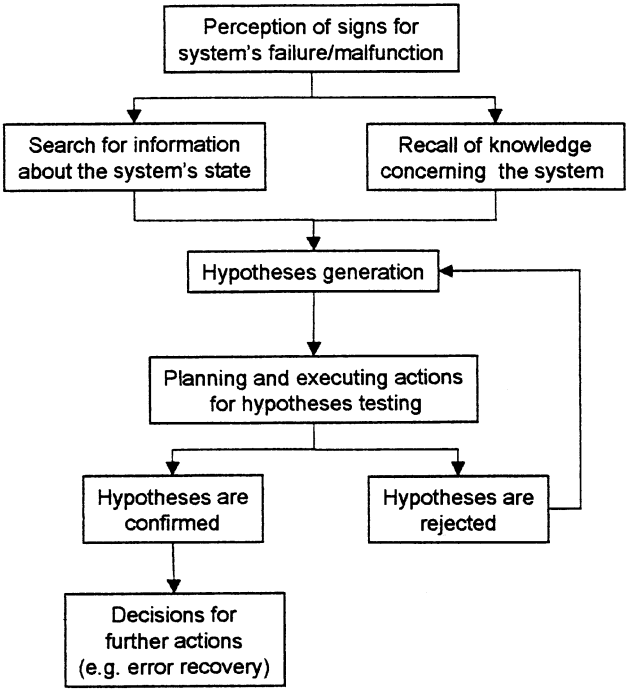

Perception of signs for system's failure/malfunction
Search for information about the system's state
Recall of knowledge concerning the system
Hypotheses generation
Planning and executing actions for hypotheses testing
Hypotheses are confirmed
Hypotheses are rejected
Decisions for further actions (e.g. error recovery)

Figure 6 Typical Processes in Fault Diagnosis.

where dynamic fault management occurs are flight deck operations, control of space systems, anesthetic management, and process control.

When experienced personnel perform fault diagnosis, they tend to use several heuristics to overcome their cognitive limitations. Although heuristics may decrease mental workload, they may often lead to cognitive biases and errors. It is worth considering some heuristics commonly used when searching for information, interpreting a situation, and making diagnostic decisions.

Failures in complex systems may give rise to large amounts of information to be processed. Experienced people tend to filter information according to its informativeness—that is, the degree to which it helps distinguish one failure from another. However, people may be biased when applying heuristics. For instance, diagnosticians may accord erroneous or equal informative value to a pattern of cues (''as-if'' bias: Johnson et al. 1973). In other cases, salient cues, such as noises, bright lights, and abrupt onsets of intensity or motion, may receive disproportionate attention (salience bias: Payne 1980).

Diagnosis may be carried out under psychological stress, which can impose limitations in entertaining a set of hypotheses about the situation (Rasmussen 1981; Mehle 1982; Lusted 1976). To overcome this cognitive limitation, experienced diagnosticians may use alternative strategies for hypothesis generation and testing. For instance, diagnosticians may start with a hypothesis that is seen as the most probable one; this probability may be either subjective, based on past experience, or communicated by colleagues or the hierarchy. Alternatively, they may start with a hypothesis associated with a high-risk failure, or a hypothesis that can be easily tested. Finally, they may start with a hypothesis that readily comes to mind (availability bias: Tversky and Kahneman 1974).

Another heuristic used in diagnosis is anchoring, whereby initial evidence provides a cognitive anchor for the diagnostician's belief in this hypothesis over the others (Tversky and Kahneman 1974). Consequently, people may tend to seek information that confirms the initial hypothesis and avoid any disconfirming evidence (Einhorn and Hogarth 1981; DeKeyser and Woods 1990). This bias is also known as cognitive tunnel vision (Sheridan 1981).

Finally, Einhorn and Hogarth (1981), Schustack and Sternberg (1981), and DeKeyser and Woods (1990) describe situations where diagnosticians tend to seek—and therefore find—information that confirms the chosen hypothesis and to avoid information or tests whose outcomes could reject it. This is known as confirmation bias. Two possible causes for the confirmation bias have been proposed: (1) people seem to have greater cognitive difficulty dealing with negative information than with positive information (Clark and Chase 1972); (2) abandon a hypothesis and reformulating a new one requires more cognitive effort than does searching for and acquiring information consistent with the first hypothesis (Einhorn and Hogarth 1981; Rasmussen 1981).

3.2. Decision Making

Decision making is a cognitive process whereby a person tries to choose a goal and a method that would stabilize the system or increase its effectiveness. In real-world situations, goals may be insufficiently defined, and thus goal setting becomes part of decision making. Furthermore, evaluation criteria for choosing among options may vary, including economic, safety, and quality considerations. In such cases, optimizing on most criteria may become the basis for a good decision. Managerial planning, political decisions, and system design are typical examples of decision making.

Traditional models of decision making have adopted a rational approach that entailed:

1. Determining the goal(s) to be achieved and the evaluation criteria
2. Examining aspects of the work environment
3. Developing alternative courses of action
4. Assessing alternatives
5. Choosing an optimal course of action

Rational models of decision making view these stages as successive; however, in real fields of practice their sequence may change (Marmaras et al. 1992; Klein 1997; Dreyfus 1997) as a function of several factors in the work environment.

Factors that could influence the difficulty of decision making include:

- The amount of information that the decision maker has to consider
- The uncertainty of available information
- The dynamic nature of the work environment
- The complexity of evaluation criteria
- The alternative courses of action that can be developed
- The time constraints imposed on the decision maker
- The risk related to possible decisions

To overcome limitations in human cognition, experienced decision makers may use a number of heuristics. As in diagnosis, heuristics may allow decision makers to cope with complexity in real-world situations but on the other hand, they may lead to cognitive biases and errors. For instance, decision makers may use information selectively and make speculative inferences based on limited data. Hayes-Roth (1979) has called this heuristic ''opportunistic thinking,'' which is similar to Simon's (1978) concept of ''satisfying and search cost''; in other words, decision makers may pursue a trade-off between seeking more information and minimizing the cost in obtaining information. Although this strategy can simplify decision making, important data may be neglected, leading to erroneous or suboptimal decisions.

To cope with complexity and time pressure, decision makers tend to acquire sufficient evidence to form a mental representation of the situation and examine a rather limited set of alternative actions (Marmaras et al. 1992). Instead of generating a complete set of alternatives at the outset and subsequently performing an evaluation based on optimizing several criteria, decision makers may start with an option that has been incompletely assessed. This heuristic could be attributed to the fact that experienced decision makers possess a repertoire of well-practiced responses accessed through recognition rather than conscious search (Simon 1978). Limited consideration of alternatives, however, can lead to ineffective practices. For instance, if the situation at hand differs in subtle ways from previous ones, suboptimal solutions may be adopted. Furthermore, in domains such as system design and managerial planing, new innovative solutions and radical departures may be of great advantage.

In dynamic systems, the distinguishing features of a situation may change over time and new events may add up. Operators should reflect on their thinking and revise their assessment of the situation or their earlier decisions in order to take account of new evidence, interruptions, and negative feedback (Weick 1983; Schon 1983; Lindblom 1980). Under stress, however, experienced decision makers may fixate on earlier decisions and fail to revise them at later stages. Thinking/acting cycles may compensate to some extent for cognitive fixation on earlier decisions. That is, initial decisions can be effected on the basis of related experiences from similar situations, but their suitability can be evaluated after the first outcomes; in this way, decision makers can undertake corrective actions and tailor earlier decisions to new circumstances.

3.3. Supporting Diagnosis and Decision Making

To overcome these weaknesses of human cognition, many engineering disciplines, such as artificial intelligence, operations research, and supervisory control, have pursued the development of stand-alone expert systems that support humans in diagnosis and decision making. Although these systems have made a significant contribution to system design, their performance seems to degrade in unfamiliar situations because humans find it rather awkward to cooperate. Operators are forced to repeat the whole diagnostic or decision-making process instead of taking over from the computer advisor. There is a need, therefore, to develop cognitive advisors that enhance the cognitive processes of operators rather than to construct computer advisors capable of independent performance (Woods and Hollnagel 1987; Roth et al. 1987).

To combat several biases related to diagnosis and decision making, support can be provided to human operators in the following ways:

- Bring together all information required to form a mental representation of the situation.
- Present information in appropriate visual forms.
- Provide memory aids.
- Design cognitive aids for overcoming biases.
- Make systems transparent to facilitate perception of different system states.
- Incorporate intelligent facilities for hypothesis testing and evaluation of options.

Ecological interfaces (Vicente and Rasmussen 1992), following the arguments presented earlier, provide a good example of artifacts that meet some of these requirements. Other cognitive aids are presented in Section 5.

Real-world situations requiring diagnosis and decision making can vary, each one presenting its own specific characteristics. As a result, artifacts and cognitive aids require an exhaustive analysis of the demands of the situation, the user requirements, and the user strategies in achieving satisfactory performance. This type of analysis, usually referred to as cognitive analysis, is valuable in putting human performance models into actual practice.

4. COGNITIVE TASK ANALYSIS (CTA)

Ergonomics interventions in terms of man–machine interfaces, cognitive aids, or training programs require a thorough understanding of the work constraints and user strategies. Work constraints can range from system constraints (e.g., time pressure and conflicting goals) to cognitive constraints or limitations and use constraints imposed by the tool itself (e.g., an interface or a computer program). On the other hand, user strategies can range from informal heuristics, retrieved from similar situations experienced in the past, to optimization strategies for unfamiliar events. The analysis of work constraints and use strategies is the main objective of cognitive task analysis (CTA).

Cognitive task analysis involves a consideration of user goals, means, and work constraints in order to identify the *what, how,* and *why* of operator's work (Rasmussen et al. 1994; Marmaras and Pavard 1999). Specifically, CTA can be used to identify:

- The problem-solving and self-regulation strategies* adopted by operators
- The problem-solving processes followed (heuristics)
- The specific goals and subgoals of operators at each stage in these processes
- The signs available in the work environment (both formal and informal signs), the information carried by them, and the significance attached to them by humans
- The regulation loops used
- The resources of the work environment that could help manage workload
- The causes of erroneous actions or suboptimal performance

CTA differs from traditional task analysis, which describes the performance demands imposed upon the human operator in a neutral fashion (Drury et al. 1987; Kirwan and Ainsworth 1992) regardless of how operators perceive the problem and how they choose their strategies. Furthermore, CTA differs from methods of job analysis that look at occupational roles and positions of specific personnel categories (Davis and Wacker 1987; Drury et al. 1987).

4.1. A Framework for Cognitive Task Analysis

Cognitive task analysis deals with how operators respond to tasks delegated to them either by the system or by their supervisors. *Tasks* is used here to designate the operations undertaken to achieve certain goals under a set of conditions created by the work system** (Leplat 1990). CTA looks at several mental activities or processes that operators rely upon in order to assess the current situation, make decisions, and formulate plans of actions. CTA can include several stages:

1. Systematic observation and recording of operator's actions in relation to the components of the work system. The observed activities may include body movements and postures, eye movements, verbal and gestural communications, etc.
2. Interviewing the operators with the aim of identifying the why and when of the observed actions.
3. Inference of operator's cognitive activities and processes.
4. Formulation of hypotheses about operator's competencies,*** by interpreting their cognitive activities with reference to work demands, cognitive constraints and possible resources to manage workload.
5. Validation of hypotheses by repeating stages 1 and 2 as required.

Techniques such as video and tape recording, equipment mock-ups, and eye tracking can be used to collect data regarding observable aspects of human performance. To explore what internal or cognitive processes underlie observable actions, however, we need to consider other techniques, such as thinking aloud while doing the job (i.e., verbal protocols) and retrospective verbalizations. For high-risk industries, event scenarios and simulation methods may be used when on-site observations are difficult or impossible.

CTA should cover a range of work situations representing both normal and degraded conditions. As the analysts develop a better image of the investigated scenario, they may become increasingly aware of the need to explore other unfamiliar situations. Cognitive analysis should also be expanded into how different operating crews respond to the same work situation. Presumably, crews may differ in their performance because of varying levels of expertise, different decision-making styles, and different coordination patterns (see, e.g., Marmaras et al. 1997).

---

* Self-regulation strategies are strategies for deciding how to adapt to different circumstances, monitor complete or interrupted tasks, and detect errors.

** The work system consists of the technological system, the workplace, the physical environment, the organizational and management system, and the socioeconomic policies.

*** *Competencies* is used here to designate the specific cognitive strategies and heuristics the operators develop and use to respond to the task's demands within the constraints imposed by a specific work environment.

The inference of cognitive activities and formulation of hypotheses concerning user competencies requires familiarity with models of human cognition offered by cognitive psychology, ethnology, psycholinguistics, and organizational psychology. Several theoretical models, such as those cited earlier, have already found practical applications in eliciting cognitive processes of expert users. Newell and Simon's (1972) human problem-solving paradigm, for instance, can be used as a background framework for inferring operator's cognitive processes when they solve problems. Hutchins's (1990; 1992) theory of distributed cognition can support analysts in identifying operator resources in managing workload. Rasmussen's (1986) ladder model of decision making can be used to examine how people diagnose problems and evaluate goals. Norman's (1988) action-cycle model can be used to infer the cognitive activities in control tasks. Finally, Reason's (1990) model of human errors can be used to classify, explain, and predict potential errors as well as underlying error-shaping factors. Human performance models, however, have a hypothetical rather than a normative value for the analyst. They constitute his or her background knowledge and may support interpretation of observable activities and inference of cognitive activities. Cognitive analysis may confirm these models (totally or partially), enrich them, indicate their limits, or reject them. Consequently, although the main scope of cognitive analysis is the design of artifacts and cognitive advisory systems for complex tasks, it can also provide valuable insights at a theoretical level.

Models of human cognition and behavior can provide practical input to ergonomics interventions when cast in the form of cognitive probes or questions regarding how operators search their environment, assess the situation, make decisions, plan their actions, and monitor their own performance. Table 1 shows a list of cognitive probes to help analysts infer these cognitive processes that underlie observable actions and errors.

TABLE 1 Examples of Cognitive Probes

Cognitive Activities	Probes
Recognition	• What features were you looking at when you recognized that a problem existed? • What was the most important piece of information that you used in recognizing the situation? • Were you reminded of previous experiences in which a similar situation was encountered?
Interpretation	• At any stage, were you uncertain about either the reliability or the relevance of the information that you had available? • Did you use all the information available to you when assessing the situation? • Was there any additional information that you might have used to assist in assessing the situation?
Decision making	• What were your specific goals at the various decision points? • Were there any other alternatives available to you other than the decision you made? • Why were these alternatives considered inappropriate? • At any stage, were you uncertain about the appropriateness of the decision? • How did you know when to make the decision?
Planning	• Are there any situations in which your plan of action would have turned out differently? • When you do this task, are there ways of working smart (e.g., combine procedures) that you have found especially useful? • Can you think of an example when you improvised in this task or noticed an opportunity to do something better? • Have you thought of any side effects of your plan and possible steps to prevent them or minimize their consequences?
Feedback and self-monitoring	• What would this result tell you in terms of your assessment of the situation or efficiency of your actions? • At this point, do you think you need to change the way you were performing to get the job done?

This need to observe, interview, test, and probe operators in the course of cognitive analysis implies that the role of operators in the conduct of analysis is crucial. Inference of cognitive activities and elicitation of their competencies cannot be realized without their active participation. Consequently, explicit presentation of the scope of analysis and commitment of their willingness to provide information are prerequisites in the conduct of CTA. Furthermore, retrospective verbalizations and online verbal protocols are central to the proposed methodology (Ericsson and Simon 1984; Sanderson et al. 1989).

CTA permits the development of a functional model of the work situation. The functional model should:

1. Describe the cognitive constraints and demands imposed on operators, including multiple goals and competing criteria for the good completion of the task; unreliable, uncertain, or excessive information to make a decision; and time restrictions.
2. Identify situations where human performance may become ineffective as well as their potential causes—e.g., cognitive demands exceeding operator capacities and strategies that are effective under normal situations but may seem inappropriate for the new one.
3. Identify error-prone situations and causes of errors or cognitive biases—e.g., irrelevant or superfluous information, inadequate work organization, poor workplace design, and insufficient knowledge.
4. Describe the main elements of operator's competencies and determine their strengths and weaknesses.
5. Describe how resources of the environment can be used to support the cognitive processes.

The functional model of the work situation can provide valuable input into the specification of user requirements, prototyping, and evaluation of cognitive aids. Specifically, based on elements 1, 2, and 3 of the functional model, situations and tasks for which cognitive aid would be desirable and the ways such aid must be provided can be determined and specified. This investigation can be made responding to questions such as:

- What other information would be useful to the operators?
- Is there a more appropriate form in which to present the information already used as well as the additional new information?
- Is it possible to increase the reliability of information?
- Could the search for information be facilitated, and how?
- Could the treatment of information be facilitated, and how?
- Could we provide memory supports, and how?
- Could we facilitate the complex cognitive activities carried out, and how?
- Could we promote and facilitate the use of the most effective diagnosis and decision-making strategies, and how?
- Could we provide supports that would decrease mental workload and mitigate degraded performance, and how?
- Could we provide supports that would decrease human errors occurrence, and how?

Cognitive aids can take several forms, including memory aids, computational tools, decision aids to avoid cognitive biases, visualization of equipment that is difficult to inspect, and situation-assessment aids. The functional model is also useful in designing man–machine interfaces to support retrieval of solutions and generation of new methods. By representing system constraints on the interface, operators may be supported in predicting side effects stemming from specific actions. On the other hand, the functional model can also be useful in specifying the competencies and strategies required in complex tasks and hence providing the content of skill training.

Furthermore, based on the information provided by elements 4 and 5 of the functional model, the main features of the human–machine interface can also be specified, ensuring compatibility with operators' competencies. The way task's objects should be represented by the system, the type of man–machine dialogues to be used, the procedures to be proposed, and generic or customizable elements of the system are examples of human–computer interface features that can be specified using the acquired data.

Close cooperation among ergonomics specialists, information technology specialists, and stakeholders in the design project is required in order to examine what system functions should be supported by available information technology, what features of the human-computer interface should be realized, and what functions should be given priority.

4.2. Techniques for Cognitive Task Analysis

CTA can be carried out using a variety of techniques, which, according to Redding and Seamster (1994), can include cognitive interviewing, analysis of verbal protocols, multi-dimensional scaling, computer simulations of human performance, and human error analysis. For instance, Rasmussen (1986) has conducted cognitive interviews to examine the troubleshooting strategies used by electronics technicians. Roth et al. (1992) have used cognitive environment simulation to investigate cognitive activities in fault management in nuclear power plants. Seamster et al. (1993) have carried out extensive cognitive task analyses to specify instructional programs for air traffic controllers. These CTA techniques have been used both to predict how users perform cognitive tasks on prototype systems and to analyze the difficulties and errors in already functioning systems. The former use is associated with the design and development of user interfaces in new systems, while the latter use is associated with the development of decision support systems or cognitive aids and training programs.

The results of CTA are usually cast in the form of graphical representations that incorporate the work demands and user strategies. For cognitive tasks that have been encountered in the past, operators may have developed well-established responses that may need some modifications but nevertheless provide a starting framework. For unfamiliar tasks that have not been encountered in the past or are beyond the design-basis of the system, operators are required to develop new methods or combine old methods in new ways. To illustrate how the results of CTA can be merged in a graphical form, two techniques are presented: hierarchical task analysis and the critical decision method.

4.2.1. Hierarchical Task Analysis

The human factors literature is rich in task analysis techniques for situations and jobs requiring rule-based behavior (e.g., Kirwan and Ainsworth 1992). Some of these techniques can also be used for the analysis of cognitive tasks where well-practiced work methods must be adapted to task variations and new circumstances. This can be achieved provided that task analysis goes beyond the recommended work methods and explores task variations that can cause failures of human performance. Hierarchical task analysis (Shepherd 1989), for instance, can be used to describe how operators set goals and plan their activities in terms of work methods, antecedent conditions, and expected feedback. When the analysis is expanded to cover not only normal situations but also task variations or changes in circumstances, it would be possible to record possible ways in which humans may fail and how they could recover from errors. Table 2 shows an analysis of a process control task where operators start up an oil refinery furnace. This is a safety-critical task because many safety systems are on manual mode, radio communications between control room and on-site personnel are intensive, side effects are not visible (e.g., accumulation of fuel in the fire box), and errors can lead to furnace explosions.

A variant of hierarchical task analysis has been used to examine several cognitive activities, such as goal setting and planning, and failures due to slips and mistakes. Variations in human performance were examined in terms of how teams in different shifts would perform the same task and how the same team would respond to changes in circumstances. A study by Kontogiannis and Embrey (1997) has used this technique to summarize findings from online observations of performance, interviews with process operators about their work methods, near-miss reviews, and critical incident analysis. The task analysis in Table 2 has provided valuable input in revising the operating procedures for start-up: the sequence of operations was reorganized, contingency steps were included for variations in circumstances, check boxes were inserted for tracking executed steps, and warnings and cautions were added to prevent human errors or help in their detection and recovery. In addition, the task analysis in Table 2 has been used to redesign the computer-based process displays so that all information required for the same task could be grouped and presented in the same screen. For instance, the oxygen content in flue gases is an indicator of the efficiency of combustion (see last row in Table 2) and should be related to the flow rates of air and fuel; this implies that these parameters are functionally related in achieving the required furnace temperature and thus should be presented on the same computer screen. The analysis of cognitive tasks, therefore, may provide input into several forms of human factors interventions, including control panel design, revision of operating procedures, and development of job aids and training.

4.2.2. Critical Decision Method

The Critical Decision Method (CDM) (Klein et al. 1989) is a retrospective cognitive task analysis based on cognitive interviews for eliciting expert knowledge, decision strategies and cues attended to, and potential errors. Applications of the CDM technique can be found in fireground command, tactical decision making in naval systems, ambulance emergency planning, and incident control in offshore oil industries. The technique relies on subject matter experts (SMEs) recalling a particularly memorable incident they have experienced in the course of their work. The sequence of events and actions are organized on a timeline that can be rearranged as SMEs remember other details of the

TABLE 2. Extract from the Analysis of Starting an Oil Refinery Furnace (a variant of hierarchical task analysis)

Goal Setting	Work Method (planning)	Antecedent Conditions (planning)	Feedback	Common Errors and Problem Detection
2. Establish flames on all burners.	Light first burner (2.1) and take safety precautions (2.2) if flame is not "healthy." Proceed with other burners.	Furnace has been purged with air and free of residual fuel	"Healthy" flames on all burners	√ Selects wrong equipment to test furnace atmosphere. √ Repeated failures to ignite burners may require shutting down the furnace.
2.1. Light selected burner.	Remove slip plate, open burner valve, and ignite burner.	Flame has not been extinguished more than twice before.	"Healthy" flame on selected burner	√ Forgets to light and insert torch before removing slip plate from fuel pipe; explosive mixture could be created in the event of a leaking block valve. √ Tries to light burner by flashing off from adjacent burners. √ Does not check and reset safety trips, which gives rise to burner failing to ignite.
2.2. Take safety precautions.	Close burner valve and inject steam.	Burner flame has been extinguished.	Steam injection for 2–3 minutes	√ Forgets to put air/fuel ratio control and furnace temperature control on manual; it can cause sub-stoichiometric firing √ Forgets to investigate problem if a burner does not light after two attempts.
3. Gradually increase furnace load to target temperature.	Adjust fuel flow (3.1) and air flow (3.2) in cycles until thermal load is achieved.	All burners remain lighted throughout operation.	Temperature of crude oil on target	√ In the event of a burner failure (e.g., flame put out), compensates by increasing fuel to other burners.
3.1. Adjust flow of fuel supply.	Monitor fuel supply indicator and adjust fuel valve.	Burners are fitted with suitable fuel diffusion plugs.	Temperature of crude oil on target	√ Does not consult criterion table for changing plugs at different fuel supplies
3.2. Adjust supply of combustion air.	Monitor oxygen in flue gases and adjust air damper to keep oxygen above limit.		Oxygen content above limit	√ Risk of explosion if oxygen below limit (may indicate unburned fuel).

Source: Kontogiannis and Embrey (1997).

incident. The next stage is to probe SMEs to elicit more information concerning each major decision point. Cognitive probes address the cues attended to, the knowledge needed to make a decision, the way in which the information was presented, the assessment made of the situation, the options considered, and finally the basis for the final choice. The third stage of the CDM technique involves comparisons between experts and novices. The participants or SMEs are asked to comment on the expected performance of a less-experienced person when faced with the same situation. This is usually done in order to identify possible errors made by less experienced personnel and potential recovery routes through better training, operating procedures, and interface design. The results of the cognitive task analysis can be represented in a single format by means of a decision analysis table.

One of the most important aspects of applying the CDM technique is selecting appropriate incidents for further analysis. The incidents should refer to complex events that challenged ordinary operating practices, regardless of the severity of the incident caused by these practices. It is also important that a "no-blame" culture exist in the organization so that participants are not constrained in their description of events and unsuccessful actions. To illustrate the use of the CDM technique, the response of the operating crew during the first half hour of the Ginna nuclear power incident is examined below, as reported in Woods (1982) and INPO (1982).

The operating crew at the Ginna nuclear plant encountered a major emergency in January 1982 due to a tube rupture in a steam generator; as a result, radioactive coolant leaked into the steam generator and subsequently into the atmosphere. In a pressurized water reactor such as Ginna, water coolant is used to carry heat from the reactor to the steam generator (i.e., the primary loop); a secondary water loop passes through the steam generator and the produced steam drives the turbine that generates electricity. A water leak from the primary to the secondary loop can be a potential risk when not isolated in time. In fact, the delayed isolation of the faulted or leaking steam generator was one of the contributory factors in the evolution of the incident. Woods (1982) uses a variant of the CDM technique to analyze the major decisions of the operating crew at the Ginna incident.

Figure 7 shows a timeline of events and human actions plotted against a simplified version of the SRK model of Rasmussen. *Detection* refers to the collection of data from control room instruments and the recognition of a familiar pattern; for instance, the initial plant symptoms leads to an initial recognition of the problem as a steam generator tube rupture (i.e., a SGTR event). However, alternative events may be equally plausible, and further processing of information at the next stage is required. Interpretation then involves identifying other plausible explanations of the problem, predicting the criticality of the situation, and exploring options for intervention. The formulation of

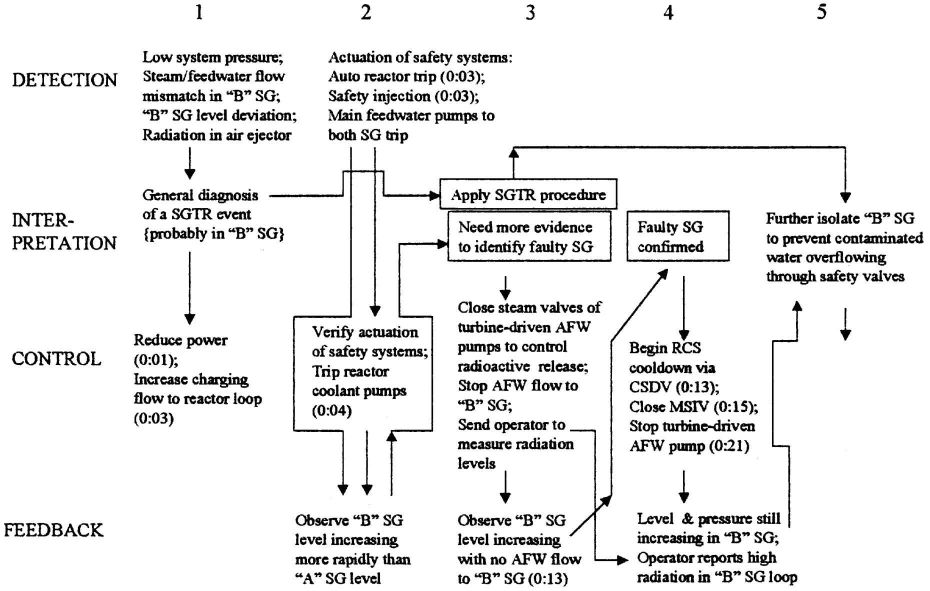

1
2
3
4
5
DETECTION
INTER-PRETATION
CONTROL
FEEDBACK
Low system pressure; Steam/feedwater flow mismatch in "B" SG; "B" SG level deviation; Radiation in air ejector
Actuation of safety systems: Auto reactor trip (0:03); Safety injection (0:03); Main feedwater pumps to both SG trip
General diagnosis of a SGTR event {probably in "B" SG}
Apply SGTR procedure
Need more evidence to identify faulty SG
Faulty SG confirmed
Further isolate "B" SG to prevent contaminated water overflowing through safety valves
Reduce power (0:01); Increase charging flow to reactor loop (0:03)
Verify actuation of safety systems; Trip reactor coolant pumps (0:04)
Close steam valves of turbine-driven AFW pumps to control radioactive release; Stop AFW flow to "B" SG; Send operator to measure radiation levels
Begin RCS cooldown via CSDV (0:13); Close MSIV (0:15); Stop turbine-driven AFW pump (0:21)
Observe "B" SG level increasing more rapidly than "A" SG level
Observe "B" SG level increasing with no AFW flow to "B" SG (0:13)
Level & pressure still increasing in "B" SG; Operator reports high radiation in "B" SG loop

Figure 7 Analysis of the Response of the Crew During the First Half Hour at the Ginna Incident (an example of the critical decision method).

specific actions for implementing a plan and executing actions are carried out at the following stage of control. The large number of actions at the control stage provides a context for the work conditions that may affect overall performance; for instance, high workload at this stage could prevent the crew from detecting emerging events. The feedback stage is similar to the detection stage because both are based on the collection of data; however, in the detection stage, observation follows an alarm, while in the feedback stage, observation is a follow-up to an action. The decision flowchart in Figure 7 has been developed on the basis of cognitive interviews with the operators involved and investigations of the operating procedures in use.

To understand the cognitive processes involved at critical decision points, analysts should employ several cognitive probes similar to those shown in Table 1. It is worth investigating, for instance, the decision of the crew to seek further evidence to identify the tube rupture in one of the two steam generators (Figure 7, column 3). At this point in time, the crew was aware that the level in the B steam generator was increasing more rapidly than the level of the A steam generator (i.e., due to tube rupture) with auxiliary feedwater flows established to both steam generators. However, the crew needed additional evidence to conclude their fault diagnosis. Was this delay in interpretation caused by insufficient plant knowledge? Did the crew interpret the consequences of the problem correctly? Were the instructions in the procedures clear? And in general, what factors influenced this behavior of the crew?

These are some cognitive probes necessary to help analysts explore the way that the crew interpreted the problem. As it appeared, the crew stopped the auxiliary feedwater flow to the B steam generator and came to a conclusion when they observed that its level was still increasing with no input; in addition, an auxiliary operator was sent to measure radioactivity in the suspect steam generator (Figure 7, column 3). Nevertheless, the crew isolated the B steam generator before feedback was given by the operator with regard to the levels of radioactivity. The cognitive interviews established that a plausible explanation for the delayed diagnosis was the high cost of misinterpretation. Isolation of the wrong steam generator (i.e., by closing the main steam isolation valve [MSIV]) would require a delay to reopen it in order to repressurize and establish this steam generator as functional. The crew also thought of a worse scenario in which the MSIV would stick in the closed position, depriving the crew of an efficient mode of cooldown (i.e., by dumping steam directly to the condensers); in the worst-case scenario, the crew would have to operate the atmospheric steam dump valves (ASDVs), which had a smaller cooling function and an increased risk of radiological release.

Other contextual factors that should be investigated include the presentation of instructions in the procedures and the goal priorities established in training regimes. Notes for prompt isolation of the faulty steam generator were buried in a series of notes placed in a separate location from the instructions on how to identify the faulty steam generator. In addition, the high cost of making an incorrect action (an error of commission) as compared to an error of omission may have contributed to this delay. In other words, experience with bad treatment of mistaken actions on the part of the organization may have prompted the crew to wait for redundant evidence. Nevertheless, the eight-minute delay in diagnosis was somewhat excessive and reduced the available time window to control the rising water level in the B steam generator. As a result, the B steam generator overflowed and contaminated water passed through the safety valves; subsequent actions (e.g., block ASDV, column 5) due to inappropriate wording of procedural instructions also contributed to this release.

Cognitive analysis of the diagnostic activity can provide valuable insight into the strategies employed by experienced personnel and the contextual factors that led to delays and errors. It should be emphasized that the crew performed well under the stressful conditions of the emergency (i.e., safety-critical event, inadequate procedures, distractions by the arrival of extra staff, etc.) given that they had been on duty for only an hour and a half. The analysis also revealed some important aspects of how people make decisions under stress. When events have high consequences, experienced personnel take into account the cost of misinterpretation and think ahead to possible contingencies (e.g., equipment stuck in closed position depriving a cooling function). Intermingling fault interpretation and contingency planning could be an efficient strategy under stress, provided that people know how to switch between them. In fact, cognitive analysis of incidents has been used as a basis for developing training programs for the nuclear power industry (Kontogiannis 1996).

The critical decision method has also been used in eliciting cognitive activities and problem-solving strategies required in ambulance emergency planning (O'Hara et al. 1998) and fire-fighting command (Klein et al. 1986). Taxonomies of operator strategies in performing cognitive tasks in emergencies have emerged, including maintaining situation assessment, matching resources to situation demands, planning ahead, balancing workload among crew members, keeping track of unsuccessful or interrupted tasks, and revising plans in the light of new evidence. More important, these strategies can be related to specific cognitive tasks and criteria can be derived for how to switch between alternative strategies when the context of the situation changes. It is conceivable that these problem-solving strategies can become the basis for developing cognitive aids as well as training courses for maintaining them under psychological stress.

5. DESIGNING COGNITIVE AIDS FOR COMPLEX COGNITIVE TASKS

Recent developments in information technology and system automation have provided challenging occasions for developing computerized artifacts aiming to support cognitive processes such as diagnosis, decision making, and planning. This enthusiasm with the capabilities of new technology has resulted in an overreliance on the merits of technology and the development of stand-alone systems capable of undertaking fault diagnosis, decision making, and planning. Recent reviews of such independent computer consultants (e.g., Roth et al. 1997), however, have found them to be difficult to use and brittle in the face of novel circumstances. Woods et al. (1990) argue that many of these systems cannot anticipate operator responses, provide unsatisfactory accounts of their own goals, and cannot redirect their line of reasoning in cases of misperception of the nature of the problem. To make cognitive aids more intelligent and cooperative, it is necessary to examine the nature of human cognition in complex worlds and its coupling with the wider context of work (Marmaras and Pavard 2000). Because of the increasing interest in developing cognitive aids for complex tasks, this section presents two case studies that address this issue in an applied context.

5.1. The Case of a Cognitive Aid for CNC Lathe Programming

The scope of this study was to design a cognitive aid for CNC lathe programming (Marmaras et al. 1997) by adopting a problem-driven approach. This approach combines an analysis of the task in terms of constraints and cognitive demands with an analysis of user strategies to cope with the problem (Marmaras et al. 1992; Woods and Hollnagel 1987). A cognitive task analysis was undertaken in the first place in order to identify the cognitive demands of the task and investigate likely problem-solving strategies leading to optimal and suboptimal solutions. On the basis of the cognitive task analysis and the results of a follow-up experiment, an information technology cognitive aid was developed for CNC lathe programming. A prototype of the cognitive aid was evaluated in a second experiment.

5.1.1. Cognitive Task Analysis

Programming a CNC lathe requires the development of a program that will guide the lathe to transform a simple cylindrical part into a complex shape. This cognitive task requires planning and codification, and it is very demanding due to several interdependent constraints, competing criteria for task efficiency, delayed feedback, and lack of memory aids for maintaining a mental image of the manufacturing process.

A cognitive analysis of the activities and strategies was conducted on the basis of real-time observations of experienced operators programming their CNC lathes and off-the-job interviews. The analysis identified three main cognitive tasks:

1. Planning the whole process of manufacturing, which entails deciding upon:
    - The order of cutting several elements of the complex shape
    - The cutting tool to be used at each stage of the cutting process
    - The movement of the tool at each stage of the cutting process (e.g., starting and ending points, type of the movement, and speed)
    - The number of iterative movements of the tool at each stage
    - The speed of rotation of the part
2. Codification of the manufacturing process to the programming language of the CNC machine. The operator has to consider the vocabulary of this language to designate the different objects and functions of the manufacturing process, as well as the grammar and syntax of the language.
3. Introduction of the program to the machine by using various editing facilities and by starting the execution of the program. It is worth noting that at this stage certain omission and syntax errors can be recognized by the computer logic and the operator may be called upon to correct them.

Table 3 summarizes the results of the cognitive analysis for the planning task. Experienced operators must take a number of constraints into account when deciding upon the most suitable strategy to manufacture the part. These include:

- The shape of the part and the constraints imposed by its material.
- The constraints of the machine tool, e.g., the rotating movement of the part, the area and movement possibilities of the cutting tool, its dimensions, and its cutting capabilities.
- The product quality constraints derived from the specification; these are affected by the speeds of the part and tools, the type of tools, and the designation of their movements.

TABLE 3 Decision Table for the Task of Planning How to Cut a Part in a CNC Lathe

Tasks	Decision Choices	Constraints	Sources of Difficulty	Strategies
Planning the manufacturing process	• Order of cutting • Cutting tools at each phase • Movement of tools • Iterative tool movements • Rotation speed of the part	• Part shape and material • Tool constraints • Product quality • Manufacturing time	• Interdependent constraints • Competing criteria for task efficiency • Lack of real-time feedback • Lack of aids for mental imagery of manufacturing process	• Serial strategy • Optimization strategy

- The manufacturing time, which is affected by the number of tool changes, the movements the tools make without cutting (idle time of the tool) and the speeds of the part and cutting tools.
- The safety considerations, such as avoiding breaking or destroying the part or the tools; safety rules concerning the area of tool movement and the speeds of the part and cutting tools.

An analysis of how experienced operators decide to use the CNC lathe in cutting parts into complex shapes revealed a variety of problem strategies ranging from serial to optimized strategies. The simplest strategy, at one end of the spectrum, involved cutting each component of the part having a different shape separately and in a serial order (e.g., from right to left). A number of heuristics have been introduced in the problem-solving process resulting in such a serial strategy, including:

- Decomposition of the part to its components (problem decomposition)
- Specification and prioritization of several criteria affecting task efficiency (criteria prioritization)
- Simplification of mental imagery of the manufacturing process (mental representation)
- Use of a unique frame of reference regarding the part to be manufactured (frames of reference minimization)

At the other end of the spectrum, a more complex strategy was adopted that relied on optimizing a number of criteria for task efficiency, such as product quality, manufacturing time, and safety considerations. A case in point is a cutting strategy that integrates more than one different shape into one tool movement, which alters the serial order of cutting the different shapes of the part. This strategy would decrease the time-consuming changes of tools and the idle time, thus minimizing the manufacturing time. Optimized cutting strategies require complex problem-solving processes, which may include:

- Adopting a holistic view of the part to be manufactured
- Considering continuously all the criteria of task efficiency
- Using dynamic mental imagery of the cutting process that includes all intermediate phases of the whole part
- Adopting two frames of reference regarding the part to be manufactured and the cutting tools in order to optimize their use

Two hypotheses have been formulated with respect to the observed differences in the performance of CNC lathe programmers: (1) very good programmers will spend more time on problem formulation before proceeding to the specification of the program than good programmers, and (2) very good programmers will adopt an optimized cutting strategy while good programmers would settle for a serial strategy. Consequently, very good programmers will adopt a more complex problem-solving process than good programmers. To test these hypotheses, an experiment was designed that would also provide valuable input into the cognitive task analysis utilized in this study. Details on the design of the experiment and the results obtained can be found in Marmaras et al. (1997).

5.1.2. *User Requirements for a Cognitive Aid Supporting CNC Lathe Programming*

The cognitive task analysis of programming the CNC lathe provided useful insights into the design of an information technology system that supports CNC programmers. The scope of this cognitive aid was twofold. On the one hand, the aid should guide programmers in using efficient cutting strategies, and on the other hand, it should alleviate the cognitive demands that are often associated with optimized problem-solving strategies. Specifically, the cognitive aid could have the following features:

- Support users in deciding upon a cutting strategy at the front-end stages of the programming process
- Generate several suggestions about efficient strategies, e.g., "try to integrate as many different shapes as possible into one movement of the cutting tool" and "avoid many changes of the cutting tool"
- Support human memory throughout the whole programming process
- Provide real-time feedback by showing programmers the effects of their intermediate decisions through real-time simulation
- Facilitate the standard actions of the whole programming process

A description of the prototype cognitive aid that was developed in accordance with the user requirements can be found in Marmaras et al. (1997). Another experiment showed that the quality of the CNC lathe programs was improved and the time required to program the lathe was decreased by approximately 28% when operators used the proposed cognitive aid. Not only did the performance of good CNC lathe programmers improve, but so did the performance of the very good programmers.

5.2. The Case of a Cognitive Aid for Managerial Planning

This section presents a study of user requirements in the design of an information technology system that supports high-level managerial planning tasks in small to medium-sized enterprises (Laios et al. 1992; Marmaras et al. 1992). High-level managerial planning tasks, usually referred to as strategic or corporate planning, involve a series of cognitive processes preceding or leading to crucial decisions about the future course of a company. Managerial planning entails:

1. *Strategies* or high-level decisions that have a long-lasting influence on the enterprise
2. *Tactics* or lower-level strategies regarding policies and actions
3. *Action plans* that specify detailed courses of action for the future

In large companies, strategic planning is an iterative, lengthy, and highly formalized process involving many managers at several levels in the organization. In contrast, in small companies, the manager-owner is the sole, most important actor and source of information on planning issues; in this case, planning is an individual cognitive process without the formalism required in larger organizations.

Elements that affect planning decisions are past, present, and expected future states of the external environment of the firm, which have a complex relationship with the internal environment. Managers may perceive these elements as threats or opportunities in the marketplace (external environment) and strengths or weaknesses of the enterprise (internal environment). Effective planning decisions should neutralize threats from the external environment and exploit its opportunities, taking into account the strengths and weaknesses of the company. At the same time, effective planning decisions should increase the strengths of the company and decrease its weaknesses.

A literature review and three case studies were undertaken in order to identify the main elements of the external and internal environments that may constrain managerial planning. The main constraints identified were:

- *Complexity/multiplicity of factors:* The external and internal environments of an enterprise include a large number of interacting factors. Market demand, intensity of competition, socioeconomic situation, labor market, and technology are examples of the external environment factors. Product quality, process technology, distribution channels, and financial position are examples of the internal environment factors.
- *Change/unpredictability:* The world in which firms operate is constantly changing. Very often these changes are difficult to predict with regard to their timing, impact, and size effects. Managers therefore face a great degree of uncertainty.
- *Limited knowledge* with respect to the final impact of planning decisions and actions. For example, what will be the sales increase resulting from an advertising campaign costing X, and what from a product design improvement costing Y?

- *Interrelation between goals and decisions:* Planning decisions made in order to achieve a concrete goal may refute—if only temporarily—some other goals. For example, renewal of production equipment aiming at increasing product quality may refute the goal of increased profits for some years.
- *Risk related to planning decisions:* The potential outcomes of planning decisions are crucial to the future of the enterprise. Inaccurate decisions may put the enterprise in jeopardy, often leading to important financial losses.

Providing valuable input to cognitive analysis was a series of structured interviews with managers from 60 small- to medium-sized enterprises in order to collect taxonomic data about typical planning tasks in various enterprises, environmental and internal factors, and personality characteristics. In other words, the aim of these interviews was to identify the different situations in which managerial planning takes place.

The scenario method was used to conduct a cognitive analysis of managerial planning because of the difficulties involved in direct observations of real-life situations. The scenario presented a hypothetical firm, its position in the market, and a set of predictions regarding two external factors: an increase in competition for two of the three types of products and a small general increase in market demand. The description of the firm and its external environment was quite general, but realistic so that managers could create associations with their own work environment and therefore rely on their own experience and competencies. For example, the scenario did not specify the products manufactured by the hypothetical firm, the production methods and information systems used, or the tactics followed by the firm. The knowledge elicited from the three case studies and the structured interviews was used to construct realistic problem scenarios.

Scenario sessions with 21 practicing managers were organized in the following way. After a brief explanation about the scope of the experiment, managers were asked to read a two-page description of the scenario. Managers were then asked to suggest what actions they would take had they owned the hypothetical firm and to specify their sequence of performance. Additional information about the scenario (e.g., financial data, analysis of profits and losses, production and operating costs) were provided in tables in case managers would like to use them. Managers were asked to think aloud and were allowed to take notes and make as many calculations as needed. The tape recordings of the scenario sessions were transcribed and verbal protocols were analyzed using a 10-category codification scheme (see Table 4). The coded protocols indicated the succession of decision-making stages and the semantic content of each stage—for example, the sort of information acquired, the specific goals to be attained, or the specific tactic chosen. Details of the verbal protocol analysis can be found in Marmaras et al. (1992). A brief presentation of the main conclusions drawn from the cognitive analysis follow, with particular emphasis on the managers' cognitive strategies and heuristics used in managerial planning.

- *Limited generation of alternative tactics:* Most managers did not generate at the outset an extensive set of alternative tactics to select the most appropriate for the situation. Instead, as

TABLE 4 Ten-Category Scheme for Interpreting the Verbal Protocols of Managers

Code	Description
C1	*Information acquisition* from the data sheets of the scenario or by the experimenter
C2	Statement about the *tactic* followed by the manager (e.g., "I will reduce the production of product C and I will focus on product A . . .")
C3	Reference to *threats or opportunities, strengths and weaknesses* of the firm (e.g., ". . . with these tactics, I will neutralize competitors penetration in the market . . .")
C4	Reference to *conditions* necessary for implementing a tactic (e.g., "I will apply tactic X only if tactic Y does not bring about the expected results . . .")
C5	Reference to the *time* a tactic would be put into effect and the duration of its implementation (e.g., "I will reduce the product price immediately . . .")
C6	Reference to *goals* (e.g., ". . . with this tactic I expect sales to grow . . .")
C7	Reference to data and *calculations* in order to form a better picture of the present planning situation
C8	Explicit reference to a *generic strategy* (e.g., "I would never reduce the prices of the product . . .")
C9	*Justification* of the selected strategy
C10	Calculations for *quantitative evaluation* of selected tactics

soon as they acquired some information and formed a representation of the situation, they defined a rather limited set of tactics that they would apply immediately (e.g., C1→C2, C1→C7→C3→C2, C1→C3→C2, C1→C3→C1→C2). This finding could be attributed to the fact that experienced managers possess an extensive repertoire of experiences accessed through recognition rather than conscious search. Optimization strategies are time consuming and difficult to sustain under the constraints of the work environment and the limited knowledge regarding the impact of decisions. In addition, the high risks associated with different decisions would probably push managers to adopt already tested tactics. Limited generation of alternative tactics, however, may lead to ineffective practices. For instance, a past solution may be inappropriate if the current situation has some subtle differences from others experienced in the past. In the case of managerial planning, new innovative solutions and radical departures may be of great importance.

- *Acting/thinking cycles:* After deciding upon a first set of actions, managers would wait for immediate feedback before proceeding to corrective actions or applying other tactics. This behavior compensates to some extent for the potential negative consequences of the previous strategy. That is, early decisions may be based on past experiences, but their suitability can be evaluated later on when feedback becomes available; thus, managers may undertake other corrective actions. However, these cycles of acting/thinking behavior may lead to delays in acting, increased costs, or money loss.
- *Limited use of information and analysis:* Most managers avoided speculating on the available predictive quantitative account data and did not make any projections in their evaluation of selected tactics. Instead, they quickly came up with ideas about what to do and stated that they would base their evaluation and future actions on specific outcomes of their initial actions. Decisions were justified by making references to environmental factors (C2→C3, C3→C2), other superordinate goals (C2→C6, C6→C2, C8→C2, C2→C8), and feedback from previous actions. This behavior may be attributed to the constraints of the environment, the uncertainty in making predictions, and the interleaving goals that may render a quantitative evaluation of tactics rather difficult or even impossible. The possible negative consequences of this behavior are that data important for planning decisions may be neglected during the planning process.
- *Lack of quantitative goals and evaluation:* The results of the analysis suggested that, in general, managers do not set themselves quantitative goals that will influence the selection and evaluation of their actions. Instead, their planning decisions seem to be based mainly on assessments of external and internal environment factors and certain generic goals related to past experiences. This observation provides evidence supporting the criticism of several authors (e.g., Lindblom 1980) with respect to the relevance of goal-led models; it also suggests that cognitive aids should not be based on optimization models of managerial planning.

Drawing on the results of the cognitive analysis, an information technology system that supports managerial planning in small- to medium-sized enterprises was specified and designed. The system took the form of an active framework that supports cognitive processes in the assessment of environmental factors and generation, selection, and evaluation of strategies and tactics. At the same time, the support system would limit some negative consequences of the managers' cognitive strategies and heuristics. With the use of appropriate screens and functions, the proposed cognitive aid should provide support in the following ways:

- Assisting managers in considering additional data in the planning process and identifying similarities to and differences from other experiences in the past. This could be done by presenting menus of external and internal factors in the planning environment and inviting managers to determine future states or evolutions of system states.
- Enriching the repertoire of planning decisions by presenting menus of candidate options and tactics relevant to the current situation.
- Supporting the acting/thinking process by providing crucial information concerning the state of external and internal environment factors at the successive action/thinking cycles and the different planning decisions made during these cycles.

The planning process proposed by the cognitive aid could be achieved through a series of steps. First, an assessment should be made of the past, present, and future states of environmental factors; this is in contrast to the setting of quantitative goals, as proposed by formal models of strategic planning. The next step involves the specification of alternative tactics and evaluation using qualitative and, optionally, quantitative criteria. Additional steps with regard to the setting of generic strategies and goals have been included in between. However, some degrees of freedom should be allowed in the choice and sequence of these steps. Although this procedure imposes a certain formalism in managerial planning (this is inevitable when using an information technology system), sufficient compatibility has been achieved with the cognitive processes entailed in managerial planning. A prototype of the proposed cognitive aid was evaluated by a number of managers of small- to medium-sized enterprises, and its functionality and usability were perceived to be of high quality.

6. CONCLUDING REMARKS

Advances in information technology and automation have changed the nature of many jobs by placing particular emphasis on cognitive tasks and, on the other hand, by increasing their demands, such as coping with more complex systems, higher volume of work, and operating close to safety limits. As a result, successful ergonomic interventions should consider the interactions in the triad of "users, tools or artifacts, work environment." These interactions are affected by the user strategies, the constraints of the work environment, and the affordances and limitations of the work tools. In this respect, techniques of cognitive task analysis are valuable because they explore how users' cognitive strategies are shaped by their experience and interaction with the work environment and the tools. This explains why a large part of this chapter has been devoted to illustrating how cognitive task analysis can be used in the context of applied work situations.

Cognitive analysis provides a framework for considering many types of ergonomic interventions in modern jobs, including the design of man–machine interfaces, cognitive tools, and training programs. However, cognitive analysis requires a good grasp of human cognition models in order to understand the mechanisms of cognition and develop techniques for eliciting them in work scenarios or real-life situations. The cognitive probes and the critical decision method, in particular, demonstrate how models of human cognition can provide direct input into cognitive task analysis. Further advances in cognitive psychology, cognitive science, organizational psychology, and ethnography should provide more insights into how experienced personnel adapt their strategies when work demands change and how they detect and correct errors before critical consequences are ensued.

The chapter has also emphasized the application of cognitive analysis in the design of information technology cognitive aids for complex tasks. The last two case studies have shown that cognitive aids should not be seen as independent entities capable of their own reasoning but as agents interacting with human users and work environments. Because unforeseen situations beyond the capabilities of cognitive aids are bound to arise, human operators should be able to take over. This requires that humans be kept in the loop and develop an overall mental picture of the situation before taking over. In other words, user-aid interactions should be at the heart of system design. Recent advantages of technology may increase the conceptual power of computers so that both users and cognitive aids can learn from each other's strategies.

REFERENCES

Allport, D. A. (1980), "Attention," in *New Directions in Cognitive Psychology,* G. L. Claxton, Ed., Routledge & Kegan Paul, London.

Card, K., Moran, P., and Newell, A. (1986), "The Model Human Processor," in *Handbook of Perception and Human Performance,* Vol. 2, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., John Wiley & Sons, New York.

Clark, H. H., and Chase, W. G. (1972), "On the Process of Comparing Sentences against Pictures," *Cognitive Psychology,* Vol. 3, pp. 472–517.

Davis, L., and Wacker, G. (1987), "Job Design," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 431–452.

DeKeyser, V., and Woods, D. D. (1990), "Fixation Errors: Failures to Revise Situation Assessment in Dynamic and Risky Systems," in *Systems Reliability Assessment,* A. G. Colombo and A. Saiz de Bustamante, Eds., Kluwer Academic, Dordrecht.

Dreyfus, H. (1997), "Intuitive, Deliberative and Calculative Models of Expert Performance," in *Naturalistic Decision Making,* C. Zambok and G. Klein, Eds., Erlbaum, Hillsdale, NJ, pp. 17–28.

Drury, C., Paramore, B., Van Cott, H., Grey, S., and Corlett, E. (1987), "Task Analysis," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 370–401.

Einhorn, H. J., and Hogarth, R. M. (1981), "Behavioral Decision Theory," *Annual Review of Psychology,* Vol. 32, pp. 53–88.

Ericsson, K. A., and Simon, H. (1984), *Protocol Analysis,* MIT Press, Cambridge, MA.

Gibson, J. J. (1979), *The Ecological Approach to Visual Perception,* Houghton Mifflin, Boston.

Green, D. M., and Swets, J. A. (1988), *Signal Detection Theory and Psychophysics,* John Wiley & Sons, New York.

Hayes-Roth, B. (1979), "A Cognitive Model of Action Planning," *Cognitive Science,* Vol. 3, pp. 275–310.

Helander, M. (1987), "Design of Visual Displays," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 507–548.

Hutchins, E. (1990), "The Technology of Team Navigation," in *Intellectual Teamwork,* J. Galegher, R. E. Kraut, and C. Edigo, Eds., Erlbaum, Hillsdale, NJ.

Hutchins, E., and Klausen, T. (1992), "Distributed Cognition in an Airline Cockpit," in *Communication and Cognition at Work,* D. Middleton and Y. Engestrom, Eds., Cambridge University Press, Cambridge.

Hutchins, E., Hollan, J., and Norman, D. (1986), "Direct Manipulation Interfaces," in *User-Centered System Design,* D. Norman and S. Draper, Eds., Erlbaum, Hillsdale, NJ, pp. 87–124.

Johnson, E. M., Cavanagh, R. C., Spooner, R. L., and Samet, M. G. (1973), "Utilization of Reliability Measurements in Bayesian Inference: Models and Human Performance," *IEEE Transactions on Reliability,* Vol. 22, pp. 176–183.

INPO (1982), "Analysis of Steam Generator Tube Rupture Events at Oconee and Ginna," Report 83-030, Institute of Nuclear Power Operations, Atlanta.

Kirwan, B., and Ainsworth, L. (1992), *A Guide to Task Analysis,* Taylor & Francis, London.

Klein, G. (1997), "An Overview of Naturalistic Decision Making Applications," in *Naturalistic Decision Making,* C. Zambok and G. Klein, Eds., Erlbaum, NJ, pp. 49–60.

Klein, G. A., Calderwood, R., and Clinton-Cirocco, A. (1986), "Rapid Decision Making on the Fire Ground," in *Proceedings of the 30th Annual Meeting of Human Factors Society,* Vol. 1 (Santa Monica, CA), pp. 576–580.

Klein, G. A., Calderwood, R., and MacGregor, D. (1989), "Critical Decision Method for Eliciting Knowledge," *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 19, pp. 462–472.

Kontogiannis, T. (1996), "Stress and Operator Decision Making in Coping with Emergencies," *International Journal of Human-Computer Studies,* Vol. 45, pp. 75–104.

Kontogiannis, T., and Embrey, D. (1997), "A User-Centred Design Approach for Introducing Computer-Based Process Information Systems," *Applied Ergonomics,* Vol. 28, No. 2, pp. 109–119.

Laios, L., Marmaras, N., and Giannakourou, M. (1992), "Developing Intelligent Decision Support Systems Through User-Centred Design," in *Methods and Tools in User-Centred Design for Information Technology,* M. Galer, S. Harker, and J. Ziengler, Eds., Elsevier Science, Amsterdam, pp. 373–412.

Leplat, J. (1990), "Relations between Task and Activity: Elements for Elaborating a Framework for Error Analysis," *Ergonomics,* Vol. 33, pp. 1389–1402.

Lindblom, C. E. (1980), "The Science of 'Muddling Thought'," in *Readings in Managerial Psychology,* H. Leavit, L. Pondby, and D. Boje, Eds., University of Chicago Press, Chicago, pp. 144–160.

Lindsay, R. W., and Staffon, J. D. (1988), "A Model Based Display System for the Experimental Breeder Reactor-II," Paper presented at the Joint Meeting of the American Nuclear Society and the European Nuclear Society (Washington, DC).

Lusted, L. B. (1976), "Clinical Decision Making," in *Decision Making and Medical Care,* F. T. De Dombal and J. Gremy, Eds., North-Holland, Amsterdam.

Marmaras, N., and Pavard, B. (1999), "Problem-Driven Approach for the Design of Information Technology Systems Supporting Complex Cognitive Tasks," *Cognition, Technology and Work,* Vol. 1, No. 4, pp. 222–236.

Marmaras, N., Lioukas, S., and Laios, L. (1992), "Identifying Competences for the Design of Systems Supporting Decision Making Tasks: A Managerial Planning Application," *Ergonomics,* Vol. 35, pp. 1221–1241.

Marmaras, N., Vassilakopoulou, P., and Salvendy, G. (1997), "Developing a Cognitive Aid for CNC-Lathe Programming through Problem-Driven Approach," *International Journal of Cognitive Ergonomics,* Vol. 1, pp. 267–289.

McCormick, E. S., and Sanders, M. S. (1987), *Human Factors in Engineering and Design,* McGraw-Hill, New York.

Mehle, T. (1982), "Hypothesis Generation in an Automobile Malfunction Inference Task," *Acta Psychologica,* Vol. 52, pp. 87–116.

Miller, G. (1956), "The Magical Number Seven Plus or Minus Two: Some Limits on our Capacity for Processing Information," *Psychological Review,* Vol. 63, pp. 81–97.

Norman, D. (1988), *The Design of Everyday Things,* Doubleday, New York.

Newell, A., and Simon, H. (1972), *Human Problem Solving,* Prentice Hall, Englewood Cliffs, NJ.

O' Hara, D., Wiggins, M., Williams, A., and Wong, W. (1998), ‘‘Cognitive Task Analysis for Decision Centred Design and Training,’’ *Ergonomics,* Vol. 41, pp. 1698–1718.

Payne, J. W. (1980), ‘‘Information Processing Theory: Some Concepts and Methods Applied to Decision Research,’’ in *Cognitive Processes in Choice and Decision Behavior,* T. S. Wallsten, Ed., Erlbaum, Hillsdale, NJ.

Rasmussen, J. (1981), ‘‘Models of Mental Strategies in Process Control,’’ in *Human Detection and Diagnosis of Systems Failures,* J. Rasmussen and W. Rouse, Eds., Plenum Press, New York.

Rasmussen, J. (1983), ‘‘Skill, Rules and Knowledge: Signals, Signs and Symbols, and other Distinctions in Human Performance Models,’’ *IEEE Transactions on Systems, Man and Cybernetics,* SMC-13, pp. 257–267.

Rasmussen, J. (1985), ‘‘The Role of Hierarchical Knowledge Representation in Decision Making and System Management,’’ *IEEE Transactions on Systems, Man and Cybernetics,* SMC-15, pp. 234–243.

Rasmussen, J. (1986), *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering,* North-Holland, Amsterdam.

Rasmussen, J., Pejtersen, M., and Goodstein, L. (1994), *Cognitive Systems Engineering,* John Wiley & Sons, New York.

Reason, J. (1990), *Human Error,* Cambridge University Press, Cambridge.

Redding, R. E., and Seamster, T. L. (1994), ‘‘Cognitive Task Analysis in Air Traffic Control and Aviation Crew Training,’’ in *Aviation Psychology in Practice,* N. Johnston, N. McDonald, and R. Fuller, Eds., Avebury Press, Aldershot, pp. 190–222.

Roth, E. M., Bennett, K. B., and Woods, D. D. (1987), ‘‘Human Interaction with an Intelligent Machine,’’ *International Journal of Man–Machine Studies,* Vol. 27, pp. 479–525.

Roth, E. M., Woods, D. D., and Pople, H. E., Jr. (1992), ‘‘Cognitive Simulation as a Tool for Cognitive Task Analysis,’’ *Ergonomics,* Vol. 35, pp. 1163–1198.

Roth E. M., Malin, J. T., and Schreckenghost, D. L. (1997), ‘‘Paradigms for Intelligent Interface Design,’’ in M. Helander, T. K. Landauer, and P. Phabhu, Eds., *Handbook of Human–Computer Interaction,* Elsevier Science, Amsterdam, pp. 1177–1201.

Sanderson, P. M., James, J. M., and Seidler, K. S. (1989), ‘‘SHAPA: An Interactive Software Environment for Protocol Analysis,’’ *Ergonomics,* Vol. 32, pp. 463–470.

Schon, D. (1983), *The Reflective Practitioner,* Basic Books, New York.

Schustack, M. W., and Sternberg, R. J. (1981), ‘‘Evaluation of Evidence in Causal Inference,’’ *Journal of Experimental Psychology: General,* Vol. 110, pp. 101–120.

Seamster, T. L., Redding, R. E., Cannon, J. R., Ryder, J. M., and Purcell, J. A. (1993), ‘‘Cognitive Task Analysis of Expertise in Air Traffic Control, *International Journal of Aviation Psychology,* Vol. 3, pp. 257–283.

Serfaty, D., Entin, E., and Hohnston, J. H. (1998), ‘‘Team Coordination Training,’’ in J. A. Cannon-Bowers and E. Salas, Eds., *Making Decisions under Stress: Implications for Individual and Team Training,* American Psychological Association, Washington, DC, pp. 221–245.

Shepherd, A. (1989), ‘‘Analysis and Training of Information Technology Tasks,’’ in *Task Analysis for Human Computer Interaction,* D. Diaper, Ed., Ellis Horwood, Chichester.

Sheridan, T. (1981), ‘‘Understanding Human Error and Aiding Human Diagnosis Behavior in Nuclear Power Plants,’’ in *Human Detection and Diagnosis of Systems Failures,* J. Rasmussen and W. Rouse, Eds., Plenum Press, New York.

Shneiderman, B. (1982), ‘‘The Future of Interactive Systems and the Emergence of Direct Manipulation,’’ *Behavior and Information Technology,* Vol. 1, pp. 237–256.

Shneiderman, B. (1983), ‘‘Direct Manipulation: A Step Beyond Programming Languages,’’ *IEEE Computer,* Vol. 16, pp. 57–69.

Simon, H. (1978), ‘‘Rationality as Process and Product of Thought,’’ *American Economic Review,* Vol. 68, pp. 1–16.

Tversky, A., and Kahneman, D. (1974), ‘‘Judgement under Uncertainty: Heuristics and Biases,’’ *Science,* Vol. 185, pp. 1124–1131.

Vicente, K., and Rasmussen, J. (1992), ‘‘Ecological Interface Design: Theoretical Foundations,’’ *IEEE Transactions on Systems, Man and Cybernetics,* SMC-22, pp. 589–606.

Weick, K. E. (1983), ‘‘Managerial Thought in the Context of Action,’’ in *The Executive Mind,* S. Srivastra, Ed., Jossey-Bass, San Francisco, pp. 221–242.

Wickens, C. (1992), *Engineering Psychology and Human Performance,* HarperCollins, New York.

Woods, D. D. (1982), ‘‘Operator Decision Behavior During the Steam Generator Tube Rupture at the Ginna Nuclear Power Station,’’ Research Report 82-1057-CONRM-R2, Westinghouse Research and Development Center, Pittsburgh.

Woods, D., and Hollnagel, E. (1987), ‘‘Mapping Cognitive Demands in Complex and Dynamic Worlds,’’ *International Journal of Man–Machine Studies,* Vol. 26, pp. 257–275.

Woods, D. D., Roth, E. M., and Bennett, K. B. (1990), ‘‘Explorations in Joint Human–Machine Cognitive Systems,’’ in *Cognition, Computing, and Cooperation,* S. P. Robertson, W. Zachary, and J. B. Black, Eds., Ablex, Norwood, NJ.

CHAPTER 40

Physical Tasks: Analysis, Design, and Operation

WALDEMAR KARWOWSKI
DAVID RODRICK
University of Louisville

1. **INTRODUCTION** **1042**
2. **ENGINEERING ANTHROPOMETRY** **1043**
    - 2.1. Description of Human Body Position 1043
    - 2.2. The Statistical Description of Anthropometric Data 1043
    - 2.3. The Method of Design Limits 1048
    - 2.4. Anthropometric Design Criteria 1048
    - 2.5. Alternative Design Procedures 1049
    - 2.6. Computer-Aided Models of Man 1049
3. **DESIGN FOR HUMAN STRENGTH** **1050**
    - 3.1. Occupational Strength Testing 1052
    - 3.2. Static vs. Dynamic Strengths 1052
    - 3.3. Computer Simulation of Human Strength Capability 1054
    - 3.4. Push–Pull Force Limits 1054
        - 3.4.1. Static Standing Forces 1055
        - 3.4.2. Dynamic Standing Forces 1055
        - 3.4.3. One-Handed Force Magnitudes 1056
        - 3.4.4. Pinch–Pull Force Magnitudes 1056
4. **STATIC EFFORTS AND FATIGUE** **1056**
    - 4.1. Design Limits for Static Work 1056
    - 4.2. Intermittent Static Work 1057
    - 4.3. Static Efforts of the Arm 1058
5. **WORKPLACE ANALYSIS AND DESIGN** **1061**
    - 5.1. Evaluation of Working Postures 1061
    - 5.2. Computer-Aided Analysis of Working Postures 1061
    - 5.3. Postural Analysis Systems 1061
        - 5.3.1. OWAS 1061
        - 5.3.2. Standard Posture Model 1062
    - 5.4. Acceptability of Working Postures 1063
    - 5.5. International Standards 1067
        - 5.5.1. Exposure Assessment of Upper-Limb Repetitive Movements 1067
        - 5.5.2. European Standard for Working Postures During Machinery Operation 1068
6. **OCCUPATIONAL BIOMECHANICS** **1068**
    - 6.1. Definitions 1068
    - 6.2. Principles of Mechanics 1069
    - 6.3. Biomechanical Analysis 1069
        - 6.3.1. Static Analysis 1069
        - 6.3.2. Lever Systems 1069
7. **DESIGN OF MANUAL MATERIALS HANDLING TASKS** **1070**
    - 7.1. Epidemiology of Low-Back Disorders 1070
    - 7.2. MMH Capacity Design Criteria 1071
    - 7.3. The Psychophysiological Approach 1071
    - 7.4. MMH Design Databases 1071
    - 7.5. Psychophysical Models 1072
    - 7.6. The Physiological Approach 1072
    - 7.7. The Biomechanical Approach 1072

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

7.8. Revised NIOSH (1991) Lifting Equation 1076

7.9. Computer Simulation of the Revised NIOSH Lifting Equation (1991) 1079

7.10. Prevention of LBDs in Industry 1080

7.10.1. Job Severity Index 1080

7.10.2. Dynamic Model for Prediction of LBDs in Industry 1081

8. WORK-RELATED MUSCULOSKELETAL DISORDERS OF THE UPPER EXTREMITY 1082

8.1. Characteristics of Musculoskeletal Disorders 1082

8.2. Definitions 1082

8.3. Conceptual Models for Development of WRMDs 1083

8.4. Causal Mechanism for Development of WUEDs 1085

8.5. Musculoskeletal Disorders: Occupational Risk Factors 1086

9. ERGONOMIC DESIGN TO REDUCE WUEDs 1087

9.1. Quantitative Models for Development of WUEDs 1087

9.2. Semiquantitative Job-Analysis Methodology for Wrist/Hand Disorders 1087

9.3. Psychophysical Models: The Maximum Acceptable Wrist Torque 1091

10. MANAGEMENT OF MUSCULOSKELETAL DISORDERS 1091

10.1. Ergonomic Guidelines 1091

10.2. Administrative and Engineering Controls 1092

11. JOB ANALYSIS AND DESIGN 1093

11.1. Risk Factors and Definitions 1093

11.2. Work Organization Risk Factors 1093

11.3. Procedures for Job Analysis and Design 1095

12. SURVEILLANCE FOR JOB ANALYSIS AND DESIGN 1095

12.1. Provisions of the ANSI Z-365 (1999) Draft Standard 1095

12.2. Musculoskeletal Disorder Surveillance 1095

12.3. Worker Reports (Case-Initiated Entry into the Process) 1095

12.4. Analysis of Existing Records and Survey (Past Case(s)-Initiated Entry into the Process) 1095

12.5. Job Surveys (Proactive Entry into the Process) 1096

12.6. ANSI Z-365 Evaluation Tools for Control of WRMDs 1096

12.7. Analysis and Interpretation of Surveillance Data 1096

13. ERGONOMIC PROGRAMS IN INDUSTRY 1097

14. PROPOSED OSHA ERGONOMICS REGULATIONS 1097

14.1. Main Provisions of the Draft Ergonomics Standard 1098

14.2. Job Hazard Analysis and Control 1099

REFERENCES 1100

ADDITIONAL READING 1108

1. INTRODUCTION

According to the Board of Certification in Professional Ergonomics (BCPE 21000), ergonomics is a body of knowledge about human abilities, human limitations, and other human characteristics that are relevant to design. Ergonomic design is the application of this body of knowledge to the design of tools, machines, systems, tasks, jobs, and environments for safe, comfortable, and effective human use. The underlying philosophy of ergonomics is to design work systems where job demands are within the capacities of the workforce. Ergonomic job design focuses on fitting the job to capabilities of workers by, for example, eliminating occurrence of nonnatural postures at work, reduction of excessive strength requirements, improvements in work layout, design of hand tools, or optimizing work/rest requirements (Karwowski 1992; Karwowski and Salvendy 1998; Karwowski and Marras 1999).

Ergonomics is seen today as a vital component of the value-adding activities of the company, with well-documented cost–benefit aspects of the ergonomics management programs (GAO 1997). A company must be prepared to accept a participative culture and utilize participative techniques in implementation of work design principles. The job design-related problems and consequent intervention should go beyond engineering solutions and include all aspects of business processes, including product design, engineering and manufacturing, quality management, and work organizational issues, along the side of task design or worker education and training (Karwowski and Salvendy 1999; Karwowski and Marras 1999; Genaidy et al. 1999).

This chapter deals primarily with work analysis and design, as well as related human performance on physical tasks. The information about cognitive and other human performance aspects can be found in other chapters of this Handbook.

2. ENGINEERING ANTHROPOMETRY

Anthropometry is an empirical science branching from physical anthropology that deals with physical dimensions of human body and its segments, such as body size and form, including location and distribution of center of mass; segment lengths and weights; range of joint movements; and strength characteristics. Anthropometric data are fundamental to work analysis and design. Engineering anthropometry focuses on physical measurements and applies appropriate methods to human subjects in order to develop engineering design requirements (Roebuck et al. 1975). Anthropometry is closely related to biomechanics because occupational biomechanics provides the criteria for the application of anthropometric data to the problems of workplace design (Pheasant 1989).

Anthropometry can be divided into two types: physical anthropometry, which deals with basic dimensions of the human body in standing and sitting positions (see, e.g., Tables 1 and 2), and functional anthropometry, which is task oriented. Both physical and functional anthropometry can be considered in either a static or dynamic sense. Static analysis implies that only the body segment lengths in fixed position will be considered in workplace design. Dynamic analysis requires that acceptability of design be evaluated with respect to the need to move the body from one position to another, as well as the reach and clearance considerations.

An example of the important dynamic data for workplace design is range of joint mobility (Table 3) which corresponds to postures illustrated in Figure 1. Very useful anthropometric data, both static and dynamic, are provided by the Humanscale (Henry Dreyfuss Associates 1981). When anthropometric requirements for the workplace are not met, biomechanical stresses, which may manifest themselves in postural discomfort, low back pain, and overexertion injury, are likely to occur (Grieve and Pheasant 1982). Inadequate anthropometric design can lead to machine safety hazards, loss of motion economy, and poor visibility. In other words, the consequences of anthropometric misfits may of be a biomechanical and perceptual nature, directly impacting worker safety, health, and plant productivity.

2.1. Description of Human Body Position

The anatomical body position depicts a person standing upright, with feet together, arms by the sides, and with palms forward. As a reference posture, this position is symmetrical with respect to so-called *mid-sagittal plane*. All planes parallel to it are also called *sagittal*. The vertical plane perpendicular to the sagittal is called the *coronal* plane. The horizontal (or *transverse*) plane is perpendicular to both the sagittal and coronal planes. Definition of planes of reference are especially important when the body is in other than the anatomical position.

According to Grieve and Pheasant (1982), terms of *relative* body position can be defined as follows. The *medial* and *lateral* positions refer to nearer to or farther from the mid-sagittal plane. The *superior* or *inferior* positions refer to nearer to or further from the top of the body. The *anterior* (*ventral*) and *posterior* (*dorsal*) positions refer to in front of or behind another structure. The *superficial* and *deep* positions refer to nearer to and farther from the body surface, respectively. Nearer to or farther from the trunk positions are called *proximal* and *distal*. Terms of body movements are defined in Table 4.

2.2. The Statistical Description of Anthropometric Data

The concept of normal distribution can used to describe random errors in the measurement of physical phenomena (Pheasant 1989). If the variable is normally distributed, the population may be completely described in terms of its mean ($\bar{x}$) and its standard deviation (s), and specific percentile (X_p) values can be calculated, where: $X_p = \bar{x} + sz$, where z (the standard normal deviate) is a factor for the percentile concerned. Values of z for some commonly used percentiles (X_p) are given in Table 5. Figure 2 depicts data from Humanscale calculated for different percentiles of U.S. females. A word of caution: anthropometric data are not necessarily normally distributed in any given population (Kroemer 1989).

2.3. The Method of Design Limits

The recommendations for workplace design with respect to anthropometric criteria can be established by the principle of *design for the extreme,* also known as the *method of limits* (Pheasant 1989). The basic idea behind this concept is to establish specific boundary conditions (percentile value of the

TABLE 1 US Civilian Body Dimensions (cm) for Ages 20–60 Years

Dimension	Men				Women			
	5th Percentile	50th Percentile	95th Percentile	SD	5th Percentile	50th Percentile	95th Percentile	SD
1. Stature (height)[f]	161.8	173.6	184.4	6.9	149.5	160.5	171.3	6.6
2. Eye height[f]	151.1	162.4	172.7	6.6	138.3	148.9	159.3	6.4
3. Shoulder (acromion height)[f]	132.3	142.8	152.4	6.1	121.1	131.1	141.9	6.1
4. Elbow height[f]	100.0	109.9	119.0	5.8	93.6	101.2	108.8	4.6
5. Knuckle height	69.8	75.4	80.4	3.2	64.3	70.2	75.9	3.5
6. Height, sitting[s]	84.2	90.6	96.7	3.7	78.2	85.0	90.7	3.5
7. Eye height, sitting[s]	72.6	78.6	84.4	3.6	67.5	73.3	78.5	3.3
8. Shoulder height, sitting[s]	52.7	59.4	65.8	4.0	49.2	55.7	61.7	3.8
9. Elbow rest height, sitting[s]	19.0	24.3	29.4	3.0	18.1	23.3	28.1	2.9
10. Knee height, sitting[f]	49.3	54.3	59.3	2.9	45.2	49.8	54.5	2.7
11. Popliteal height, sitting[f]	39.2	44.2	48.8	2.8	35.5	39.8	44.3	2.6
12. Thigh clearance height[s]	11.4	14.4	17.7	1.7	10.6	13.7	17.5	1.8
13. Chest depth	21.4	24.2	27.6	1.9	21.4	24.2	29.7	2.5
14. Elbow–fingertip distance	44.1	47.9	51.4	2.2	38.5	42.1	56.0	2.2
15. Buttock–knee distance, sitting	54.0	59.4	64.2	3.0	51.8	56.9	62.5	3.1
16. Buttock (popliteal distance, sitting)	44.2	49.5	54.8	3.0	43.0	48.1	53.5	3.1
17. Forward reach, functional	76.3	82.5	88.3	5.0	64.0	71.0	79.0	4.5
18. Breadths								
19. Elbow-to-elbow breadth	35.0	41.7	50.6	4.6	31.5	38.4	49.1	5.4
20. Hip breadth, sitting	30.8	35.4	40.6	2.8	31.2	36.4	43.7	3.7
21. Head dimensions	14.4	15.42	16.4	0.59	13.6	14.54	15.5	0.57
22. Head circumference	53.8	56.8	59.3	1.68	52.3	54.9	57.7	1.63
23. Interpupillary distance	5.5	6.20	6.8	0.39	5.1	5.83	6.5	0.44
24. Hand dimensions								
25. Hand length	17.6	19.05	20.6	0.93	16.4	17.95	19.08	1.04
26. Breadth, metacarpal	8.2	8.88	9.8	0.47	7.0	7.66	8.4	0.41
27. Circumference, metacarpal	19.9	21.55	23.5	1.09	16.9	18.36	19.9	0.89
28. Thickness, meta III	2.4	2.76	3.1	0.21	2.5	2.77	3.1	0.18
29. Weight (kg)	56.2	74.0	97.1	12.6	46.2	61.1	89.9	13.80

Adapted from Kroemer 1989, with permission from Taylor & Francis Ltd., London, http://www.tandf.co.uk.

[f] above floor, [s] above seat surface.

TABLE 2 Anthropometric Estimates for Elderly People (cm)

	Men				Women			
Dimension	5th Percentile	50th Percentile	95th Percentile	SD	5th Percentile	50th Percentile	95th Percentile	SD
1. Stature	151.5	164.0	176.5	7.7	140.0	151.5	163.0	7.0
2. Eye height	141.0	153.5	166.0	7.6	130.5	142.0	153.5	6.9
3. Shoulder height	122.5	134.5	146.5	7.2	113.0	123.5	134.0	6.5
4. Elbow height	93.5	102.5	112.0	5.7	86.0	94.5	103.0	5.2
5. Hip height	78.5	87.5	96.5	5.5	70.0	78.0	86.0	4.9
6. Knuckle height	64.0	71.5	78.5	4.5	61.0	68.0	74.5	4.1
7. Fingertip height	55.0	62.0	69.0	4.2	51.5	59.0	66.0	4.3
8. Sitting height	78.5	85.0	92.0	4.2	71.0	78.5	86.5	4.8
9. Sitting eye height	67.5	74.0	80.5	4.0	61.0	68.5	75.5	4.5
10. Sitting shoulder height	49.5	55.5	61.5	3.7	44.5	51.5	58.5	4.2
11. Sitting elbow height	16.0	21.5	27.0	3.4	15.0	20.5	25.5	3.2
12. Thigh thickness	12.0	14.5	17.5	1.7	10.5	14.0	17.0	1.9
13. Buttock–knee length	51.0	56.5	62.0	3.4	49.0	54.5	60.0	3.4
14. Buttock–popliteal length	41.0	47.0	53.0	3.6	40.5	46.0	51.5	3.4
15. Knee height	45.5	51.5	57.0	3.5	43.0	48.0	53.0	3.0
16. Popliteal height	36.5	41.5	47.0	3.2	33.0	38.0	43.0	3.1
17. Shoulder breadth (bideltoid)	38.0	43.0	48.0	3.1	32.5	37.0	41.5	2.7
18. Shoulder breadth (biacromial)	33.5	36.5	40.0	2.0	30.5	33.5	37.0	2.0
19. Hip breadth	29.0	34.0	39.5	3.2	28.5	35.5	42.5	4.3
20. Chest (bust) depth	21.5	25.5	29.0	2.3	20.5	25.5	30.5	3.0
21. Abdominal depth	23.0	29.0	35.5	3.9	20.5	16.0	32.0	3.5
22. Shoulder–elbow length	30.5	34.5	38.0	2.2	28.0	31.0	34.5	2.0
23. Elbow–fingertip length	41.0	45.0	48.5	2.3	37.0	40.5	44.0	2.2
24. Upper limb length	67.0	73.5	80.0	3.9	60.5	66.5	72.5	3.6
25. Shoulder-grip length	57.0	62.5	68.5	3.5	51.0	56.5	62.0	3.3
26. Head length	17.0	18.5	20.0	0.8	15.5	17.0	18.0	0.8
27. Head breadth	7.5	13.5	15.5	0.7	12.5	13.5	14.5	0.6
28. Hand length	16.0	18.0	19.5	1.1	14.5	16.5	18.0	1.0
29. Hand breadth	7.5	8.0	9.0	0.5	6.5	7.0	8.0	0.5
30. Foot length	22.5	25.0	27.5	1.5	20.0	22.5	24.5	1.4
31. Foot breadth	8.0	9.0	10.5	0.7	7.5	8.5	9.5	0.6
32. Span	154.0	169.0	184.0	9.1	138.0	151.5	164.5	8.0
33. Elbow span	80.5	89.0	97.5	5.2	72.0	80.0	88.0	4.8
34. Vertical grip reach (standing)	177.0	191.5	206.0	8.8	164.0	177.0	190.0	8.0
35. Vertical grip reach (sitting)	106.5	117.5	128.0	6.6	98.5	108.5	118.0	6.0
36. Forward grip reach	67.5	73.5	79.5	3.8	60.5	66.0	72.0	3.5

Adapted from Pheasant 1986.

TABLE 3 Range of Joint Mobility Values Corresponding to Postures in Figure 1

Movement	Mean	S.D.	5 Percentile	95 Percentile
Shoulder flexion	188	12	168	208
Shoulder extension	61	14	38	84
Shoulder abduction	134	17	106	162
Shoulder adduction	48	9	33	63
Shoulder medial rotation	97	22	61	133
Shoulder lateral rotation	34	13	13	55
Elbow flexion	142	10	126	159
Forearm supination	113	22	77	149
Forearm pronation	77	24	37	117
Wrist flexion	90	12	70	110
Wrist extension	113	13	92	134
Hip abduction	53	12	33	73
Hip adduction	31	12	11	51
Hip medial rotation (prone)	39	10	23	56
Hip lateral rotation (prone)	34	10	18	51
Hip medial rotation (sitting)	31	9	16	46
Hip lateral rotation (sitting)	30	9	15	45
Knee flexion, voluntary (prone)	125	10	109	142
Knee flexion, forearm (prone)	144	9	129	159
Knee flexion, voluntary (standing)	113	13	92	134
Knee flexion forced (kneeling)	159	9	144	174
Knee medial rotation (sitting)	35	12	15	55
Knee lateral rotation (sitting)	43	12	23	63
Ankle flexion	35	7	23	47
Ankle extension	38	12	18	58
Foot inversion	24	9	9	39
Foot eversion	23	7	11	35

Adapted from Chaffin and Andersson, *Occupational Biomechanics,* 3rd Ed. Copyright © 1999. Reprinted by permission of John Wiley & Sons, Inc., New York.

[a] Measurement technique was photography. Subjects were college-age males.

Data are in angular degrees.

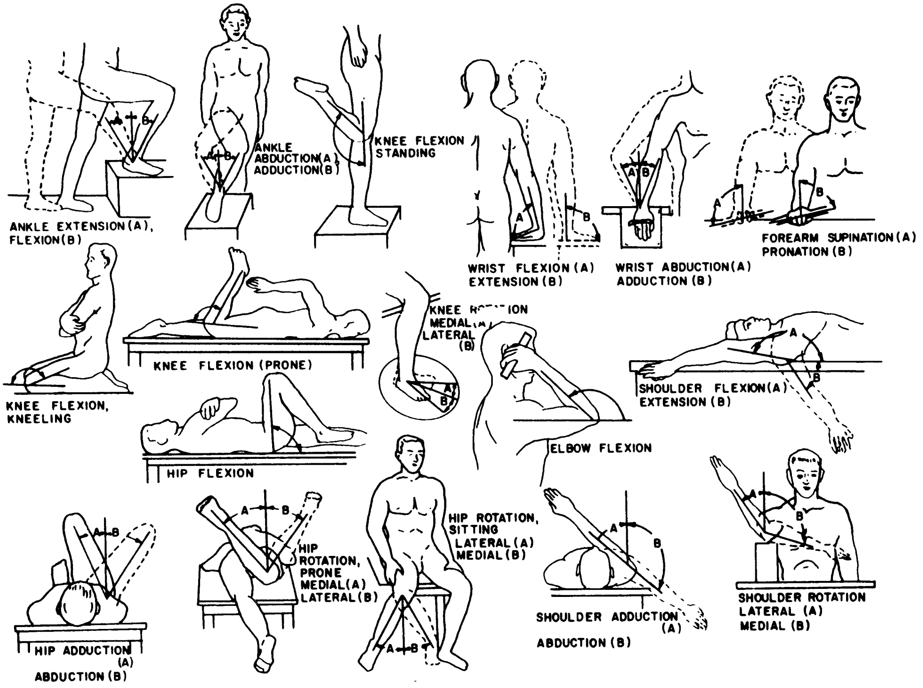

ANKLE EXTENSION (A), FLEXION (B)
A
B
ANKLE ABDUCTION (A) ADDUCTION (B)
KNEE FLEXION STANDING
WRIST FLEXION (A) EXTENSION (B)
WRIST ABDUCTION (A) ADDUCTION (B)
FOREARM SUPINATION (A) PRONATION (B)
KNEE FLEXION, KNEELING
KNEE FLEXION (PRONE)
KNEE ROTATION MEDIAL (A) LATERAL (B)
SHOULDER FLEXION (A) EXTENSION (B)
HIP FLEXION
ELBOW FLEXION
HIP ADDUCTION (A) ABDUCTION (B)
HIP ROTATION, PRONE MEDIAL (A) LATERAL (B)
HIP ROTATION, SITTING LATERAL (A) MEDIAL (B)
SHOULDER ADDUCTION (A) ABDUCTION (B)
SHOULDER ROTATION LATERAL (A) MEDIAL (B)

Figure 1 Illustration of Joint Mobility. (Adapted from Chaffin et al., *Occupational Biomechanics,* 3rd Ed. Copyright © 1999. Reprinted by permission of John Wiley & Sons, Inc., New York.)

TABLE 4 Terms of Movements

Body Part	Movement	Plane of Reference		
		Sagittal	Coronal	Transverse
Trunk[1]	Bend forward	Flexion		
	Bend backwards	Extension		
	Bend sideways		Lateral flexion	
	Rotate about the Long axis			Axial rotation
Shoulder[2] girdle	Raise shoulders	Elevation		
	Lower shoulders	Depression		
	Draw forward	Protraction		
	Draw backward	Retraction		
Shoulder and hip joints	Raise arm or thigh forward	Flexion (extension[a])		
	Raise arm or thigh to side		Abduction (adduction[a])	
	Rotate arm or leg along its long axis			Lateral or outward rotation (medial or inward rotation[a])
Elbow and knee joints	Bend from the fully straightened position	Flexion (extension[a])		
Wrist joint (combined with carpal joints)	Bend palm upwards	Flexion (extension[a])		
	Move hand away from trunk to the side of radius		Adduction or radial deviation (adduction or ulnar deviation[a])	
Forearm rotation	Towards:			
	palm up position	Supination		
	palm down position	Pronation		

Adapted from Grieve and Pheasant (1982).
[a] Opposite movement.

TABLE 5 Commonly Used Values of z and X_p

X_p	z	X_p	z
1	−2.33	99	2.33
2.5	−1.96	97.5	1.96
5	−1.64	95	1.64
10	−1.28	90	1.28
15	−1.04	85	1.04
20	−0.84	80	0.84
25	−0.67	75	0.67
30	−0.52	70	0.52
40	−0.25	60	0.25
50	−0.00	50	0.00

relevant human characteristic) that, if satisfied, will also accommodate the rest of the expected user population. The NIOSH's (1991) recommended weight limit concept is an example of application of the method of limits or design for the extreme principles to the design of manual lifting tasks. Such design is based on the expected human characteristics, where the limiting users are the weakest of the worker population.

2.4. Anthropometric Design Criteria

The basic anthropometric criteria for workplace design are clearance, reach, and posture (Pheasant 1986). Typically, clearance problems refer to design of space needed for the knees, availability of space for wrist support, or safe passageways around and between equipment. If the clearance problems are disregarded, they may lead to poor working postures and hazardous work layouts. Consideration of clearance requires designing for the largest user, typically by adapting the 95th percentile values of the relevant characteristics for male workers. Typical reach problems in industry include consid-

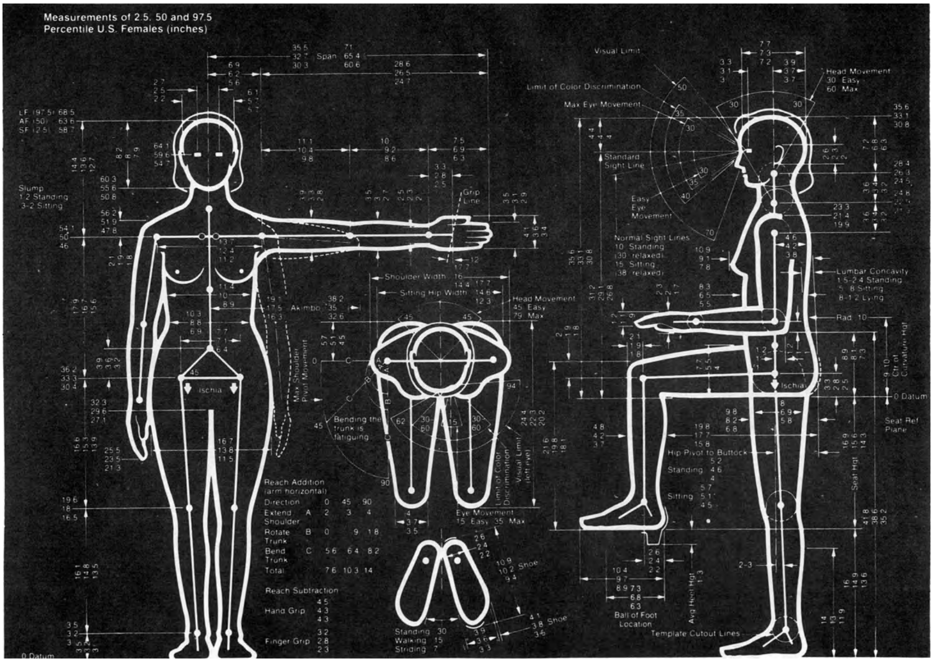

Measurements of 2.5, 50 and 97.5
Percentile U.S. Females (inches)
Span 35.5 32.7 30.3 71 65.4 60.6
LF (97.5) 68.5
AF (50) 63.6
SF (2.5) 58.7
Slump
1.2 Standing
3-2 Sitting
Grip Line
Shoulder Width 16 14.4
Sitting Hip Width 14.6 12.3
Akimbo 38.2 35 32.6
Max Shoulder Pivot Movement
Bending the trunk is fatiguing
Ischia
Reach Addition (arm horizontal)
Direction 0 45 90
Extend Shoulder A 2 3 4
Rotate Trunk B 0 9 1.8
Bend Trunk C 5.6 6.4 8.2
Total 7.6 10.3 14
Reach Subtraction
Hand Grip 4.5 4.3 4.3
Finger Grip 3.2 2.8 2.3
0 Datum
Standing 30
Walking 15
Striding 7
Eye Movement 15 Easy 35 Max
Head Movement 45 Easy 79 Max
Limit of Color Discrimination
Visual Limit (left eye)
Shoe
Visual Limit
Limit of Color Discrimination
Max Eye Movement
Head Movement 30 Easy 60 Max
Standard Sight Line
Easy Eye Movement
Normal Sight Lines
10 Standing
(30 relaxed)
15 Sitting
(38 relaxed)
Lumbar Concavity
1.5-2.4 Standing
.5-.8 Sitting
.8-1.2 Lying
Rad 10
Ctr of Curvature Hgt
0 Datum
Seat Ref Plane
Seat Hgt
Hip Pivot to Buttock
Standing 5.2 4.6 4
Sitting 5.7 5.1 4.5
Ball of Foot Location
Avg Heel Hgt 1.3
Template Cutout Lines

Figure 2 Illustration of Design Data Presented in Humanscale. (Reproduced with permission from Henry Dreyfuss Associates 1984)

eration of the location of controls and accessibility of control panels in the workplace. The procedure for solving the reach problems is similar to the one used for solving the clearance problems. This time, however, the limiting user will be a smaller member of the population and the design will be usually based upon the 5th percentile value of the relevant characteristic for female workers.

Both the clearance and the reach criteria are *one-tailed constraints,* that is, they impose the limits in one direction only (Pheasant 1989). The clearance criterion points out when an object is too small. It does not, however, indicate when an object is too large. In some design problems, such as safeguarding of industrial machinery, the conventional criteria of clearance and reach are often reversed.

2.5. Alternative Design Procedures

An alternative to single-percentile anthropometric design models has been presented by Robinette and McConville (1981). They point out that single-percentile models are inappropriate for both theoretical and practical reasons. As discussed by Kroemer (1989), there are two other methods that can be used to develop the analogues of the human body for the design purposes. One method is to create models that represent the extreme ends of the body size range called the subgroup method. The other method is the regression-based procedure. which generates design values that are additive. The estimated U.S. civilian body dimensions published by Kroemer (1981) are given in Table 1. A useful and correct general procedure for anthropometric design was recently proposed by Kroemer et al. (1986). This procedure consists of the following steps:

- *Step 1:* Select those anthropometric measures that directly relate to defined design dimensions. Example: hand length related to handle size.
- *Step 2:* For each of these pairings, determine independently whether the design must fit either only one given percentile of the body dimension, or if a range along that body dimension must be fitted. The height of a seat should be adjustable to fit persons with short and with long lower legs.
- *Step 3:* Combine all selected dimensions in a careful drawing, mock-up, or computer model to ascertain that all selected design values are compatible with each other. For example: the required leg room clearance height needed for sitting persons with long lower legs may be very close to the height of the working surface, determined from elbow height.
- *Step 4:* Determine whether one design will fit all users. If not, several sizes or adjustment must be provided to fit all users.

2.6. Computer-Aided Models of Man

In order to facilitate the application of anthropometric data and biomechanical analysis in workplace design, several computer-based models of man have been developed. These computer-aided tools

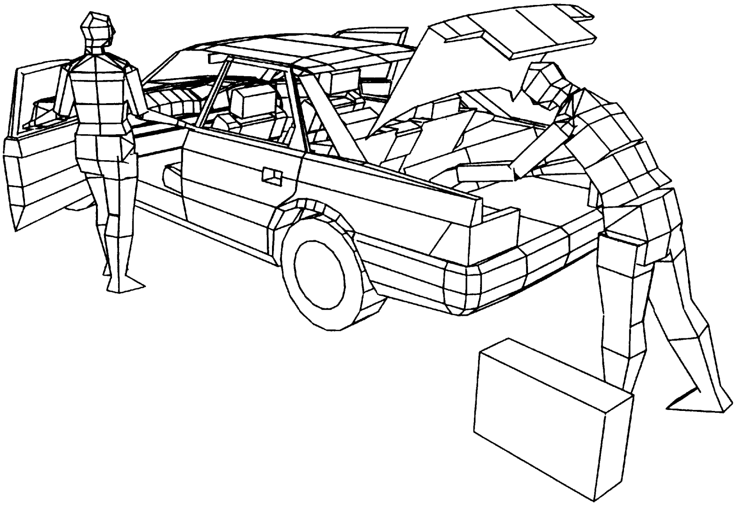


Figure 3 Illustration of Workplace Design Using SAMMIE System. (Reproduced with permission from SAMMIE CAD, Ltd.)

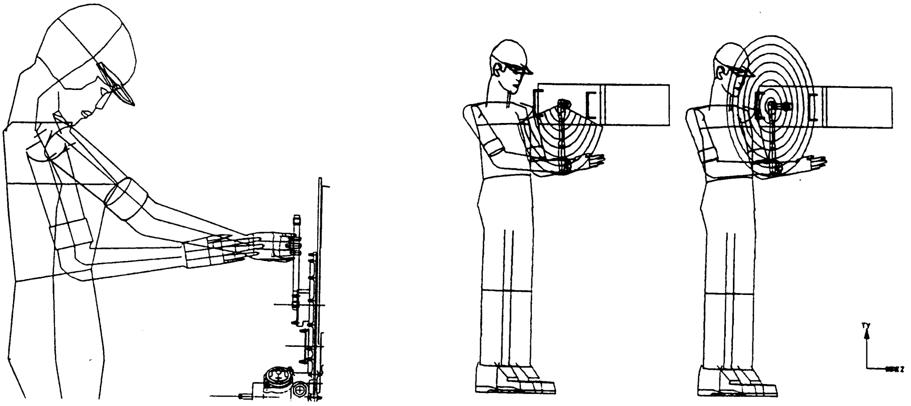


Figure 4 Illustration of the CREW CHIEF Design Capabilities: Left—removing a recessed bolt from a jet engine; right—modification of ratchet wrench interaction with handles on a box. (Reproduced with permission from McDaniel, copyright by Taylor & Francis, 1990)

should make the analysis and application of biomechanical principles at work less complicated and more useful. For a review of the state of the art in ergonomic models of anthropometry, human biomechanics and operator–equipment interfaces, see Kroemer et al. (1988). Other developments in computer-aided ergonomics, specifically computer models of man and computer-assisted workplace design, are discussed by Karwowski et al. (1990). According to Schaub and Rohmert (1990), man model development originated with SAMMIE (System for Aiding Man–Machine Interaction Evaluation) in England (see Figure 3) (Porter et al. 1995). Examples of computer models developed in the United States include BOEMAN (Ryan 1971) for aircraft design, COMBIMAN and CREW CHIEF (McDaniel 1990) (see Figure 4), Deneb/ERGO (Nayar 1995) and JACK (Badler et al. 1995)

Other computer-aided man models developed in Europe include ERGOMAN (France), OSCAR (Hungary), ADAPTS (Netherlands), APOLINEX (Poland), and WERNER, FRANKY, and ANYBODY (Germany). A comprehensive 3D man model for workplace design, HEINER, was developed by Schaub and Rohmert (1990). Advances in applied artificial intelligence made it possible to develop knowledge-based expert systems for ergonomic design and analysis (Karwowski et al. 1987; Jung and Freivalds 1990). Examples of such models include SAFEWORK (Fortin et al. 1990), ERGONEXPERT (Rombach and Laurig1990), and ERGOSPEC (Brennan et al. 1990). Other models, such as CAD-video somotograph (Bullinger and Lorenz 1990) and AutoCAD-based anthropometric design systems (Grobelny 1990), or ergonomic databases (Landau et al. 1990), were also developed.

The computer-aided systems discussed above serve the purpose of biomechanical analysis in workplace design. For example, COMBIMAN, developed in the Human Engineering Division of Armstrong Laboratory since 1975, is both illustrative and analytic software. It allows the analysis of physical accessibility (reach and fit capabilities), strength for operating controls, and visibility accessibility. CREW CHIEF, a derivative of COMBIMAN, also allows the user with similar analyses. Another important development is Deneb's ERGO, a system capable of rapid prototyping of human motion, analyzing human joint range of motion, reach, and visual accessibility. In a recent study by Schaub et al. (1997), the authors revised the models and methods of ERGOMAN and reported added capabilities to predict maximum forces/moments of relevant posture, evaluate stress of human body joints, and carry out a general risk assessment. Probably the most advanced and comprehensive computer-aided digital human model and design/evaluation system today is JACK, from Transform Technologies (2000).

3. DESIGN FOR HUMAN STRENGTH

Knowledge of human strength capabilities and limitations can be used for ergonomic design of jobs, workplaces, equipment, tools, and controls. Strength measurements can also be used for worker preemployment screening procedures (Chaffin et al. 1978; Ayoub 1983). Human strengths can be assessed under static (isometric) or dynamic conditions (Kroemer 1970; Chaffin et al. 1977). Dynamic strengths can be measured isotonically, isokinetically, and isoinertially. Isometric muscle strengths are the capacity of muscles to produce force or moment force by a single maximal voluntary exertion; the body segment involved remains stationary and the length of the muscle does not change. In

TABLE 6 Static Strengths Demonstrated by Workers when Lifting, Pushing, and Pulling with Both Hands on a Handle Placed at Different Locations Relative to the Midpoint between the Ankles on Floor

	Location of Handle (cm)[a]		Male Strengths (N)			Female Strengths (N)		
Test Description	Vertical	Horizontal	Sample Size	Mean	SD	Sample Size	Mean	SD
Lift—leg partial squat	38	0	673	903	325	165	427	187
Lift—torso stooped over	38	38	1141	480	205	246	271	125
Lift—arms flexed	114	38	1276	383	125	234	214	93
Lift—shoulder high and arms out	152	51	309	227	71	35	129	36
Lift—shoulder high and arms flexed	152	38	119	529	222	20	240	84
Lift—shoulder high and arms close	152	25	309	538	156	35	285	102
Lift—floor level, close (squat)	15	25	309	890	245	35	547	182
Lift—floor level, out (stoop)	15	38	170	320	125	20	200	71
Push down—waist level	118	38	309	432	93	35	325	71
Pull down—above shoulders	178	33	309	605	102	35	449	107
Pull in—shoulder level, arms out	157	33	309	311	80	35	244	53
Pull in—shoulder level, arms in	140	0	205	253	62	52	209	62
Push—out waist level, stand erect	101	35	54	311	195	27	226	76
Push—out chest level, stand erect	124	25	309	303	76	35	214	49
Push-out—shoulder level, lean forward	140	64	205	418	178	52	276	120

Adapted from Chaffin et al., *Occupational Biomechanics,* 3rd Ed. Copyright © 1999. Reprinted by permission of John Wiley & Sons, Inc., New York.

[a] The location of the handle is measured in midsagittal plane, vertical from the floor and horizontal from the midpoint between the ankles.

dynamic muscular exertions, body segments move and the muscle length changes (Ayoub and Mital 1989). The static strengths demonstrated by industrial workers on selected manual handling tasks are shown in Table 6. Maximum voluntary joint strengths are depicted in Table 7.

3.1. Occupational Strength Testing

The main goal of worker selection is to screen the potential employee on the basis of his or her physical capability and match it with job demands. In order to evaluate an employee's capability, the following criteria should be applied when selecting between alternative screening methods (NIOSH 1981):

1. Safety in administering
2. Capability of giving reliable, quantitative values
3. Relation to specific job requirements
4. Practicality
5. Ability to predict the risk of future injury or illness

Isometric strength testing has been advocated as a means to predict the risk of future injuries resulting from jobs that require a high amount of force. Chaffin et al. (1977) reported that both frequency and severity rates of musculoskeletal problems were about three times greater for workers placed in jobs requiring physical exertion above that demonstrated by them in isometric strength tests when compared with workers placed in jobs having exertion requirements well below their demonstrated capabilities. The literature on worker selection has been reviewed by NIOSH (1981), Ayoub (1983), Chaffin et al. (1999), and Ayoub and Mital (1989). Typical values for the static strengths are shown in Figure 5.

3.2. Static vs. Dynamic Strengths

The application of static strength exertion data has limited value in assessing workers' capability to perform dynamic tasks that require application of force through a range of motions (Ayoub and Mital 1989). Mital et al. (1986) found that the correlation coefficients between simulated job dynamic strengths and maximum acceptable weight of lift in horizontal and vertical planes were substantially higher than those between isometric strengths and weights lifted. Two new studies offer design data based on dynamic strengths (Mital and Genaidy 1989; Mital and Faard 1990).

TABLE 7 Maximum Voluntary Joint Strengths (Nm)

	Range of Moments (Nm) of Subjects from Several Studies			
Joint strength	Joint Angle (degrees)	Men	Women	Variation with Joint Angle +
Elbow flexor	90	50–120	15–85	Peak at about 90°
Elbow extensor	90	25–100	15–60	Peak between 50° and 100°
Shoulder flexor	90	60–100	25–65	Weaker at flexed angles
Shoulder extensor	90	40–150	10–60	Decreases rapidly at angles less than 30°
Shoulder adductor	60	104	47	As angle decreases, strength increases then levels at 30° to −30°
Trunk flexor	0	145–515	85–320	Pattern differ among authors
Trunk extensor	0	143	78	Increases with trunk flexion
Trunk lateral flexor	0	150–290	80–170	Decreases with joint flexion
Hip extensor	0	110–505	60–130	Increases with joint flexion
Hip abductor	0	65–230	40–170	Increases as angle decreases
Knee flexor	90	50–130	35–115	In general, decreases some disagreement with this, depending on hip angle
Knee extensor	90	100–260	70–150	Minima at full flexion and extension
Ankle plantarflexor	0	75–230	35–130	Increases with dorsiflexion
Ankle dorsiflexor	0	35–70	25–45	Decreases from maximum plantar flexion to maximum dorsiflexion

Adapted from Tracy 1990.

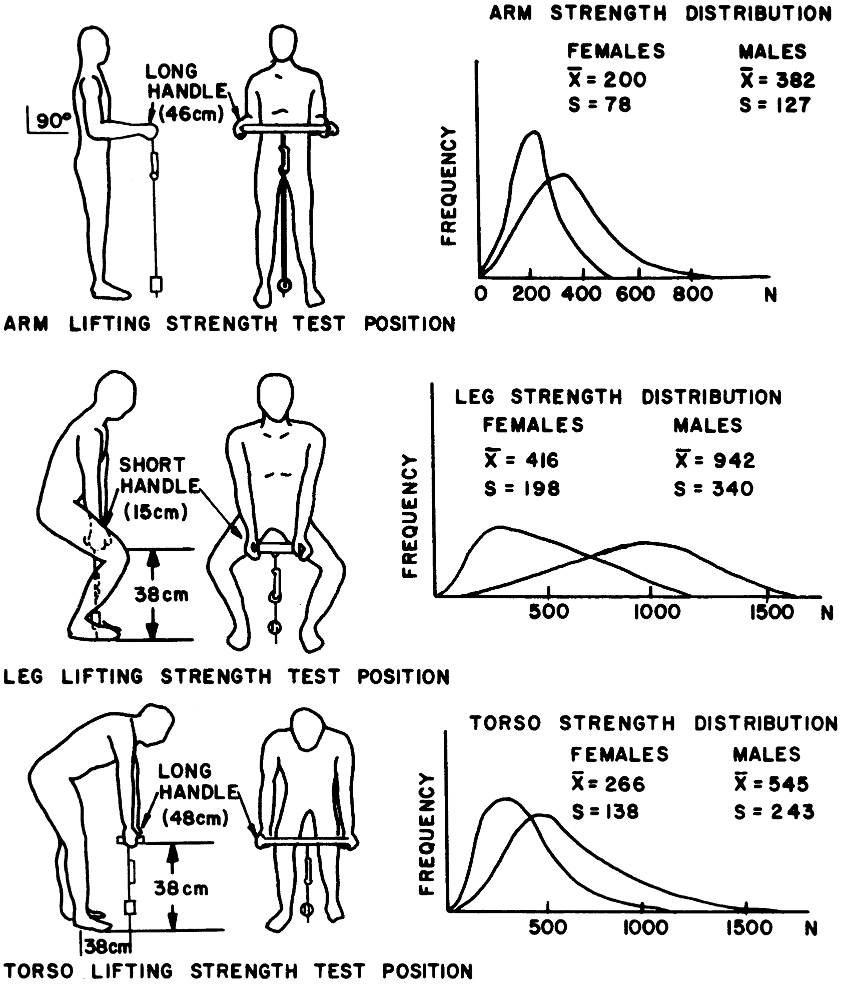


Figure 5 Results of Static Strength Measurement. (Adapted from Chaffin et al., *Occupational Biomechanics,* 3rd Ed. Copyright © 1999. Reprinted by permission of John Wiley & Sons, Inc., New York.)

The study by Mital and Genaidy (1989) provides isokinetic strengths of males and females for infrequent vertical exertions in 15 different working postures. This study showed that dynamic strength exertions of females are approximately half those of the male exertions, not about 67%, as in the case of isometric strength exertions. Mital and Faard (1990) investigated the effects of reach distance, preferred arm orientation, and sitting and standing posture on isokinetic force exertion capability of males in the horizontal plane. The results indicated that peak isokinetic strengths increased with the reach distance and were strongly influenced by the arm orientation. Also, peak isokinetic exertions were substantially greater than static strength when subjects were allowed to exert at freely chosen speed.

Karwowski and Mital (1986) and Karwowski and Pongpatanasuegsa (1988) tested the additivity assumption of isokinetic lifting and back extension strengths (the additivity assumption states that strength of a team is equal to the sum of individual members' strengths). They found that, on average, the strength of two-person teams is about 68% of the sum of the strengths of its members. For three-member teams, male and female teams generate only 58% and 68.4% of the sum of the strengths of its members, respectively. For both genders, the isokinetic team strengths were much lower than static team strengths.

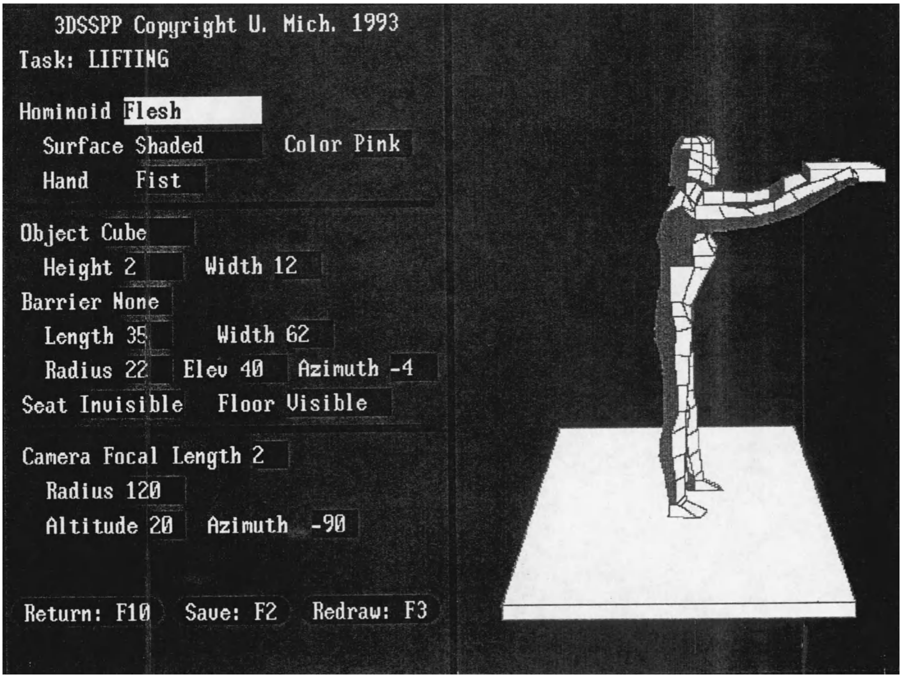
3.3. Computer Simulation of Human Strength Capability

The worker strength exertion capability in heavy manual tasks can be simulated with the help of a microcomputer. Perhaps the best-known microcomputer system developed for work design and analysis concerning human strength is the Three Dimensional Static Strength Program (3D SSPP), developed by the Center for Ergonomics at the University of Michigan and distributed through the Intellectual Properties Office (University of Michigan, 1989). The program can aid in the evaluation of the physical demands of a prescribed job, and is useful as a job design/redesign and evaluation tool. Due to its static nature, the 3D SSPP model assumes negligible effects of accelerations and momentums and is applicable only to slow movements used in manual handling tasks. It is claimed that the 3D SSPP results correlate with average population static strengths at $r = 0.8$, and that the program should not be used as the sole determinant of worker strength performance (University of Michigan, 1989). In their last validation study, Chaffin and Erig (1991) reported that if considerable care is taken to ensure exactness between simulated and actual postures, the prediction error standard deviation would be less than 6% of the mean predicted value. However, 3D SSPP does not allow simulation of dynamic exertions.

The body posture, in 3D SSPP, is defined through five different angles about the joints describing body link locations. The input parameters, in addition to posture data, include percentile of body height and weight for both male and female populations, definition of force parameters (magnitude and direction of load handled in the sagittal plane), and the number of hands used. The output from the model provides the estimation of the percentage values of the population capable of exerting the required muscle forces at the elbow, shoulder, lumbosacral (L5/S1), hip, knee and ankle joints, and calculated back compression force on L5/S1 in relation to NIOSH action limit and maximum permissible limit. The body balance and foot/hip potential is also considered. An illustration of the model output is given in Figure 6.

3.4. Push–Pull Force Limits

Safe push–pull force exertion limits may be interpreted as the maximum force magnitudes that people can exert without injuries (for static exertions) or CTD (for repeated exertions) of the upper extremities under a set of conditions.



3DSSPP Copyright U. Mich. 1993
Task: LIFTING
Hominoid Flesh
Surface Shaded   Color Pink
Hand   Fist
Object Cube
Height 2   Width 12
Barrier None
Length 35   Width 62
Radius 22   Elev 40   Azimuth -4
Seat Invisible   Floor Visible
Camera Focal Length 2
Radius 120
Altitude 20   Azimuth -90

Figure 6 Illustration of Results from the 3D Static Strength Prediction Program of the University of Michigan.

3.4.1. Static Standing Forces

Because many factors influence the magnitude of a static MVC force, it would be wise not to recommend a single value for either push or pull force limits. After reviewing several studies Imrhan (1999), has concluded that average static two-handed MVC push forces have ranged from about 400–620 N in males and 180–335 N in females without bracing of the body, and pull forces from about 310–370 N in males and 180–270 N in females.

3.4.2. Dynamic Standing Forces

Dynamic push forces have ranged from 170 to 430 N in males and 200 to 290 N in females, and pull forces from 225 to 500 N in males and 160 to 180 N in females. As a result of series of researches by Snook and his colleagues (Snook et al. 1970; Snook and Ciriello 1974a; Snook 1978; Ciriello and Snook 1978, 1983; Ciriello et al. 1990), by utilizing psychophysical methodology, Snook and Ciriello (1991) have published the most useful guidelines on maximum initial or sustained push–pull force limits. Partial reproductions of the final four tables are given in Tables 8–11. The forces in are stated as a function of other work-related independent variables for both males and females. These are as follows:

1. Distance of push/pull: 2.1, 7.6, 15.2, 30.5, 45.7, and 61.0 m.
2. Frequency of push/pull: each distance has force limits for one exertion per 8 hr., 30 min, 5 min, and 2 min.
3. Height (vertical distance from floor to hands: 144, 95, 64 cm for males and 135, 89, and 57 for females.
4. The percentage of workers: (10, 25, 50, 75, and 90%) who are capable of sustaining the particular force during a typical 8-hr job.

TABLE 8 Maximum Acceptable Forces of Pull for Females (kg)

		2.1 m Pull							45.7 m Pull				
		6	12	1	2	5	30	8	1	2	5	30	8
Height	Percent	s		min				h	min				h
		Initial Forces											
	90	13	16	17	18	20	21	22	12	13	14	15	17
	75	16	19	20	21	24	25	26	14	16	17	18	20
135	50	19	22	24	25	28	29	31	17	18	20	21	24
	25	21	25	28	29	32	33	35	19	21	23	24	27
	10	24	28	31	32	36	37	39	22	24	25	27	31
	90	15	17	19	20	22	23	24	13	14	15	17	19
	75	17	20	22	23	26	27	28	16	17	18	20	22
57	50	20	24	26	27	30	32	33	18	20	22	23	26
	25	23	27	30	31	35	36	38	21	23	25	27	30
	10	26	31	34	35	39	40	43	24	26	28	30	34
		Sustained Forces											
	90	6	9	10	10	11	12	15	6	6	7	7	9
	75	8	12	13	14	15	16	20	8	9	9	9	12
135	50	10	16	17	18	19	21	25	10	11	11	12	16
	25	13	19	21	21	23	25	31	12	13	14	14	19
	10	15	22	24	25	27	29	36	14	15	16	17	23
	90	5	8	9	9	10	11	13	5	6	6	6	8
	75	7	11	12	12	13	14	18	7	8	8	8	11
57	50	9	14	15	16	17	18	23	9	10	10	11	14
	25	11	17	18	19	21	22	27	11	12	12	13	19
	10	13	20	21	22	24	26	32	12	14	14	15	20

From Snook and Ciriello (1991) with permission from Taylor & Francis Ltd., London, http://www.tandf.co.uk.
Height = vertical distance from floor to hand–object (handle) contact
Percent = percentage of industrial workers capable of exerting the stated forces in work situations.

TABLE 9 Maximum Acceptable Forces of Pull for Females (kg)

		2.1 m Pull							45.7 m Pull				
		6	12	1	2	5	30	8	1	2	5	30	8
Height	Percent	s		min				h	min				h
Initial Forces													
135	90	14	15	17	18	20	21	22	12	13	14	15	17
	75	17	18	21	22	24	25	27	15	16	17	19	21
	50	20	22	25	26	29	30	32	18	19	21	22	25
	25	24	25	29	30	33	35	37	20	22	24	26	29
	10	26	28	33	34	38	39	41	23	25	27	29	33
57	90	11	12	14	14	16	17	18	11	12	12	13	15
	75	14	15	17	17	19	20	21	13	14	15	16	18
	50	16	17	20	21	23	24	25	15	17	18	19	22
	25	19	20	23	24	27	28	30	18	19	21	22	25
	10	21	23	26	27	30	31	33	20	22	23	25	28
Sustained Forces													
135	90	6	8	10	10	11	12	14	5	5	5	6	8
	75	9	12	14	14	16	17	21	7	8	8	8	11
	50	12	16	19	20	21	23	28	9	10	11	11	15
	25	16	20	24	25	27	29	36	11	13	13	14	19
	10	18	23	28	29	32	34	42	14	15	16	17	22
57	90	5	6	8	8	9	9	12	5	5	5	6	7
	75	7	9	11	12	13	14	17	7	7	8	8	11
	50	10	13	15	16	17	18	23	9	10	10	11	15
	25	12	16	19	20	22	23	29	11	13	13	14	19
	10	15	19	23	23	26	28	34	13	15	16	16	22

From Snook and Ciriello (1991) with permission from Taylor & Francis Ltd., London, http://www.tandf.co.uk.
Height = vertical distance from floor to hand–object (handle) contact
Percent = percentage of industrial workers capable of exerting the stated forces in work situations.

3.4.3. One-Handed Force Magnitudes

One-handed forces vary considerably among studies with similar variables and within individual studies depending on test conditions or variables. Generalizations about recommended forces, therefore, are not easy to make. Average static standing-pull forces have ranged from 70 to 134 N and sitting forces from 350 to 540 N. Dynamic pull forces, in almost all studies, have ranged from 170 to 380 N in females and from 335 to 673 N in males when sitting. Average pull forces in males while lying down prone have ranged from 270 to 383 N and push forces from 285 to 330 N (Hunsicker and Greey 1957).

3.4.4. Pinch–Pull Force Magnitudes

Pinching and pulling with one hand while stabilizing the object with the other hand has been observed in male adults to yield forces of 100, 68, and 50 N when using the lateral, chuck, and pulp pinches, respectively (Imrhan and Sundararajan 1992; Imrhan and Alhaery 1994).

4. STATIC EFFORTS AND FATIGUE

4.1. Design Limits for Static Work

Static efforts at work are often fatiguing and cannot be sustained over a long period of time (Rohmert 1960; Monod and Scherrer 1965; Pottier et al. 1969 and Monod 1972). Figure 7 illustrates the relationship between a percentage of the maximum voluntary contraction used and the time duration. This relationship has been determined for arm, leg, and trunk muscles by Rohmert (1960), for an upperlimb pulling action by Caldwell and Smith (1963), and for biceps brachii, triceps brachii, the middle finger flexor, and quadriceps femoris by Monod and Scherrer (1965). The results of these

TABLE 10 Maximum Acceptable Forces of Pull for Males (kg)

		2.1 m Pull							45.7 m Pull				
		6	12	1	2	5	30	8	1	2	5	30	8
Height	Percent	s		min				h	min				h
Initial Forces													
144	90	14	16	18	18	19	19	23	10	11	13	13	16
	75	17	19	22	22	23	24	28	12	14	16	16	20
	50	20	23	26	26	28	28	33	15	16	19	19	24
	25	24	27	31	31	32	33	39	17	19	22	22	28
	10	26	30	34	34	36	37	44	20	22	25	25	31
64	90	22	25	28	28	30	30	36	16	18	21	21	26
	75	27	30	34	34	37	37	44	19	22	25	25	31
	50	32	36	41	41	44	44	53	23	26	30	30	37
	25	37	42	48	48	51	51	61	27	30	35	35	43
	10	42	48	54	54	57	58	69	30	34	39	39	49
Sustained Forces													
144	90	8	10	12	13	15	15	18	6	7	8	9	10
	75	10	13	16	17	19	20	23	7	9	10	11	14
	50	13	16	20	21	23	24	28	9	11	12	14	17
	25	15	20	24	25	28	29	34	11	13	15	17	20
	10	17	22	27	28	32	33	39	12	14	17	19	23
64	90	11	14	17	18	20	21	25	8	9	11	12	15
	75	14	19	23	23	26	27	32	10	12	14	16	19
	50	17	23	28	29	32	34	40	13	15	17	20	23
	25	20	27	33	35	39	40	48	15	18	21	24	28
	10	23	31	38	40	45	46	54	17	20	24	27	32

From Snook and Ciriello (1991). Used with permission by Taylor & Francis Ltd., London, http://www.tandf.co.uk.
Height = vertical distance from floor to hand–object (handle) contact
Percent = percentage of industrial workers capable of exerting the stated forces in work situations.

three studies indicate that the limit time approaches infinity at a force of 8–10% maximum voluntary contraction and converges to zero at 100% of the maximum strength.

As discussed by Kahn and Monod (1989), the maximum duration of static effort or the maximum maintenance time (limit time) varies inversely with the applied force and may be sustained for a long time if the force does not exceed 15–20% of the maximum voluntary contraction (MVC) of the muscle considered. The relation between force and limit time has been defined by Monod (1965) as:

$$t = \frac{k}{(F - f)^n}$$

where t is the limit time (min), F the relative force used (%), f the force (%) for which t tends to infinity (called the critical force), and k and n are constants. Rohmert (1960) subsequently proposed a more elaborate equation:

$$t = -1.5 + \left(\frac{2.1}{F}\right) - \left(\frac{0.6}{F^2}\right) + \left(\frac{0.1}{F^3}\right)$$

In both cases the maximum maintenance time is linked to the force developed by a hyperbolic relation, which applies to all muscles.

4.2. Intermittent Static Work

In the case of intermittent static efforts during which the contraction phases are separated by rest periods of variable absolute and relative duration, the maximum work time, given the relative force used and the relative duration of contraction, can be predicted as follows (Rohmert 1973):

TABLE 11 Maximum Acceptable Forces of Pull for Males (kg)

		2.1 m Pull							45.7 m Pull				
		6	12	1	2	5	30	8	1	2	5	30	8
Height	Percent	s		min				h	min				h
		Initial Forces											
	90	20	22	25	25	26	26	31	13	14	16	16	20
	75	26	29	32	32	34	34	41	16	18	21	21	26
144	50	32	36	40	40	42	42	51	20	23	26	26	33
	25	38	43	47	47	50	51	61	24	27	32	32	39
	10	44	49	55	55	58	58	70	28	31	36	36	45
	90	19	22	24	24	25	26	31	12	14	16	16	20
	75	25	28	31	31	33	33	40	16	18	21	21	26
64	50	31	35	39	39	41	41	50	20	22	26	26	32
	25	38	42	46	46	49	50	59	24	27	31	31	39
	10	43	48	53	53	57	57	68	27	31	36	36	44
		Sustained Forces											
	90	10	13	15	16	18	18	22	7	8	10	11	13
	75	13	17	21	22	24	25	30	10	11	13	15	18
144	50	17	22	27	28	31	32	38	12	14	17	19	23
	25	21	27	33	34	38	40	47	15	18	21	24	28
	10	25	31	38	40	45	46	54	18	21	24	28	33
	90	10	13	16	16	18	19	23	7	8	9	11	13
	75	4	18	21	22	25	26	31	9	11	12	14	17
64	50	18	23	28	29	32	33	39	12	14	16	18	22
	25	22	28	34	35	39	41	48	14	17	20	23	27
	10	26	32	39	41	46	48	56	17	20	23	26	31

From Snook and Ciriello (1991). Used with permission by Taylor & Francis Ltd., London, http://www.tandf.co.uk.
Height = vertical distance from floor to hand–object (handle) contact
Percent = percentage of industrial workers capable of exerting the stated forces in work situations.

$$t = \frac{k}{(F - f)^{np}}$$

where p is the static contraction time as a percentage of the total time. Rohmert (1962) had devised another method for estimating the minimum duration of rest periods required to avoid fatigue during intermittent static work:

$$t_r = 18\left(\frac{t}{t_{\max}}\right)^{1.4} \times \left(\frac{F}{F_{\max}} - 0.15\right) \times 100\%$$

where t_r is the rest time as a percentage of t, which is the duration of contraction (min). Kahn and Monod (1989) concluded that the main causal factor in the onset of fatigue due to static efforts (isometrically contracting muscles) is local muscle ischemia. Furthermore, the onset of local muscle fatigue can be delayed if changes in recovery time are sufficient to allow restoration of normal blood flow through the muscle.

4.3. Static Efforts of the Arm

As of this writing, only limited guidelines regarding the placement of objects that must be manipulated by the arm have been proposed (Chaffin et al. 1999). Figure 8 depicts the effect of horizontal reach on shoulder muscle fatigue endurance times. This figure illustrates that the workplace must be designed to allow for the upper arm to be held close to the torso. In general, any load-holding tasks should be minimized by the use of fixtures and tool supports. Strasser et al. (1989) showed experimentally that the local muscular strain of the hand–arm–shoulder system is dependent upon the direction of horizontal arm movements. Such strain, dependent on the direction of repetitive manual movements, is of great importance for workplace layout. The authors based their study on the premise

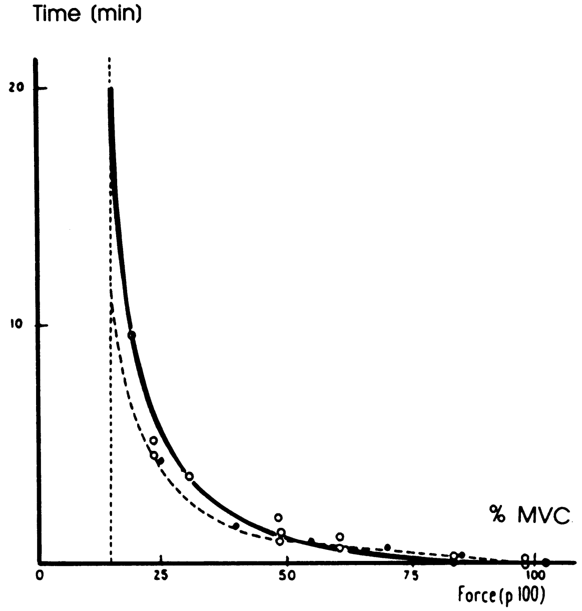


Figure 7 Isometric Muscle Endurance Time as a Function of Percentage of Muscle Strength. (Reproduced with permission from Kahn and Monod, copyright © by Taylor & Francis, http:// www.tandf.co.uk 1989.)

that in order to avoid unnecessary strain imposed by unfavorable postures and working directions in repetitive material-handling tasks, patterns of static and dynamic musculoskeletal loads need to be determined. Figure 9 shows the results of this normalization procedure applied to the experimental data. The static components of the EA values within an angle range between 110° and 200° are also

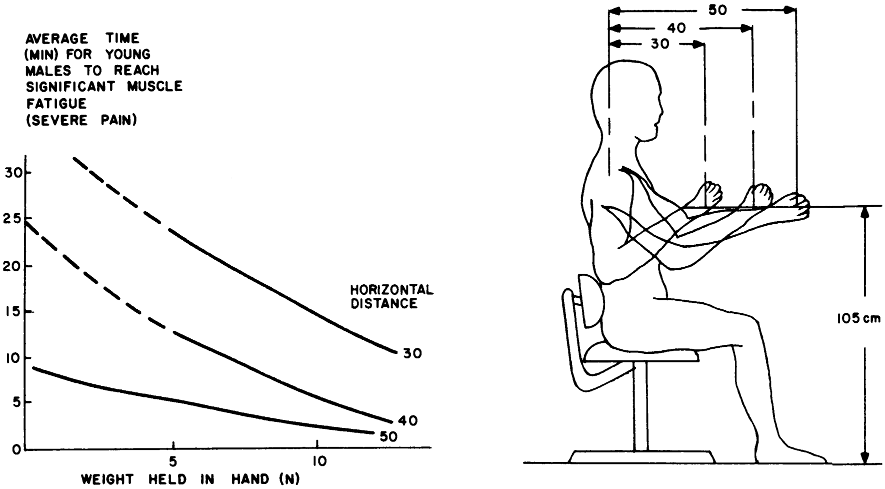


Figure 8 Endurance Time of Shoulder Muscles at Different Arm Reach Postures. (Adapted from Chaffin et al., *Occupational Biomechanics,* 3rd Ed. Copyright © 1999. Reprinted by permission of John Wiley & Sons, Inc., New York.)

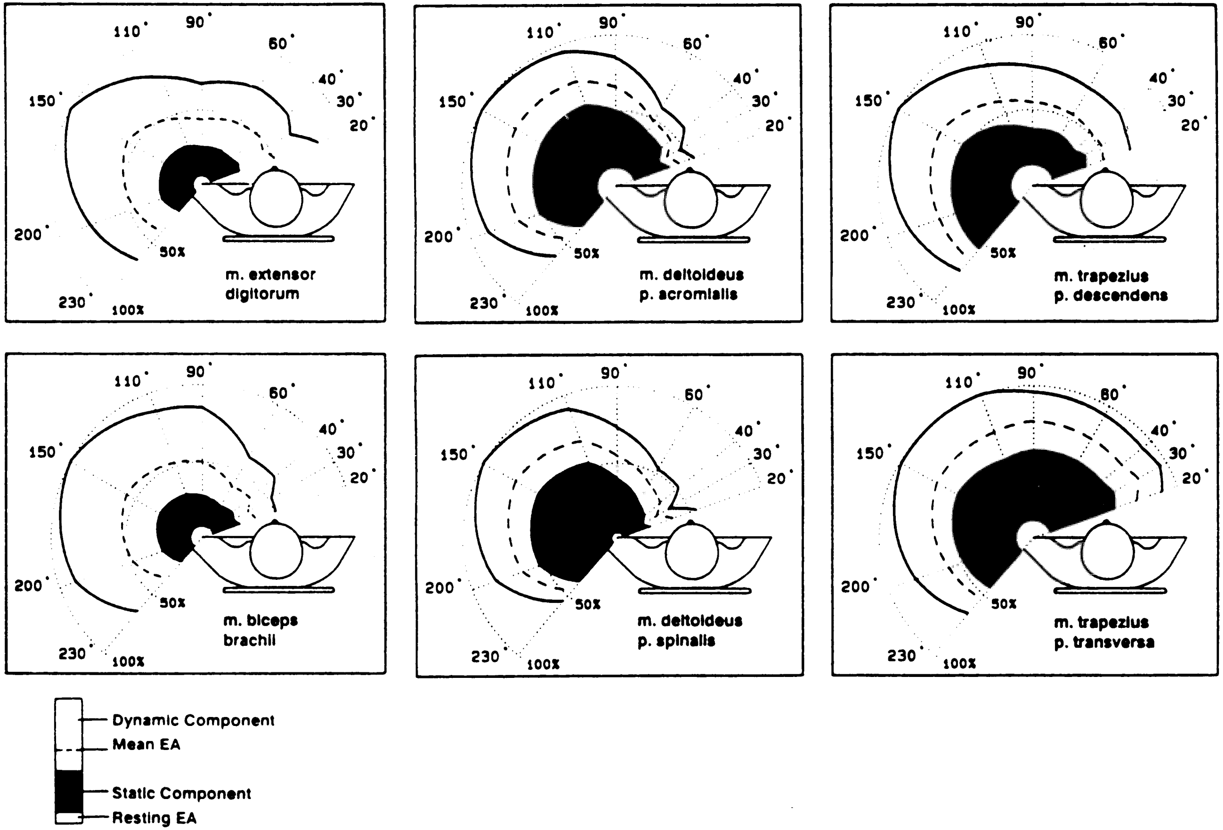


Figure 9 Illustration of Changes in Muscular Effort with the Direction of Contraction at Horizontal Plane. (Reproduced with permission from Strasser et al., copyright © by Taylor & Francis, http://www.tandf.co.uk. 1989.)

significantly higher than those of 20°, 30°, 40°, 60°, and 230°. With regard to the musculature of the shoulder region represented by two recordings of the trapezius (right part of Figure 9), an essentially lower dependence of the muscular strain on the direction of the repetitive horizontal arm movements was observed.

5. WORKPLACE ANALYSIS AND DESIGN

5.1. Evaluation of Working Postures

The posture of the human body at work is influenced by several factors, including workstation layout (heights of the workplace, orientation of tools and work objects), hand tool design, work methods and work habits, visual control and force exertion requirements, and anthropometric characteristics of the workers (Chaffin et al. 1999; Grandjean 1980; Habes and Putz-Anderson, 1985; Corlett et al. 1986; Kilbom et al. 1986; Keyserling 1986; Wallace and Buckle 1987). Poor and unnatural (i.e., not-neutral) working postures have been associated with the onset of fatigue, bodily discomforts and pains, and musculoskeletal disorders (Tichauer 1978; Karhu et al. 1977; and Keyserling et al. 1988). Keyserling (1990) discusses the scientific evidence of such associations. For example, it was shown that trunk flexion, lateral bending, or twisting increases muscle stress and intervetrabral disc pressure, while prolonged sitting or forward bending leads to increased risk of low back pain and muscle fatigue (Chaffin 1973; Schultz et al. 1982; Kelsey and Hochberg 1988). Prolonged elevation of the arms may cause tendonitis (Hagberg 1984), while shoulder extension and elevation may lead to thoracic outlet syndrome (Armstrong 1986). Also, strong association was found between poor neck posture and cervicobrachial disorders (Jonsson et al. 1988).

5.2. Computer-Aided Analysis of Working Postures

Ergonomics provides useful guidelines for evaluation of working postures, especially with respect to identification and quantification of postural stresses and their relation to posture-related work injury. The ultimate goal of such analysis is to improve the workplace design by reducing postural stresses imposed upon the body to the acceptable (safe) levels. Some of the methods used in the past to systematically evaluate work postures by using computerized or semicomputerized techniques are reported by Karhu et al. (1977); Corlett et al. (1979); Holzmann (1982); Keyserling (1986); Pearcy et al. (1987); and Wangenheim and Samuelson (1987). Snijders et al. (1987) introduced devices for measurement of forward bending, lateral bending, and torsion continuously. Ferguson et al. (1992) used a lumbar motion monitor to measure the back motion during asymmetric lifting tasks. The Ovako Working Posture Analysis System (OWAS), which uses predefined standard postures, was first reported by Karhu et al. (1977). The posture targeting technique (1988) and RULA (1996), developed by Corlett et al. (1979), are based on the recording of the positions of the head, trunk, and upper and lower arms.

5.3. Postural Analysis Systems

5.3.1. OWAS

OWAS (the Ovako Working Posture Analyzing System), first reported by Karhu et al. (1977), identifies the most common work postures for the back, arms, and legs and estimates the weight of the loads handled or the extent of the strength (effort). A rating system categorizes 72 different postures in terms of discomfort caused and the effect on health. Back postures are defined as either straight, bent, straight and twisted, or bent and twisted. No specificity (in terms of number of degrees) is provided. This categorization results in the specification of four action categories. The observed posture combinations are classified according to the OWAS method into ordinal scale action categories. The four action categories described here are based on experts' estimates on the health hazards of each work posture or posture combination in the OWAS method on the musculoskeletal system:

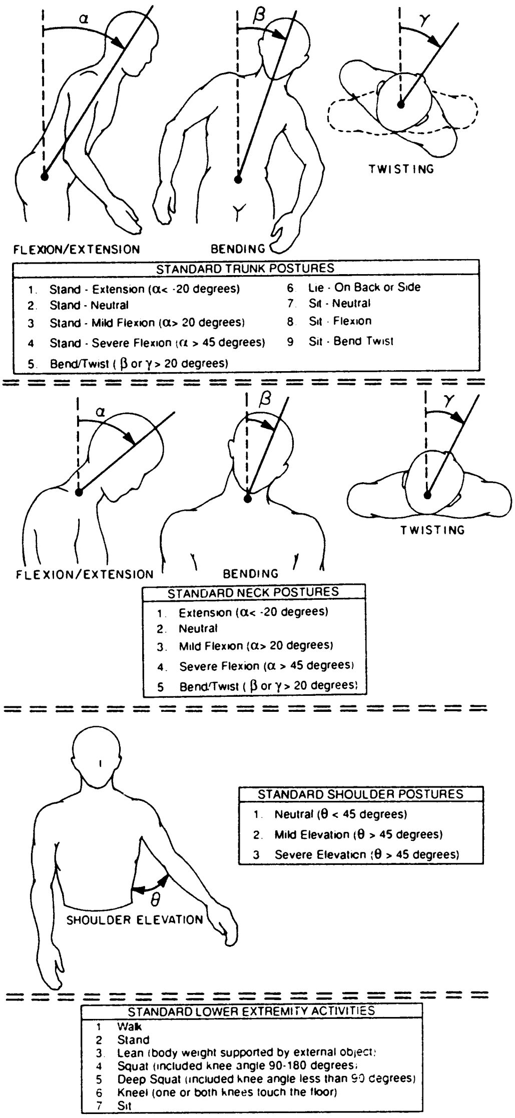
1. Work postures are considered usually with no particular harmful effect on the musculoskeletal system. No actions are needed to change work postures.
2. Work postures have some harmful effect on the musculoskeletal system. Light stress, no immediate action is necessary, but changes should be considered in future planning.
3. Work postures have a distinctly harmful effect on the musculoskeletal system. The working methods involved should be changed as soon as possible.
4. Work postures have an extremely harmful effect on the musculoskeletal system. Immediate solutions should be found to reduce these postures.

OWASCA, a computer-aided visualizing and training software for work posture analysis, was developed using OWAS. OWASCA is intended as OWAS training software (Vayrynen et al. 1990).

The system is also suitable for visualizing the work postures and for the basic analysis of the postures and their loads. The posture is presented with parametric vector using 2D graphics, OWAS codes, and texts. The posture of the back, arms and legs, posture combination, force or effort used, additional postures, and action categories can be studied interactively step by step. The required OWAS skills can be tested by OWASCA. The program shows a random work posture, and the user is asked to identify it. OWASCA describes the errors and gives the numbers of test postures and correct answers (Mattila et al. 1993).

5.3.2. Standard Posture Model

A standard system for analyzing and describing postures of the trunk, neck, shoulders, and lower extremities during dynamic work was developed at the University of Michigan (Keyserling 1990) (see Figure 10). Neutral joint postures and their deviations were also defined (Table 12). The postural



STANDARD TRUNK POSTURES	
1. Stand - Extension (α< -20 degrees)	6. Lie - On Back or Side
2. Stand - Neutral	7. Sit - Neutral
3. Stand - Mild Flexion (α> 20 degrees)	8. Sit - Flexion
4. Stand - Severe Flexion (α > 45 degrees)	9. Sit - Bend Twist
5. Bend/Twist (β or γ > 20 degrees)	

STANDARD NECK POSTURES
1. Extension (α< -20 degrees)
2. Neutral
3. Mild Flexion (α> 20 degrees)
4. Severe Flexion (α > 45 degrees)
5. Bend/Twist (β or γ > 20 degrees)

STANDARD SHOULDER POSTURES
1. Neutral (θ < 45 degrees)
2. Mild Elevation (θ > 45 degrees)
3. Severe Elevation (θ > 45 degrees)

STANDARD LOWER EXTREMITY ACTIVITIES
1 Walk
2 Stand
3 Lean (body weight supported by external object)
4 Squat (included knee angle 90-180 degrees)
5 Deep Squat (included knee angle less than 90 degrees)
6 Kneel (one or both knees touch the floor)
7 Sit

Figure 10 Illustration of Standard Posture Categories. (Reproduced with permission from Keyserling, copyright © by Taylor & Francis, 1990.)

TABLE 12 Classification of Working Postures That Deviate from Neutral

Body Segment	Neutral Posture	Deviated Posture Description
Trunk	Vertical with no axial twisting	Extended (bent backward more than 20°)
		Mildly flexed (bent forward between 20 and 40°)
		Severely flexed (bent forward more than 45°)
		Bent sideways or twisted more than 20° from the neutral position
Neck		Extended (bent backward more than 20°)
		Mildly flexed (bent forward between 20 and 45°)
		Severely flexed (bent forward more than 90°)
		Bent sideways or twisted more than 20°
Shoulder		Neutral (included angle less than 45°)
		Mild elevation (included angle between 45 and 90°)
		Severe elevation (included angle more than 45°)
Lower extremities	(standard postures)	Walk (locomotion while maintaining an upright posture)
		Stand (free standing with no support other than the feet)
		Lean (weight partially supported by an external object such as a wall or railing)
		Squat (knees bent with the included angle between thigh and calf 90–180°)
		Deep squat (included angle between thigh and calf less than 90°)
		Kneel (one or both knees contact the floor and support part of the body weight)
		Sit (body weight primarily supported by the ischial tuberosities)

Adapted from Keyserling 1990.

analysis involves three steps. First, a continuous video recording of the job is obtained. Second, a sequential description of the major tasks required to perform the job is done in a laboratory, with the job being broken into fundamental work elements and their times measured. The third and final step involves collection of the postural data using the common time scale developed from the fundamental work elements. Postural changes are keyed into the system through the preassigned keys corresponding to specific postures. The value of each posture and the time of postural change for a given joint are recorded and stored in a computer. Based on the above data, the system generates a posture profile for each joint, consisting of the total time spent on each standard posture during the work cycle, the range of times spent in each standard posture, the frequency of posture use, and so on. The system can also provide a graph showing postural changes over time for any of the body joints (segments) of interest.

5.4. Acceptability of Working Postures

Analysis of posture must take into consideration not only the spatial elements of the posture, that is, how much is the person flexed, laterally bent, or rotated (twisted), but how long these postures are maintained. Milner et al. (1986) pointed out where an individual is working to the limits of endurance capacity, it has been found that full recovery is not possible within a rest period 12 times the maximum holding time. Full recovery is possible as long as the holding time is a small percentage of maximum handling time.

Bonney et al. (1990) studied tolerability of certain postures and showed that complex postures requiring movement in more than one direction are more uncomfortable than simple postures. Lateral bending produced more discomfort than either flexed or rotated postures and appears to be the least well-tolerated posture. Rotation by itself does not cause significant discomfort. This finding is consistent with epidemiological results of Kelsey and Golden (1988), who hypothesized that twisting along may not produce enough stress to bring about a detectable increase in risk.

Corlett and Manenica (1980) derived estimates for maximum handling times for various postures when performing a no-load task. These recommendations are as follows:

1. Slightly bent forward postures (approx. 15–20°) = 8 min
2. Moderately bent forward posture (approx. 20–60 °) = 3–4 min
3. Severely bent forward postures (greater than about 60°) = approx. 2 min

Colombini et al. (1985) presented criteria on which posture assessments should be based. Postures considered tolerable include (1) those that do not involve feelings of short-term discomfort and (2) those that do not cause long-term morpho-functional complaints. Short-term discomfort is basically the presence of a feeling of fatigue and/or pain affecting any section of the asteo-arthromuscular and ligamentous apparatus appearing in periods lasting minutes, hours, or days.

Miedema et al. (1997) derived the maximum holding times (MHT) of 19 standing postures in terms of percent of shoulder height and percent of arm reach. They also classified such working postures into three categories, depending on the mean value of the MHT: (1) comfortable; (2) moderate; and (3) uncomfortable postures (see Table 13).

Recently, Kee and Karwowski (2001) presented data for the joint angles of isocomfort (JAI) for the whole body in sitting and standing postures, based on perceived joint comfort measures. The JAI value was defined as a boundary indicating joint deviation from neutral (0°), within which the perceived comfort for different body joints is expected to be the same. The JAI values were derived for nine verbal categories of joint comfort using the regression equations representing the relationships between different levels of joint deviation and corresponding comfort scores for each joint motion. The joint angles with marginal comfort levels for most motions around the wrist, elbow, neck, and ankle were similar to the maximum range-of-motion (ROM) values for these joints. However, the isocomfort joint angles with the marginal comfort category for the back and hip motions were much smaller than the maximum ROM values for these joints.

There were no significant differences in percentage of JAI in terms of the corresponding maximum ROM values between standing and sitting postures. The relative marginal comfort index, defined as the ratio between joint angles for marginal comfort and the corresponding maximum ROM values, for hip was the smallest among all joints. This was followed in increasing order of the index for lower back and for shoulder, while the index values for elbow were the largest. This means that hip motions are less comfortable than any other joint motion, while elbow motions are the most comfortable. The relative good comfort index exhibited much smaller values of joint deviation, with most

TABLE 13 MHT and Postural Load Index for the 18 Postures[a]

Posture Categories	MHT (min)	Postural Load Index
Comfortable postures (MHT > 10 min)		
SH/AR = 75/50	37.0	3
75/25	18.0	3
100/50	17.0	3
50/25	14.0	0
125/50	12.0	8
50/50	12.0	0
Moderate postures (5 min ≤ MHT ≤ 10 min)		
100/25	10.0	7
100/100	9.0	5
75/100	9.0	4
125/100	8.0	8
75/75	6.0	12
50/100	6.0	10
100/75	5.5	8
50/75	5.3	5
Uncomfortable postures (MHT < 5 min)		
25/25	5.0	10
25/50	4.0	10
150/50	3.5	13
25/75	3.3	10
25/100	3.0	13

[a] Posture was defined after Miedema et al. 1997 in terms of % of shoulder height (SH)/% of arm reach (AR).

TABLE 14 Summary of ISO/CEN Draft Standards and Work Items

Ergonomic guiding principles
ISO 6385: 1981-06-00
ENV 26385: 1990-06-00
Ergonomic principles of the design of work systems
EN 614-1: 1995-02-00
Safety of machinery—Ergonomic design principles—Part 1: Terminology and general principles
prEN 614-2: 1997-12-00
Safety of machinery—Ergonomic design principles—Part 2: Interactions between the design of machinery and work tasks

Anthropometry
EN 547-1: 1996-12-00
ISO/DIS 15534-1:1998-04-00
Safety of machinery—Human body measurements—Part 1: Principles for determining the dimensions required for openings for whole body access into machinery for mobile machinery.
EN 547-2: 1996-12-00
ISO/DIS 15534-2:1998-04-00
Safety of machinery—Human body measurements—Part 2: Principles for determining the dimensions required for access openings
EN 547-3: 1996-12-00
ISO/DIS 15534-3:1998-04-00
Safety of machinery—Human body measurements—Part 3: Anthropometric data
ISO 7250: 1996-07-00
EN ISO 7250: 1997-07-00
Basic human body measurements for technological design (ISO 7250:1996)
ISO/DIS 14738: 1997-12-00
prEN ISO 14738: 1997-12-00
Safety of machinery—Anthropometric requirements for the design of workstations at machinery

Ergonomics—Computer manikins, body templates
Under preparation
Selection of persons for testing of anthropometric aspects of industrial products and designs
Under preparation:
Safeguarding crushing points by means of a limitation of the active forces
Under preparation:

Ergonomics—Reach envelopes
Under preparation:
Anthropometric database
Document scope:
The European Standard establishes an anthropometric database for all age groups to be used as the basis for the design of work equipment, workplaces, and workstations at machinery.
Under preparation:
Notation system of anthropometric measurements used in the European Standards EN 547 Part 1 to Part 3

Biomechanics
prEN 1005-1: 1998-12-00
Safety of machinery—Human physical performance—Part 1: Terms and Definitions
prEN 1005-2: 1998-12-00
Safety of machinery—Human physical performance—Part 2: Manual handling of machinery and component parts of machinery
prEN 1005-3: 1998-12-00
Safety of machinery—Human physical performance—Part 3: Recommended force limits for machinery operation
prEN 1005-4: 1998-11-00
Safety of machinery—Human physical performance—Part 4: Evaluation of working postures in relation to machinery
Under preparation:
Safety of machinery—Human physical performance—Part 5: Risk assessment for repetitive handling at high frequency
ISO/DIS 11226; 1999-02-00
Ergonomics—Evaluation of working postures
ISO/DIS 11228-1: 1998-08-00
Ergonomics—Manual handling—Part 1: Lifting and carrying

TABLE 14 (*Continued*)

Under preparation:
Ergonomics—Manual handling—Part 2: Pushing and pulling
Document scope: To be defined
Under preparation:
Ergonomics—Manual Handling—Part 3: Handling, at high repetition of low loads
Document scope: To be defined
Under preparation:
Ergonomic design of control centers—Part 4: Workstation layout and dimensions

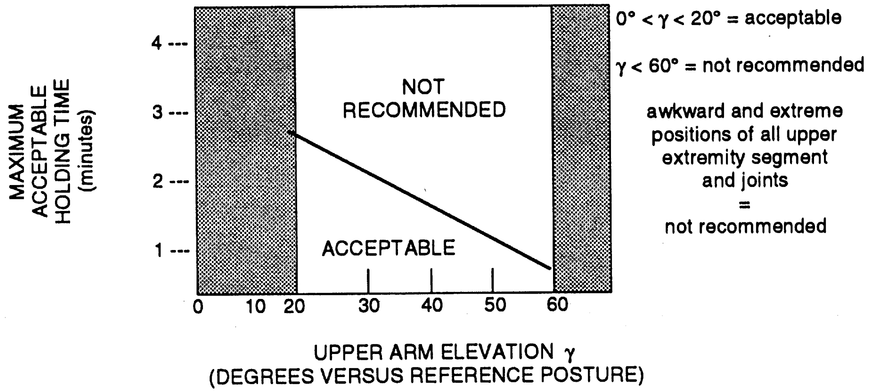
Ergonomics of operability of mobile machines
Under preparation:
Ergonomic design principles for the operability of mobile machinery
Under preparation:

Integrating ergonomic principles for machinery design
Under preparation:
Safety of machinery—Guidance for introducing ergonomic principles and for the drafting of the ergonomics' clauses

The holding time for upper arm elevation can beevaluated as follows.

HOLDING TIME	ACCEPTABLE	NOT RECOMMENDED
> maximum acceptable holding time*		X
≤ maximum acceptable holding time*	X	

- Refer to diagram shown below.

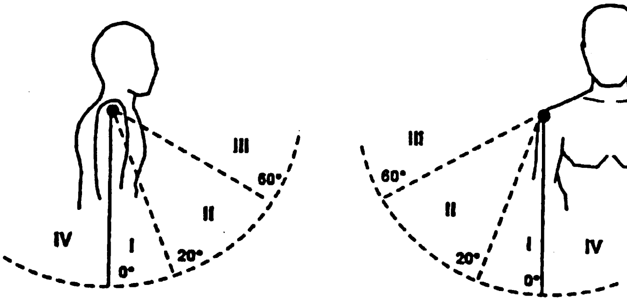


Note: An adequate recovery time should be provided following the holding time for a certain upper arm elevation.

Figure 11 Evaluation of Working Postures (ISO/DIS 11226, 1998).

UPPER ARM ELEVATION

STEP 1 – refer to figure and table below



Upper arm elevation

Evaluation of upper arm elevation

	Static posture	Movement	
		low frequency (<2/minute)	high frequency (≥2/minute)
I*	acceptable	ACCEPTABLE	acceptable
II	conditonally acceptable (step 2A)	acceptable	conditionally acceptable (step 2C)
III	not acceptable	conditionally acceptable (step 2B)	not acceptable
IV	not acceptable	conditionally acceptable (step 2B)	not acceptable

Static posture and high frequency movements (≥ 2 / minute), awkward and extreme positions of all upper extremity segments and joints
=
not acceptable

* It is recommended to strive for working postures with the upper arms hanging down.

STEP 2:

a) acceptable if there is full arm support; if there in so full arm support, acceptability depends on duration of the posture and period of recovery;
b) not acceptable if the machine may be used for long durations;
c) not acceptable if frequency ≥10 / minute and/or if the machine may be used for long durations.

Figure 12 Evaluation of Working Postures in Relation to Machinery (CEN prEN 1005—4, 1997): Upper arm elevation.

index values of less than 40.0. The presented data about joint angles of isocomfort can be used as design guidelines for postural comfort in a variety of human–machine systems.

5.5. International Standards

A list of international standards in the area of anthropometry and biomechanics being developed by ISO is shown in Table 14.

5.5.1. Exposure Assessment of Upper-Limb Repetitive Movements

Recently, Colombini et al. (1999) reported the findings of an international expert group working under auspices of the Technical Committee on Musculoskeletal Disorders of the International Ergonomics Association (IEA) and endorsed by the International Commission on Occupational Health (ICOH). This report provides a set of definitions, criteria, and procedures for assessment of working conditions with respect to exposure of the upper extremities. The document includes two important international standards: Evaluation of Working Postures (ISO/DIS 11226 1998), presented in Figure 11, and Evaluation of Working Postures in Relation to Machinery (CEN prEN 1005—4, 1997): Upper arm elevation, shown in Figure 12.

5.5.2. European Standards for Working Postures During Machinery Operation

The draft proposal of the European (CEN/TC122, WG4/SG5) and the international (ISO TC159/SC3/WG2) standardization document (1993) ''Safety of machinery—human physical performance, Part 4: Working postures during machinery operation,'' specifies criteria of acceptability of working postures vs. exposure times. These criteria are discussed below.

5.5.2.1. General Design Principles Work task and operation should be designed with sufficient physical and mental variation so that the physical load is distributed over various postures and patterns of movements. Designs should accommodate the full range of possible users. To evaluate whether working postures during machinery operation are acceptable, designers should perform a risk assessment, that is, an evaluation of the actual low-varying (static) postures of body segments. The lowest maximum acceptable holding time for the various body segment postures should be determined.

5.5.2.2. Assessment of Trunk Posture An asymmetric trunk posture should be avoided (no trunk axial rotation or trunk lateral flexion). Absence of normal lumbar spine lordosis should be avoided. If the trunk is inclined backward, full support of the lower and upper back should be provided. The forward trunk inclination should be less than 60°, on the condition that the holding time be less than the maximum acceptable holding time for the actual forward trunk inclination, as well as that adequate rest is provided after action (muscle fitness should not be below 80%).

5.5.2.3. Assessment of Head Posture An asymmetric head posture should be avoided (no axial rotation or lateral flexion of the head with respect to the trunk). The head inclination should not be less than trunk inclination (no neck extension). The head inclination should not be larger than the trunk inclination for more than 25° (no extreme neck flexion). If the head is inclined backward, full head support should be provided. The forward head inclination should be less than 25° (if full trunk support is provided), forward inclination should be between less than 85°, on the condition that the holding time should be less than the maximum acceptable holding time for the actual forward head inclination as well as that adequate rest is provided.

5.5.2.4. Assessment of Upper Extremity Posture *Shoulder and upper arm posture.* Upper arm retroflexion and upper-arm adduction should be avoided. Raising the shoulder should be avoided. The upper-arm elevation should be less than 60°, on the condition that the holding time be less than the maximum acceptable holding time for the actual upper-arm elevation as well as that adequate rest be provided after action (muscle fitness should not be below 80%).

Forearm and hand posture. Extreme elbow flexion or extension, extreme forearm pronation or supination, and extreme wrist flexion or extension should be avoided. The hand should be in line with the forearm (no ulnar/radial deviation of the wrist).

6. OCCUPATIONAL BIOMECHANICS

6.1. Definitions

As reviewed by Karwowski (1992), occupational biomechanics is a ''study of the physical interaction of workers with their tools, machines and materials so as to enhance the worker's performance while minimizing the risk of future musculoskeletal disorders'' (Chaffin et al. 1999). There are six methodological areas, or contributing disciplines, important to the development of current knowledge in biomechanics:

1. Kinesiology, or study of human movement, which includes kinematics and kinetics
2. Biomechanical modeling methods, which refer to the forces acting on the human body while a worker is performing well-defined and rather common manual task
3. Mechanical work-capacity evaluation methods in relation to physical capacity of the worker and job demands
4. Bioinstrumentation methods (performance data acquisition and analysis)
5. Classification and time-prediction methods that allow for detailed time analysis of the human work activities and implementation of biomechanics principles to fit the workplace to the worker

6.2. Principles of Mechanics

Biomechanics considers safety and health implications of mechanics, or the study of the action of forces, for the human body (its anatomical and physiological properties) in motion (at work) and at rest. Mechanics, which is based on Newtonian physics, consists of two main areas: statics or the study of the human body at rest or in equilibrium, and dynamics or the study of the human body in motion. Dynamics is further subdivided into two main parts, kinematics and kinetics. Kinematics is concerned with the geometry of motion, including the relationships among displacements, velocities, and accelerations in both translational and rotational movements, without regard to the forces involved. Kinetics, on the other hand, is concerned with forces that act to produce the movements.

The types of forces acting on the human body at any given time are gravitational forces, external forces and ground reaction forces, and muscle forces (Winter 1979). Gravitational forces act downward, through the center of mass of body segments. The external forces are due to the body segment weights and external workload. The forces generated by the muscles are usually expressed as net muscle moments acting upon given joints. It should be noted that body (segment) weight is a (gravitational) force, while body mass is a measure of inertia.

6.3. Biomechanical Analysis

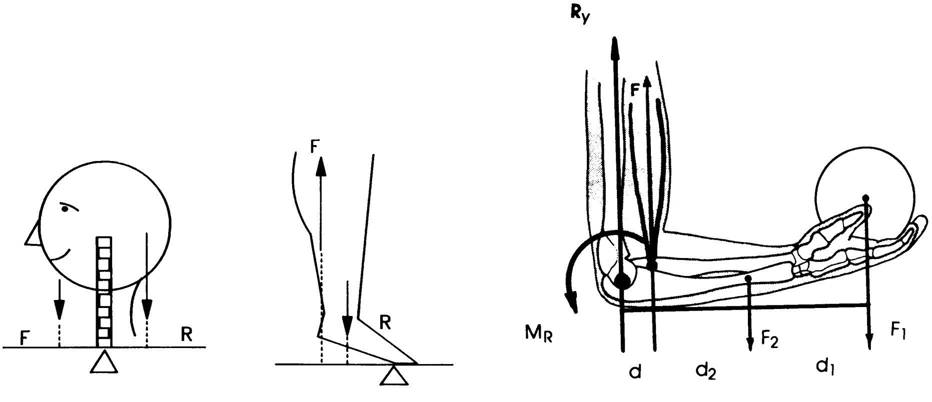
6.3.1. Static Analysis

Static analysis requires consideration of forces and moments of force acting on the body at rest (static equilibrium). A magnitude of the moment of force at the point of rotation is equal to the product of force and the perpendicular distance from the force action line to that point. The moment **(M)** equals force **(F)** times moment arm **(d)**, with unit of measurement (N·m). The static analysis ignores the effects of accelerations, momentum, and friction and is adequate only for analysis of static postures.

The body will be in a static equilibrium state when at rest or dynamic equilibrium when in motion with constant velocity. The translational equilibrium (first condition) of the body (segment) is present when the vector sum of all the forces acting on a body simultaneously is zero ($^{\Sigma}F = 0$). The rotational equilibrium (second condition) of the body is present when the sum of moments about joint is zero ($^{\Sigma}M = 0$). In other words, for the body to be at rest (zero velocity), the sum of all clockwise moments must be equal to the sum of all counterclockwise moments.

6.3.2. Lever Systems

Skeletal muscles of the human body produce movements of the bones about the joints by pulling at their anatomical attachments on the bones across the joints in order to counteract the external forces. This is possible due to three types of lever systems, composed of bones that serve as levers and joints that serve as points of pivot or fulcrums. The lever system consists of an effort force and the resistance force acting at different locations and distances with respect to the fulcrum. The force arm



(*a*) (*b*) (*c*)

Figure 13 (a) Example of the first-class lever system. (b) Example of the second-class lever system. (c) Example of the third-class lever system and single-segment static analysis. (Adapted from Troup and Edwards, reproduced with permission of the Controller of Her Majesty's Stationery Office) and the resistance arm are defined as the distance from the fulcrum to the effort force or resistance force, respectively. Mechanical advantage (MA) of the lever system is defined as the ratio between the force arm distance (df) and the resistance arm distance (dr) where MA = df/dr.

In the first class of lever systems, the fulcrum is located between the effort and resistance. Examples include the triceps muscle action of the ulna when the arm is abducted and held over the head, and the splenium muscles, acting to extend the head across the atlanto-occipital joints (Williams and Lissner 1977). The second class of lever systems is one where resistance is located between the fulcrum and the effort, providing for mechanical advantage greater than one. An example of such a lever system is the distribution of forces in the lower leg when raising one's heel off the ground (see Figure 13). In the third class of lever systems, the effort is located between the fulcrum and the resistance, and consequently the mechanical advantage is always less than one, that is, to balance the resistance, the magnitude of effort must be greater than the magnitude of resistance. Many bone lever systems in the human body, for example, the system involved in forearm flexion as illustrated in Figure 13(*c*), are third-class systems.

7. DESIGN OF MANUAL MATERIALS HANDLING TASKS

7.1. Epidemiology of Low-Back Disorders

As reviewed by Ayoub et al. (1997), manual materials-handling (MMH) tasks, which include unaided lifting, lowering, carrying, pushing, pulling, and holding activities, are the principal source of compensable work injuries affecting primarily the low back in the United States (NIOSH 1981; National Academy of Sciences 1985; *Federal Register* 1986; Bigos et al. 1986; Battié et al. 1990). These include a large number of low-back disorders (LBDs) that are due to either cumulative exposure to manual handling of loads over a long period of time or to isolated incidents of overexertion when handling heavy objects (BNA 1988; National Safety Council 1989; Videman et al. 1990). Overexertion injuries in 1985 for the United States accounted for 32.7% of all accidents: lifting objects (15.1%), carrying, holding, etc. (7.8%), and pulling or pushing objects (3.9%). For the period of 1985–1987, back injuries accounted for 22% of all cases and 32% of the compensation costs.

In the United States, about 28.2% of all work injuries involving disability are caused by overexertion, lifting, throwing, folding, carrying, pushing, or pulling loads that weigh less than 50 lb (National Safety Council 1989). The analysis by industry division showed that the highest percent of such injuries occurred in service industries (31.9%), followed by manufacturing (29.4%), transportation and public utility (28.8%), and trade (28.4%). The total time lost due to disabling work injuries was 75 million workdays, while 35 million days were lost due to other accidents. The total work accident cost was $47.1 billion; the average cost per disabling injury was $16,800. Spengler et al. (1986) reported that while low-back injuries comprised only 19% of all injuries incurred by the workers in one of the largest U.S. aircraft companies, they were responsible for 41% of the total injury costs. It is estimated that the economic impact of back injuries in the United States may be as high as 20 billion annually, with compensation costs exceeding $6 billion per year (BNA 1988).

Major components of the MMH system, related risk factors for low-back pain (LBP), and LBDs include worker characteristics, material/container characteristics, task/workplace characteristics, and work practice characteristics (Karwowski et al. 1997). A wide spectrum of work- and individual-related risk factors have been associated with the LBP and LBDs (Riihimäki 1991). However, the precise knowledge about the extent to which these factors are etiologic and the extent to which they are symptom precipitating or symptom aggravating is still limited. Kelsey and Golden (1988) reported that the risk of lumbar disk prolapse for workers who lift more than 11.3 kg (25 lb) more than 25 times a day is over three times greater than for workers who lift lower weights. The OSHA (1982) study also revealed very important information regarding workers perception of the weights lifted at the time of injury. Among the items perceived by the workers as factors contributing to their injuries were lifting too-heavy objects (reported by 36% of the workers) and underestimation of weight of objects before lifting (reported by 14% of the workers).

An important review of epidemiological studies on risk factors of LBP using five comprehensive publications on LBP was made by Hildebrant (1987). A total of 24 work-related factors were found that were regarded by at least one of the reviewed sources as risk indicators of LBP. These risk factors include the following categories:

1. *General:* heavy physical work, work postures in general
2. *Static workload:* static work postures in general, prolonged sitting, standing or stooping, reaching, no variation in work posture
3. *Dynamic workload:* heavy manual handling, lifting (heavy or frequent, unexpected heavy, infrequent, torque), carrying, forward flexion of trunk, rotation of trunk, pushing/pulling
4. *Work environment:* vibration, jolt, slipping/falling
5. *Work content:* monotony, repetitive work, work dissatisfaction

Many of the cross-sectional studies have shown that LBP is related to heavy manual work (Riihimäki 1991). MMH activities, including those that involve sudden body motions, have been associated with LBP (Svensson and Andersson 1983; Frymoyer et al. 1983, Hansson 1989; Bigos et al. 1991). Among the office workers, LBP was most commonly attributed to lifting tasks (Lloyd et al. 1986). An increased risk of herniated disc was also reported in jobs involving heavy lifting combined with body twisting and forward bending (Kelsey et al. 1984). Manual lifting tasks are often associated with adopting nonneutral trunk postures, which also have been related to LBP (Frymoyer et al. 1980; Maeda et al. 1980; Keyserling et al. 1988; Riihimäki et al 1989).

In addition to physical factors, several studies identified a variety of psychological and psychosocial risk factors of LBP that are related to work environment (Damkot et al. 1984; Magora 1973; Svensson and Andersson 1983, 1989; Bigos et al. 1991). However, as pointed out by Riihimäki (1991), since most of the studies have been retrospective in nature, it is difficult to determine whether these factors are antecedents or consequences of back pain (Kelsey et al. 1988), and whether these factors play a role in the etiology of LBDs or only affect the perception of symptoms and sickness behavior.

7.2. MMH Capacity Design Criteria

Workers who perform heavy physical work are subjected not only to forces and stresses from the immediate physical environment but also to mechanical forces generated from within the body. As a result of these forces and stresses, a strain is produced on the worker's musculoskeletal system as well as on other systems such as the cardiopulmonary system. One of the most important issues in the application of ergonomics to work design is to reduce the stresses imposed on the musculoskeletal and cardiopulmonary systems (Ayoub and Mital 1989). Several approaches have been used by different investigators to establish safe handling limits, including the psychophysical approach, the physiological approach, and the biomechanical approach.

7.3. The Psychophysical Approach

The psychophysical approach relies on the worker's perceived exertion to quantify his or her tolerance level, thereby establishing the maximum acceptable weights or forces (MAW/F) for different MMH activities (e.g., maximum acceptable weight of lift [MAWL]). Psychophysics deals with the relationship between human sensation and their physical stimuli. Borg (1962) and Eisler (1962) found that the perception of both muscular effort and force obey the psychophysical function, where sensation magnitude (S) grows as a power function of the stimulus (I). Stevens (1975) reported the relationship between the strength of the sensation (S) and the intensity of its physical stimulus (I) by the power function:

$$S = k \times I^n$$

where: S = strength of sensation
I = intensity of physical stimulus
k = constant
n = slope of the line, which represents the power function when plotted on log–log coordinates

The use of psychophysics in the study of MMH tasks requires the worker to adjust the weight, force, or frequency in a handling situation until they feel it represents their MAW/F (Asfour et al. 1984; Gamberale and Kilböm 1988; Garg and Banagg 1988; Legg and Myles 1985; Mital et al. 1989; Snook and Ciriello 1991). Psychophysical limits usually refer to weights or forces (MAW/F), although maximum acceptable frequencies have also been established (e.g., Nicholson and Legg 1986; Snook and Ciriello 1991). Despite the relative simplicity of the psychophysical method to determine acceptable limits for manual lifting, which makes this approach quite popular (Karwowski et al 1999), caution should be exercised with respect to interpretation and usability of the currently available design limits and databases.

7.4. MMH Design Databases

One can use already available databases such as the one reported by Snook and Ciriello (1991). Another database was reported by Mital (1992) for symmetrical and asymmetrical lifting, and other databases include work by Ayoub et al. (1978) and Mital (1984). Using such available data replaces conducting a study for every work task and group of workers. Tables provided by the various investigators can be used to estimate the MAW/F for a range of job conditions and work populations. The databases provided in tabular format often make allowances for certain task, workplace, and/or worker characteristics. The use of databases begins with the determination of the various characteristics with which the database is stratified.

7.5. Psychophysical Models

Another method to estimate MAW/F is regression models based on the psychophysical data. Most of these models predict MAWL. The design data presented here are based upon the database of Snook and Ciriello (1991). Table 15 provides Snook and Ciriello's (1991) two-handed lifting data for males and females, as modified by Mital et al. (1993). Those values that were modified have been identified. The data in these tables were modified so that a job severity index value of 1.5 is not exceeded, which corresponds to 27.24 kg. Likewise, a spinal compression value that, on average, provides a margin of safety for the back of 30% was used for the biomechanical criterion, yielding a maximum load of 27.24 for males and 20 kg for females. Finally, the physiological criterion of energy expenditure was used. The limits selected were 4 kcal/min for males and 3 kcal/min for females for an 8-hr working day (Mital et al. 1993). The design data for maximal acceptable weights for two-handed pushing/pulling tasks, maximal acceptable weights for carrying tasks, and maximal acceptable holding times can be found in Snook and Ciriello (1991) and Mital et al. (1993). The maximal acceptable weights for manual handling in unusual postures are presented by Smith et al. (1992).

7.6. The Physiological Approach

The physiological approach is concerned with the physiological response of the body to MMH tasks. During the performance of work, physiological changes take place within the body. Changes in work methods, performance level, or certain environmental factors are usually reflected in the stress levels of the worker and may be evaluated by physiological methods. The basis of the physiological approach to risk assessment is the comparison of the physiological responses of the body to the stress of performing a task with levels of permissible physiological limits. Many physiological studies of MMH tended to concentrate on whole body indicators of fatigue such as heart rate, energy expenditure, blood lactate, or oxygen consumption as a result of the workload. Mital et al. (1993) arrived at the same conclusion as Petrofsky and Lind, that is, that the physiological criteria for lifting activities for males should be approximately 4 kcal/min and 3 kcal/min for females.

The energy cost of manual handling activities can be estimated based on the physiological response of the body to the load, that is, by modeling the physiological cost using work and worker characteristics. The estimates obtained from such models are then compared to the literature recommendations of permissible limits. Garg et al. (1978) report metabolic cost models. Although currently in need of update, they still provide a more comprehensive and flexible set of physiological cost models as a function of the task variables. The basic form of the Garg et al. model is:

$$E_{\text{job}} = \left(\sum_{i=1}^{N_p} (\mathrm{E}_{\text{post-}i} \times T_i) + \sum_{i=1}^{N_t} \mathrm{E}_{\text{task-}i} \right) \div T$$

where: E_{job} = average energy expenditure rate of the job (kcal/min)
$E_{\text{post-}i}$ = metabolic energy expenditure rate due to maintenance of ith posture (kcal/min)
T_i = time duration of ith posture (min)
N_p = total number of body postures employed in the job
$E_{\text{task-}i}$ = net metabolic energy expenditure of the ith task in steady state (kcal)
N_t = total number of tasks in the given job
T = time duration of the job (min)

Different models require different input data, but typically most of these models involve input information regarding task type, load weight/force, load size, height, frequency, and worker characteristics, which include body weight and gender.

7.7. The Biomechanical Approach

The biomechanical approach focuses on the establishment of tissue tolerance limits of the body, especially the spine (e.g., compressive and shear force limits tolerated by the lumbar spine). The levels of stresses imposed on the body are compared to permissible levels of biomechanical stresses, measured by, for example, peak joint moments, peak compressive force on the lumbar spine, and peak shear forces on the lumbar spine. Other measures include mechanical energy, average and integrated moments or forces over the lifting, and MMH activity times (Andersson 1985; Gagnon and Smyth 1990; Kumar 1990). Methods used to estimate the permissible level of stress in biomechanics for MMH include strength testing, lumbar tissue failure, and the epidemiological relationship between biomechanical stress and injury.

Tissue failure studies are based on cadaver tissue strength. Generally, the research has focused on the ultimate compressive strength of the lumbar spine. Studies and literature reviews by Brinckmann et al. (1989) and Jäger and Luttmann (1991) indicate that the ultimate compressive strength of

TABLE 15 Recommended Weight of Lift (kg) for Male (Female) Industrial Workers for Two-Handed Symmetrical Lifting for Eight Hours

		Floor to 80 cm Height							
		Frequency of Lift							
Cont. Size		1/8 hr	1/30 min	1/5 min	1/min	4/min	8/min	12/min	16/min
75 cm	90	17 (12)	14 (9)	14 (8)	11(7)	9 (7)	7 (6)	6 (5)	4.5 (4)
	75	24 (14)	21 (11)	20 (10)	16 (9)	13 (9)	10.5 (8)	9 (7)	7 (6)
	50	27* (17)	27* (13)	27 (12)	22 (11)	17 (10)	14 (9)	12 (8)	9.5 (7)
	25	27* (20*)	27* (15)	27* (14)	27* (13)	21 (12)	17.5 (11)	15 (9)	12 (7)
	10	27* (20*)	27* (17)	27* (16)	27* (14)	25 (14)	20.5 (13)	18 (11)	14.5 (9)
49 cm	90	20 (13)	17 (9)	16 (8)	13 (8)	10 (8)	7 (7)	7 (6)	6.5 (5)
	75	27* (16)	24 (12)	24 (10)	19 (10)	14 (9)	10 (8)	10 (7)	9 (6)
	50	27* (19)	27* (14)	27* (13)	26 (12)	19 (11)	15 (10)	12.5 (9)	10 (8)
	25	27* (20*)	27* (17)	27* (15)	27* (14)	24 (13)	18.5 (11)	15 (10)	12 (8)
	10	27* (20*)	27* (19)	27* (17)	27* (15)	28 (15)	22 (13)	17.5 (11)	15 (9)
34 cm	90	23 (15)	19 (11)	19 (10)	15 (9)	11 (9)	7 (8)	7 (7)	6.5 (7)
	75	27* (19)	27* (14)	27* (13)	22 (12)	17 (11)	10 (9)	10 (8)	9.5 (7)
	50	27* (20*)	27* (17)	27* (16)	27* (14)	22 (13)	15 (11)	14 (10)	12 (8)
	25	27* (20*)	27* (20*)	27* (18)	27* (17)	27* (15)	20 (13)	17 (12)	14 (10)
	10	27* (20*)	27* (20*)	27* (20*)	27* (19)	27* (18)	25 (15)	21 (13)	15 (11)
75 cm	90	15 (10)	13 (7.5)	13 (6.5)	10 (6)	8 (6)	6 (5)	6 (4)	4 (3)
	75	22 (12)	20 (9)	19 (8)	14.5 (7.5)	12 (7.5)	10 (6.5)	9 (6)	7 (5)
	50	27* (14)	25 (11)	24 (10)	20 (9)	15 (8)	13 (7.5)	11 (6.5)	9 (6)
	25	27* (17)	27* (12.5)	27* (11.5)	24.5 (11)	18 (10)	15 (9)	12 (7.5)	11 (6.5)
	10	27* (19)	27* (14)	27* (13)	27* (11.5)	22 (11.5)	19 (11)	16 (9)	13 (8)
49 cm	90	18 (11)	16 (7.5)	15 (6.5)	12.5 (6.5)	9 (6.5)	6 (6)	6 (5)	5 (4)
	75	27 (13)	22.5 (10)	22.5 (8)	18 (8)	14 (7.5)	10 (6.5)	9 (6)	8 (5)
	50	27* (16)	27* (11.5)	27* (11)	24 (10)	18 (9)	14 (8)	12 (7.5)	10 (6.5)
	25	27* (17)	27* (14)	27* (12.5)	27* (11.5)	22 (11)	18 (9.5)	14 (8)	11 (7)
	10	27* (19)	27* (16)	27* (14)	27* (12.5)	27 (12.5)	21 (11)	17 (9)	14 (7.5)
34 cm	90	22 (12.5)	18 (9)	18 (8)	14 (7.5)	11 (7.5)	6 (6.5)	6 (6)	5 (5)
	75	27* (16)	26 (11.5)	25 (11)	21 (10)	16 (9)	10 (8)	9 (6.5))	8 (5.5)
	50	27* (19)	27* (14)	27* (13)	27* (11.5)	22 (11)	14 (9.5)	12 (8)	10 (7)
	25	27* (20*)	27* (17)	27* (15)	27* (14)	27 (12.5)	20 (11)	14 (10)	11 (9)
	10	27* (20*)	27* (19)	27* (17)	27* (16)	27* (15)	21 (13)	17 (11)	14 (9)

TABLE 15 (*Continued*)

		Floor to 80 cm Height							
		Frequency of Lift							
Cont. Size		1/8 hr	1/30 min	1/5 min	1/min	4/min	8/min	12/min	16/min
	90	15 (9)	12 (6)	12 (6)	9.5 (5)	8 (5)	6 (4.5)	5 (4)	3 (3)
	75	21 (11)	18 (8)	17 (7)	14 (7)	11 (7)	9 (6)	8 (5)	6 (4.5)
75 cm	50	27* (12.5)	24 (10)	23 (9)	19 (8)	15 (7)	12 (7)	10 (6)	8 (5.5)
	25	27* (15)	27* (11)	27* (10)	24 (10)	18 (9)	14 (8)	12 (7)	9 (6)
	10	27* (17)	27* (12.5)	27* (12)	27* (10)	22 (10)	18 (10)	15 (8)	12 (7)
	90	17 (10)	15 (7)	14 (6)	11 (6)	9 (6)	6 (5.5)	6 (4.5)	4 (3.5)
	75	24 (12)	21 (9)	21 (7)	16 (7)	12 (7)	9 (6)	9 (5)	7 (4.5)
49 cm	50	27* (14)	27* (10)	27* (10)	22 (9)	16 (8)	14 (7)	12 (7)	10 (6)
	25	27* (15)	27* (12)	27* (11)	27* (10)	20 (10)	17 (8.5)	14 (7)	11 (6.5)
	10	27* (17)	27* (14)	27* (12)	27* (11)	23 (11)	20 (10)	17 (8)	14 (7)
	90	20 (11)	16 (8)	16 (7)	13 (7)	9 (7)	6 (6)	6 (5)	4 (4.5)
	75	27* (14)	24 (10)	24 (10)	19 (9)	15 (8)	9 (7)	9 (6)	7 (5)
34 cm	50	27* (17)	27* (12)	27* (12)	26 (10)	19 (10)	14 (8.5)	12 (7)	10 (6)
	25	27* (20)	27* (15)	27* (13.5)	27* (12)	23 (11)	20 (10)	14 (9)	11 (8)
	10	27* (20*)	27* (17)	27* (15)	27* (14)	27* (13.5)	24 (12)	17 (10)	14 (8)
	90	19 (13)	18 (11)	16 (10)	15 (9)	13 (8)	7 (6)	6 (6)	5 (5)
	75	25 (15)	23 (13)	21 (12)	20 (11)	17 (9)	8 (7)	8 (7)	7 (6)
75 cm	50	27* (17)	27* (15)	26 (14)	25 (13)	21 (11)	12 (9)	11 (9)	9 (8)
	25	27* (20)	27* (17)	27* (16)	27* (14)	26 (12)	17 (11)	13 (10)	12 (9)
	10	27* (20*)	27* (19)	27* (17)	27* (16)	27* (14)	23 (12.5)	20 (11)	16 (9.5)
	90	19 (13)	18 (11)	16 (10)	15 (9)	13 (8)	7 (6)	6 (6)	5 (5)
	75	25 (15)	23 (13)	21 (12)	20 (11)	17 (9)	8 (7)	8 (7)	7 (6)
49 cm	50	27* (17)	27* (15)	26 (14)	25 (13)	21 (11)	12 (9)	11 (9)	9 (8)
	25	27* (20)	27* (17)	27* (16)	27* (14)	26 (12)	17 (11)	13 (10)	12 (9)
	10	27* (20*)	27* (19)	27* (17)	27* (16)	27* (14)	23 (12.5)	20 (11)	16 (9.5)
	90	22 (14)	20 (12)	18 (11)	17 (10)	14 (9)	7 (7)	6 (6.5)	5 (6.5)
	75	27* (17)	26 (14)	23 (13)	22 (12)	18 (11)	8 (8.5)	8 (8.5)	7 (8)
34 cm	50	27* (19)	27* (17)	27* (15)	27* (14)	23 (13)	12 (11)	11 (10)	9 (8.5)
	25	27* (20*)	27* (19)	27* (17)	27* (16)	27 (14)	17 (13.5)	13 (11.5)	12 (11)
	10	27* (20*)	27* (20*)	27* (19)	27* (18)	27* (16)	24 (14.5)	21 (13)	16 (11.5)

75 cm	90	16 (11)	15 (9.5)	13 (9)	12 (8)	11 (7)	7 (5)	6 (5)	5 (4.5)
	75	22 (13)	20 (11)	18 (10.5)	17 (9.5)	15 (8)	8 (6)	8 (6)	6 (5)
	50	27* (15)	25 (13)	23 (12)	21 (11)	19 (10)	12 (8)	11 (8)	8 (7)
	25	27* (17.5)	27* (15)	27 (14)	26 (12)	23 (10.5)	17 (10)	13 (9)	11 (8)
	10	27* (19)	27* (17)	27* (15)	27* (14)	27 (12)	22 (11)	18 (10)	13 (8)
49 cm	90	16 (11)	15 (9.5)	13 (9)	12 (8)	11 (7)	7 (6)	6 (5)	5 (4.5)
	75	22 (13)	20 (11)	18 (10.5)	17 (9.5)	15 (8)	8 (6)	8 (6)	6 (5)
	50	27* (15)	25 (13)	23 (12)	21 (11)	19 (10)	12 (8)	11 (8)	8 (7)
	25	27* (17.5)	27* (15)	27 (14)	26 (12)	23 (10.5)	17 (10)	13 (9)	11 (8)
	10	27* (19)	27* (17)	27* (15)	27* (14)	27 (12)	22 (11)	18 (10)	13 (8)
34 cm	90	18 (12)	17 (10.5)	15 (10)	14 (9)	12 (8)	7 (6)	6 (6)	5 (6)
	75	24 (15)	22 (12)	20 (11)	19(10.5)	16 (10)	8 (7.5)	8 (7.5)	7 (7)
	50	27* (17)	27* (15)	25 (13)	24 (12)	20 (11)	12 (10)	11 (9)	9 (7.5)
	25	27* (19)	27* (17)	27* (15)	27* (14)	24 (12)	20 (11)	16 (10)	12 (10)
	10	27* (20*)	27* (19)	27* (17)	27* (16)	27* (14)	22 (13)	18 (11)	13 (10)
75 cm	90	15 (9)	14 (8)	12 (7)	12 (7)	9 (7)	7 (5)	6 (4)	4 (3)
	75	20 (11)	18 (9)	15 (9)	15 (8)	12 (8)	9 (6)	8 (5)	6 (4)
	50	25 (13)	23 (11)	20 (10)	19 (9)	16 (9)	12 (8)	10 (7)	7 (6)
	25	27* (14)	27 (12)	25 (11)	23 (10)	19 (10)	15 (9)	12 (8)	10 (7)
	10	27* (16)	27* (14)	27* (13)	27 (12)	22 (11)	17 (10)	13 (9)	12 (8)
49 cm	90	18 (10)	16 (9)	14 (8)	14 (7)	11 (7)	7 (5)	7 (4)	5 (3)
	75	23 (12)	21 (10)	19 (9)	18 (9)	14 (8)	9 (6)	8 (5)	6 (4)
	50	27* (14)	27 (12)	24 (11)	23 (10)	18 (9)	12 (8)	10 (7)	9 (6)
	25	27* (15)	27* (13)	27* (12)	27* (11)	21 (10)	15 (9)	12 (8)	10 (7)
	10	27* (17)	27* (15)	27* (14)	27* (13)	25 (11)	17 (10)	13 (9)	11 (8)
34 cm	90	20 (12)	18 (11)	17 (10)	16 (9)	13 (8)	7 (6)	6 (6)	5 (6)
	75	26 (14)	24 (12)	22 (11)	21 (11)	17 (9)	9 (7)	8 (7)	8 (7)
	50	27* (17)	27* (14)	27* (13)	26 (12)	21 (11)	12 (9)	11 (9)	10 (8)
	25	27* (19)	27* (16)	27* (15)	27* (14)	25 (12)	15 (11)	14 (10)	13 (9)
	10	27* (20*)	27* (18)	27* (16)	27* (15)	27* (14)	17 (12)	16 (11)	15 (9.5)

Adapted from Mital et al. 1993b.

* Weight limited by biomechanical design criterion (3930 N spinal compression for males, 2689 N for females).

Weight limited by physiological design criterion (4 kcal / min for males, 3 kcal / min for females).

()—Values in parentheses are for females

hr = hours; min. = minute(s)

cadaver lumbar segments varies from approximately 800 N to approximately 13,000 N. Jäger and Luttman (1991) report a mean failure for compression at 5,700 N for males with a standard deviation of 2600 N. For females, this failure limit was found to be 3900 N with a standard deviation of approximately 1500 N. In addition, several factors influence the compressive strength of the spinal column, including age, gender, specimen cross-section, lumbar level, and structure of the disc or vertebral body. The ultimate compressive strength of various components of the spine can be estimated with following regression model (Jäger and Luttman 1991):

$$\begin{aligned}\text{Compressive strength (kN)} = {} & (7.65 + 1.18\,G) - (0.502 + 0.382\,G)\,A \\ & + (0.035 + 0.127\,G)\,C - 0.167\,L - 0.89\,S\end{aligned}$$

where: G = gender (0 for female; 1 for male)
A = decades (e.g., 30 years = 3, 60 = 6)
L = lumbar level (0 for L5/S1; incremental values for each lumbar disc or vertebra)
C = cross-section (cm^2)
S = structure (0 for disc; 1 for vertebra)

It should be noted that statically determined tolerances may overestimate compressive tolerances (Jäger and Luttman 1992b). Modeling studies by Potvin et al. (1991) suggest that erector spinae oblique elements could contribute about 500 N sagittal shear to leave only 200 N sagittal shear for discs and facets to counter. According to Farfan (1983), the facet joints are capable of absorbing 3,100 N to 3,600 N while the discs support less than 900 N.

Due to the complexity of dynamic biomechanical models, assessment of the effects of lifting on the musculoskeletal system has most frequently been done with the aid of static models. Many lifting motions, which are dynamic in nature, appear to have substantial inertia components. McGill and Norman (1985) also compared the low-back moments during lifting when determined dynamically and statically. They found that the dynamic model resulted in peak L4/L5 moments 19% higher on the average, with a maximum difference of 52%, than those determined from the static model. Given the complexity of the human body and the simplicity of the biomechanical models, values from these models can only be estimates and are best used for comparison purposes rather than suggesting absolute values (Delleman et al. 1992).

7.8. Revised NIOSH (1991) Lifting Equation

The 1991 revised lifting equation has been expanded beyond the previous guideline and can be applied to a larger percentage of lifting tasks (Waters et al. 1993). The recommended weight limit (RWL) was designed to protect 90% of the mixed (male/female) industrial working population against LBP. The 1991 equation is based on three main components: standard lifting location, load constant, and risk factor multipliers. The standard lifting location (SLL) serves as the 3D reference point for evaluating the parameters defining the worker's lifting posture. The SLL for the 1981 Guide was defined as a vertical height of 75 cm and a horizontal distance of 15 cm with respect to the midpoint between the ankles. The horizontal factor for the SLL was increased from 15 cm to 25 cm displacement for the 1991 equation. This was done in view of recent findings that showed 25 cm as the minimum horizontal distance in lifting that did not interfere with the front of the body. This distance was also found to be used most often by workers (Garg and Badger 1986; Garg 1989).

The load constant (LC) refers to a maximum weight value for the SLL. For the revised equation, the load constant was reduced from 40 kg to 23 kg. The reduction in the load constant was driven in part by the need to increase the 1981 horizontal displacement value from a 15-cm to a 25-cm displacement for the 1991 equation (noted above in item 1). Table 16 shows definitions of the relevant terms utilized by the 1991 equation. The RWL is the product of the load constant and six multipliers:

$$\text{RWL (kg)} = LC \times HM \times VM \times DM \times AM \times FM \times CM$$

The multipliers (M) are defined in terms of the related risk factors, including the horizontal location (HM), vertical location (VM), vertical travel distance (DM), coupling (CM), frequency of lift (FM), and asymmetry angle (AM). The multipliers for frequency and coupling are defined using relevant tables. In addition to lifting frequency, the work duration and vertical distance factors are used to compute the frequency multiplier (see Table 17). Table 18 shows the coupling multiplier (CM), while Table 19 provides information about the coupling classification.

The horizontal location (H) is measured from the midpoint of the line joining the inner ankle bones to a point projected on the floor directly below the midpoint of the hand grasps (i.e., load center). If significant control is required at the destination (i.e., precision placement), then H should be measured at both the origin and destination of the lift. This procedure is required if there is a need to: (1) regrasp the load near the destination of the lift, (2) momentarily hold the object at the

TABLE 16 Terms of the 1991 NIOSH Equation

Multiplier	Formula (cm)
Load constant	$LC = 23$ kg
Horizontal	$HM = 25/H$
Vertical	$VM = 1 - (0.003 \mid V - 75 \mid)$
Distance	$DM = 0.82 + 4.5/D$
Asymmetry	$AM = 1 - 0.0032\,A$
Frequency	FM (see Table 147)
Coupling	CM (see Table 148)

H = the horizontal distance of the hands from the midpoint of the ankles, measured at the origin & destination of the lift (cm).
V = the vertical distance of the hands from the floor, measured at the origin and destination of the lift (cm).
D = the vertical travel distance between the origin and destination of the lift (cm).
A = the angle of asymmetry—angular displacement of the load from the sagittal plane, measured at the origin and destination of the lift (degrees).
F = average frequency of lift (lifts/minute).
C = load coupling, the degree to which appropriate handles, devices, or lifting surfaces are present to assist lifting and reduce the possibility of dropping the load.

From Waters et al. 1993. Reprinted with permission by Taylor & Francis Ltd., London, http://www.tandf.co.uk.

destination, or (3) position or guide the load at the destination. If the distance is less than 10 in (25 cm), then *H* should be set to 10 in (25 cm).

The vertical location (*V*) is defined as the vertical height of the hands above the floor and is measured vertically from the floor to the midpoint between the hand grasps, as defined by the large middle knuckle. The vertical location is limited by the floor surface and the upper limit of vertical reach for lifting (i.e., 70 in or 175 cm).

The vertical travel distance variable (*D*) is defined as the vertical travel distance of the hands between the origin and destination of the lift. For lifting tasks, *D* can be computed by subtracting the vertical location (*V*) at the origin of the lift from the corresponding *V* at the destination of the

TABLE 17 Frequency Multipliers for the 1991 Lifting Equation

Frequency Lifts (min)	Continuous Work Duration					
	< 8 hours		< 2 hours		< 1 hour	
	$V < 75$	$V > 75$	$V < 75$	$V > 75$	$V < 75$	$V > 75$
0.2	0.85	0.85	0.95	0.95	1.00	1.00
0.5	0.81	0.81	0.92	0.92	0.97	0.97
1	0.75	0.75	0.88	0.88	0.94	0.94
2	0.65	0.65	0.84	0.84	0.91	0.91
3	0.55	0.55	0.79	0.79	0.88	0.88
4	0.45	0.45	0.72	0.72	0.84	0.84
5	0.35	0.35	0.60	0.60	0.80	0.80
6	0.27	0.27	0.50	0.50	0.75	0.75
7	0.22	0.22	0.42	0.42	0.70	0.70
8	0.18	0.18	0.35	0.35	0.60	0.60
9	–	0.15	0.30	0.30	0.52	0.52
10	–	0.13	0.26	0.26	0.45	0.45
11	–	–	–	0.23	0.41	0.41
12	–	–	–	0.21	0.37	0.37
13	–	–	–	–	–	0.34
14	–	–	–	–	–	0.31
15	–	–	–	–	–	0.28

From Waters et al. 1993. Reprinted with permission by Taylor & Francis Ltd., London, http://www.tandf.co.uk.

TABLE 18 The Coupling Multipliers for the 1991 Lifting Equation

Couplings	$V < 75$ cm	$V = 75$ cm
Good	1.00	1.00
Fair	0.95	1.00
Poor	0.90	0.90

From Waters et al. 1993. Reprinted with permission by Taylor & Francis Ltd., London, http://www.tandf.co.uk.

lift. For lowering tasks, D is equal to V at the origin minus V at the destination. The variable (D) is assumed to be at least 10 in (25 cm) and no greater than 70 in (175 cm). If the vertical travel distance is less than 10 in (25 cm), then D should be set to 10 in (25 cm).

The asymmetry angle A is limited to the range of 0° to 135°. If $A > 135°$, then AM is set equal to zero, which results in a RWL of 0. The asymmetry multiplier (AM) is $1 - (0.0032A)$. The AM has a maximum value of 1.0 when the load is lifted directly in front of the body and a minimum value of 0.57 at 135° of asymmetry.

The frequency multiplier (FM) is defined by (1) the number of lifts per minute (frequency), (2) the amount of time engaged in the lifting activity (duration), and (3) the vertical height of the lift from the floor. Lifting frequency (F) refers to the average number of lifts made per minute, as measured over a 15-min period. Lifting duration is classified into three categories: short duration,

TABLE 19 Coupling Classification

Good coupling

1. For containers of optimal design, such as some boxes, crates, etc., a "Good" hand-to-object coupling would be defined as handles or hand-hold cutouts of optimal design
2. For loose parts or irregular objects that are not usually containerized, such as castings, stock, and supply materials, a "Good" hand-to-object coupling would be defined as a comfortable grip in which the hand can be easily wrapped around the object.

Fair coupling

1. For containers of optimal design, a "Fair" hand-to-object coupling would be defined as handles or hand-hold cut-outs of less than optimal design.
2. For containers of optimal design with no handles or hand-hold cutouts or for loose parts or irregular objects, a "Fair" hand-to-object coupling is defined as a grip in which the hand can be flexed about 90°.

Poor coupling

1. Containers of less than optimal design or loose parts or irregular objects that are bulky, hard to handle, or have sharp edges.
2. Lifting nonrigid bags (i.e., bags that sag in the middle).

Notes:

1. An optimal handle design has 0.75–1.5 in (1.9–3.8 cm) diameter, ≥ 4.5 in (11.5 cm) length, 2 in (5 cm) clearance, cylindrical shape, and a smooth, nonslip surface.
2. An optimal hand-hold cutout has the following approximate characteristics: ≥ 1.5 in (3.8 cm) height, 4.5 in (11.5 cm) length, semioval shape, ≥ 2 in (5 cm) clearance, smooth nonslip surface, and ≥ 0.25 in (0.60 cm) container thickness (e.g., double-thickness cardboard).
3. An optimal container design has ≤ 16 in (40 cm) frontal length, ≥ 12 in (30 cm) height, and a smooth nonslip surface.
4. A worker should be capable of clamping the fingers at nearly 90° under the container, such as required when lifting a cardboard box from the floor.
5. A container is considered less than optimal if it has a frontal length > 16 in (40 cm), height > 12 in (30 cm), rough or slippery surfaces, sharp edges, asymmetric center of mass, unstable contents, or requires the use of gloves. A loose object is considered bulky if the load cannot easily be balanced between the hand grasps.
6. A worker should be able to wrap the hand comfortably around the object without causing excessive wrist deviations or awkward postures, and the grip should not require excessive force.

After Waters et al. 1994. Courtesy of the U.S. Department of Health and Human Services.

moderate duration, and long duration. These categories are based on the pattern of continuous work-time and recovery-time (i.e., light work) periods.

A continuous work-time period is defined as a period of uninterrupted work. Recovery time is defined as the duration of light work activity following a period of continuous lifting. Short duration defines lifting tasks that have a work duration of one hour or less, followed by a recovery time equal to 1.2 times the work time. Moderate duration defines lifting tasks that have a duration of more than one hour, but not more than two hours, followed by a recovery period of at least 0.3 times the work time. Long duration defines lifting tasks that have a duration of between two and eight hours, with standard industrial rest allowances (e.g., morning, lunch, and afternoon rest breaks).

The lifting index (LI) provides a relative estimate of the physical stress associated with a manual lifting job and is equal to the load weight divided by the RWL. According to Waters et al. (1994), the RWL and LI can be used to guide ergonomic design in several ways:

1. The individual multipliers can be used to identify specific job-related problems. The general redesign guidelines related to specific multipliers are shown in Table 20.
2. The RWL can be used to guide the redesign of existing manual lifting jobs or to design new manual lifting jobs.
3. The LI can be used to estimate the relative magnitude of physical stress for a task or job. The greater the LI, the smaller the fraction of workers capable of safely sustaining the level of activity.
4. The LI can be used to prioritize ergonomic redesign. A series of suspected hazardous jobs could be rank ordered according to the LI and a control strategy could be developed according to the rank ordering (i.e., jobs with lifting indices about 1.0 or higher would benefit the most from redesign).

The 1991 equation should not be used if any of the following conditions occur: lifting/lowering with one hand; lifting/lowering for over eight hours, lifting/lowering while seated or kneeling; lifting/lowering in a restricted workspace, lifting/lowering unstable objects; lifting/lowering while carrying, pushing, or pulling; lifting/lowering with wheelbarrows or shovels; lifting/lowering with high-speed motion (faster than about 30 in/sec); lifting/lowering with unreasonable foot/floor coupling (<0.4 coefficient of friction between the sole and the floor); lifting/lowering in an unfavorable environment (i.e., temperature significantly outside 66–79°F (19–26°C) range; relative humidity outside 35–50% range).

7.9. Computer Simulation of the Revised NIOSH Lifting Equation (1991)

One way to investigate the practical implications of the 1991 lifting equation for industry is to determine the likely results of the equation when applying a realistic and practical range of values

TABLE 20 General Design/Redesign Suggestions for Manual Lifting Tasks

If HM is less than 1.0	Bring the load closer to the worker by removing any horizontal barriers or reducing the size of the object. Lifts near the floor should be avoided; if unavoidable, the object should fit easily between the legs.
If VM is less than 1.0	Raise/lower the origin/destination of the lift. Avoid lifting near the floor or above the shoulders.
If DM is less than 1.0	Reduce the vertical distance between the origin and the destination of the lift.
If AM is less than 1.0	Move the origin and destination of the lift closer together to reduce the angle of twist, or move the origin and destination further apart to force the worker to turn the feet and step, rather than twist the body.
If FM is less than 1.0	Reduce the lifting frequency rate, reduce the lifting duration, or provide longer recovery periods (i.e., light work period).
If CM is less than 1.0	Improve the hand-to-object coupling by providing optimal containers with handles or hand-hold cutouts, or improve the hand-holds for irregular objects.
If the RWL at the destination is less than at the origin	Eliminate the need for significant control of the object at the destination by redesigning the job or modifying the container/object characteristics.

As recommended by Waters et al. 1994. Courtesy of the U.S. Department of Health and Human Services.

for the risk factors (Karwowski 1992). This can be done using modern computer simulation techniques in order to examine the behavior of the 1991 NIOSH equation under a broad range of conditions. Karwowski and Gaddie (1995) simulated the 1991 equation using SLAM II (Pritsker 1986), a simulation language for alternative modeling, as the product of the six independent factor multipliers represented as attributes of an entity flowing through the network. For this purpose, probability distributions for all the relevant risk factors were defined and a digital simulation of the revised equation was performed.

As much as possible, the probability distributions for these factors were chosen to be representative of the real industrial workplace (Ciriello et al. 1990; Brokaw 1992; Karwowski and Brokaw 1992; Marras et al. 1993). Except for the vertical travel distance factor, coupling, and asymmetry multipliers, all factors were defined using either normal or lognormal distributions. For all the factors defined as having lognormal distributions, the procedure was developed to adjust for the required range of real values whenever necessary. The SLAM II computer simulation was run for a total of 100,000 trials, that is, randomly selected scenarios that realistically define the industrial tasks in terms of the 1991 equation. Descriptive statistical data were collected for all the input (lifting) factors, the respective multipliers, and the resulting recommended weight limits. The input factor distributions were examined in order to verify the intended distributions.

The results showed that for all lifting conditions examined, the distribution of recommended weight limit values had a mean of 7.22 kg and a standard deviation of 2.09 kg. In 95% of all cases, the RWL was at or below the value of 10.5 kg (about 23.1 lb). In 99.5% of all cases, the RWL value was at or below 12.5 kg (27.5 lb). That implies that when the LI is set to 1.0 for task design or evaluation purposes, only 0.5% of the (simulated) industrial lifting tasks would have the RWLs greater than 12.5 kg. Taking into account the lifting task duration, in the 99.5% of the simulated cases, the RWL values were equal to or were lower than 13.0 kg (28.6 lb) for up to one hour of lifting task exposure, 12.5 kg (or 27.5 lb) for less than two hours of exposure, and 10.5 kg (23.1 lb) for lifting over an eight-hour shift.

From a practical point of view, these values define simple and straightforward lifting limits, that is, the threshold RWL values (TRWL) that can be used by practitioners for the purpose of immediate and easy-to-perform risk assessment of manual lifting tasks performed in industry. Because the 1991 equation is designed to ensure that the RWL will not exceed the acceptable lifting capability of 99% of male workers and 75% of female workers, this amounts to protecting about 90% of the industrial workers if there is a 50/50 split between males and females. The TRWL value of 27.5 lb can then be used for immediate risk assessment of manual lifting tasks performed in industry. If this value is exceeded, then a more thorough examination of the identified tasks, as well as evaluation of physical capacity of the exposed workers, should be performed.

7.10. Prevention of LBDs in Industry

The application of ergonomic principles to the design of MMH tasks is one of the most effective approaches to controlling the incidence and severity of LBDs (Ayoub et al. 1997). The goal of ergonomic job design is to reduce the ratio of task demands to worker capability to an acceptable level (see Figure 14). The application of ergonomic principles to task and workplace design permanently reduces stresses. Such changes are preferable to altering other aspects of the MMH system, such as work practices. For example, worker training may be ineffective if practices trained are not reinforced and refreshed (Kroemer 1992), whereas altering the workplace is a lasting physical intervention

7.10.1. Job Severity Index

The job severity index (JSI) is a time- and frequency-weighted ratio of worker capacity to job demands. Worker capacity is predicted with the models developed by Ayoub et al. (1978), which use isometric strength and anthropometric data to predict psychophysical lifting capacity. JSI and each of the components are defined below.

$$\mathrm{JSI} = \sum_{i=1}^{n} \frac{\mathrm{hours}_i \times \mathrm{days}_i}{\mathrm{hours}_t \times \mathrm{days}_t} \sum_{j=1}^{m_i} \left[\frac{F_j}{F_i} \times \frac{\mathrm{WT}_j}{\mathrm{CAP}_j} \right]$$

where: n =number of task groups

hours_i = exposure hours/day for group i
days_i = exposure days/week for group i
hours_t = total hours/day for job
days_t = total days/week for job
m_i = number of tasks in group i
WT_j = maximum required weight of lift for task j

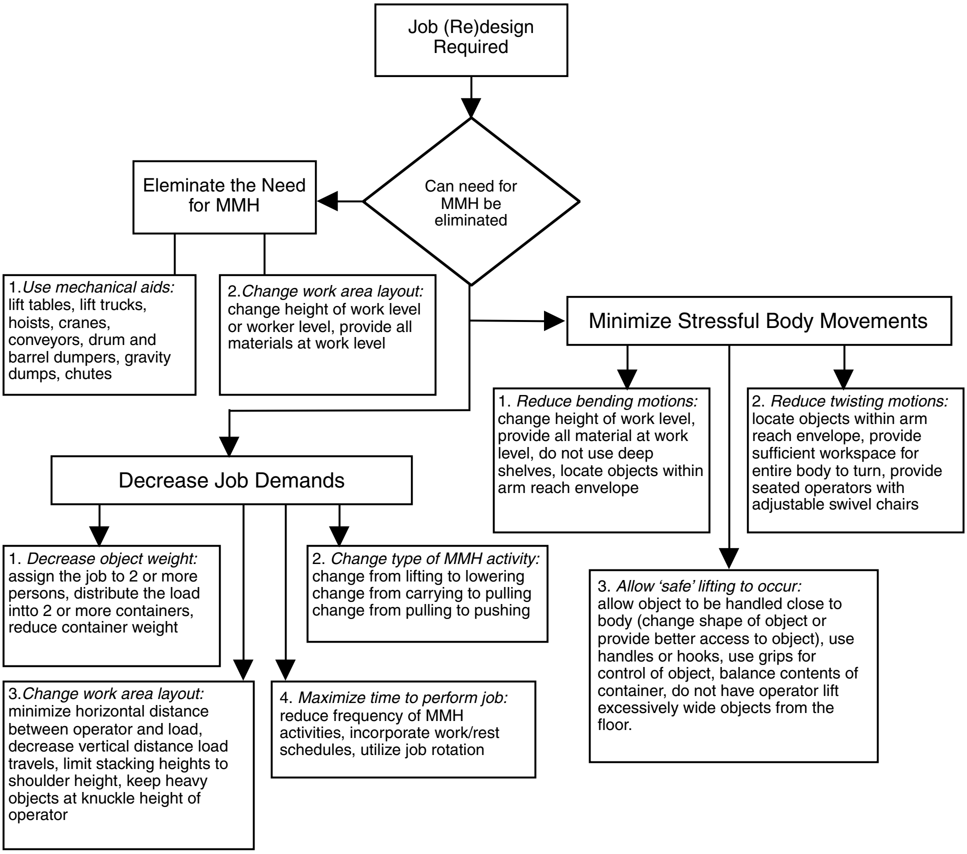

Job (Re)design Required
Can need for MMH be eliminated
Eleminate the Need for MMH
1.*Use mechanical aids:* lift tables, lift trucks, hoists, cranes, conveyors, drum and barrel dumpers, gravity dumps, chutes
2.*Change work area layout:* change height of work level or worker level, provide all materials at work level
Minimize Stressful Body Movements
1. *Reduce bending motions:* change height of work level, provide all material at work level, do not use deep shelves, locate objects within arm reach envelope
2. *Reduce twisting motions:* locate objects within arm reach envelope, provide sufficient workspace for entire body to turn, provide seated operators with adjustable swivel chairs
3. *Allow 'safe' lifting to occur:* allow object to be handled close to body (change shape of object or provide better access to object), use handles or hooks, use grips for control of object, balance contents of container, do not have operator lift excessively wide objects from the floor.
Decrease Job Demands
1. *Decrease object weight:* assign the job to 2 or more persons, distribute the load intto 2 or more containers, reduce container weight
2. *Change type of MMH activity:* change from lifting to lowering change from carrying to pulling change from pulling to pushing
3.*Change work area layout:* minimize horizontal distance between operator and load, decrease vertical distance load travels, limit stacking heights to shoulder height, keep heavy objects at knuckle height of operator
4. *Maximize time to perform job:* reduce frequency of MMH activities, incorporate work/rest schedules, utilize job rotation

Figure 14 Summary of Ergonomic Approaches to MMH Task (re)design. (Adapted from Ayoub 1982 and Ayoub et al. 1983)

CAP_j = the adjusted capacity of the person working at task j
F_j = lifting frequency for task j
F_i = total lifting frequency for group i

$$= \sum_{j=1}^{m_i} F_j$$

Liles et al. (1984) performed a field study to determine the relationship between JSI and the incidence and severity of LBDs. A total of 453 subjects was included in the study. The results of the field study indicated that both incidence and severity of recordable back injuries rose rapidly at values of JSI greater than 1.5. The denominator for the incidence and severity rates is 100 full-time employees, that is, 200,000 exposure hours. JSI can be reduced to a desirable level by increasing worker capacity (e.g., selecting a worker with higher capacity) or altering task and job parameters to reduce JSI to an acceptable level.

7.10.2. Dynamic Model for Prediction of LBDs in Industry

Marras et al. (1993) performed a retrospective study to determine the relationships between workplace factors and trunk motion factors and LBD occurrence. A logistic regression analysis was performed to provide a model used to estimate the probability of high-risk LBD membership. High-risk jobs were defined as jobs having incidence rates of 12 or more injuries per 200,000 hours of exposure. The regressors included in the model were lift rate (lifts/hr), average twisting velocity (deg/sec), maximum moment (Nm), maximum sagittal flexion (degrees), and maximum lateral velocity (deg/sec). The above model can be used to guide workplace design changes because the probability of high-risk LBD membership can be computed before and after design changes. For example, maximum moment could be reduced by decreasing the load weight or the maximum horizontal distance between the load and the lumbar spine, and the associated decrease in high-risk membership probability can be estimated. The model is considerably different from the models discussed above in that LBD risk is not assumed to be related to individual capacity.

8. WORK-RELATED MUSCULOSKELETAL DISORDERS OF THE UPPER EXTREMITY

8.1. Characteristics of Musculoskeletal Disorders

The National Institute of Occupational Safety and Health (NIOSH 1997) states that musculoskeletal disorders, which include disorders of the back, trunk, upper extremity, neck, and lower extremity are one of the 10 leading work-related illnesses and injuries in the United States. Praemer et al, (1992) report that work-related upper-extremity disorders (WUEDs), which are formally defined by the Bureau of Labor Statistics (BLS) as cumulative trauma illnesses, account for 11.0 % of all work-related musculoskeletal disorders (illnesses). For comparison, occupational low-back disorders account for more than 51.0% of all WRMDs. According to BLS (1995), the cumulative trauma illnesses of upper extremity accounted for more than 60% of the occupational illnesses reported in 1993. These work-related illnesses, which include hearing impairments due to occupational noise exposure, represent 6.0% of all reportable work-related injuries and illnesses (Marras 1996).

As reviewed by Karwowski and Marras (1997), work-related musculoskeletal disorders currently account for one-third of all occupational injuries and illnesses reported to the Bureau of Labor Statistics (BLS) by employers every year. These disorders thus constitute the largest job-related injury and illness problem in the United States today. According to OSHA (1999), in 1997 employers reported a total of 626,000 lost workday disorders to the BLS, and these disorders accounted for $1 of every $3 spent for workers' compensation in that year. Employers pay more than $15–20 billion in workers' compensation costs for these disorders every year, and other expenses associated with MSDs may increase this total to $45–54 billion a year.

Such statistics can be linked to several occupational risk factors, including the increased production rates leading to thousands of repetitive movements every day, widespread use of computer keyboards, higher percentage of women and older workers in the workforce, better record keeping of reportable illnesses and injuries on the job by employers, greater employee awareness of WUEDs and their relation to the working conditions, and a marked shift in social policy regarding recognition and compensation of the occupational injuries and ilnesses.

8.2. Definitions

Work-related musculoskeletal disorders (WRMDs) are those disorders and diseases of the musculoskeletal system which have a proven or hypothetical work related causal component (Kuorinka and Forcier 1995). Musculoskeletal disorders are pathological entities in which the functions of the musculoskeletal system are disturbed or abnormal, while diseases are pathological entities with observable impairments in body configuration and function. Although WUEDs are a heterogeneous group of disorders, and the current state of knowledge does not allow for a general description of the course of these disorders, it is possible nevertheless to identify a group of so-called generic risk factors, including biomechanical factors, such as static and dynamic loading on the body and posture, cognitive demands, and organizational and psychosocial factors, for which there is an ample evidence of work-relatedness and higher risk of developing the WUEDs.

Generic risk factors, which typically interact and cumulate to form cascading cycles, are assumed to be directly responsible for the pathophysiological phenonmena that depend on location, intensity, temporal variation, duration, and repetitivness of the generic risk factors (Kuorinka and Forcier 1995). It is also proposed that both insufficient and exccessive loading on the musculoskeletal system have deleterious effects and that the pathophysiological process is dependent upon individual characteristics with respect to body responses, coping mechanisms, and adaptation to risk factors.

Musculoskeletal disorders can be defined by combining the separate meanings for each word (Putz-Anderson 1993). *Cumulative* indicates that these disorders develop gradually over periods of time as a result of repeated stresses. The cumulative concept is based on the assumption that each repetition of an activity produces some trauma or wear and tear on the tissues and joints of the particular body part. The term *trauma* indicates bodily injury from mechanical stresses, while *disorders* refer to physical ailments. The above definition also stipulates a simple cause-and-effect model for CTD development. According to such a model, because the human body needs sufficient intervals of rest time between episodes of repeated strains to repair itself, if the recovery time is insufficient, combined with high repetition of forceful and awkward postures, the worker is at higher risk of developing a CTD. In the context of the generic model for prevention shown in Figure 15, the above

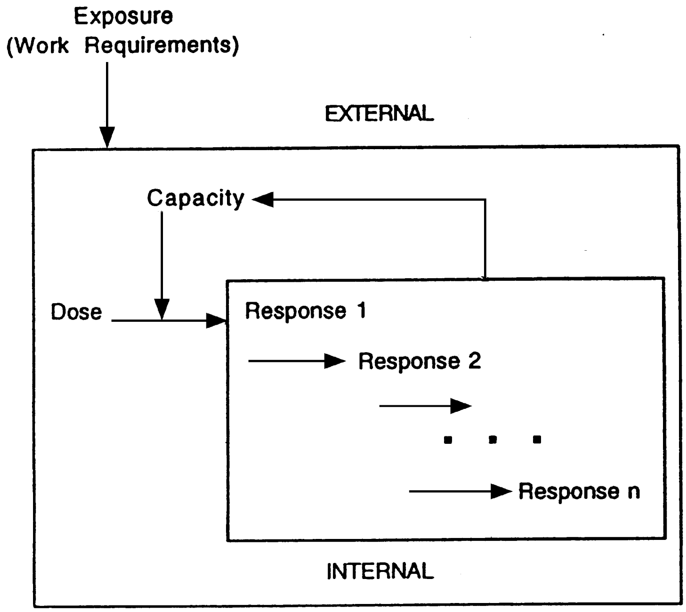


Figure 15 A Conceptual Model for Development of WMSDs proposed by Armstrong et al. (1993). Reproduced with permission from Finnish Institute of Occupational Health.

definition is primarily oriented towards biomechanical risk factors for WUEDs and is, therefore, incomplete. Table 21 presents a summary of the potential risk factors for work-related musculoskeletal disorders.

8.3. Conceptual Models for Development of WRMDs

According to the World Health Organization (WHO 1985), an occupational disease is a disease for which there is a direct cause-and-effect relationship between hazard and disease (e.g., asbestos–

TABLE 21 Potential Risk Factors for Development of Work-Related Musculoskeletal Disorders

1. Physical strength requirements
2. Biomechanical stressors (dynamic and static)
3. Endurance requirement and physiological costs of work
4. Motion factors:
    a. repetitive movement rates
    b. reach distances (functional, extended, and repetitive)
    c. motion times and efficiency
5. Postural factors:
    a. characteristics and range of motion
    b. joint deviations
    c. static loading
    d. awkward postures
6. Work duration factors:
    a. rate/work/rest ratios
    b. stressful task assignments
7. Work organization demands:
    a. work pace/time pressures
    b. machine/team pacing
    c. overtime demands
    d. monotony of work

asbestosis). Work-related diseases are defined as multifactorial when the work environment and the performance of work contribute significantly to the causation of disease (WHO 1985). Work-related diseases can be partially caused by adverse work conditions. However, personal characteristics, environmental, and sociocultural factors are also recognized as risk factors for these diseases.

The scientific evidence of work-relatedness of musculoskeletal disorders has been firmly established by numerous epidemiologic studies conducted over the last 25 years of research in the field (NIOSH 1997). It has also been noted that the incidence and prevalence of musculoskeletal disorders in the reference populations were low, but not zero, most likely indicating the nonwork-related causes of these disorders. It was also documented that such variables as cultural differences, psychosocial and economic factors, which may influence one's perception and tolerance of pain and consequently affect the willingness to report musculoskeletal problems, may have significant impact on the progressions from disorder to work disability (WHO 1985; Leino 1989).

Armstrong et al. (1993) developed a conceptual model for the pathogenesis of work-related musculoskeletal disorders. The model is based on the set of four cascading and interacting state variables of exposure, dose, capacity, and response, which are measures of the system state at any given time. The response at one level can act as dose at the next level (see Figure 15). Furthermore, it is assumed that a response to one or more doses can diminish or increase the capacity for responding to successive doses. This conceptual model for development of WRMDs reflects the multifactorial nature of work-related upper-extremity disorders and the complex nature of the interactions between exposure, dose, capacity, and response variables. The proposed model also reflects the complexity of interactions among the physiological, mechanical, individual, and psychosocial risk factors.

In the proposed model, exposure refers to the external factors (i.e., work requirements) that produce the internal dose (i.e., tissue loads and metabolic demands and factors). Workplace organization and hand tool design characteristics are examples of such external factors that can determine work postures and define loads on the affected tissues or velocity of muscular contractions. Dose is defined by a set of mechanical, physiological, or psychological factors that in some way disturb an internal state of the affected worker. Mechanical disturbance factors may include tissue forces and deformations produced as a result of exertion or movement of the body.

Physiological disturbances are such factors as consumption of metabolic substrates or tissue damage, while the psychological disturbance factors are those related to, for example, anxiety about work or inadequate social support. Changes in the state variables of the worker are defined by the model as responses. A response is an effect of the dose caused by exposure. For example, hand exertion can cause elastic deformation of tendons and changes in tissue composition and/or shape, which in turn may result in hand discomfort. The dose–response time relationship implies that the effect of a dose can be immediate or the response may be delayed for a long periods of time.

The proposed model stipulates that system changes (responses) can also result in either increased dose tolerance (adaptation) or reduced dose tolerance lowering the system capacity. Capacity is defined as the worker's ability (physical or psychological) to resist system destabilization due to various doses. While capacity can be reduced or enhanced by previous doses and responses, it is assumed that most individuals are able to adapt to certain types and levels of physical activity.

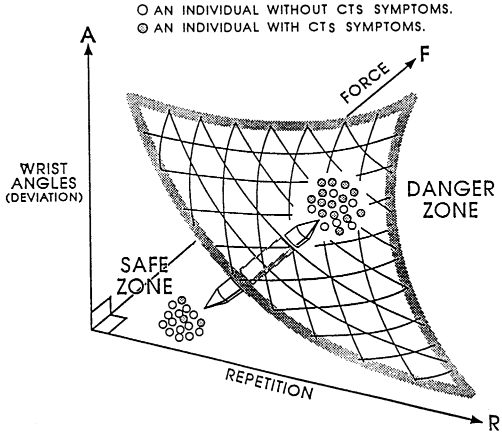

○ AN INDIVIDUAL WITHOUT CTS SYMPTOMS.
◎ AN INDIVIDUAL WITH CTS SYMPTOMS.
A
WRIST ANGLES (DEVIATION)
FORCE
F
DANGER ZONE
SAFE ZONE
REPETITION
R

Figure 16 Conceptual CTS Model. (Adapted from Tanaka and McGlothlin 1999, reprinted with permission from Elsevier Science.)

Muscles, for example, can develop increased aerobic or anaerobic metabolic capacity. Furthermore, muscular responses are characterized in the model as a series of cascading mechanical and physiological events. The local changes (system responses), such as deformation and the yielding of connective tissues within the muscle, are conveyed to the central nervous system by sensory afferent nerves and cause corresponding sensations to effort and discomfort, often referred to as perceived fatigue.

The main purpose of the dose-response model is to account for the factors and processes that result in WRMDs in order to specify acceptable limits with respect to work design parameters for a given individual. The proposed model should be useful in the design of studies on the etiology and pathomechanisms of work-related musculoskeletal disorders, as well as in the planning and evaluation of preventive programs. The model should complement the epidemiologic studies, which focus on associations between the top and bottom of the cascade with physical workload, psychological demands, and environmental risk factors of work at one end and the manifestations of symptoms, diseases, or disabilities at the other.

Recently, Tanaka and McGlothlin (1999) updated their 3D heuristic dose–response model for repetitive manual work risk factors using the epidemiologic finding. Their earlier model for the postulated relationships between the risk factors for carpal tunnel syndrome (CTS) (for description see Karwowski and Marras 1997) was modified by including the time exposure factor. This was based on examination of prevalence of CTS data for 1988 from the National Health Review Survey (NHIS) and the Occupational Health Supplement (OHS). The authors found that compared to the nonexposed population, the prevalence (P) of CTS among the people exposed to bending/twisting of the hands/wrists many times an hour increased by several times regardless of the length of daily hours exposed. The prevalence of CTS was then defined as follows:

$$P = k \times I \times T = k \{aF \times bR \times e^{cA}\} T$$

where: P = prevalence of CTS
I = intensity
K = constant
a, b, c = coefficients
F = force
R = repetition
A = joint angles
T = time duration

The proposed model indicates that high-intensity work (involving high force, high repetition, and/or high joint deviation) should not be performed for a long period of time, but low-intensity work may be performed for a longer period. The 3D representation of the relationships between the risk factors for CTS is illustrated in Figures 16 and 17.

8.4. Causal Mechanism for Development of WUEDs

As reviewed by Armstrong et al. (1993), work-related muscle disorders are likely to occur when a muscle is fatigued repeatedly without sufficient allowance for recovery. An important factor in development of such disorders is motor control of the working muscle. Hägg (1991) postulates that the recruitment pattern of the motor neurons can occur according to the size principle, where the small units are activated at low forces. Given that the same units can be recruited continuously during a given work task, even if the relative load on the muscle is low, the active low-threshold motor units can work close to their maximal capacity and consequently maybe at a high risk of being damaged. It has also been shown that muscle tension due to excessive mental load can cause an overload on some specific muscle fibers (Westgaard and Bjørkland 1987). Karwowski et al. (1994) showed that cognitive aspects of computer-related task design affect the postural dynamics of the operators and the related levels of perceived postural discomfort. Finally, Edwards (1988) hypothesizes that occupational mucle pain might be a consequence of a conflict between motor control of the postural activity and control needed for rhythmic movement or skilled manipulations. In other words, the primary cause of work-related muscular pain and injury may be altered motor control, resulting in imbalance between harmonious motor unit recruitment relaxation of mucles not directly involved in the activity.

As discussed by Armstrong et al. (1993), poor ergonomic design of tools with respect to weight, shape, and size can impose extreme wrist positions and high forces on the worker's musculoskeletal system. Holding heavier objects requires an increased power grip and high tension in the finger flexor tendons, causing increased pressure in the carpal tunnel. Furthermore, the tasks that induce hand and arm vibration cause an involuntary increase in power grip through a reflex of the strength receptors. Vibration can also cause protein leakage from the blood vessels in the nerve trunks and result in

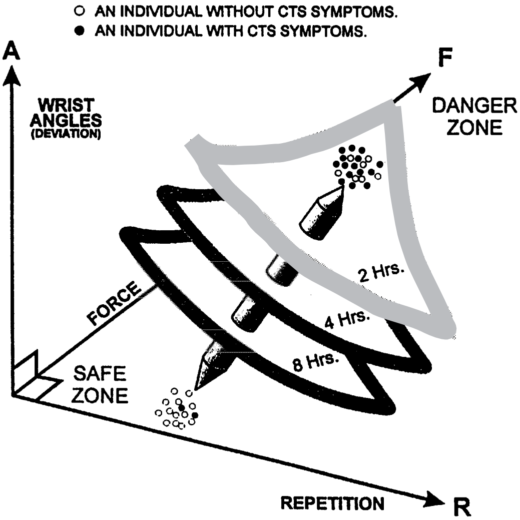


Figure 17 Three-Dimensional Illustration of the conceptual CTS Model with Time Exposure Factor. (Adapted from Tanaka and McGlothlin 1999, reprinted with permission from Elsevier Science.)

TABLE 22 Relationship between Physical Stresses and WRMD Risk Factors (ANSI Z-365)

Physical Stress	Magnitude	Repetition rate	Duration
Force	Forceful exertions and motions	Repetitive exertions	Sustained exertions
Joint angle	Extreme postures and motions	Repetitive motions	Sustained postures
Recovery	Insufficient resting level	Insufficient pauses or breaks	Insufficient rest time
Vibration	High vibration level	Repeated vibration exposure	Long vibration exposure
Temperature	Cold temperature	Repeated cold exposure	Long cold exposure

edema and increased pressure in the nerve trunks and therefore also result in edema and increased pressure in the nerve (Lundborg et al. 1987).

8.5. Musculoskeletal Disorders: Occupational Risk Factors

A risk factor is defined as an attribute or exposure that increases the probability of a disease or disorder (Putz-Anderson, 1988). Biomechanical risk factors for musculoskeletal disorders include repetitive and sustained exertions, awkward postures, and application of high mechanical forces. Vibration and cold environments may also accelerate the development of musculoskeletal disorders. Typical tools that can be used to identify the potential for development of musculoskeletal disorders include conducting work-methods analyses and checklists designed to itemize undesirable work site conditions or worker activities that contribute to injury. Since most of manual work requires the active use of the arms and hands, the structures of the upper extremities are particularly vulnerable to soft tissue injury. WUEDs are typically associated with repetitive manual tasks with forceful exertions, such as those performed at assembly lines, or when using hand tools, computer keyboards and other devices, or operating machinery. These tasks impose repeated stresses to the upper body, that is, the muscles, tendons, ligaments, nerve tissues, and neurovascular structures. There are three basic types of WRDs to the upper extremity: tendon disorder (such as tendonitis), nerve disorder (such as carpal tunnel syndrome), and neurovascular disorder (such as thoracic outlet syndrome or vibration–Raynaud's syndrome). The main biomechanical risk factors of musculoskeletal disorders are presented in Table 22.

9. ERGONOMICS DESIGN TO REDUCE WUEDs

In order to reduce the extent of work-related musculoskeletal injuries, progress in four methodologic areas is expected (NIOSH 1986):

1. Identifying accurately the biomechanical hazards
2. Developing effective health-promotion and hazard-control interventions
3. Changing management concepts and operational policies with respect to expected work performance
4. Devising strategies for disseminating knowledge on control technology and promoting their application through incentives

From the occupational safety and health perspective, the current state of ergonomics knowledge allows for management of musculoskeletal disorders in order to minimize human suffering, potential for disability, and the related workers' compensation costs. Ergonomics can help to:

1. Identify working conditions under which musculoskeletal disorders might occur
2. Develop engineering design measures aimed at elimination or reduction of the known job risk factors
3. Identify the affected worker population and target it for early medical and work intervention efforts

The musculoskeletal disorders-related job risk factors, which often overlap, typically involve a combination of poorly designed work methods, workstations, and hand tools and high production demands. Furthermore, while perfect solutions are rarely available, the job redesign decisions may often require some design trade-offs (Putz-Anderson 1992). In view of the above, the ergonomic intervention should allow:

1. Performing a thorough job analysis to determine the nature of specific problems
2. Evaluating and selecting the most appropriate intervention(s)
3. Developing and applying conservative treatment (implementing the intervention), on a limited scale if possible
4. Monitoring progress
5. Adjust or refining the intervention as needed

9.1. Quantitative Models for Development of WUEDs

It is generally recognized that force, repetition, posture, recovery time, duration of exposure, static muscular work, use of the hand as a tool, and type of grasp are important factors in the causation of WUEDs (Armstrong et al. 1987; Keyserling et al. 1993). Additional job factors that may increase the risk of WUEDs, in combination with the other factors, include cold temperature, use of gloves, and use of vibrating tools. Given the above knowledge, even if limited and in need of more comprehensive validation, it is currently possible to develop quantitative methodologies for ergonomics practitioners in order to discriminate between safe and hazardous jobs in terms of workers being at increased risk of developing the WUEDs. Such models are described below.

9.2. Semiquantitative Job-Analysis Methodology for Wrist/Hand Disorders

Moore and Garg (1995) developed a semiquantitative job analysis methodology (SJAM) for identifying industrial jobs associated with distal upper-extremity (wrist/hand) disorders. An existing body of knowledge and theory of the physiology, biomechanics, and epidemiology of distal upper-extremity disorders was used for that purpose. The proposed methodology involves the measurement or estimation of six task variables:

1. Intensity of exertion
2. Duration of exertion per cycle
3. Efforts per minute
4. Wrist posture
5. Speed of exertion
6. Duration of task per day

An ordinal rating is assigned for each of the variables according to the exposure data. The proposed strain index is the product of these six multipliers assigned to each of the variables.

The strain index methodology aims to discriminate between jobs that expose workers to risk factors (task variables) that cause WUEDs and jobs that do not. However, the strain index is not designed to identify jobs associated with an increased risk of any single specific disorder. It is anticipated that jobs identified as in the high-risk category by the strain index will exhibit higher levels of WUEDs among workers who currently perform or historically performed those jobs that are believed to be hazardous. Large-scale studies are needed to validate and update the proposed methodology. The strain index has the following limitations in terms of its application:

1. There are some disorders of the distal upper extremity that should not be predicted by the strain index, such as hand–arm vibration syndrome (HAVS) and hypothenar hammer syndrome.
2. The strain index has not been developed to predict increased risk for distal upper-extremity disorders to uncertain etiology or relationship to work. Examples include ganglion cysts, osteoarthritis, avascular necrosis of carpal bones, and ulnar nerve entrapment at the elbow.
3. The strain index has not been developed to predict disorders outside of the distal upper extremity, such as disorders of the shoulder, shoulder girdle, neck, or back.

The following major principles have been derived from the physiological model of localized muscle fatigue:

1. The primary task variables are intensity of exertion, duration of exertion, and duration of recovery.
2. Intensity of exertion refers to the force required to perform a task one time. It is characterized as a percentage of maximal strength.
3. Duration of exertion describes how long an exertion is applied. The sum of duration of exertion and duration of recovery is the cycle time of one exertional cycle.
4. Wrist posture, type of grasp, and speed of work are considered via their effects of maximal strength.
5. The relationship between strain on the body (endurance time) and intensity of exertion is nonlinear.

The following are the major principles derived from the epidemiological literature:

1. The primary task variable associated with an increased prevalence or incidence of distal upper-extremity disorders are intensity of exertion (force), repetition rate, and percentage of recovery time per cycle.
2. Intensity of exertion was the most important task variable in two of the three studies explicitly mentioned. The majority (or all) of the morbidity was related to disorders of the muscle–tendon unit. The third study, which considered only CTS, found that repetition was more important than forcefulness (Silverstein et al. 1987).
3. Wrist posture may not be an independent risk factor. It may contribute to an increased incidence of distal upper-extremity disorders when combined with intensity of exertion.
4. The roles of other task variables have not been clearly established epidemiologically; therefore, one has to rely on biomechanical and physiological principles to explain their relationship to upper-extremity disorders, if any.

Moore and Garg (1994) compared exposure factors for jobs associated with WUEDs to jobs without prevalence of such disorders. They found that the intensity of exertion, estimated as a percentage of maximal strength and adjusted for wrist posture and speed of work, was the major discriminating factor. The relationship between the incidence rate for distal upper-extremity disorder and the job risk factors was defined as follows:

$$\mathrm{IE} = \frac{30 \times F^2}{\mathrm{RT}^{0.6}}$$

where: IR = incidence rate (per 100 workers per year)
F = intensity of exertion (%MS)
RT = recovery time (percentage of cycle time)

The proposed concept of the strain index is a semiquantitative job analysis methodology that results in a numerical score that is believed to correlate with the risk of developing distal upper-extremity disorders. The SI score represents the product of six multipliers that correspond to six task variables. These variables:

TABLE 23 Rating Criteria for Strain Index

Rating	Intensity of Exertion	Duration of Exertion (% of Cycle)	Efforts/Minute	Hand–Wrist Posture	Speed of Work	Duration per Day (h)
1	Light	<10	<4	Very good	Very slow	≥ 1
2	Somewhat hard	10–29	4–8	Good	Slow	1–2
3	Hard	30–49	9–14	Fair	Fair	2–4
4	Very hard	50–79	15–19	Bad	Fast	2–8
5	Near maximal	≤ 80	≤ 20	Very bad	Very fast	≤ 8

Adapted from Moore and Garg 1995.

1. Intensity of exertion
2. Duration of exertion
3. Exertions per minute
4. Hand–wrist posture
5. Speed of work
6. Duration of task per day

These ratings, applied to model variables, are presented in Table 23. The multipliers for each task variable related to these ratings are shown in Table 24. The strain index score as the product of all six multipliers is defined as follows:

$$\begin{aligned}\text{Strain index (SI)} &= (\text{intensity of exertion multiplier})\\ &\times (\text{duration of exertion multiplier})\\ &\times (\text{exertions per minute multiplier})\\ &\times (\text{posture multiplier}) \times (\text{speed of work multiplier})\\ &\times (\text{duration per day multiplier})\end{aligned}$$

Intensity of exertion, the most critical variable of SI, is an estimate of the force requirements of a task and is defined as the percentage of maximum strength required to perform the task once. As such, the intensity of exertion is related to physiological stress (percentage of maximal strength) and biomechanical stresses (tensile load) on the muscle–tendon units of the distal upper extremity. The intensity of exertion is estimated by an observer using verbal descriptors and assigned corresponding rating values (1, 2, 3, 4, or 5). The multiplier values are defined based on the rating score raised to a power of 1.6 in order to reflect the nonlinear nature of the relationship between intensity of exertion and manifestations of strain according to the psychophysical theory. The multipliers for other task variables are modifiers to the intensity of exertion multiplier.

Duration of exertion is defined as the percentage of time an exertion is applied per cycle. The terms cycle and cycle time refer to the exertional cycle and average exertional cycle time, respectively. Duration of recovery per cycle is equal to the exertional cycle time minus the duration of exertion per cycle. The duration of exertion is the average duration of exertion per exertional cycle (calculated by dividing all durations of a series of exertions by the number of observed exertions). The percentage

TABLE 24 Multiplier Table for Strain Index

Rating	Intensity of Exertion	Duration of Exertion (% of Cycle)	Efforts/Minute	Hand–Wrist Posture	Speed of Work	Duration per Day (h)
1	1	0.5	0.5	1.0	1.0	0.25
2	3	1.0	1.0	1.0	1.0	0.50
3	6	1.5	1.5	1.5	1.0	0.75
4	9	2.0	2.0	2.0	1.5	1.00
5	13	3.0[a]	3.0	3.0	2.0	1.50

Adapted from Moore and Garg 1995.

[a] If duration of exertion is 100%, then the efforts/minute multiplier should be set to 3.0.

TABLE 25 An Example to Demonstrate the Procedure for Calculating SI Score

	Intensity of Exertion	Duration of Exertion (% of Cycle)	Efforts/Minute	Posture	Speed of Work	Duration per Day (h)
Exposure dose	Somewhat hard	60%	12	Fair	Fair	4–8
Ratings	2	4	3	3	3	4
Multiplier	3.0	2.0	1.5	1.5	1.0	1.0
SI Score = 3.0 × 2.0 × 1.5 × 1.5 × 1.0 × 1.0 = 13.5						

Adapted from Moore and Garg 1995.

duration of exertion is calculated by dividing the average duration of exertion per cycle by the average exertional cycle time, then multiplying the result by 100. (See equation below.) The calculated percentage duration of exertion is compared to the ranges and assigned the appropriate rating. The corresponding multipliers are identified using Table 23.

$$\%\text{duration of exertion} = \frac{(\text{average duration of exertion per cycle})}{(\text{average exertional cycle time})}$$

Efforts per minute is the number of exertions per minute (i.e., repetitiveness) and is synonymous with frequency. Efforts per minute are measured by counting the number of exertions that occur during a representative observation period (as described for determining the average exertional cycle time). The measured results are compared to the ranges shown in Table 23 and given the corresponding ratings. The multipliers are defined in Table 24.

Posture refers to the anatomical position of the wrist or hand relative to neutral position and is rated qualitatively using verbal anchors. As shown in Table 23, posture has four relevant ratings. Postures that are "very good" or "good" are essentially neutral and have multipliers of 1.0. Hand or wrist postures progressively deviate beyond the neutral range to extremes, graded as "fair," "bad," and "very bad."

Speed of work estimates perceived pace of the task or job and is subjectively estimated by a job analyst or ergonomics team. Once a verbal anchor is selected, a rating is assigned.

Duration of task per day is defined as a total time that a task is performed per day. As such, this variable reflects the beneficial effects of task diversity such as job rotation and the adverse effects of prolonged activity such as overtime. Duration of task per day is measured in hours and assigned a rating according to Table 23.

Application of the strain index involves five steps:

1. Collecting data
2. Assigning rating values
3. Determining multipliers
4. Calculating the SI score
5. Interpreting the results

TABLE 26 Maximum Acceptable Forces for Female Wrist Flexion (Power Grip) (N)

Percentage of Population	Repetition Rate				
	2/min	5/min	10/min	15/min	20/min
90	14.9	14.9	13.5	12.0	10.2
75	23.2	23.2	20.9	18.6	15.8
50	32.3	32.3	29.0	26.0	22.1
25	41.5	41.5	37.2	33.5	28.4
10	49.8	49.8	44.6	40.1	34.0

Adapted with permission from Snook et al. 1995. Copyright © by Taylor & Francis Ltd., London, http://www.tandf.co.uk.

TABLE 27 Maximum Acceptable Forces for Female Wrist Flexion (Pinch Grip) (N)

Percentage of Population	Repetition Rate				
	2/min	5/min	10/min	15/min	20/min
90	9.2	8.5	7.4	7.4	6.0
75	14.2	13.2	11.5	11.5	9.3
50	19.8	18.4	16.0	16.0	12.9
25	25.4	23.6	20.6	20.6	16.6
10	30.5	28.3	24.6	24.6	19.8

Adapted with permission from Snook et al. 1995. Copyright © by Taylor & Francis Ltd., London, http://www.tandf.co.uk.

The values of intensity of exertion, hand–wrist posture, and speed of work are estimated using the verbal descriptors in Table 23. The values of percentage duration of exertion per cycle, efforts per minute, and duration per day are based on measurements and counts. Theses values are then compared to the appropriate column in Table 24 and assigned a rating. The calculations of SI are shown in Table 25.

9.3. Psychophysical Models: The Maximum Acceptable Wrist Torque

Snook et al. (1995) used the psychophysical approach to determine the maximum acceptable forces for various types and frequencies for repetitive wrist motion, grips, and repetition rates that would not result in significant changes in wrist strength, tactile sensitivity, or number of symptoms reported by the female subjects. Three levels of wrist motion were used:

1. Flexion motion with a power grip
2. Flexion motion with a pinch grip
3. Extension motion with a power grip

The dependent variables were maximum acceptable wrist torque, maximum isometric wrist strength, tactile sensitivity, and symptoms. The maximum acceptable wrist torque (MAWT) was defined as the number of Newton meters of resistance set in the brake by the participants (averaged and recorded every minute). The data for maximum acceptable wrist torques for the two-days-per-week exposure were used to estimate the maximum acceptable torques for different repetitions of wrist flexion (power grip) and different percentages of the population. This was done by using the adjusted means and coefficients of variation from the two-days-per-week exposure. The original torque values were converted into forces by dividing each torque by the average length of the handle lever (0.081 m).

The estimated values for the maximum acceptable forces for female wrist flexion (power grip) are shown in Table 26. Similarily, the estimated maximum acceptable forces were developed for wrist flexion (pinch grip, see Table 27) and wrist extension (power grip, see Table 28). The torques were converted into forces by dividing by 0.081 m for the power grip and 0.123m for the pinch grip. Snook et al. (1995) note that the estimated values of the maximum acceptable wrist torque do not apply to any other tasks and wrist positions than those that were used in the study.

TABLE 28 Maximum Acceptable Forces for Female Wrist Extension (Power Grip) (N)

Percentage of Population	Repetition Rate				
	2/min	5/min	10/min	15/min	20/min
90	8.8	8.8	7.8	6.9	5.4
75	13.6	13.6	12.1	10.9	8.5
50	18.9	18.9	16.8	15.1	11.9
25	24.2	24.2	21.5	19.3	15.2
10	29.0	29.0	25.8	23.2	18.3

Adapted with permission from Snook et al. 1995. Copyright © by Taylor & Francis Ltd., London, http://www.tandf.co.uk.

10. MANAGEMENT OF MUSCULOSKELETAL DISORDERS

10.1. Ergonomic Guidelines

Most of the current guidelines for control of musculoskeletal disorders at work aim to reduce the extent of movements at the joints, reduce excessive force levels, and reduce the exposure to highly repetitive and stereotyped movements. For example, some of the common methods to control for wrist posture, which is believed one of the risk factors for carpal tunnel syndrome, are altering the geometry of tool or controls (e.g., bending the tool or handle), changing the location/positioning of the part, and changing the position of the worker in relation to the work object. In order to control for the extent of force required to perform a task, one can reduce the force required through tool and fixture redesign, distribute the application of force, or increase the mechanical advantage of the (muscle) lever system.

It has been shown that in the dynamic tasks involving upper extremities, the posture of the hand itself has very little predictive power for the risk of musculoskeletal disorders. Rather, it is the velocity and acceleration of the joint that significantly differentiate the musculoskeletal disorders risk levels (Schoenmarklin and Marras 1990). This is because the tendon force, which is a risk factor of musculoskeletal disorders, is affected by wrist acceleration. The acceleration of the wrist in a dynamic task requires transmission of the forearm forces to the tendons. Some of this force is lost to friction against the ligaments and bones in the carpal tunnel. This frictional force can irritate the tendons' synovial membranes and cause tenosynovitis or carpal tunnel syndrom (CTS). These new research results clearly demonstrate the importance of dynamic components in assessing CTD risk of highly repetitive jobs.

With respect to task repetitivness, it is believed today that jobs with a cycle time of less than 30 seconds and a fundamental cycle that exceeds 50% of the total cycle (exposure) time lead to increased risk of musculoskeletal disorders. Because of neurophysiological needs of the working muscles, adequate rest pauses (determined based on scientific knowledge on the physiology of muscular fatigue and recovery) should be scheduled to provide relief for the most active muscles used on the job. Furthermore, reduction in task repetition can be achieved by, for example, by task enlargement (increasing variety of tasks to perform), increase in the job cycle time, and work mechanization and automation.

The expected benefits of reduced musculoskeletal disorders problems in industry are improved productivity and quality of work products, enhanced safety and health of employees, higher employee morale, and accommodation of people with alternative physical abilities. Strategies for managing musculoskeletal disorders at work should focus on prevention efforts and should include, at the plant level, employee education, ergonomic job redesign, and other early intervention efforts, including engineering design technologies such as workplace reengineering, active and passive surveillance. At the macro-level, management of musculoskeletal disorders should aim to provide adequate occupational health care provisions, legislation, and industry-wide standardization.

10.2. Administrative and Engineering Controls

The recommendations for prevention of musculoskeletal disorders can be classified as either primarily administrative, that is, focusing on personnel solutions, or engineering, that is, focusing on redesigning tools, workstations, and jobs (Putz-Anderson 1988). In general, administrative controls are those actions to be taken by the management that limit the potentially harmful effects of a physically stressful job on individual workers. Administrative controls, which are focused on workers, refer to modification of existing personnel functions such as worker training, job rotation, and matching employees to job assignments.

Workplace design to prevent repetitive strain injury should be directed toward fulfilling the following recommendations:

1. Permit several different working postures.
2. Place controls, tools, and materials between waist and shoulder height for ease of reach and operation.
3. Use jigs and fixtures for holding purposes.
4. Resequence jobs to reduce the repetition.
5. Automate highly repetitive operations.
6. Allow self-pacing of work whenever feasible.
7. Allow frequent (voluntary and mandatory) rest breaks.

The following guidelines should be followed (for details see Putz-Anderson 1988):

1. Make sure the center of gravity of the tool is located close to the body and the tool is balanced.
2. Use power tools to reduce the force and repetition required.
3. Redesign the straight tool handle; bend it as necessary to preserve the neutral posture of the wrist.
4. Use tools with pistol grips and straight grips, respectively, where the tool axis in use is horizontal and vertical (or when the direction of force is perpendicular to the workplace).
5. Avoid tools that require working with the flexed wrist and extended arm at the same time or call for the flexion of distal phalanges (last joints) of the fingers.
6. Minimize the tool weight; suspend all tools heavier than 20 N (or 2 kg of force) by a counterbalancing harness.
7. Align the tool's center of gravity with the center of the grasping hand.
8. Use special-purpose tools that facilitate fitting the task to the worker (avoid standard off the-shelf tools for specific repetitive operations).
9. Design tools so that workers can use them with either hand.
10. Use power grip where power is needed and precision grip for precise tasks.
11. The handles and grips should be cylindrical or oval with a diameter of 3.0–4.5 cm (for precise operations the recommended diameter is 0.5–1.2 cm).
12. The minimum handle diameter should be 10.0 cm, and 11.5–12.0 cm is preferable.
13. A handle span of 5.0–6.7 cm can be used by male and female workers.
14. Triggers on power tools should be at least 5.1 cm wide, allowing their activation by two or three fingers.
15. Avoid form-fitting handles that cannot be easily adjusted.
16. Provide handles that are nonporous, nonslip and nonconductive (thermally and electrically).

11. JOB ANALYSIS AND DESIGN

According to ANSI Z-365 (1995), job analysis and design serve two common purposes:

1. To identify potential work-related risk factors associated with musculoskeletal disorders after they are reported
2. To assist in identifying work-related factors associated with musculoskeletal disorders before they occur

Detailed job analysis consists of analyzing the job at the element or micro-level. These analyses involve breaking down the job into component actions, measuring and quantifying risk factors, and identifying the problems and conditions contributing to each risk factor. Job surveys, on the other hand, are used for establishing work relatedness, prioritizing jobs for further analysis, or proactive risk factors surveillance. Such survey methods may include facility walk-throughs, worker interviews, risk-factor checklists, and team problem-solving approaches.

11.1. Risk Factors and Definitions

The risk factors are job attributes or exposures that increase probability of the occurrence of work-related musculoskeletal disorders. The WRMD risk factors are present at varying levels for different jobs and tasks. It should be noted that these risk factors are not necessarily causation factors of WRMDs. Also, the mere presence of a risk factor does not necessarily mean that a worker performing a job is at excessive risk of injury. (The relationship between physical stresses and WRMD risk factors is shown in Table 22.) Generally, the greater the exposure to a single risk factor or combination of factors, the greater the risk of a WRMD. Furthermore, the more risk factors that are present, the higher the risk of injury. According to ANSI Z-365 (1995), this interaction between risk factors may have a multiplicative rather than an additive effect. However, these risk factors may pose minimal risk of injury if sufficient exposure is not present or if sufficient recovery time is provided. It is known that changes in the levels of risk factors will result in changes in the risk of WRMDs. Therefore, a reduction in WRMD risk factors should reduce the risk for WRMDs. Figure 18 shows the flow chart for the ergonomics rule for control of MSDs at the workplace proposed by OSHA (2000).

11.2. Work Organization Risk Factors

The mechanisms by which poor work organization could increase the risk for WUEDs include modifying the extent of exposure to other risk factors (physical and environmental) and modifying the

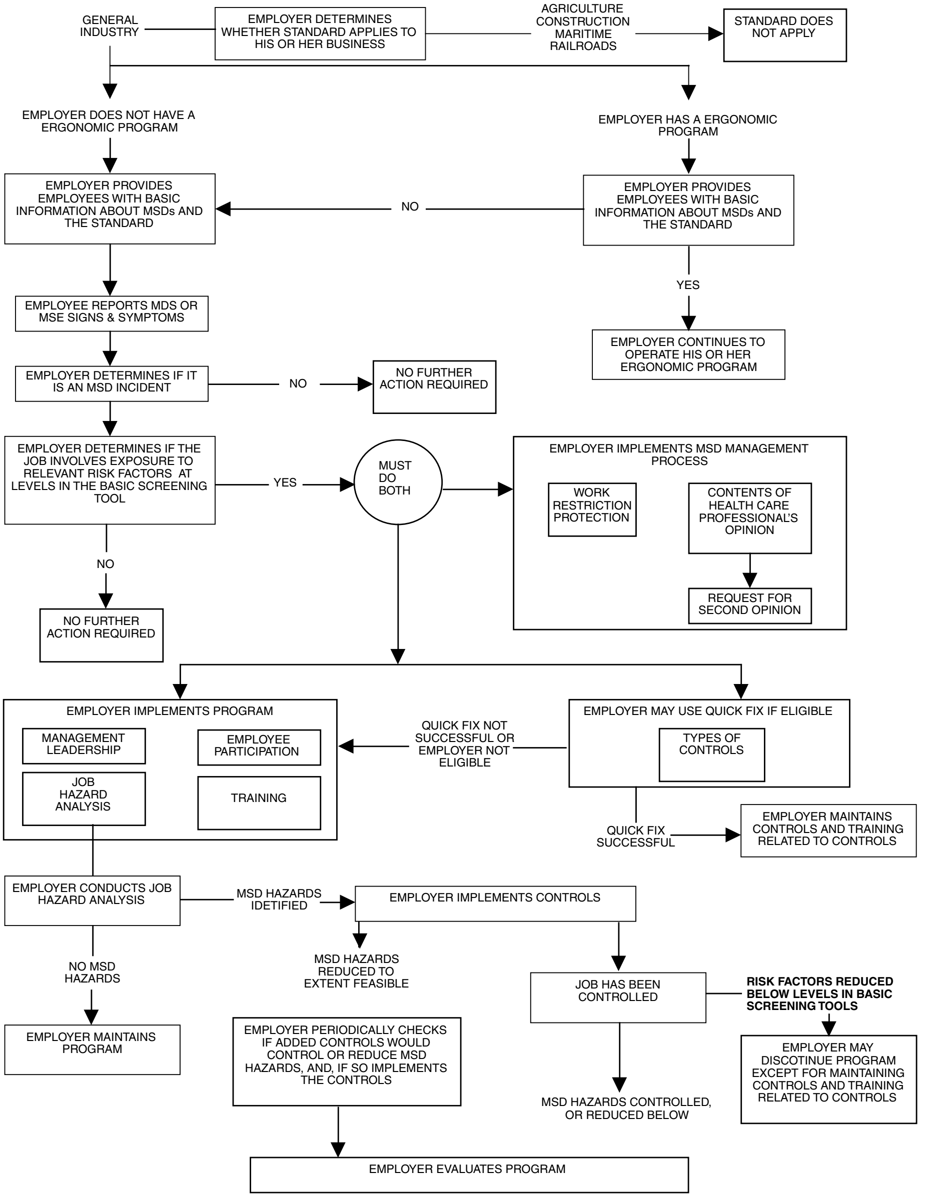

GENERAL INDUSTRY
EMPLOYER DETERMINES WHETHER STANDARD APPLIES TO HIS OR HER BUSINESS
AGRICULTURE CONSTRUCTION MARITIME RAILROADS
STANDARD DOES NOT APPLY
EMPLOYER DOES NOT HAVE A ERGONOMIC PROGRAM
EMPLOYER HAS A ERGONOMIC PROGRAM
EMPLOYER PROVIDES EMPLOYEES WITH BASIC INFORMATION ABOUT MSDs AND THE STANDARD
NO
EMPLOYER PROVIDES EMPLOYEES WITH BASIC INFORMATION ABOUT MSDs AND THE STANDARD
YES
EMPLOYEE REPORTS MDS OR MSE SIGNS & SYMPTOMS
EMPLOYER CONTINUES TO OPERATE HIS OR HER ERGONOMIC PROGRAM
EMPLOYER DETERMINES IF IT IS AN MSD INCIDENT
NO
NO FURTHER ACTION REQUIRED
EMPLOYER DETERMINES IF THE JOB INVOLVES EXPOSURE TO RELEVANT RISK FACTORS AT LEVELS IN THE BASIC SCREENING TOOL
YES
MUST DO BOTH
EMPLOYER IMPLEMENTS MSD MANAGEMENT PROCESS
WORK RESTRICTION PROTECTION
CONTENTS OF HEALTH CARE PROFESSIONAL'S OPINION
REQUEST FOR SECOND OPINION
NO
NO FURTHER ACTION REQUIRED
EMPLOYER IMPLEMENTS PROGRAM
MANAGEMENT LEADERSHIP
EMPLOYEE PARTICIPATION
JOB HAZARD ANALYSIS
TRAINING
QUICK FIX NOT SUCCESSFUL OR EMPLOYER NOT ELIGIBLE
EMPLOYER MAY USE QUICK FIX IF ELIGIBLE
TYPES OF CONTROLS
QUICK FIX SUCCESSFUL
EMPLOYER MAINTAINS CONTROLS AND TRAINING RELATED TO CONTROLS
EMPLOYER CONDUCTS JOB HAZARD ANALYSIS
MSD HAZARDS IDETIFIED
EMPLOYER IMPLEMENTS CONTROLS
NO MSD HAZARDS
MSD HAZARDS REDUCED TO EXTENT FEASIBLE
JOB HAS BEEN CONTROLLED
RISK FACTORS REDUCED BELOW LEVELS IN BASIC SCREENING TOOLS
EMPLOYER MAINTAINS PROGRAM
EMPLOYER PERIODICALLY CHECKS IF ADDED CONTROLS WOULD CONTROL OR REDUCE MSD HAZARDS, AND, IF SO IMPLEMENTS THE CONTROLS
EMPLOYER MAY DISCOTINUE PROGRAM EXCEPT FOR MAINTAINING CONTROLS AND TRAINING RELATED TO CONTROLS
MSD HAZARDS CONTROLLED, OR REDUCED BELOW
EMPLOYER EVALUATES PROGRAM

Figure 18 The Proposed OSHA Ergonomics Rule Flow Chart for Control of MSDs at the Workplace. (Modified after OSHA 2000)

stress response of the individual, thereby increasing the risk associated with a given level of exposure (ANSI 1995). Specific work organization factors that have been shown to fall into at least one of these categories include (but are not limited to):

1. Wage incentives
2. Machine-paced work
3. Workplace conflicts of many types

4. Absence of worker decision latitude
5. Time pressures and work overload
6. Unaccustomed work during training periods or after return from long-term leave

11.3. Procedures for Job Analysis and Design

Job analysis should be performed at a sufficient level of detail to identify potential work-related risk factors associated with WRMDs and include the following steps:

1. Collection of the pertinent information for all jobs and associated work methods
2. Interview of the representative sample of the affected workers
3. Breakdown of the jobs into tasks or elements
4. Description of the component actions of each task or element
5. Measurement and quantification of WRMD risk factors
6. Identification of the risk factors for each task or element
7. Identification of the problems contributing to risk factors
8. Summary of the problem areas and needs for intervention for all jobs and associated new work methods

12. SURVEILLANCE FOR JOB ANALYSIS AND DESIGN

12.1. Provisions of the ANSI Z-365 (1999) Draft Standard

Working Draft ANSI Z-365 includes sections on:

1. *Surveillance* of musculoskeletal disorders, including:
   - (a) Worker reports
   - (b) Analysis of existing records and surveys
   - (c) Job surveys/proactive entry into the process
2. *Job analysis and design,* including:
   - (a) Description of work
   - (b) Definitions of risk factors
   - (c) Risk-factor measurement, quantification, and interaction

12.2. Musculoskeletal Disorder Surveillance

As discussed in ANSI Z-365 (1999), surveillance is defined as the ongoing systematic collection, analysis, and interpretation of health and exposure data in the process of describing and monitoring work-related cumulative trauma disorders. Surveillance is used to determine when and where job analysis is needed and where ergonomic interventions may be warranted.

A surveillance system can be used in any workplace to evaluate cumulative trauma disorders (musculoskeletal disorders) in the working population. Surveillance is defined as ‘‘the ongoing systematic collection, analysis and interpretation of health and exposure data in the process of describing and monitoring a health event. Surveillance data are used to determine the need for occupational safety and health action and to plan, implement and evaluate ergonomic interventions and programs’’ (Klaucke 1988). Health and job risk-factor surveillance provide employers and employees with a means of systematically evaluating musculoskeletal disorders and workplace ergonomic risk factors by monitoring trends over time. This information can be used for planning, implementing, and continually evaluating ergonomic interventions. Therefore, incidence (rate of new cases), prevalence (rate of existing cases), and parameters that may be used in estimating severity must be defined.

12.3. Worker Reports (Case-Initiated Entry into the Process)

Follow-up to medical/first aid reports or worker symptoms consists of collecting and following employee medical reports through the medical management process.

12.4. Analysis of Existing Records and Survey (Past Case(s)-Initiated Entry into the Process)

Analysis of existing records and surveys consists of reviewing existing databases, principally collected for other purposes, to identify incidents and patterns of work-related cumulative trauma disorders. It can help determine and prioritize the jobs to be further analyzed using job analysis. There are three types of existing records and survey analyses:

1. Initial analysis of upper-limb WRMDs reported over the last 24–36 months
2. Ongoing trend analysis of past cases
3. Health surveys

12.5. Job Surveys (Proactive Entry into the Process)

The aim of proactive job surveys is to identify specific jobs and processes that may put employees at risk of developing WRMDs. Job surveys are typically performed after the jobs identified by the previous two surveillance components have been rectified. Job surveys of all jobs or a sample of representatives should be performed. Analysis of existing records will be used to estimate the potential magnitude of the problem in the workplace. The number of employees in each job, department, or similar population will be determined first. Then the incidence rates will be calculated on the basis of hours worked, as follows:

$$\text{Incidence (new case) rate (IR)} = \frac{\text{\# of new cases during time} \times 200{,}000}{\text{Total hours worked during time}}$$

This is equivalent to the number of new cases per 100 worker years. Workplace-wide incidence rates (IRs) will be calculated for all cumulative trauma disorders and by body location for each department, process, or type of job. (If specific work hours are not readily available, the number of full-time equivalent employees in each area multiplied by 2000 hours will be used to obtain the denominator.) Severity rates (SRs) traditionally use the number of lost workdays rather than the number of cases in the numerator. Prevalence rates (PRs) are the number of existing cases per 200,000 hours or the percentage of workers with the condition (new cases plus old cases that are still active).

12.6. ANSI Z-365 Evaluation Tools for Control of WRMDs

Some of the research evaluation tools defined by the ANSI Z-365 Draft Standard for the purpose of surveillance and job analysis include the following:

1. Proactive job survey (checklist #1)
2. Quick check risk factor checklist (checklist #2)
3. Symptom survey (questionnaire)
4. Posture discomfort survey
5. History of present illness recording form

12.7. Analysis and Interpretation of Surveillance Data

Surveillance data can be analyzed and interpreted to study possible associations between the WRMD surveillance data and the risk-factor surveillance data. The two principal goals of the analysis are to help identify patterns in the data that reflect large and stable differences between jobs or departments and to target and evaluate intervention strategies. This analysis can be done on the number of existing WRMD cases (cross-sectional analysis) or on the number of new WRMD cases in a retrospective and prospective fashion (retrospective and prospective analysis).

The simplest way to assess the association between risk factors and WRMDs is to calculate odds ratios (Table 29). To do this, the prevalence data obtained in health surveillance are linked with the data obtained in risk-factor surveillance. The data used can be those obtained with symptom ques-

TABLE 29 Examples of Odds Ratio Calculations for a Firm of 140 Employees

		WRMDs		
		Present	Not Present	Total
Risk factor				
	Present	15 (*A*)	25 (*B*)	40 (*A* + *B*)
	Not present	15 (*C*)	85 (*D*)	100 (*C* + *D*)
	Total	30 (*A* + *C*)	110 (*B* + *D*)	140 (N)

Number in each cell indicates the count of employees with or without WRMD and the risk factor.
Odds ratio (OR) = $(A \times D)/(B \times C) = (15 \times 85)/(25 \times 15) = 3.4$

tionnaires (active level 1 health surveillance) and risk-factor checklists (level 1 active risk-factor surveillance). Each risk factor could be examined in turn to see whether it has an association with the development of WRMDs. In the example shown here, one risk factor at a time is selected (overhead work for more than four hours).

Using the data obtained in surveillance the following numbers of employees are counted:

- Employees with WRMDs and exposed to more than four hours of overhead work (15 workers)
- Employees with WRMDs and not exposed to more than four hours of overhead work (15 workers)
- Employees without WRMDs and exposed to more than four hours of overhead work (25 workers)
- Employees without WRMDs and not exposed to more than four hours of overhead work (85 workers)

The overall prevalence rate (PR), that is, rate of existing cases, for the firm is 30/140, or 21.4%. The prevalence rate for those exposed to the risk factor is 37.5% (15/40) compared to 15.0% (15/100) for those not exposed. The risk of having a WRMD depending on exposures to the risk factor, the odds ratio, can be calculated using the number of existing cases of WRMD (prevalence). In the above example, those exposed to the risk factor have 3.4 times the odds of having the WRMD than those not exposed to the risk factor. An odds ratio of greater than 1 indicates higher risk. Such ratios can be monitored over time to assess the effectiveness of the ergonomics program in reducing the risk of WRMDs, and a variety of statistical tests can be used to assess the patterns seen in the data.

13. ERGONOMICS PROGRAMS IN INDUSTRY

An important component of musculoskeletal disorders management efforts is development of a well-structured and comprehensive ergonomic program. According to Alexander and Orr (1992), the basic components of such a program should include:

1. Health and risk-factor surveillance
2. Job analysis and improvement
3. Medical management
4. Training
5. Program evaluation

An excellent program must include participation of all levels of management, medical, safety and health personnel, labor unions, engineering, facility planners, and workers and contain the following elements:

1. Routine (monthly or quarterly) reviews of the OSHA log for patterns of injury and illness and using of special computer programs to identify problem areas.
2. Workplace audits for ergonomic problems are a routine part of the organization's culture (more than one audit annually for each operating area). Problems identified in this manner are dealt with quickly.
3. List maintained of most critical problems—jobs with job title clearly identified. Knowledge of these problem jobs is widespread, including knowledge by management and the workers.
4. Use of both engineering solutions and administrative controls and seeking to use engineering solutions for long-term solutions.
5. Design engineering is aware of ergonomic considerations and actively builds them into new or reengineered designs. People are an important design consideration.
6. Frequent refresher training for the site-appointed ergonomists in ergonomics and access to short courses and seminars.

14. PROPOSED OSHA ERGONOMICS REGULATIONS

The National Research Council/National Academy of Sciences of the United States recently concluded that there is a clear relationship between musculoskeletal disorders and work and between ergonomic interventions and a decrease in such disorders. According to the Academy, research demonstrates that specific interventions can reduce the reported rate of musculoskeletal disorders for workers who perform high-risk tasks (National Research Council 1998). The effective and universal standard for dealing with the work-related hazards should significantly reduce the risk to WRMDS to employees.

The high prevalence of work-related musculoskeletal disorders, has motivated the Occupational Safety and Health Administration (OSHA) to focus on standardization efforts. Recently, OSHA announced the initiation of rulemaking under Section 6(b) of the Occupational Safety and Health Act of 1970, 29 U.S.C. 655, to amend Part 1910 of Title 29 of the Code of Federal Regulations and requested information relevant to preventing, eliminating, and reducing occupational exposure to ergonomic hazards.

According to OSHA (2000), the proposed standard is needed to bring this protection to the remaining employees in general industry workplaces that are at significant risk of incurring a work-related musculoskeletal disorder but are currently without ergonomics programs. A substantial body of scientific evidence supports OSHA's effort to provide workers with ergonomic protection. This evidence strongly supports two basic conclusions: (1) there is a positive relationship between work-related musculoskeletal disorders and workplace risk factors, and (2) ergonomics programs and specific ergonomic interventions can reduce these injuries.

14.1. Main Provisions of the Draft Ergonomics Standard

The standard applies to employers in general industry whose employees work in manufacturing jobs or manual handling jobs or report musculoskeletal disorders (MSDs) that meet the criteria of the standard (see Figure 18). The standard applies to the following jobs:

1. Manufacturing jobs. Manufacturing jobs are production jobs in which employees perform the physical work activities of producing a product and in which these activities make up a significant amount of their work time;
2. Manual handling jobs. Manual handling jobs are jobs in which employees perform forceful lifting/lowering, pushing/pulling, or carrying. Manual handling jobs include only those jobs in which forceful manual handling is a core element of the employee's job; and
3. Jobs with a musculoskeletal disorder. Jobs with an MSD are those jobs in which an employee reports an MSD that meets all of these criteria:
    - **(a)** The MSD is reported after the effective date;
    - **(b)** The MSD is an ''OSHA recordable MSD,'' or one that would be recordable if the employer was required to keep OSHA injury and illness records; and
    - **(c)** The MSD also meets the screening criteria.

The proposed standard covers only those OSHA-recordable MSDs that also meet these screening criteria:

1. The physical work activities and conditions in the job are reasonably likely to cause or contribute to the type of MSD reported; and
2. These activities and conditions are a core element of the job and/or make up a significant amount of the employee's work time.

The standard applies only to the jobs specified in Section 1910.901, not to the entire workplace or to other workplaces in the company. The standard does not apply to agriculture, construction, or maritime operations. In the proposed standard, a full ergonomics program consists of these six program elements:

1. Management leadership and employee participation
2. Hazard information and reporting
3. Job hazard analysis and control
4. Training
5. MSD management
6. Program evaluation

According to the standard, the employer must:

1. Implement the first two elements of the ergonomics program (management leadership and employee participation, and hazard information and reporting) even if no MSD has occurred in those jobs.
2. Implement the other program elements when either of the following occurs in those jobs (unless one eliminates MSD hazards using the quick fix option

- **(a)** A covered MSD is reported; or
- **(b)** Persistent MSD symptoms are reported plus:
  - **(i)** The employer has knowledge that an MSD hazard exists in the job;
  - **(ii)** Physical work activities and conditions in the job are reasonably likely to cause or contribute to the type of MSD symptoms reported; and
  - **(iii)** These activities and conditions are a core element of the job and/or make up a significant amount of the employee's work time.

In other jobs in general industry, the employer should comply with all of the program elements in the standard when a covered MSD is reported (unless the MSD hazards are eliminated using the quick fix option). The employer should do the following to quick fix a problem job:

1. Promptly make available the MSD management
2. Consult with employee(s) in the problem job about the physical work activities or conditions of the job they associate with the difficulties, observe the employee(s) performing the job to identify whether any risk factors are present, and ask employee(s) for recommendations for eliminating the MSD hazard
3. Put in quick fix controls within 90 days after the covered MSD is identified and check the job within the next 30 days to determine whether the controls have eliminated the hazard
4. Keep a record of the quick fix controls
5. Provide the hazard information the standard requires to employee(s) in the problem job within the 90-day period

The employer should set up the complete ergonomics program if either the quick fix controls do not eliminate the MSD hazards within the quick fix deadline (120 days) or another covered MSD is reported in that job within 36 months.

The employer should demonstrate management leadership of your ergonomics program. Employees (and their designated representatives) must have ways to report MSD signs and MSD symptoms, get responses to reports; and be involved in developing, implementing, and evaluating each element of your program. The employer should not have policies or practices that discourage employees from participating in the program or from reporting MSDs signs or symptoms. The employer also should:

1. Assign and communicate responsibilities for setting up and managing the ergonomics program so managers, supervisors, and employees know what you expect of them and how you will hold them accountable for meeting those responsibilities
2. Provide those persons with the authority, resources, information, and training necessary to meet their responsibilities
3. Examine your existing policies and practices to ensure that they encourage and do not discourage reporting and participation in the ergonomics program
4. Communicate periodically with employees about the program and their concerns about MSDs

According to the proposed standard, the employees (and their designated representatives) must have a way to report MSD signs and symptoms; prompt responses to their reports; access to the standard and to information about the ergonomics program; and ways to be involved in developing, implementing, and evaluating each element of the ergonomics program.

The employer should set up a way for employees to report MSD signs and symptoms and get prompt responses. The employer should evaluate employee reports of MSD signs and symptoms to determine whether a covered MSD has occurred. The employer should periodically provide information to employees that explains how to identify and report MSD signs and symptoms. The employer should also provide this information to current and new employees about common MSD hazards, the signs and symptoms of MSDs and the importance of reporting them early, how to report the signs and symptoms, and a summary of the requirements of the standard.

14.2. Job Hazard Analysis and Control

According to the Draft Standard, the employer should analyze the problem job to identify the ergonomic risk factors that result in MSD hazards. The employer should eliminate the MSD hazards, reduce them to the extent feasible, or materially reduce them using the incremental abatement process in the standard. If the MSD hazards only pose a risk to the employee with the covered MSD, the job hazard analysis and control can be limited to that individual employee's job. In such a case, the employer should:

1. Include in the job-hazard analysis all of the employees in the problem job or those who represent the range of physical capabilities of employees in the job.
2. Ask the employees whether performing the job poses physical difficulties and, if so, which physical work activities or conditions of the job they associate with the difficulties.
3. Observe the employees performing the job to identify which of the physical work activities, workplace conditions, and ergonomic risk factors are present.
4. Evaluate the ergonomic risk factors in the job to determine the MSD hazards associated with the covered MSD. As necessary, evaluate the duration, frequency, and magnitude of employee exposure to the risk factors.

The proposed engineering controls include physical changes to a job that eliminate or materially reduce the presence of MSD hazards. Examples of engineering controls for MSD hazards include changing, modifying, or redesigning workstations, tools, facilities, equipment, materials, and processes. Administrative controls are changes in the way that work in a job is assigned or scheduled that reduce the magnitude, frequency, or duration of exposure to ergonomic risk factors. Examples of administrative controls for MSD hazards include employee rotation, job task enlargement, alternative tasks, and employer-authorized changes in work pace.

Finally, it should be noted that the OSHA's Final Ergonomic Program Standard took effect on January 16, 2001.

REFERENCES

Alexander, D. C., and Orr, G. B. (1992), ''The Evaluation of Occupational Ergonomics Programs,'' in *Proceedings of the Human Factors Society 36th Annual Meeting,* pp. 697–701.

Andersson, G. B. J. (1985), ''Permissible Loads: Biomechanical Considerations,'' *Ergonomics,* Vol. 28, No. 1, pp. 323–326.

ANSI Z-365 Draft (1995), ''Control of Work-Related Cumulative Trauma Disorders, Part I: Upper Extremities,'' April 17.

Armstrong, T. (1986), ''Ergonomics and Cumulative Trauma Disorders,'' *Hand Clinics,* Vol. 2, pp. 553–565.

Armstrong, T. J., Buckle, P., Fine, L. J., Hagberg, M., Jonsson, B., Kilbom, A., Kuorinka, I., Silverstein, B. A., Sjogaard, G., and Viikari-Juntura, E. (1993), ''A Conceptual Model for Work-Related Neck and Upper-Limb Musculoskeletal Disorders,'' *Scandinavian Journal of Work and Environmental Health,* Vol. 19, pp. 73–84.

Armstrong, T. J., and Lifshitz, Y. (1987), ''Evaluation and Design of Jobs for Control of Cumulative Trauma Disorders,'' in *Ergonomic Interventions to Prevent Musculoskeletal Injuries in Industry,* American Conference of Governmental Industrial Hygienists, Lewis, Chelsea, MI.

Asfour, S. S., Genaidy, A. M., Khalil, T. M., and Greco, E. C. (1984), ''Physiological and Psychophysical Determination of Lifting Capacity for Low Frequency Lifting Tasks,'' in *Trends in Ergonomics/Human Factors I,* North-Holland, Cincinnati, pp. 149–153.

Ayoub, M. A. (1982), ''Control of Manual Lifting Hazards: II. Job Redesign,'' *Journal of Occupational Medicine,* Vol. 24, No. 9, pp. 668–676.

Ayoub, M. M., Bethea, N. J., Deivanayagam, S., Asfour, S. S., Bakken, G. M., Liles, D., Mital, A., and Sherif, M. (1978), ''Determination and Modelling of Lifting Capacity,'' Final Report HEW [NIOSH] Grant No. 5R010H-000545-02, National Institute of Occupational Safety and Health, Washington, DC.

Ayoub, M. M. (1983), ''Design of a Pre-Employment Screening Program,'' in *Ergonomics of Workstation Design,* T. O. Kvalseth, Ed., Butterworths, London.

Ayoub, M. M., and Mital, A. (1989), *Manual Material Handling,* Taylor & Francis, London.

Ayoub, M. M., Dempsey, P. G., and Karwowski, W. (1997), ''Manual Materials Handling,'' in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1085–1123.

Badler, N. L., Becket, W. M., and Webber, B. L. (1995), ''Simulation and Analysis of Complex Human Tasks for Manufacturing,'' in *Proceedings of SPIE—The International Society for Optical Engineering,* Vol. 2596, pp. 225–233.

Battié, M. C., Bigos, S. J., Fisher, L. D., Spengler, D. M., Hansson, T. H., Nachemson, A. L., and Wortley, M. D. (1990), ''The Role of Spinal Flexibility in Back Pain Complaints within Industry: A Prospective Study,'' *Spine,* Vol. 15, pp. 768–773.

Bigos, S. J., Spengler, D. M., Martin, N. A., et al. (1986), ''Back Injuries in Industry: A Retrospective Study: II. Injury Factors,'' *Spine,* Vol. 11, pp. 246–251.

Bigos, S. J., Battié, M. C., Spengler, D. M., Fisher, L. D., Fordyce, W. E., Hansson, T. H, Nachemson, A. L., and Wortley, M. D. (1991), "A Prospective Study of Work Perceptions and Psychosocial Factors Affecting the Report of Back Injury," *Spine,* Vol. 16, pp. 1–6.

Board of Certification in Professional Ergonomics (2000), website www.bcpe.org.

Bonney, R., Weisman, G., Haugh, L. D., and Finkelstein, J. (1990), "Assessment of Postural Discomfort," in *Proceedings of the Human Factors Society,* pp. 684–687.

Borg, G. A. V. (1962), *Physical Performance and Perceived Exertion,* Lund, Gleerup.

Brennan, L., Farrel, L., and McGlennon, D. (1990), "ERGOSPEC System for Workstation Design," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 117–127.

Brinckmann, P., Biggemann, M., and Hilweg, D. (1988), "Fatigue Fractures of Human Lumbar Vertebrae," *Clinical Biomechanics,* Vol. 3 (Supplement 1), 1988.

Brinckmann, P., Biggemann, M., and Hilweg, D. (1989), "Prediction of the Compressive Strength of Human Lumbar Vertebrae," *Clinical Biomechanics,* Vol. 4 (Suppl. 2).

Brokaw, N. (1992), "Implications of the Revised NIOSH Lifting Guide of 1991: A Field Study," M.S. Thesis, Department of Industrial Engineering, University of Louisville.

Bullinger, H. J., and Lorenz, D. (1990), "CAD—Video Somotography: A Method for the Anthropometric Design of Workplaces," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 128–137.

Bureau of Labor Statistics (BLS) (1995), *Occupational Injuries and Illness in the United States by Industry,* Department of Labor, Washington, DC.

Bureau of National Affairs (BNA) (1988), Special Report, "Back Injuries: Costs, Causes, Cases and Prevention," Bureau of National Affairs, Washington, DC.

Caldwell, L. S., and Smith, R. P. (1963), "Pain and Endurance of Isometric Contractions," *Journal of Engineering Psychology,* Vol. 5, pp. 25–32.

Chaffin, D. B. (1973), "Localized Muscle Fatigue—Definition and Measurement," *Journal of Occupational Medicine,* Vol. 15, p. 346.

Chaffin, D. B., Herrin, G. D., Keyserling, W. M., and Foulke, J. A. (1977), "Pre-Employment Strength Testing," DHEW (NIOSH), Publication No. 77-163, National Institute of Occupational Safety and Health.

Chaffin, D. B., Herrin, G. D., and Keyserling, W. M. (1978), "Pre-Employment Strength Testing," *Journal of Occupational Medicine,* Vol. 20, No. 6, pp. 403–408.

Chaffin, D. B., and Erig, M. (1991), "Three-Dimensional Biomechanical Statis Strength Prediction Model Sensitivity to Postural and Anthropometric Inaccuracies," *IIE Transactions,* Vol. 23, No. 3, pp. 215–227.

Chaffin, D. B., Andersson, G. B. J., and Martin, B. J. (1999), *Occupational Biomechanics,* 3rd Ed., John Wiley & Sons, New York.

Ciriello, V. M., and Snook, S. H. (1978), "The Effect of Size, Distance, Height, and Frequency on Manual Handling Performance," in *Proceedings of the Human Factors Society's 22nd Annual Meeting* (Detroit), pp. 318–322.

Ciriello, V. M., and Snook, S. H. (1983), "A Study of Size, Distance, Height, and Frequency Effects on Manual Handling Tasks," *Human Factors,* Vol. 25, pp. 473–483.

Ciriello, V. M., Snook, S. H., Blick, A. C., and Wilkinson, P. L. (1990), "The Effects of Task Duration on Psychophysically-Determined Maximum Acceptable Weights and Forces," *Ergonomics,* Vol. 33, No. 2, pp.187–200.

Colombini, D., Occhipinti, E., Molteni, G., Grieco, A., Pedotti, A., Boccardi, S., Frigo, C., and Menoni, D. (1985), "Posture Analysis," *Ergonomics,* Vol. 28, No. 1, pp. 275–284.

Colombini, D., Occhipinti, E., Delleman, N., Fallentin, N., Kilbom, A., and Grieco, A. (1999), "Exposure Assessment of Upper Limb Repetitive Movements: A Consensus Document." Report prepared by the Technical Committee on Musculoskeletal Disorders of the International Ergonomics Association (IEA) and endorsed by the International Commission on Occupational Health (ICOH).

Corlett, E. N., and Manenica, I. (1980), "The Effects and Measurement of Working Postures," *Applied Ergonomics,* Vol. 11, No. 1, pp. 7–16.

Corlett, E. N., Madeley, S. J., and Manenica, J. (1979), "Posture Targeting: A Technique for Recording Work Postures," *Ergonomics,* Vol. 24, pp. 795–806.

Corlett, N., Wilson, J., and Manenica, I., Eds. (1986), *The Ergonomics of Working Postures: Models, Methods and Cases,* Taylor & Francis, London.

Damkot, D. K., Pope, M. H., Lord, J., and Frymoyer, J. W. (1984), "The Relationship Between Work History, Work Environment and Low-Back Pain in Men," *Spine,* Vol. 9, pp. 395–399.

Delleman, N. J., Drost, M. R., and Huson, A. (1992), "Value of Biomechanical Macromodels as Suitable Tools for the Prevention of Work-Related Low Back Problems," *Clinical Biomechanics,* Vol. 7, pp. 138–148.

Edwards, R. H. T. (1988), "Hypothesis of Peripheral and Central Mechanisms Underlying Occupational Muscle Pain and Injury," *European Journal of Applied Physiology,* Vol. 57, pp. 275–282.

Eisler, H. (1962), "Subjective Scale of Force for a Large Muscle Group," *Journal of Experimental Psychol.,* Vol. 64, No. 3, pp. 253–257.

Farfan, H. F. (1983), "Biomechanics of the Lumbar Spine," in *Managing Low Back Pain,* W. H. Kirkaldy-Willis, Ed., Churchill Livingstone, New York, pp. 9–21.

Federal Register (1986), Vol. 51, No. 191, October 2, pp. 35–41.

Ferguson, S. A., Marras, W. S., and Waters, T. R. (1992), "Quantification of Back Motion During Asymmetric Lifting," *Ergonomics,* Vol. 40, No. 7/8, pp. 845–859.

Fortin, C., Gilbert, R., Beuter, A., Laurent, F., Schiettekatte, J., Carrier, R., and Dechamplain, B. (1990), "SAFEWORK: A Microcomputer-Aided Workstation Design and Analysis, New Advances and Future Developments," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds.,Taylor & Francis, London, pp. 157–180.

Frymoyer, J. W., Pope, M. H., Constanza, M. C., Rosen, J. C., Goggin, J. E., and Wilder, D. G. (1980), "Epidemiologic Studies of Low Back Pain," *Spine,* Vol. 5, pp. 419–423.

Frymoyer, J. W., Pope, M. H., Clements, J. H., Wilder, D. G., MacPherson, B., and Ashikaga, T. (1983), "Risk Factors in Low Back Pain: An Epidemiological Survey," *Journal of Bone Joint Surgery,* Vol. 65-A, pp. 213–218.

Gagnon, M., and Smyth, G. (1990), "The Effect of Height in Lowering and Lifting Tasks: A Mechanical Work Evaluation," in *Advances in Industrial Ergonomics and Safety II,* B. Das, Ed., Taylor & Francis, London, pp. 669–672.

Gamberale, F., and Kilbom, A. (1988), "An Experimental Evaluation of Psychophysically Determined Maximum Acceptable Workload for Repetitive Lifting Work," in *Proceedings of the 10th Congress of the International Ergonomics Association,* A. S. Adams, R. R. Hall, B. J. McPhee, and M. S. Oxenburgh, Eds., Taylor & Francis, London, pp. 233–235.

GAO (1997), "Worker Protection: Private Sector Ergonomics Programs Yield Positive Results," GAO/HEHS-97-163, U.S. General Accounting Office, Washington, DC.

Garg, A. (1989), "An Evaluation of the NIOSH Guidelines for Manual Lifting with Special Reference to Horizontal Distance," *American Industrial Hygiene Association Journal,* Vol. 50, No. 3, pp. 157–164.

Garg, A., and Badger, D. (1986), "Maximum Acceptable Weights and Maximum Voluntary Strength for Asymmetric Lifting," *Ergonomics,* Vol. 29, No. 7, pp. 879–892.

Garg, A., and Banaag, J. (1988), "Psychophysical and Physiological Responses to Asymmetric Lifting," in *Trends in Ergonomics/Human Factors V,* F. Aghazadeh, Ed., North-Holland, Amsterdam, pp. 871–877.

Garg, A., Chaffin, D. B., and Herrin, G. D. (1978), "Prediction of Metabolic Rates for Manual Materials Handling Jobs," *American Industrial Hygiene Association Journal,* Vol. 39, No. 8, pp. 661–675.

Garg, A., Chaffin, D. B., and Freivalds, A. (1982), "Biomechanical Stresses from Manual Load Lifting: Static vs. Dynamic Evaluation," *Institute of Industrial Engineers Transactions,* Vol. 14, pp. 272–281.

Genaidy, A., Karwowski, W., and Christensen, D. (1999), "Principles of Work System Performance Optimization: A Business Ergonomics Approach," *Human Factors and Ergonomics in Manufacturing,* Vol. 9, No. 1, pp. 105–128.

Grandjean, E. (1980), *Fitting the Task to the Man,* Taylor & Francis, London, pp. 41–62.

Grieve, D., and Pheasant, S. (1982), "Biomechanics," in *The Body at Work,* W. T. Singleton, Ed., Taylor & Francis, London, pp. 71–200.

Habes, D. J., and Putz-Anderson, V. (1985), "The NIOSH Program for Evaluating Biomechanical Hazards in the Workplace," *Journal of Safety and Research,* Vol. 16, pp. 49–60.

Hagberg, M. (1984), "Occupational Musculoskeletal Stress and Disorders of the Neck and Shoulder: A Review of Possible Pathophysiology," *International Archives of Occupational and Environmental Health,* Vol. 53, pp. 269–278.

Hägg, G. (1991), "Static Work Loads and Occupational Myalgia: A New Explanation Model," in *Electromyographical Kinesiology,* P. Anderson, D. Hobart, and J. Danoff, Eds., Elsevier Science, New York, pp. 141–144.

Hansson, T. (1989), *Ländryggsbesvär och arbete* Arbetsmiljöfonden, Stockholm.

Henry Dreyfuss Associates (1981), *HUMANSCALE,* MIT Press, Cambridge, MA.

Hildebrandt, V. H. (1987), "A Review of Epidemiological Research on Risk Factors of Low Back Pain," in *Musculoskeletal Disorders at Work,* P. W. Buckle, Ed., Taylor & Francis, London.

Holzmann, P. (1982), "ARBAN: A New Method for Analysis of Ergonomic Effort," *Applied Ergonomics,* Vol. 13, pp. 82–86.

Hunsicker, P. A., and Greey, G. (1957), "Studies in Human Strength," *Research Quarterly,* Vol. 28, No. 109.

Imrhan, S. N. (1999), "Push–Pull Force Limits," in *The Occupational Ergonomics Handbook,* W. Karwowski and W. Marras, Eds., CRC Press, Boca Raton, FL, pp. 407–420.

Imrhan, S. N., and Alhaery, M. (1994), "Finger Pinch-Pull Strengths: Large Sample Statistics," in *Advances in Industrial Ergonomics and Safety VI,* F. Aghazadeh, Ed., Taylor & Francis, London, pp. 595–597.

Imrhan, S. N., and Sundararajan, K. (1992), "An Investigation of Finger Pull Strength," *Ergonomics,* Vol. 35, No. 3, pp. 289–299.

Jäger, M., and Luttmann, A. (1991), "Compressive Strength of Lumbar Spine Elements Related to Age, Gender, and Other Influencing Factors," in *Electromyographical Kinesiology,* P. A. Anderson, D. J. Hobart, and J. V. Dainoff, Eds., Elsevier, Amsterdam, pp. 291–294.

Jäger, M., and Luttmann, A. (1992a), "The Load on the Lumbar Spine During Asymmetrical Bi-Manual Materials Handling," *Ergonomics,* Vol. 35, No. 7/8, pp. 783–805.

Jäger, M., and Luttmann, A. (1992b), "Lumbosacral Compression for Uni- and Bi-Manual Asymmetrical Load Lifting," in *Advances in Industrial Ergonomics and Safety IV,* S. Kumar, Ed., Taylor & Francis, London, pp. 839–846.

Jonsson, B., Persson, J., and Kilbom, A. (1988), "Disorders of the Cervicobrachial Region among Female Workers in the Electronics Industry," *International Journal of Industrial Ergonomics,* Vol. 3, pp. 1–12.

Jung, E. S., and Freivalds, A. (1990), "Development of an Expert System for Designing Workplaces in Manual Materials Handling Jobs," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 279–298.

Kahn, J. F., and Monod, H. (1989), "Fatigue Induced by Static Work," *Ergonomics,* Vol. 32, pp. 839–846.

Karhu, O., Kansi, P., and Kuorinka, I. (1977), "Correcting Working Postures in Industry: A Practical Method for Analysis," *Applied Ergonomics,* Vol. 8, pp. 199–201.

Karwowski, W. (1992), "Occupational Biomechanics," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1005–1046.

Karwowski, W., and Brokaw, N. (1992), "Implications of the Proposed Revisions in a Draft of the Revised NIOSH Lifting Guide (1991) for Job Redesign: A Field Study," in *Proceedings of the 36th Annual Meeting of the Human Factors Society* (Atlanta), pp. 659–663.

Karwowski, W., and Gaddie, P. (1995), "Simulation of the 1991 Revised NIOSH Manual Lifting Equation," in *Proceeding of the Human Factors and Ergonomics Society Annual Meeting* (San Diego), pp. 699–701.

Karwowski, W., and Marras, W. S. (1997), "Cumulative Trauma Disorders," in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1124–1173.

Karwowski, W., and Marras, W. S., Eds., (1999), *The Occupational Ergonomics Handbook,* CRC Press, Boca Raton, FL.

Karwowski, W., and Pongpatanasuegsa, N. (1988), "Testing of Isometric and Isokinetic Lifting Strengths of Untrained Females in Teamwork," *Ergonomics,* Vol. 31, pp. 291–301.

Karwowski, W., and Salvendy, G., Eds., (1998), *Ergonomics in Manufacturing: Raising Productivity through Workplace Improvement,* SME Press and IIE Press, MI.

Karwowski, W., and Salvendy, G., Eds., (1999), *Design of Work and Organization,* John Wiley & Sons, New York.

Karwowski, W. et al., (1994), "Integrating People, Organizations, and Technology in Advanced Manufacturing: A Position Paper Based on the Joint View of Industrial Managers, Engineers, Consultants, and Researchers," *International Journal of Human Factors in Manufacturing,* Vol. 4, No. 1, pp. 1–19.

Karwowski, W., Wogalter, M., and Dempsey, P.G., Eds., (1997), *Ergonomics and Musculoskeletal Disorders,* Human Factors and Ergonomics Society, Santa Monica, CA.

Karwowski, W., Gaddie, P., Jang, R., and GeeLee, W. (1999), "A Population-Based Load Threshold Limit (LTL) for Manual Lifting Tasks Performed by Males and Females," in *The Occupational Ergonomics Handbook,* W. Karwowski and W. S. Marras, Eds., CRC Press, Boca Raton, FL, pp. 1063–1074.

Karwowski, W., Genaidy, A. M., and Asfour, S. S., Eds. (1990), *Computer-Aided Ergonomics,* Taylor & Francis, London.

Kee, D., and Karwowski, W. (2001), "Ranking Systems for Evaluation of Joint Motion Stressfulness Based on Perceived Discomforts," *Ergonomics* (forthcoming).

Kelsey, J. L., and Golden, A. L. (1988), Occupational and Workplace Factors Associated with Low Back Pain," *Occupational Medicine: State of the Art Reviews,* Vol. 3, No. 1, pp. 7–16.

Kelsey, J. L., and Golden, A. L. (1988), "Occupational and Workplace Factors Associated with Low Back Pain," in *Back Pain in Workers,* R. A. Deyo, Ed., Hanley & Belfus, Philadelphia.

Kelsey, J., and Hochberg, M. (1988), "Epidemiology of Chronic Musculoskeletal Disorders," *Annual Review of Public Health,* Vol. 9, pp. 379–401.

Kelsey, J. L., Githens, P. B., White, A. A., Holford, R. R., Walter, S. D., O'Connor, T., Astfeld, A. M., Weil, U., Southwick, W. O., and Calogero, J. A. (1984), "An Epidemiologic Study of Lifting and Twisting on the Job and Risk for Acute Prolapsed Lumbar Intervertebral Disc," *Journal of Orthopaedic Research,* Vol. 2, pp. 61–66.

Keyserling, W. M. (1986), "Postural Analysis of the Trunk and Shoulder in Simulated Real Time," *Ergonomics,* Vol. 29, No. 4, pp. 569–583.

Keyserling, W. M. (1990), "Computer-Aided Posture Analysis of the Trunk, Neck, Shoulders and Lower Extremities," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 261–272.

Keyserling, W. M., Punnett, L., and Fine, L. J. (1988), "Trunk Posture and Low Back Pain: Identification and Control of Occupational Risk Factors," *Applied Industrial Hygiene,* Vol. 3, pp. 87–92.

Keyserling, W. M., Stetson, D. S., Silverstein, B. A., and Brouver, M. L. (1993), "A Checklist for Evaluating Risk Factors Associated with Upper Extremity Cumulative Trauma Disorders," *Ergonomics,* Vol. 36, No. 7, pp. 807–831.

Klaucke, D. N., Buehler, J. W., Thacker, S. B., Parrish, R. G., Trowbridge, R. L., and Berkelman, R. L. (1988), "Guidelines for Evaluating Surveillance System," *Morbidity and Mortality Weekly,* Vol. 37, Suppl. 5, pp. 1–18.

Kroemer, K. H. E. (1970), "Human Strength: Terminology, Measurement, and Interpretation of Data," *Human Factors,* Vol. 12, pp. 297–313.

Kroemer, K. H. E. (1989), "Engineering Anthropometry," *Ergonomics,* Vol. 32, No. 7, pp. 767–784.

Kroemer, K. H. E. (1992), "Personnel Training for Safer Material Handling," *Ergonomics,* Vol. 35, No. 9, pp. 1119–1134.

Kroemer, K. H. E., Kroemer, H. J., and Kroemer-Elbert, K. E. (1986), *Engineering Physiology,* Elsevier, Amsterdam.

Kumar, S. (1990), "Cumulative Load as a Risk Factor for Back Pain," *Spine,* Vol. 15, No. 12, pp. 1311–1316.

Kuorinka, I., and Forcier, L., Eds. (1995), *Work Related Musculoskeletal Disorders* (*WMSDs*)*: A Reference Book for Prevention,* Taylor & Francis, London.

Landau, K., Brauchler, R., Brauchler, W., Landau, A., and Bobkranz, R. (1990), "Job Analysis and Work Design Using Ergonomic Databases," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 197–212.

Legg, S. J., and Myles, W. S. (1985), "Metabolic and Cardiovascular Cost, and Perceived Effort over an 8 Hour Day When Lifting Loads Selected by the Psychophysical Method," *Ergonomics,* Vol. 28, pp. 337–343.

Leino, P. (1989), "Symptoms of Stress Production and Musculoskeletal Disorders," *Journal of Epidemiology and Community Health,* Vol. 43, pp. 293–300.

Liles, D. H., Deivanayagam, S., Ayoub, M. M., and Mahajan, P. (1984), "A Job Severity Index for the Evaluation and Control of Lifting Injury," *Human Factors,* Vol. 26, pp. 683–693.

Lloyd, M. H., Gauld, S., and Soutar, C. A. (1986), "Epidemiologic Study of Back Pain in Miners and Office Workers," *Spine,* Vol. 11, pp. 136–140.

Lundborg, G., Dahlin, L. B., Danielsen, N., and Kanje, M. (1990), "Vibration Exposure and Nerve Fibre Damage," *Journal of Hand Surgery,* Vol. 15A, pp. 346–351.

Maeda, K., Hunting, W., and Grandjean, E. (1980), "Localized Fatigue in Accounting Machine Operators," *Journal of Occupational Medicine,* Vol. 22, pp. 810–816.

Magora, A. (1973), "Investigation of the Relation Between Low Back Pain and Occupation: IV. Physical Requirements: Bending Rotation, Reaching and Sudden Maximal Effort," *Scandinavian Journal of Rehabilitative Medicine,* Vol. 5, pp. 186–190.

Marras, W. S. (1997), "Biomechanics of the Human Body," in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York.

Marras, W. S., Lavender, S. A., Leurgans, S. E., Rajulu, S. L., Allread, W. G., Fathallah, F. A., and Ferguson, S. A. (1993), "The Role of Dynamic Three-Dimensional Trunk Motion in Occupationally-Related Low Back Disorders: The Effects of Workplace Factors, Trunk Position and Trunk Motion Characteristics on Risk of Injury," *Spine,* Vol. 18, No. 5, pp. 617–628.

Mattila, M., Karwowski, W., and Vikki, M. (1993), "Analysis of Working Postures in Hammering Tasks at Building Construction Sites Using the Computerized OWAS-Method," *Applied Ergonomics,* Vol. 24, No. 6, pp. 405–412.

McDaniel, J. W. (1990), "Models for Ergonomic Analysis and Design: COMBIMAN and CREW CHIEF," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 138–156.

McGill, S. M., and Norman, R. W. (1985), "Dynamically and Statically Determined Low Back Moments During Lifting," *Journal of Biomechanics,* Vol. 18, No. 12, pp. 877–885.

Miedema, M. C., Douwes, M., and Dul, J. (1997), "Recommended Maximum Holding Times for Prevention of Discomfort of Static Standing Postures," *International Journal of Industrial Ergonomics,* Vol. 19, pp. 9–18.

Milner, N. P., Corlett, E. N., and O'Brien, C. (1986), "A Model to Predict Recovery from Maximal and Submaximal Isometric Exercise," in *Ergonomics of Working Postures,* N. Corlett, J. Wilson, and I. Manenica, Eds., Taylor & Francis, London.

Mital, A. (1984), "Comprehensive Maximum Acceptable Weight of Lift Database for Regular 8-Hour Work Shifts," *Ergonomics,* Vol. 27, No. 11, pp. 1127–1138.

Mital, A. (1992), "Psychophysical Capacity of Industrial Workers for Lifting Symmetrical and Asymmetrical Loads Symmetrically and Asymmetrically for 8 Hour Work Shifts," *Ergonomics,* Vol. 35, No. 7/8, pp. 745–754.

Mital, A., and Faard, H. F. (1990), "Effects of Posture, Reach, Distance and Preferred Arm Orientation on Isokinetic Pull Strengths in the Horizontal Plane," Technical Report, Ergonomics and Engineering Controls Laboratory, University of Cincinnati, Cincinnati.

Mital, A., and Genaidy, A. (1989), "Isokinetic Pull-Up Profiles of Men and Women in Different Working Postures," *Clinical Biomechanics,* Vol. 4, pp. 168–172.

Mital, A., Genaidy, A. M., and Brown, M. L. (1989), "Predicting Maximum Acceptable Weights of Symmetrical and Asymmetrical Loads for Symmetrical and Asymmetrical Lifting," *Journal of Safety Research,* Vol. 20, No. 1, pp. 1–6.

Mital, A., Garg, A., Karwowski, W., Kumar, S., Smith, J. L., and Ayoub, M. M. (1993a), "Status in Human Strength Research and Application," *IIE Transactions,* Vol. 25, No. 6, pp. 57–69.

Mital, A., Nicholson, A. S., and Ayoub, M. M. (1993b), *A Guide to Manual Materials Handling,* Taylor & Francis, London.

Monod, H. (1965), "The Work Capacity of a Synergistic Muscular Group," *Ergonomics,* Vol. 8, pp. 329–338.

Monod, H. A. (1972), "A Contribution to the Study of Static Work," Medical Thesis, Paris.

Monod, H., and Scherrer, J. (1965), "The Work Capacity of a Synergic Muscular Group," *Ergonomics,* Vol. 8, pp. 329–338.

Moore, J. S., and Garg, A. (1995), "The Strain Index: A Proposed Method to Analyze Jobs for Risk of Distal Upper Extremity Disorders," *American Industrial Hygiene Association Journal,* Vol. 56, pp. 443–458.

National Academy of Sciences (1985), *Injury in America,* National Academy Press, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1981), "Work Practices Guide for Manual Lifting," Technical Report No. 81-122, Cincinnati, U.S. Government Printing Office, Washington, D.C.

National Institute for Occupational Safety and Health (NIOSH) (1986), "Proposed National Strategies for the Prevention of Leading Work-Related Diseases and Injuries, Part I," DHEW (NIOSH), PB87-114740, NIOSH.

National Institute for Occupational Safety and Health (NIOSH) (1997), *Musculoskeletal Disorder and Workplace Factors. A Critical Review of Epidemiologic Evidence for Work-Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back,* DHHS (NIOSH) Publication No. 97-141, U.S. Department of Health and Human Services, Washington, DC.

National Research Council (1998), ''Work-Related Musculoskeletal Disorders: A Review of the Evidence,'' National Academy Press, Washington, DC., website www.nap.edu/books/0309063272/html/index.html.

National Safety Council (1989), *Accident Facts 1989,* National Safety Council, Chicago.

Nayar, N. (1995), ''Deneb/ERGO—A Simulation Based Human Factors Tool,'' in *Proceedings of the Winter Simulation Conference.*

Nicholson, L. M., and Legg, S. J. (1986), ''A Psychophysical Study of the Effects of Load and Frequency upon Selection of Workload in Repetitive Lifting,'' *Ergonomics,* Vol. 29, No. 7, pp. 903–911.

Occupational Safety and Health Administration (OSHA) (1982), ''Back Injuries Associated with Lifting,'' Bulletin 2144, U.S. Government Printing Office, Washington, DC.

Occupational Safety and Health Administration (OSHA) (2000), ''Final Ergonomic Program Standard:Regulatory Text,'' website www.osha-slc.gov/ergonomics-standard/regulatory/index.html.

Pearcy, M. J., Gill, J. M., Hindle, J., and Johnson, G. R. (1987), ''Measurement of Human Back Movements in Three Dimensions by Opto-Electronic Devices,'' *Clin. Biomech,* Vol. 2, pp. 199–204.

Pheasant, S. (1986), *Bodyspace: Anthropometry, Ergonomics and Design,* Taylor & Francis, London.

Pheasant, S. (1989), ''Anthropometry and the Design of Workspaces,'' in *Evaluation of Human Work,* J. R. Wilson, and N. Corlett, Eds., Taylor & Francis, London, pp. 455–471.

Porter, J. M., Freer, M., Case, K., and Bonney, M. c. (1995), ''Computer Aided Ergonomics and Workspace Design,'' in J. R. Wilson and E. N. Corlett, Eds., *Evaluation of Human Work: A Practical Ergonomics Methodology,* Taylor & Francis, London.

Pottier, M., Lille, F., Phuon, M., and Monod, H. (1969), ''Etude de la contraction Statique Intermittente,'' *Le Travail Humain,* Vol. 32, pp. 271–284 (in French).

Potvin, J., Norman, R. W., and McGill, S. M. (1991), ''Reduction in Anterior Shear Forces on the L4/L5 Disc by the Lumbar Musculature,'' *Clinical Biomechanics,* Vol. 6, pp. 88–96.

Praemer, A., Furner, S., and Rice, D. P. (1992), *Musculoskeletal Conditions in the United States,* American Academy of Orthopaedic Surgeons, Park Ridge, IL.

Pritsker, A. A. B. (1986), *Introduction to Simulation and SLAM II,* 3d Ed., John Wiley & Sons, New York.

Putz-Anderson, V., Ed. (1988), *Cumulative Trauma Disorders: A Manual for Musculoskeletal Diseases for the Upper Limbs,* Taylor & Francis, London.

Putz-Anderson, V., Ed. (1993), *Cumulative Trauma Disorders: A Manual for Musculoskeletal Diseases for the Upper Limbs,* Taylor & Francis, London.

Riihimäki, H. (1991), ''Low-Back Pain, Its Origin and Risk Indicators,'' *Scandinavian Journal of Work, Environment and Health,* Vol. 17, pp. 81–90.

Riihimäki, H., Tola, S., Videman, T., and Hänninen, K. (1989), ''Low-Back Pain and Occupation: A Cross-Sectional Questionnaire Study of Men in Machine Operating, Dynamic Physical Work and Sedentary Work,'' *Scandinavian Journal of Work, Environment and Health,* Vol. 14, pp. 204–209.

Robinette, K. M. M., and McConville, J. T. (1981), ''An Alternative to Percentile Models,'' SAE Technical Paper #810217, Society of Automotive Engineers, Warrendale, PA.

Roebuck, J. A., Kroemer, K. H. E., and Thomson, W. G. (1975), *Engineering Anthropometry Methods,* John Wiley & Sons, New York.

Rohmert, W. (1960), ''Ermittlung von Erholungspausen für Statische Arbeit des Menschen,'' *Internationele Zeitschrift für Angewandte Physiologie Einschliesslich Arbeitsphysiologie,* Vol. 18, pp. 123–164.

Rohmert, W. (1973), ''Problems of Determination of Rest Allowances, Part 2,'' *Applied Ergonomics,* Vol. 4, pp. 158–162.

Rombach, V., and Laurig, W. (1990), ''ERGON-EXPERT: A Modular Knowledge-Based Approach to Reduce Health and Safety Hazards in Manual Materials Handling Tasks,'' in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 299–309.

Ryan, P. W. (1971), ''Cockpit Geometry Evaluation, Phase II, Vol. II,'' Joint Army–Navy Aircraft Instrument Research Report 7012313, Boeing, Seattle, WA.

Schaub, K., and Rohmert, W. (1990), ''HEINER Helps to Improve and Evaluate Ergonomic Design,'' in *Proceedings to the 21st International Symposium on Automotive Technology and Automation,* Vol. 2, pp. 999–1016.

Schaub, K., Landau, K., Menges, R., and Grossmann, K. (1997), "A Computer-Aided Tool for Ergonomic Workplace Design and Preventive Health Care," *Human Factors and Ergonomics in Manufacturing,* Vol. 7, No. 4, pp. 269–304.

Schoenmarklin, R. W., and Marras, W. S. (1991), "Quantification of Wrist Motion and Cumulative Disorders in Industry," *in Proceedings of the Human Factors Society 35th Annual Meeting,* Human Factors and Ergonomic Society, Santa Monica, CA.

Schultz, A., Andersson, G. B. J., Ortengren, R., Haderspeck, K., and Nathemson, A. (1982), "Loads on Lumbar Spine: Validation of a Biomechanical Analysis by Measurements of Intradiscal Pressures and Myoelectric Signals," *Journal of Bone and Joint Surgery,* Vol. 64-A, pp. 713–720.

Silverstein, B. A., Fine, L. J., and Armstrong, T. J. (1987), "Occupational Factors and Carpal Tunnel Syndrome," *American Journal of Industrial Medicine,* Vol. 11, pp. 343–358.

Smith, J. L., Ayoub, M. M., and McDaniel, J. W. (1992), "Manual Materials Handling Capabilities in Non-Standard Postures," *Ergonomics,* Vol. 35, pp. 807–831.

Snijders, C. J., van Riel, M. P. J. V., and Nordin, M. (1987), "Continuous Measurements of Spine Movements in Normal Working Situations over Periods of 8 Hours or More," *Ergonomics,* Vol. 30, No. 4, pp. 639–653.

Snook, S. H. (1978), "The Design of Manual Handling Tasks," *Ergonomics,* Vol. 21, pp. 963–985.

Snook, S. H., and Ciriello, V. M. (1974), "Maximum Weights and Workloads Acceptable to Female Workers," *Journal of Occupational Medicine,* Vol. 16, pp. 527–534.

Snook, S. H., and Ciriello, V. M. (1991), "The Design of Manual Handling Tasks: Revised Tables of Maximum Acceptable Weights and Forces," *Ergonomics,* Vol. 34, No. 9, pp. 1197–1213.

Snook, S. H., Irvine, C. H., and Bass, S. F. (1970), "Maximum Weights and Work Loads Acceptable to Male Industrial Workers," *American Industrial Hygiene Association Journal,* Vol. 31, pp. 579–586.

Snook, S. H., Vaillancourt, D. R., Ciriello, V. M., and Webster, B. S. (1995), "Psychophysical Studies of Repetitive Wrist Flexion and Extension," *Ergonomics,* Vol. 38, pp. 1488–1507.

Spengler, D. M. J., Bigos, S. J., Martin, N. A., Zeh, J., Fisher, L., and Nachemson, A. (1986), "Back Injuries in Industry: A Retrospective Study," *Spine,* Vol. 11, pp. 241–256.

Stevens, S. S. (1975), *Psychophsics: Introduction to its Perceptual, Neural, Social Prospects,* John Wiley & Sons, New York.

Strasser, H., Keller, E., Müller, K. W., and Ernst, J. (1989), "Local Muscular Strain Dependent on the Direction of Horizontal Arm Movements," *Ergonomics,* Vol. 32, pp. 899–910.

Svensson, H.-O., and Andersson, G. B. J. (1983), "Low-Back Pain in Forty to Forty-Seven Year Old Men: Work History and Work Environment," *Scandinavian Journal of Rehabilitative Medicine,* Vol. 8, pp. 272–276.

Svensson, H.-O., and Andersson, G. B. J. (1989), "The Relationship of Low-Back Pain, Work History, Work Environment and Stress: A Retrospective Cross-Sectional Study of 38- to 64-Year-Old Women," *Spine,* Vol. 14, pp. 517–522.

Tanaka, S., and McGlothlin, J. D. (1993), "A Conceptual Quantitative Model for Prevention of Work-Related Carpal Tunnel Syndrome (CTS)," *International Journal of Industrial Ergonomics,* Vol. 11, pp. 181–193.

Tichauer, E. R. (1978), *The Biomechanical Basis of Ergonomics,* John Wiley & Sons, New York.

Tracy, M. F. (1990), Biomechanical Methods in Posture Analysis, in Evaluation of Human Work, J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London.

Troup, J. D., and Edwards, F. C. (1985), *Manual Handling and Lifting,* HMSO, London.

University of Michigan (1989), *User Manual 3D Static Strength Prediction Program,* University of Michigan, Ann Arbor, MI.

Vayrynen, S., Ojanen, K., Pyykkonen, M., Peuranemi, A., Suurnakki, T., and Kemppainen, M. (1990), "OWASCA: Computer-Aided Visualizing and Training Software for Work Posture Analysis," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 273–278.

Videman, T., Nurminen, M., and Troup, J. D. (1990), "Lumbar Spinal Pathology in Cadaveric Material in Relation to History of Back Pain, Occupation, and Physical Loading," *Spine,* Vol. 15, pp. 728–740.

Wallace, M., and Buckle, P. (1987), "Ergonomic Aspects of Neck and Upper Limb Disorders," in *International Reviews of Ergonomics,* 1, D. Osborne, Ed., Taylor & Francis, London, pp. 173–200.

Wangenheim, M., and Samuelson, B. (1987), "Automatic Ergonomic Work Analysis," *Applied Ergonomics,* Vol. 18, pp. 9–15.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks," *Ergonomics,* Vol. 36, No. 7, pp. 749–776.

Waters, T. R., Putz-Anderson, V., and Garg, A. (1994), "Application Manual for the Revised NIOSH Lifting Equation," U.S. Department of Health and Human Services, Cincinnati.

Westgaard, R. H., and Bjørkland, R. (1987), "Generation of Muscle Tension Additional to Postural Muscle Load," *Ergonomics,* Vol. 30, No. 6, pp. 196–203.

Williams, M., and Lissner, H. R. (1977), *Biomechanics of Human Motion,* 2nd Ed., W. B. Saunders, Philadelphia.

Winter, D. A. (1979), *Biomechanics of Human Movement,* John Wiley & Sons, New York.

World Health Organization (WHO) (1985), "Identification and Control of Work-Related Diseases," Technical Report No. 174, WHO, Geneva, pp. 7–11.

ADDITIONAL READING

Aaras, A., Westgaard, R. H., and Stranden, E., "Postural Angles as an Indicator of Postural Load and Muscular Injury in Occupational Work Situations," *Ergonomics,* Vol. 31, 1988, pp. 915–933.

Ayoub, M. M., Mital, A., Bakken, G. M., Asfour, S. S., and Bethea, N. J., "Development of Strength and Capacity Norms for Manual Materials Handling Activities: The State of the Art," *Human Factors,* Vol. 22, No. 3, 1980, pp. 271–283.

Bonney, M. C., and Case, K., "The Development of SAMMIE for Computer-Aided Workplace and Work Task Design," in *Proceedings of the 6th Congress of the International Ergonomics Association,* Human Factors Society, Santa Monica, CA, 1976.

Brinckmann, P., Biggemann, M., and Hilweg, D. (1988), "Fatigue Fractures of Human Lumbar Vertebrae," *Clinical Biomechanics,* Vol. 3 (Supplement 1), 1988.

Burdorf, A., Govaert, G., and Elders, L. (1991), "Postural Load and Back Pain of Workers in the Manufacturing of Prefabricated Elements," *Ergonomics,* Vol. 34, 1991, pp. 909–918.

Chaffin, D. B., "A Computerized Biomechanical Model: Development of and Use in Studying Gross Body Actions," *Journal of Biomechanics,* Vol. 2, 1969, pp. 429–441.

Chaffin, D. B., "Human Strength Capability and Low-Back Pain," *Journal of Occupational Medicine,* Vol. 16, No. 4, 1974, pp. 248–254.

Chaffin, D. B., and Park, K. S., "A Longitudinal Study of Low Back Pain as Associated with Occupational Weight Lifting Factors," *American Industrial Hygiene Association Journal,* Vol. 34, No. 12, 1973, pp. 513–525.

Fernandez, J. E., and Ayoub, M. M., "The Psychophysical Approach: The Valid Measure of Lifting Capacity," in *Trends in Ergonomics/Human Factors V,* F. Aghazadeh, Ed., North-Holland, Amsterdam, 1988, pp. 837–845.

Garg, A., and Chaffin, D. B., "A Biomechanical Computerized Simulation of Human Strength," *IIE Transactions,* Vol. 14, No. 4, 1975, pp. 272–281.

Genaidy, A., and Karwowski, W.,"The Effects of Body Movements on Perceived Joint Discomfort Ratings in Sitting and Standing Postures," *Ergonomics,* Vol. 36, No. 7, 1993, pp. 785–792.

Genaidy, A. M., and Karwowski, W., "The Effects of Neutral Posture Deviations on Perceived Joint Discomfort Ratings in Sitting and Standing Postures," *Ergonomics,* Vol. 36, No. 7, 1993, pp. 785–792.

Genaidy, A., Al-Shedi, A., and Karwowski, W., "Postural Stress Analysis in Industry," *Applied Ergonomics,* Vol. 25, No. 2, 1994, pp. 77–87.

Genaidy, A., Barkawi, H., and Christensen, D., "Ranking of Static Non-neutral Postures around the Joints of the Upper Extremity and the Spine," *Ergonomics,* Vol. 38, No. 9, pp. 1851–1858.

Genaidy, A., Karwowski, W., Christensen, D., Vogiatzis, C., and Prins, A., "What Is 'Heavy'?" *Ergonomics,* Vol. 41, No. 4, 1998, pp. 320–432.

Gescheider, G. A. (1985), *Psychophysics: Method, Theory, and Application,* 2nd Ed., Erlbaum, London.

Grobelny, J., "Anthropometric Data for a Driver's Workplace Design in the AutoCAD System," in *Computer-Aided Ergonomics,* W. Karwowski, A. Genaidy, and S. S. Asfour, Eds., Taylor & Francis, London, pp. 80–89.

Herrin, G. D., Chaffin, D. B., and Mach, R. S., "Criteria for Research on the Hazards of Manual Materials Handling," in *Workshop Proceedings,* Contract CDC-99-74-118, U.S. Department of Health and Human Services (NIOSH), Cincinnati, 1974.

Hidalgo, J., Genaidy, A., Karwowski, W., Christensen, D., Huston, R., and Stambough, J. "A Cross-Validation of the NIOSH Limits for Manual Lifting," *Ergonomics,* Vol. 38, No. 12, 1995, pp. 2455–2464.

Hidalgo, J., Genaidy, A., Karwowski, W., Christensen, D., Huston, R., and Stambough, J. (1997), "A Comprehensive Lifting Model: Beyond the NIOSH Lifting Equation," *Ergonomics,* Vol. 40, No. 9, pp. 916–927.

Hsiao, H., and Keyserling, W. M., "Evaluating Posture Behavior During Seated Tasks," *International Journal of Industrial Ergonomics,* Vol. 8, 1991, pp. 313–334.

Hwang, C. L., and Yoon, K., *Multiple Attribute Decision Making: Methods and Applications,* Springer, New York, 1981.

Jäger, M., and Luttmann, A., "Biomechanical Analysis and Assessment of Lumbar Stress During Load Lifting Using a Dynamic 19-Segment Human Model," *Ergonomics,* Vol. 32, No. 1, 1989, pp. 93–112.

Jiang, B. C., Smith, J. L., and Ayoub, M. M., "Psychophysical Modelling of Manual Materials Handling Capacities Using Isoinertial Strength Variables," *Human Factors,* Vol. 28, No. 6, 1986, pp. 691–702.

Karwowski, W., Lee, W. G, Jamaldin, B., Gaddie, P., and Jang, R., "Beyond Psychophysics: A Need for Cognitive Modeling Approach to Setting Limits in Manual Lifting Tasks," *Ergonomics,* Vol. 42, No. 1, 1999, pp. 40–60.

Kee, D., "Measurement on Range of Two Degrees of Freedom Motion for Analytic Generation of Workspace," *Journal of the Ergonomics Society of Korea,* Vol. 15, 1997, pp. 15–24.

Keyserling, W. M., Herrin, G. D., Chaffin, D. B., Armstrong, T. J., and Foss, M. L., "Establishing an Industrial Strength Testing Program," *American Industrial Hygiene Association Journal,* Vol. 41, No. 10, 1980, pp. 730–736.

Kilbom, A., Persson, J., and Jonsson, B., "Risk Factors for Work-Related Disorders of the Neck and Shoulder—with Special Emphasis on Working Postures and Movements," in *The Ergonomics of Working Postures,* E. N. Corlett, J. R. Wilson, and I. Manenica, Eds., Taylor & Francis, London, 1986, pp. 243–257.

Kroemer, K. H. E., "COMBIMAN—*COM*puterized *BI*omechanical *MAN*-Model," Technical Report, AMRL-TR-72-16, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH, 1973.

Kroemer, K. H. E., "Engineering Anthropometry: Designing the Workplace to Fit the Human," in *Proceedings of the Annual Conference of the American Institute of Industrial Engineers* (Norcross, GA, 1981), pp. 119–126.

Kroemer, K., Kroemer, H., and Kroemer-Elbert, K., *Ergonomics: How to Design for Ease and Efficiency,* Prentice Hall, Englewood Cliffs, NJ, 1994.

Lawrence, J. S., "Rheumatism in Coal Miners: Part III. Occupational Factors," *British Journal of Industrial Medicine,* Vol. 12, 1955, pp. 249–261.

Leamon, T. B., "Research to Reality: A Critical Review of the Validity of Various Criteria for the Prevention of Occupationally Induced Low Back Pain Disability," *Ergonomics,* Vol. 37, No. 12, 1994, pp. 1959–1974.

Legg, S. J., and Haslam, D. R., "Effect of Sleep Deprivation on Self-Selected Workload," *Ergonomics,* Vol. 27, No. 4, 1984, pp. 389–396.

Marras, W. S., and Schoenmarklin, R. W., "Wrist Motions in Industry," *Ergonomics,* Vol. 36, No. 4, 1993, pp. 341–351.

Martin, J. B., and Chaffin, D. B., "Biomechanical Computerized Simulation of Human Strength in Sagittal Plane Activities," *AIIE Transactions,* Vol. 4, No. 1, 1972, pp. 19–28.

Morrissey, S. J., Bittner, A. C., and Arcangeli, K. K., "Accuracy of a Ratio-Estimation Method to Set Maximum Acceptable Weights in Complex Lifting Tasks," *International Journal of Industrial Ergonomics,* Vol. 5, 1990, pp. 169–174.

Murrell, K., *Ergonomics,* Chapman & Hall, London, 1969.

Park, K. S., and Chaffin, D. B., "A Biomechanical Evaluation of Two Methods of Manual Load Lifting," *AIIE Transactions,* Vol. 6, No. 2, 1974, pp. 105–113.

Priel, V. C., "A Numerical Definition of Posture," *Human Factors,* Vol. 16, 1974, pp. 576–584.

Rohmert, W., *Untersuchungen über Muskelermüdung und Arbeitsgestaltung,* Beuth-Vertrieb, Berlin, 1962.

Ryan, G. A., "Musculo-skeletal Symptoms in Supermarket Workers," *Ergonomics,* Vol. 32, 1989, pp. 359–371.

Sanders, M., and McCormick, E., *Human Factors In Engineering Design,* McGraw-Hill, New York, 1987.

Shoaf, C., Genaidy, A., Karwowski, W., Waters, T., and Christensen, D., ‘‘Comprehensive Manual Handling Limits for Lowering, Pushing, Pulling and Carrying Activities,’’ *Ergonomics,* Vol. 40, No. 11, 1997, pp. 1183–1200.

Snook, S. H., ‘‘The Costs of Back Pain in Industry,’’ *Occupational Medicine: State of the Art Reviews,* Vol. 3, January–March, 1988, pp. 1–5.

Stevens, S. S., Mack, J. D., and Stevens, S. S., ‘‘Growth of Sensation on Seven Continua as Measured by Force of Hand-Grip,’’ *Journal of Experimental Psychology,* Vol. 59, 1960, pp. 60–67.

Taboun, S. M., and Dutta, S. P., ‘‘Energy Cost Models for Combined Lifting and Carrying Tasks,’’ *International Journal of Industrial Ergonomics,* Vol. 4, No. 1, 1989, pp. 1–17.

Webb Associates, *Anthropometric Source Book,* Vol. 1, Ch. 6, NASA Ref. 1024, 1978.

CHAPTER 41

Ergonomics in Digital Environments

ULRICH RASCHKE
Engineering Animation, Inc.

LISA M. SCHUTTE
Engineering Animation, Inc.

DON B. CHAFFIN
The University of Michigan

1. INTRODUCTION 1112

2. DIGITAL HUMAN FIGURES 1112
  2.1. Kinematic Representation 1112
  2.2. Anthropometry 1113
    2.2.1. Anthropometric Databases 1113
    2.2.2. Accommodation Methods 1113
  2.3. Human Figure Posturing 1115
    2.3.1. Coupled Joints 1115
    2.3.2. Inverse Kinematics 1115
  2.4. Motion/Animation 1116

3. HUMAN PERFORMANCE MODELS 1116
  3.1. Strength 1116
  3.2. Fatigue/Metabolic Energy Requirements 1118
  3.3. Low-Back Injury Risk 1119
  3.4. Comfort 1120
  3.5. Motion Timing 1120

4. ERGONOMIC ANALYSIS IN DIGITAL ENVIRONMENTS 1120
  4.1. Workplace Analysis 1120
    4.1.1. Setting up the Workplace Environment 1120
    4.1.2. Identify Test Population Anthropometry 1120
    4.1.3. Accurately Posture (or Animate) the Figures at the Workplace 1121
    4.1.4. Service Analysis 1121
  4.2. Product Design 1121
    4.2.1. Accommodation 1122
    4.2.2. Definition of the Test Population Anthropometry 1122
    4.2.3. Usability 1123

5. IMMERSIVE VIRTUAL REALITY 1124
  5.1. Motion-Tracking Technologies 1125

6. HUMAN SIMULATION CHALLENGES 1125
  6.1. Performance Models 1126
    6.1.1. Performance Factors 1126
    6.1.2. Variation Modeling 1126
  6.2. Human Motion Control 1126
    6.2.1. Modeling Motion Data 1127
    6.2.2. Multiple-Figure Interactions 1127

7. CONCLUSION 1127

REFERENCES 1127

ADDITIONAL READING 1129

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

The design of workplaces and products continues to migrate from paper to the computer, where analysis accuracy, visualization, and collaboration utilities allow designs to be realized much faster and better than ever before. As the pace of this development accelerates with the increased capabilities of the software design tools, less time is spent on physical prototyping, allowing for shortened time-to-market for new products. Ergonomists, who in the past used the physical prototypes to perform human factors analyses, are now challenged to move the analysis into the virtual domain using new tools and methods. Usability, maintainability, physical ergonomic assessments, psychological perception, and procedural training are some of the human factors issues that might benefit from analysis prior to the first physical incarnation of the design. While this represents a challenge for the ergonomists, it provides an opportunity to effect change in the designs much earlier than was typically possible in the past and to take advantage of the dramatically reduced cost of design alterations in the early design phases. Commercial pressures that leverage the cost benefits offered by complete ‘‘in-tube’’ design are driving a rapid development of the available computer technologies. Human simulation technology is no exception. Contemporary human modeling software is assimilating a variety of human modeling knowledge, including population anthropometry descriptions and physical capability models. Companies are deploying these human modeling products to allow their ergonomists and designers to populate digital representations of products and workplaces efficiently with virtual human figures and ask meaningful questions regarding the likely performance of actual people in those environments. Identification of ergonomic design problems early in the design phase allows time-consuming and expensive reworking of the manufacturing process or design to be avoided.

Computerized human modeling itself has been evolving over some time. Perhaps the first attempt to develop a computer-integrated tool for performing reach tasks was performed by Vetter and Ryan for the Boeing Aircraft company in the late 1960s. This effort was referred to as the ‘‘First Man’’ program, which later became ‘‘Boeman.’’ This software was later expanded by the USAF Aerospace Medical Research Laboratory Crew Systems’ Interface Division, which added the ability to simulate a variety of male and female anthropometric dimensions while seated in different types of aircraft, culminating in the software COMBIMAN. In the 1980s, this software was further developed at AMRL to address maintenance tasks, adding performance models of lifting, pulling, and pushing on various tools and objects placed in the hands, and became CrewChief. During this same time in Europe, a wide variety of models were developed, perhaps the most widely known being SAMMIE (System for Aiding Man–Machine Interaction Evaluation), developed by Case, Porter, and Bonney at Nottingham and Loughborough Universities in the United Kingdom. SAMMIE was conceived as a very general model for assessing reach, interference, and sight-line issues within a CAD environment. The details of these developments are described in greater depth elsewhere (e.g., Chaffin 2000; Bubb 1999; Badler 1993). Perhaps as a testament to the rapid development in this field, new human models that are integrated in modern CAD, 3D visualization, and automation simulation products are now the most popular and seeing the most rapid development and deployment. These include Deneb Ergo, EAI Jack, Genicom Safeworks, TecMath Ramsis, and Tecnomatix RobCAD Man.

This chapter reviews the foundation of contemporary human modeling technology for physical ergonomics and presents examples of how digital humans are currently used in industry. The chapter concludes with a discussion of the current development efforts in the area of human modeling.

2. DIGITAL HUMAN FIGURES

2.1. Kinematic Representation

Human models are varied in both their complexity and construction. Any mathematical representation of human structure, physiology, or behavior can be considered to be a human model. For example, complex models of human musculoskeletal dynamics are commonly used to study motor control issues (Winters and Woo 1990). These models are typically quite detailed to allow the dynamic effects of individual muscle activation and contraction, and hypothesized neural control strategies, to be investigated. Moreover, this detail is typically focused on one part of the body, for example the lower extremity for gait analysis, or the upper limbs for investigation of movement control. In contrast, simple, sometimes incomplete, human forms are used in the investigation of cognitive models, wherein the human form acts as an agent to effect changes in its world. The pedagogical agent developed at the University of Southern California Information Sciences Institute (Johnson et al. 2000) is an example. These models focus on the cognitive rather than motor processes and simulate the interactions among multiple humans.

For physical ergonomics investigations in digital environments, the human models need to mirror our structure, shape, and size in sufficient detail to allow the figures to assume realistically the observed postures of actual individuals performing similar tasks. Such models typically consist of an underlying kinematic linkage system that closely parallels our own skeletal structure and an attached geometric shell that duplicates our surface shape.

Today's human models have kinematic linkages that include from 30 to 148 degrees of freedom, depending on the detail provided in the hands, feet, shoulder, and spine. The joints are constructed to move like our own joints, with the appropriate number of degrees of freedom, and typically also have physiological limits on the range of motion. In more detailed models, the shoulder and spine are modeled to behave naturally, with the links moving in concert as the particular segment is manipulated. For example, the shoulder complex consisting of the sternoclavicular, acromioclavicular, and glenohumeral joints is modeled to move in a realistic pattern when the upper arm is adjusted, moving the elevation and fore–aft position of the shoulder as the arm moves through its range of motion.

2.2. Anthropometry

The internal skeletal structure and surface topography of a digital human figure influence both the qualitative and quantitative use of the figures. As an engineering tool, the accuracy of the internal link structure affects the dimensional measures made between the environment and the human figure, such as head clearance and reach. The ability of the figure to take on physiologic surface topography directly adds to the perception of reality when one is viewing a simulation. While both of these aspects are important, to date more effort has been concentrated on the accurate scaling of the link lengths in commercial human modeling. This bias is in part motivated by the large amount of traditional 1D anthropometric data available (e.g., stature, sitting height, shoulder breadth), in contrast to the largely nonexistent 3D surface contour data available. Secondly, a driving factor of human modeling in visualization environments has been to produce a system that works in near real time (Badler et al. 1993). The complexity of the figure surface description presents a burden on the real-time performance, so a balance is sought in which there is sufficient surface detail for visual reality without undue computational load. As computer hardware technology improves, the ability to add to this surface complexity is afforded.

2.2.1. Anthropometric Databases

Of the many anthropometric databases available, one of the most widely used is the U.S. Army 1988 Anthropometric Survey (ANSUR) (Gordon et al. 1988). The ANSUR study was performed by the U.S. military to provide a representative anthropometric database of the U.S. military personnel. This database has a demographic representation that matches the U.S. army, which is known to differ from the gender, racial, age, and conditioning distributions of the population as a whole. Nevertheless, the statistical measures of the ANSUR data have been estimated to be within 3% of the general U.S. civilian population (Roebuck 1995). This study contains 132 standard anthropometric measurements from approximately 9000 military personnel, of which a sample of 1774 men and 2208 females were selected to represent accurately the military population demographics. Documents that contain the individual subject data as well as the summary statistics are publicly available, so publishers of human modeling software can independently develop statistical models for figure scaling and boundary manikin generation.

Another anthropometric database available is the National Health and Nutrition Examination Survey III (NHANES 1994), which contains the dimensions of 33,994 persons ages 2 months and older, of which 17,752 are age 18 and older. While the 21 measures of this database do not provide enough information to define adequately the dimensions of most contemporary human models, the database currently represents the most comprehensive and representative database for the U.S. population. These publicly available data contain weighting information based on the most recent U.S. census (1988–1994). The census weighting data allow U.S. representative population statistics to be computed for any population selections based on gender, race, ethnicity, and age. While both the ANSUR and NHANES data describe single dimension measures taken between anthropometric landmarks, a new anthropometric survey has been initiated to provide a database of population 3D body shapes. The CAESAR project (Civilian American and European Surface Anthropometric Resource) will scan approximately 6000 individuals in the United States and Europe. These data are in the form of both traditional anthropometric measures and new 3D data from whole body laser scanners, that provide a highly detailed data cloud describing the shape of the subject surface contour (Figure 1).

Both children- and nationality-specific anthropometric databases are also available, although these databases have not been adopted to the same degree as those previously mentioned due to their limited international availability and data restrictions (Table 1).

2.2.2. Accommodation Methods

One of the advantages digital ergonomics can bring to the development process is the ability to investigate accommodation issues early in the design process. In the past, physical mockups were created and evaluated using a large subject population to arrive at accommodation metrics. This approach is both expensive and time consuming and does not lend itself to rapid evaluation of design alternatives. In the digital space, a population of figures can be used to investigate many of the same

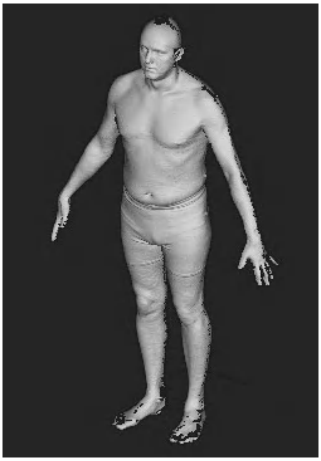


Figure 1 The CAESAR Data are Collected with Laser Body Scanners That Produce a 3D Point Cloud of Surface Topography. These data will allow accurate measures in such applications as accommodation analysis and clothing fit.

issues of clearance, visibility, reach, and comfort. Defining the sizes of the manikins to be used in the process is one of the first steps of these analyses.

To perform an accommodation study, the user defines a percentage of the population that he or she wishes to accommodate in their design or workplace and then scales representative figures using data from applicable anthropometric databases to represent the extreme dimensions of this accommodation range. As few as one measure, often stature, may be judged to be important to the design and used in the creation of the representative figures. For other applications, such as cockpit design, multiple measures, such as head clearance, eye height, shoulder breadth, leg length, and reach length, may all affect the design simultaneously. Several methods are in use to investigate the accommodation status of a design and are described below.

2.2.2.1. Monte Carlo Simulation The Monte Carlo approach randomly samples subjects from an anthropometric database to create a representative sample of the population and processes these figures through the design. Recording the success or failure of the design to accommodate each

TABLE 1 Sample of Recent Anthropometric Databases Used by Contemporary Human Models

Name	Dimensions	N (population and ages)	Survey Year	Availability
ANSUR 88	132	9,000 (U.S. Army)	1987	Summary statistics and individual subject data
NHANES III	21	33,994 (U.S. civilian ages 2 mo to 99 years)	1988–1994	Individual weighted subject data
CPSC Children	87	4,127 (U.S. children ages 2 weeks–18 years)	1975–1977	Summary statistics and individual subject data
CAESAR—3D	44 Traditional and 3D surface	∼6,000 (U.S. and European)	Present	Planned: individual 3D landmark locations, summary statistics, 3D point cloud data
HQL—Japan	178	40,000 (Japanese ages 7–99 years)	1992–1994	Summary statistics and portions of individual subject data
KRISS—Korea	84	8,886 (Korean ages 6–50 years)	1992 (1997)	Dimensional statistics with display software

individual of the sample allows an indication of the percentage accommodation to be derived. This method is computationally fairly expensive because it requires that a large number of figures be generated and tested for a meaningful analysis. Also, because the distribution of sizes follows a bell-shaped distribution, many more people are close to the average than to the extremes, which results in many figures of fairly similar dimensions needlessly tested. This can make this approach somewhat inefficient.

2.2.2.2. Boundary Manikins In contrast, the boundary manikin approach can be used for multiple dimensional analysis (Zehner et al. 1993; Bittner et al. 1986). The statistics of principal components (factor analysis) can be used to identify bounding multidimensional ellipsoids that contain a portion of the population. For example a 95% hyperellipsoid can be found that defines the dimensional ranges for all variables of interest within which 95% of the population can be expected. This approach can reduce the number of manikins that need to be tested in a design to a practical number, which is beneficial in cases where the computer rendering speed or the effort to manually posture the figure is significant.

2.2.2.3. Whole-Population Analysis Both the Monte Carlo and boundary manikin approaches attempt to reduce the number of subjects that are run through the analysis while still providing statistical validity to the results. However, as computer technology improves and as models to posture the manikins realistically in the environment become available, it becomes not unreasonable to run several thousand figures through a design automatically. Because this approach does not use statistical models of the data but instead uses the measured data directly, the unexplained variability that is not captured by the statistical data modeling is avoided. This approach still requires substantial run time and is currently not a popular approach to accommodation analysis.

2.3. Human Figure Posturing

As mentioned briefly in the previous anthropometric discussion, figure posturing is a critical component, along with anthropometry, in accommodation studies. While automatic posturing methods based on empirical studies are becoming available and will be discussed in later sections, there are more fundamental lower-level tools that have been developed for general manipulation of figures in the virtual environments. Because contemporary figures may have over 100 DOF, adjustment of each joint angle individually is unworkably tedious.

2.3.1. Coupled Joints

Fortunately, the human skeleton, while infinitely adjustable, is held together by muscles and connective tissues that constrain the movement of certain joints to work as motion complexes. Examples include the shoulder, spine, and hands. While there are obvious freakish exceptions, most people cannot voluntarily dislocate their shoulders or move the carpal bones in the digits independently. Modelers take advantage of these relationships to build movement rules for these joints such that many degrees of freedom can be adjusted easily with a few angles that are defined in common human factors parlance. For example, the EAI Jack human model has 54 DOF in the spine below the neck, which can be manipulated using three standard angle definitions: flexion, extension, and axial rotation. Similarly, as described earlier, the shoulder comprises a clavicle link that moves in concert with the arm as it is manipulated, mimicking the natural kinematics of this joint. Such coupled joint complexes greatly simplify the posturing of high-degree of freedom human figures.

2.3.2. Inverse Kinematics

Even with the substantial reduction in degrees of freedom that coupled joints bring, there are still far too many degrees of freedom remaining in a contemporary figure for rapid posturing in production use. To address this, human modelers have drawn from the robotics field the concept of inverse kinematics (IK) or specifying joint kinematics based on a desired end-effector position. Inverse kinematics operates on a linked chain of segments, for example the torso, shoulder, arm, forearm, and wrist, and, given the location of the distal segment (i.e., hand), solves all of the joint postures along this chain based on some optimization criteria. For human models, these criteria include that the joints do not separate and that the joint angles remain within their physiological range of motion. Using inverse kinematics, the practitioner is able to grab the virtual figure's hand in the 3D visualization environment and manipulate its position in real time while the rest of the figure modifies its posture (i.e., torso, shoulder, arm) to satisfy the requested hand position. While the IK methods can be made to respect the physiologic range of motion limitations inherent to the joints, they tend not to have the sophistication always to select the most likely or physiologically reasonable postures. This is especially problematic when the number of joints in the joint linkage is large. If the number of degrees of freedom is too great, there is unlikely to be just one unique posture that satisfies the specified end-effector position. For specific cases, this is being addressed with empirical-based posturing models, which are discussed in greater detail below. However, even with the caveat that IK sometimes selects inappropriate postural solutions, it is currently the most popular and rapid method of general postural manipulation in 3D environments.

2.4. Motion/Animation

While static posturing is often sufficient to analyze many ergonomic issues, such as reach, vision, clearance, and joint loading, often figure motion in the form of an animation is important. Examples include simulated training material, managerial presentations, and analyses that depend on observations of a person performing an entire task cycle, for example when assessing the efficiency a workplace layout. Realistically controlling figure motion is without question one of the most challenging aspects of human modeling. Humans are capable of an almost infinite number of different movements to accomplish the same task. Indeed, people may use several postural approaches during a single task, for example to get better leverage on a tool or gain a different vantage point for a complex assembly. This incredible postural flexibility makes it very difficult for human modeling software to predict which motions a worker will use to perform a task. Most current animation systems circumvent this dilemma by requiring the user to specify the key postures of the figure during the task. The software then transitions between these postures, driving the joint angles to change over time such that motions conform to correct times. A variety of mechanisms are used to perform the posture transitions, from predefined motion rules to inverse kinematics. Depending on the system, the level of control given to the user to define and edit the postures also varies, with some products making more assumptions than others. While built-in rules offer utility to the novice user, the inflexibility imposed by the system automatically selecting task postures can be restrictive and a source of frustration to the advanced user. In addition, the level of fidelity required in the motion varies greatly depending on the application. For applications such as the validation of a factory layout or animation of a procedure for training or communication purposes, a human motion simulation that simply looks reasonable may be sufficient. However, if sophisticated biomechanical analyses are to be run on the simulated motion, it may be necessary to generate motions that are not only visually reasonable but also obey physiologic rules. These include accurate joint velocities and accelerations, realistic positioning of the center of mass relative to the feet, and accurate specification of externally applied forces.

3. HUMAN PERFORMANCE MODELS

Human capability analysis is one of the primary motivations for simulation. Commercial modelers have implemented performance models from the academic literature into their software, taking advantage of the human figure sophistication and real-time visualization technologies. A review of the commonly available performance models reveals that independent research groups largely developed them. The development diversity is reflected in the variety of inputs that these tools require in order to provide an ergonomic assessment. This represents a challenge to the modelers as they work to seamlessly integrate these assessment tools into their animation and simulation offerings. Some tools lend themselves well to integration, such as those that can capture all required information from posture, figure, and load mass characteristics. Typically these are the tools that have as their foundation biomechanical models from which the inputs are derived. Others, which were originally intended to be used with a checklist approach, are more challenging in that they often require complex questions to be answered that are straightforward for a trained ergonomist but quite complex for a computer simulation system (Table 2).

Most often, simulation engineers expect to ask human performance questions of their simulation without having to redescribe the simulation in the language of the tool. Thus, ideally, the tools are implemented such that they can derive all the necessary information from the simulation directly. Less ideally, the engineer performing the assessment must explicitly identify details of the simulation using tool specific descriptors.

3.1. Strength

Strength assessments are a typical human performance analysis, regardless of whether the application involves manual handling tasks, serviceability investigations, or product operation. Questions of strength can be posed in a variety of ways. Designers may want to know the maximum operating force for a lever, dial, or wheel such that their target demographic will have the strength to operate it. Or the engineer may create a job design and might ask what percentage of the population would be expected to have the strength to perform the required tasks of the job. Strength data might also be used to posture virtual human figures by using an algorithm that optimally adjusts the individual joints of the manikin to produce most effectively the required forces for a task.

A large amount of strength data has been collected over the past quarter century. Most of these have been in the form of maximal voluntary exertions (MVEs) of large muscle groups. Subjects are placed in special strength testing devices (e.g., strength chairs) to isolate individual muscle groups, and standard methods controlling for repeatability and duration of effort are then used to capture the

TABLE 2 Partial List of Performance Tools Available in High-End Human Simulation Tools[a]

Performance Model	Data Source	Input Parameters	Integration Issues
NIOSH lifting equation	Waters et al. 1993	Posture and lift begin and end, object weight, hand coupling	Must determine start and end of lift. Must identify hand coupling
Low-back injury risk assessment	See Chaffin et al. 1999	Joint torques, postures	Suitable
Strength assessment	University of Michigan static strength equations	Body posture, hand loads	Suitable
	Burandt 1978 and Schultetus 1987	Body posture, hand loads	Table lookup
	Ciriello and Snook 1991	Task description, hand coupling, gender	Table lookup
	CrewChief	Gender, body size, posture, task condition	Suitable
Fatigue analysis	Rohmert 1973a, b; Laurig 1973	Joint torques, strength equations	Suitable
Metabolic energy expenditure	Garg et al. 1978	Task descriptions, gender, load description	Must identify the type of task motion (i.e., lift, carry, arm work, etc.).
Rapid upper limb assessment	McAtamney and Corlett 1993	Posture assessment, muscle use, force description	Must identify muscle use and force descriptions.
Ovako working posture	Karhu et al. 1977	Posture assessment	Suitable
Comfort	Variety of sources, including Dreyfuss 1993; Rebiffé 1966; Krist 1994	Posture assessment	Suitable

[a] The tools require different types of input that often cannot be accurately deduced from an animation sequence, requiring tool specific user input.

strength levels accurately. Strength can also be assessed while taking into account a subject's perception of the effort. These data, in which subjects' impression of the load strain is included, are called psychophysical strength data. They differ from the maximal voluntary exertions in that they are more task specific. Subjects are asked to identify the load amount they would be comfortable working with over the duration of a work shift. Typically, these are laboratory studies in which mockups of tasks very close to the actual work conditions are created and subjects are given an object to manipulate to which weight can be added. The worker has no knowledge of the weight amount (bags of lead shot in false bottoms are often used), and experimental techniques are employed to converge on the weight that the subject feels comfortable manipulating over the course of a workday. The data of Ciriello and Snook (1991) are of this nature. Finally, a few dynamic strength data have been collected. These data are complex to collect and for this reason are also the most scarce. Specific dynamic strength-testing equipment is required to control for the many variables that affect dynamic strength, including the movement velocity and posture. As a consequence of the data-collection limitations, these data are also quite restricted in terms of the range of conditions in which they can be applied, such as the prediction of dynamic strength capability for ratchet wrench push and pull operations (Pandya et al. 1992). Lately, the rise of cumulative trauma injuries for the lower arm, wrist, and hand has created a need for strength data specifically pertaining to the hand and fingers and estimates of hand fatigue. A extensive amount of data is available on grip strengths in various grip postures, but these data, because they do not adequately describe the distribution of exertion loads on the individual digits, do not lend themselves well to the estimation of hand fatigue issues. This problem is compounded by the challenge of accurately posturing the hands in digital models. There are many bones and joints that allow the complex movement of the fingers, most of which are included in contemporary human models. For example, the Jack human model has a total of 69 segments and 135 DOF, of which 32 segments and 40 DOF are in the hands alone. While a solution to the manipulation of these many degrees of freedom is presented in the section describing motion-tracking technologies, models that are able to analyze the postures and gripping conditions are still needed before hand fatigue issues can be addressed.

Whole body strength data in contemporary human models are available in a range of forms, from simple data lookup tables to statistical equations that are used in conjunction with biomechanical models to drive a measure of population strength capability. In the United States, perhaps the most widely used strength data are from the University of Michigan Center for Ergonomics (3DSSPP) and Liberty Mutual Insurance Co. (Ciriello and Snook 1991). Within the defense industry, the CrewChief strength data are also popular because the modeled strengths were obtained from military-related maintenance tasks. In Europe, particularly Germany, the data of Burandt and Schultetus are often used. As mentioned previously, a few of these data were obtained without the intent to incorporate them into human models. Instead, the data are presented in tables indexed by such factors as loading condition and gender. Those data that were collected and described with a focus toward eventual human model inclusion tend to be formulated such that all relevant information needed for a strength assessment can be deduced from the human model mass, loading, and posture information. These strength models now are very attractive to the human modeling community because they afford the real-time assessment of strength issues during a simulation without the user having to identify data-specific parameters or conditions (Table 2).

As discussed above, the availability of dynamic strength data is limited. The narrow scope of applications to which these data can be applied restricts their attractiveness to human modelers and the user community. An interesting associated note regarding these dynamic data and human models is that even if these data were more complete, the difficulty in accurately determining movement velocities from simulations would affect their use. Unless the virtual human motions are defined via motion-capture technology, the designer's guess of the movement speeds is fraught with error. Even under conditions in which actual motion capture data are used to animate the virtual figures, the velocities are derived by differentiation of the position information, which can result in noisy and unreliable input to the dynamic strength predictors. However, because people undeniably move during work and dynamic strength capability can differ greatly from static, this is clearly an area that will likely see research and technological attention in the near future.

3.2. Fatigue/Metabolic Energy Requirements

Once a simulation of a virtual human performing a task has been created, questions regarding the fatigue of the worker are commonplace. Can the worker be expected to perform at this cycle rate, or do we have to decrease the line rate or otherwise reengineer the task to avoid worker fatigue? Research suggests that whole-body fatigue increases the risk of musculoskeletal injury through premature decrease in strength. Unfortunately, the available empirical data are largely inadequate to predict a worker's fatigue level accurately. The lack of data can be explained by the large number of variables that affect fatigue, including exertion level, dynamism of the exertion, muscle temperature, previous exertion levels, the muscle groups involved, and individual conditioning. Nevertheless, two approaches are currently in modeling practice to provide at least some level of quantitative fatigue assessment for a work task. The strongest of these from a data perspective is the use of empirical metabolic energy prediction equations, in particular the equations published by Garg et al. (1978). These equations model a series of typical industrial materials-handling motions, such as walking, lifting, carrying, reaching, and arm work. Based on motion-specific parameters and load amount, the mathematical models provide an estimate of the energy consumption in kcal/min. These data were validated on a broad range of manual handling activities and were shown to predict actual energy-consumption rates well. The energy-consumption rate can be compared with accepted standards for exertion levels, such as the NIOSH 1991 recommended limits. The guideline put forth in the development of the NIOSH 1991 lifting equation recommends a limit of 33% of the maximum aerobic capacity of 9.5 kcal/min for healthy individuals performing whole body lifts over an eight-hour day (3.1 kcal/min). For work that involves mostly the arms, NIOSH recommends a reduction of 30% from this level or approximately 2.1 kcal/min (Waters et al. 1993). If the simulated task is found to require a higher energy-consumption rate than the recommended limit, it is assumed that the task is fatiguing and must be modified.

One of the challenges for modelers in using these energy-expenditure models is in the ability to deduce automatically which equations apply to the particular motion under simulation and then to provide the appropriate equation parameters. Some models include these data as a separate tool wherein the user explicitly defines the simulation in terms of the motion sets defined and modeled by Garg et al. A criticism of the approach regardless of implementation is that the granularity of the analysis is large, making it difficult to identify the particular body area that is fatigued, and that the data do not provide information on a broad enough range of activities.

In contrast to this approach, a variety of endurance equations may be used to estimate the amount of time static exertions can be held (see Van Dieën and Vrielink 1994). These equations describe the amount of time subjects can perform static exertions at various levels of effort relative to their maximum capability. Relevant to industrial work, some of these include the effects of interspersed rest periods (Sjogaard et al. 1988). Equations to describe the amount of time required to recover from these exertions were published by Rohmert (1973a, b) and Laurig (1973). If the estimated amount of time needed to recover from an exertion exceeds the amount of time available during a job cycle, then fatigue is assumed to accumulate. The endurance relations are applied to each body area separately, requiring an estimate of exertion level, or percentage of maximum capability, at these areas. While the original subjects were strength tested to derive their strength capability, these data are not available for workers in general and an estimate of strength capability must be used. One solution is to use biomechanically based strength models. A task simulation is analyzed with regard to the postures adopted by the virtual worker, and an estimate is given to the amount of time the worker spends in each of the postural conditions. The level of exertion required is estimated utilizing the strength equations, and this information is input to the endurance equations to provide the recovery time estimate.

While the methodologies for using these endurance data within the modeling tools have been implemented and are in use, the endurance data themselves are limited, as mentioned earlier. Gender and age effects are not taken into account, nor are most of the multitude of other factors that influence fatigue. Only the exertion level under static conditions is considered. However, the need to predict quantitative assessments of worker fatigue in simulations is high enough that users of human models look for ways to obtain a metric of fatigue, working around the limitations of the foundation data. Toward this end, joint use of the energy expenditure equations, endurance equations, and stress analysis using the strength tools will currently provide the best estimate of the task injury potential from fatigue.

3.3. Low-Back Injury Risk

Low-back injury is estimated to cost the U.S. industry tens of billions annually through compensation claims, lost workdays, reduced productivity, and retraining needs (NIOSH 1997; Cats-Baril and Frymoyer 1991; Frymoyer et al. 1983). Approximately 33% of all workers' compensation costs are for musculoskeletal disorders. Experience has shown that these injuries can be avoided with the proper ergonomic intervention. Biomechanical models available can be used for job analysis either proactively, during the design phase, or reactively in response to injury incidence, to help identify the injurious situations. The most common types of injury-assessment analyses performed using human models include low-back compression force analysis and strength analysis.

Low-back pain has been well researched over the past 20 years, including epidemiological studies that have identified spinal compression force as one of the significant predictors of low-back injury. In response, sophisticated biomechanical models have been developed to estimate this compression force accurately, taking into account not only the weight of the object and the body segments but also internal forces generated by the musculature and connective tissues as they balance the external loads (e.g., Nussbaum et al. 1997; Raschke et al. 1996; Van Dieën 1997). These internal contributions to the spinal forces can be an order of magnitude larger than the applied loads. NIOSH has recommended guidelines against which the predicted compression forces can be compared and job-design decisions can be made.

3.4. Comfort

Assessment of worker comfort using digital models can be based on both posture and performance model analysis. However, since comfort is influenced by a wide variety of interacting factors, these tools are largely insufficient to quantify the perception of comfort with accuracy. Most comfort studies performed to date have been centered around a specific task, such as VDT operation or vehicle driving (Rebiffé 1966; Grandjean 1980; Porter and Gyi 1998; Krist 1994; Dreyfuss 1993). Within the boundaries of these tasks, subjects are observed in different postures and asked to report on their comfort via a questionnaire. The joint angles are measured and correlated with the comfort rating to arrive at a postural comfort metric. Because these data are mostly collected under specific, often seated, task conditions, some caution is required to apply these to the analysis of comfort in standing postures such as materials-handling operations. In addition to the posture-based comfort assessment, a variety of the performance tools can be used to help with the assessment of comfort, including strength capability, fatigue, and posture duration information. Certainly the multifactorial nature of the comfort assessment makes it challenging, and perhaps for this reason it is seldomly used in the analysis of physical tasks.

3.5. Motion Timing

A typical simulation question regards the estimated time it will require a person to perform a task. Digital human models can draw on a wealth of predetermined time data available. Motion-timing data are collected in studies where elemental motions (walk, turn, reach, grasp, etc.) are observed performed by skilled operators in the workplace, and timed using a stopwatch. The motion time data are then published in an easily indexed form with associated movement codes. The best known of these is the methods time measurement (MTM-1) system published by the MTM Association. This system has the advantage that it has a large number of elemental motions defined, allowing for a precise partitioning of the work motions within a job task and subsequent accurate assessment of the movement time. One drawback of this high resolution is that considerable overhead is required to break the motion into the elemental movements. To address this, the MTM association as well as other groups have published grosser movement times, which combine several elemental movements into one. Several of the human modeling solutions now provide simulation solutions that can define movement duration with input from these movement time systems.

4. ERGONOMIC ANALYSIS IN DIGITAL ENVIRONMENTS

The large cost of worker injury, in both social and economic terms, has motivated considerable research in the development of models that predict potentially injurious situations in the workplace. According to the Bureau of Labor Statistics (1999), 4 out of 10 injuries and illnesses resulting in time away from work in 1997 were sprains or strains. In the following sections, the key steps in a human modelings based ergonomic assessment are outlined.

4.1. Workplace Analysis

4.1.1. Setting up the Workplace Environment

The first step to using the human simulation technology typically involves the construction of the work area to be analyzed. While pressures for integration of the CAD, process simulation, and human modeling solutions is paramount in the marketplace, at present the geometry data are mostly created in an external piece of software to the human simulation tool. This work cell layout, part and tooling geometry is mostly imported to the human modeling software from these external systems via standard file formats (e.g., IGES, VRML, STL). If the digital form of these data is not available, as may be the case in an analysis of an older existing workplace, primitive geometry-creation tools available in the human simulation environment can be used to mock up the critical parts.

4.1.2. Identify Test Population Anthropometry

Most companies have design criteria that define the percentage of the population that must be accommodated by their product and manufacturing designs. For example, all individuals ranging from a small female (5% in stature) to a large male (95% in stature) might be prescribed. Often only the extremes are tested, but a more comprehensive practice includes a figure with average-sized proportions as well because it may help to identify unexpected stature-dependent postural effects. Under more complex types of analyses that may include other ergonomic factors such as reach, the range of anthropometric dimensions comprising the digital figures (known as a cadre family) can be selected through sophisticated multidimensional statistical methods such as the principle component analysis (PCA) mentioned earlier. The added anthropometric ranges of the figure dimensions will help to test for the effects of multiple criteria (e.g., low-back compression and reach) concurrently.

4.1.3. Accurately Posture (or Animate) the Figures at the Workplace

Research has demonstrated that the biomechanical models that predict injury risk are quite sensitive to posture (Chaffin and Erig 1991). This makes it important to pose the digital figures such that they realistically represent the actual postures or movements required by a worker. If the workplace under analysis exists, still photography or video recordings of the workers performing these tasks can be used to guide the engineer to pose the digital figures accurately. Conversely, if the workplace is still under design, the engineer may have to draw on his or her human movement intuition, or perhaps even virtual reality body tracking methods (described in Section 5), to posture the figures realistically. However, new research efforts that are expected to provide posture-prediction methodologies to aid designers with this process are underway. Currently the posturing task is left largely as the responsibility of the simulation engineer.

Depending on the human performance tool, the postural information required for an assessment may require a static posture at an instance in time, or multiple key postures at different times in the task. For example, the NIOSH lifting guide (NIOSH 1991) requires starting and ending postures of a lift to arrive at an assessment of the lift conditions. In contrast, analysis tools based on biomechanical models, such as low-back injury risk-assessment tools, can analyze loading conditions continuously for each posture throughout the simulation.

Once the geometry has been populated with the correct humans and these have been postured or animated to reflect the task, the host of ergonomic assessment tools discussed in Section 3 can be applied.

A typical manufacturing analysis includes the investigation of design for assembly and service, in which the question is asked whether the task can be performed. Can the person reach into the opening while holding the part and assemble the pieces? Can the object be reached by short and tall workers? Is there sufficient clearance for the part and the hands? Can the worker have an unobstructed view of the assembly so that it can be performed accurately? Will the worker have sufficient strength to perform the assembly task, or will it require potentially injurious exertions?

4.1.4. Service Analysis

The application of human models to the analysis of maintenance and service operations is one of the original targets of 3D human modeling. Particularly in the military and aerospace industry, the issues surrounding rapid serviceability motivated the development and use of this technology. One specific modern example can be found in the serviceability analysis of aircraft engines in the commercial airline industry, where the very dense nature of the engine packaging and the economics of service downtime make questions of how parts can be extracted for maintenance critical. These questions must be asked while the engine is still under design in CAD to avoid expensive reworks later on. Very complex software has been created to find collision-free extraction paths for virtual parts, both as validation that the part can be extracted from the surroundings and to provide training for maintenance personnel on how to perform the extraction operation. Unfortunately, these methodologies to date have not included the human, so the challenge posed to the human modeling publishers is to determine whether the part can actually be held, and extracted, with sufficient clearance for the part, fingers, and arm. Depending on the complexity of the environment, this is an incredibly difficult problem to solve without user input, and to date no solution is available that finds a solution in a reasonable amount of time. To address this, human models can be used in conjunction with immersive technologies in which the design engineer moves the virtual part in the environment with an avatar (virtual human) representing their arm and hand in the scene (see Section 5). Collision-detection capabilities of the human modeling software are used to identify if a free path can be found. This technology is now being evaluated to identify serviceability issues prior to the first physical build, and also to provide task timing estimates (cost) of performing a particular service operation (Figure 2).

4.2. Product Design

The availability of human modeling technology during the product-design phase expands the range of analyses that can be performed prior to a physical prototype construction. In the past, SAE recommended practices, or “J-standards,” were among the limited tools available for benchmarking and design. These tools, derived from empirical studies of people in vehicles, provide population response models that describe such functional information as reach, eye location, and head clearance. However, these data are presented as statistical summaries of the population response, which do not maintain information on the response of any particular individual. The SAE eye-ellipse zone, for example, provides an ellipsoid that defines a region where the eye locations of a specific portion of the

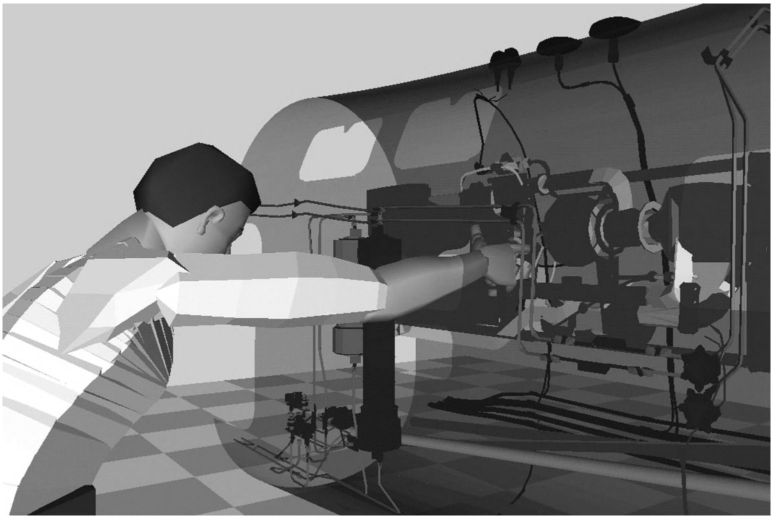


Figure 2 Serviceability Analysis of a Design Can Be Performed Prior to a Prototype Build. Here the Jack human figure is used to analyze the maintenance access to an electrical box inside of the aircraft nose cone. Eye views and collision detection can help the designer during the evaluation process. (*Courtesy EMBRAER*)

population can be expected. The specific location where a small or tall person's eyes might fall within this distribution is not defined (Figure 3). For this reason, when the behavior of a specifically sized individual or group is required, such as when a design is targeted to a specific demographic, human modeling tools can be used to answer these questions.

4.2.1. Accommodation

Once a design proposal is in place, accommodation questions can be posed. The process generally mirrors that for workplace analysis, with a few modifications.

4.2.2. Definition of the Test Population Anthropometry

Most often the accommodation needs for product design are more involved than during manufacturing ergonomic analysis because more anthropometric dimensions typically need to be taken into account. For example, the product design may have to accommodate individuals with a variety of sitting eye heights, shoulder breadths, and arm lengths. As mentioned in the sections describing anthropometric methods, statistical methods such as factor analysis (principal components) can be used to select a family of figures or boundary manikins that will adequately test the range of these multiple dimensions.

4.2.2.1. Figure Posturing Posturing a figure within the digital environment can impact the design analysis dramatically. As evidence of the importance of this issue, various posture-prediction methodologies have been developed in different industries. Pioneering work at Boeing led to a posture prediction method for the aerospace industry (Ryan and Springer 1969). In the late 1980s, a consortium of German automotive manufactures and seat suppliers sponsored the development of driver posture-prediction methodologies for the RAMSIS CAD manikin (Seidl 1993). Most recently, a global automotive industrial consortium sponsored new and more comprehensive methodologies to predict the postures of drivers and passengers through the ASPECT program (Reed 1998). These latest methods have been made available to modelers for inclusion in their software, allowing for sophisticated accommodation studies in automotive environments. Data for posture prediction in heavy truck and earth-moving equipment environments are still needed.

The boundary manikins are postured in the environment and tested for clearance, reach, vision, and comfort issues. Measurements from these boundary manikins can be used to establish design zones with which product design decisions can be made. A common technique for considering reachability, for example, is to generate zones representing the space where the boundary manikins

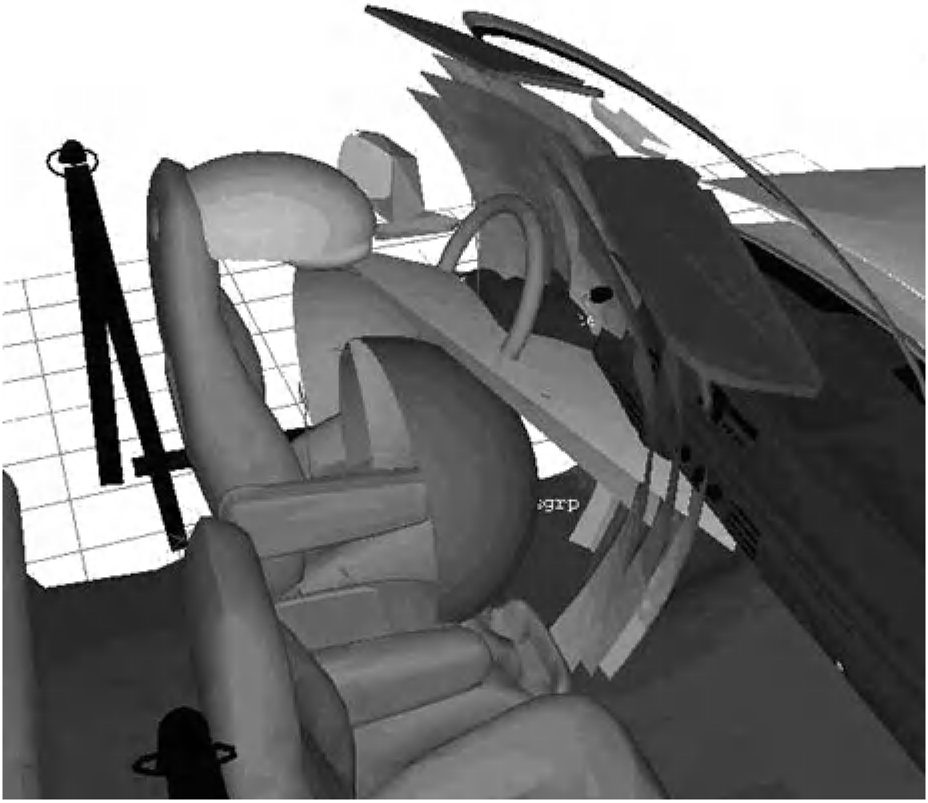


Figure 3 The SAE-J Standards Are Used Worldwide in Benchmarking and Design. The statistical summary descriptions describe posturing and functional information for a predefined population of male and female drivers. Design questions regarding a specifically sized individual cannot be asked using these zones but are suited for human simulation (see text).

can reach under different circumstances. Such reach zones can be used to make design decisions about where controls should or shouldn't be placed. An alternative approach to the sequential test of boundary manikins involves the simultaneous display of important landmark data in the form of data clouds (Figure 4). With this technique, the 3D locations of landmarks are collected either from actual subjects sitting in a mockup or from digital boundary manikins in the electronic environment. The complete set of landmarks from all the test figures is subsequently displayed as a data cloud and used in the design process. This technique provides a concise view of the population anthropometric variability, making it well suited to guide control placement and adjustment range analyses.

4.2.2. Usability

Usability can be considered part of the accommodation process. Traditionally, accommodation considered mostly the analysis of fit, and not necessarily the functional aspects of operating the product. With digital human models, this boundary is rapidly disappearing as anthropometric accommodation, vision, performance models, reach, and movement can all be analyzed using the same tool. Traditional methods such as the SAE J-standards are based on static symmetric zones for analysis, but designers are increasingly interested in taking advantage of the increased functionality modern human models provide. For automotive analyses, these more involved analyses include investigations of reach under various constraint conditions, vision coverage using mirror views, vision obscuration during head check procedures, psychological effects of roominess, obscuration, and collision. Research into natural human motion prediction provides guidelines on layout to accommodate maximally the natural motion of the operator.

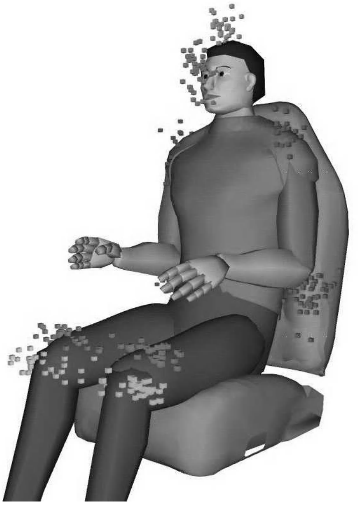


Figure 4 The Anthropometric Landmark Data from Multiple Subjects Is Displayed in the Form of Data Clouds. The simultaneous view of the data allows for rapid assessment of control location and adjustment questions. (*Courtesy Dr. Jerry Duncan, John Deere Corporation*)

5. IMMERSIVE VIRTUAL REALITY

Many of the predictive technologies available for human modeling currently do not adequately answer the questions designers pose to their human modeling solutions. This limitation is especially pronounced in the areas of natural complex motion simulation and cognitive perception. As mentioned in previous sections, a designer might ask how a person will move to perform an operation, or ask if there is sufficient clearance for a person to grasp a part within the confines of surrounding parts. Situations that require nontypical, complex motions currently cannot be answered adequately with the movement prediction algorithms available. Only through the use of immersive virtual reality technology that allows the mapping of a designer's movements to an avatar in the virtual scene can these complex movement situations be adequately and efficiently analyzed. Similarly, cognitive models providing subjective perceptions of an environment, such as feelings of spaciousness, control, and safety, are not currently available, yet a designer looking through the eyes of a digital human can assess these emotions of the virtual environment under design. For these reasons, immersive virtual reality (VR) is increasingly being used in both design, and manufacturing applications. Early in a design cycle when only CAD models are available, VR can allow the design to be experienced by designers, managers, or even potential users. Such application allows problems to be identified earlier in the design cycle and can reduce the need for physical prototypes. Immersive VR usually includes a combination of motion tracking and stereo display to give the user the impression of being immersed in a 3D computer environment. Auditory and, increasingly, haptic technology are also available to add realism to the user's perspective.

Virtual reality does not necessarily require a digital human model. Simply tracking a subject's head motion is sufficient to allow the stereo view to reflect the subject's view accurately and thus provide the user with a sense of being present in the computerized world. However, the addition of a full human model, tracking the subject's body and limb movements in real time, allows for additional realism because the user can see a representation of themselves in the scene. Additional analysis is also possible with full-body motion tracking. For example, collisions between limbs and the objects in the scene can be detected so that reach and fit can be better assessed. This capability is especially useful in design for maintainability or design for assembly applications. Another area where the full body tracking can be useful is for a designer to gain experience in interacting with the design from the perspective of a very short or very tall person. By scaling the environment in proportion to the increase or decrease in anthropometric size he or she wishes to experience, the designer can evaluate such issues as clearance, visibility, and reachability of extreme-sized individuals without actually having to recruit a subject pool of these people. Real-time application of the tracked motions to the virtual human also gives observers a realistic third-person view of the human motion in relation to the design geometry.

In addition to the qualitative assessments provided by the engineer's subjective perceptions of interacting with the design, quantitative assessments are possible. Analysis tools, such as those described in Section 3, can often be run in real time while immersed. For example, an engineer can perform a virtual operation using virtual tools, while in real time the performance tools evaluate the postures, forces, and motions to derive performance metrics such as population strength capability, low-back strain, fatigue, or postural affects. The designer gets immediate feedback as to the specific actions that are likely to put the worker at an elevated risk of injury without exposing the test subject to unsafe loading conditions. The design surrounding these actions can then be assessed and modified to reduce the injury risk, all while in the digital design space. Alternatively, motions can be captured and then played back for human performance analysis or presentation purposes.

Such quantitative analyses may be performed in the context of a full immersive VR application or may simply make use of the same human motion-tracking and capture technology to assist in the generation of accurate human postures. For example, a dataglove with posture-sensing electronics incorporated can be a valuable tool with which to obtain accurate hand postures while avoiding the tedium of trying of manipulate each individual finger joint irrespective of the actual application.

5.1. Motion-Tracking Technologies

A number of technologies are available for tracking human motions, including piezioelectric strain gages, magnetic and optical. Such human motion-tracking technologies have long been used for scientific and clinical applications (e.g., Chao 1978; Davis et al. 1991). In recent years, real-time forms of these technologies have become feasible and made VR possible. In addition to VR applications, such real-time technologies have found application in the entertainment industry, enabling quick generation of realistic human movements for computer games, 3D animation, and movie special effects.

Data gloves, such as Virtual Technology, Inc.'s Cyberglove (www.virtex.com), and other such devices measure relative motion between two body segments using either fiberoptic or strain gage-based technologies. The location of the segment in space is not reported. This limitation has ramifications for how these devices are used in human models. For example, the data gloves can measure the amount of finger flexion and splay, yet these gloves do not provide information about where the hand is located relative to the body or in the scene. For this reason, they cannot be used in isolation in such applications as maintenance part extraction, where the orientation and position of the hand is equally as important as the hand posture. This global positioning information can however be captured using whole-body-tracking technologies.

Both magnetic and optical motion-tracking devices are used to capture the global spatial position of the body in space. Magnetic systems are composed of a transmitter that emits an electric field and sensors that can detect their position and orientation (six DOF) in this field. The magnetic sensors are attached to body segments (e.g., the hand, forearm, arm, torso) to determine the relative positions of adjacent body segments. These data are then used to animate a digital human figure. Magnetic systems until recently were the only systems that could track multiple segments in real time and thus are very popular for immersive applications. However, metallic objects in the environment can affect the magnetic fields emitted by these systems. The field distortion caused by metal in the surroundings, including structural metal in the floor, walls, and ceiling, can cause measurement inaccuracies. In contrast, video-based methods use retroreflecting or LED markers placed on the subject and cameras in the environment to triangulate the position of the markers. Multiple markers can be arranged on segments to derive both position and orientation of individual segments. Although multiple markers are required to obtain the same position and orientation information as one magnetic sensor, these markers are typically passive (simply balls covered with reflective material) and so do not encumber the motion of the subject as dramatically as the wires of magnetic systems. The downside of optical motion-tracking technology is that it is necessary for every marker to be seen by at least two (and preferably more) cameras. Placement of cameras to meet this requirement can be a challenge, especially in enclosed spaces such as a vehicle cab.

Examples of commercially available magnetic systems include the Ascension MotionStar (www.ascension-tech.com) and Polhemus FastTrak (www.polhemus.com). Examples of optical systems include those sold by Vicon Motion Systems (www.vicon.com), Qualysis AB (www.qualysis.com), and Motion Analysis Corp. (www.motionanalysis.com).

6. HUMAN SIMULATION CHALLENGES

As human modeling becomes an integral part of the design process, the need for visual realism and analysis sophistication also increases. For better or worse, the visual appearance of human figures plays an important role in the acceptance of the technology and the perceived confidence of the results. Efforts in several areas focus on the increased realism of the human skin form. For performance reasons, current commercial human models are "skinned" using polygonal segment representations that are either completely static or pseudostatic. The figures are composed of individual segments, such as feet, lower and upper legs, pelvis, and torso. The segments are drawn as a collection of static polygons arranged to give the segment its anthropomorphic shape. Prudent selection of the shape at the ends of the segments allows joints to travel through the physiological range of motion without the creation of gaps. Pseudostatic skinning solutions "stitch" polygons between the nodes of adjacent segments in real time to avoid the skin breaking apart at the joints. These solutions can be made to look very realistic and are adequate for most ergonomic assessments and presentations. However, they do not model the natural tissue deformation that occurs at the joints throughout the range of motion. This is visually most noticeable at complex joints, such as the shoulder joint, or quantitatively at joints to which measurements are taken, such as the popliteal region of the knee. To better model these areas, a variety of methods have been described in the literature that deform the surface polygons according to parametric descriptions or underlying muscle deformation models (e.g., Scheepers et al. 1997). However, these methods have generally not been commercially adopted because they are computationally expensive and mostly unnecessary from the ergonomic analysis standpoint. Nevertheless, as the computer hardware capability increases and the availability of highly detailed whole-body-surface scans elevates the expected level of visual realism, these deformation techniques will become more prevalent.

6.1. Performance Models

6.1.1. Performance Factors

Performance models used in current commercial models are largely an amalgamation of models and data available in the ergonomics and human factors literature. As mentioned in the review of the performance models, the presentation of most of these research findings was originally not intended for integration into real-time simulation environments. The studies from which these data were derived also did not address some of the more contemporary ergonomic issues, such as the performance limitations of the elderly, cumulative trauma, shoulder injury, and movement modeling.

The aging population is elevating the need to have more specific performance models for this demographic. Questions of functional limitations resulting from decreased strength, reaction time, and joint range of motion all affect the design, both of products and workplaces. In the automotive design space, ingress/egress capability is an example of a task that may be influenced by these limitations. In the workplace, questions of strength and endurance need to be addressed. Cumulative trauma prediction presents a particular academic challenge because the etiology of the injury is largely unknown. Biomechanical factors clearly play a role but to date do not provide sufficient predictive power upon which to base a risk-assessment tool. At best, conditions associated with an increased likelihood of cumulative trauma can be flagged. Similarly, shoulder fatigue and injury prediction is not developed to the point where models incorporated into modeling software can accurately predict the injurious conditions. The significant social and economic cost of low-back injury has motivated considerable research in low-back modeling over the past 20 years. The research findings have resulted in sophisticated models and quantitative design guidelines and have allowed manufacturing organizations to reduce dramatically the incidence rates of low-back pain. Shoulder injury and cumulative trauma now need the same level of investment to mature the available data in these areas.

6.1.2. Variation Modeling

Even with the sophistication of the currently available biomechanical models, human model users are becoming increasingly interested in asking questions of these tools for which there are insufficient data. One such example is describing the expected population variability within the performance of a task. Each person will perform actions in a slightly different way, and these variations are not represented in models that describe an "average" response. However, human modeling simulation software is ideally suited to visualize this variability between people (i.e., data clouds). Future human performance and movement models may have this variability modeled so that it can be displayed in the human modeling environment.

6.2. Human Motion Control

One of the significant advantages contemporary human modeling tools provide in ergonomic assessments is the ability to assemble simulations of the workers performing their tasks. Simulations can add value for task-timing information, workcell layout optimization, training, and technical presentations. If we are confident of the motion realism, we can apply the posture-sensitive ergonomic assessment tools to help identify the situations with the greatest injury risk potential. Considerable effort has been spent searching for methods that accurately predict how humans move under different task and environmental conditions (Raschke et al. 1998). Dynamic simulation (Hodgkins and Pollard 1997; Popovic and Witkins 1999), statistical models (Faraway 1997), warping techniques (Bruderlin 1995; Witkins et al. 1995) and optimization (Chao and Rim 1973; Pandy and Zajac 1991) have all been applied to this problem. However, many of the methods for simulating human motion and behavior are computationally intensive and do not lend themselves to real-time solution. While some methods show promise, no single method for modeling human motion has yet proven to be concise, flexible, and accurate. Modeling human movements accurately in constrained surroundings and when obstacles need to be avoided presents additional challenges.

6.2.1. Modeling Motion Data

Simulating human movements, whatever method is applied, requires a detailed understanding of how people really move. Much detailed research has been conducted in understanding lifting and arm movements, and the subject continues to be extensively studied (e.g., Chaffin et al. 2000). However, the wide variety of ways that humans can move and the flexibility we have to do the same task using different postural approaches create a challenge for trying to generalize these measurements to use in human modeling tools. It is one thing to measure the time sequence of joint angles involved in a typical lift task, but it is quite another to try to use these data to simulate accurately how people perform a lift under different conditions. Start and end conditions, the size, shape, or weight of the object being lifted, and obstacles that need to be avoided all influence the motion. Clearly a great deal of detailed data describing how we move under different circumstances is needed.

6.2.2. Multiple-figure Interactions

Humans do not necessarily work in isolation. Many tasks involve more than one human interacting with another. Two people carrying a single large load or manipulating the same object and one person slowing down or speeding up to avoid running into one another are just a few examples. All the motion-control challenges associated with modeling the movement of a single individual apply and are magnified when multiple individuals, each contributing differently to the task, are involved.

6.2.2.1. Interactive "Smart" Avatars Ultimately an accurate representation of humans needs to model not only how they move but how they think and make decisions about what movements to do and how they react to a given situation. Such "smart" humans would obviously aid in the generation of a complex motion sequence involving several humans and have application to the development of workplace simulations. However, at this point in time, the development of intelligent human's agents has been motivated by applications such as interactive training simulations (Badler et al. 1999), pedagogical agents (Johnson et al. 2000), intelligent characters in computer games (Funge et al. 1999), and conversational agents (Cassell and Vilhjalmsson 1999) and have not yet been applied to any great extent to workplace simulations.

7. CONCLUSION

Digital human modeling is being actively used in industries around the world to reduce the need for physical prototypes and create better and safer designs faster than was previously possible. Contemporary human modeling software tools are actively assimilating a variety of previously disconnected human modeling knowledge, including population anthropometry descriptions and physical capability models. The large amount of ergonomic and anthropometric knowledge integrated into these solutions makes them efficient tools to answer a wide variety of human factors questions of designs. At the same time, the global nature of these tools is serving to consolidate and expose research findings from around the world and steering academic research direction and focusing the presentation of the results for model inclusion. While there are many areas that can be explored using the current offering of modeling solutions, many interesting challenges remain as we work to make virtual humans as lifelike as technology and our knowledge of humans allow.

REFERENCES

Badler, N. I., Phillips, C. B., and Webber, B. L. (1993), *Simulating Humans: Computer Graphics Animation and Control,* Oxford University Press, New York.

Badler, N. I., Palmer, M. S., and Bindiganavale, R. (1999), "Animation Control for Real-Time Virtual Humans," *Communications of the ACM,* Vol. 42, No. 8, pp. 65–73.

Bittner, A. C., Wherry, R. J., and Glenn F. A. (1986), "CADRE: A Family of Manikins for Workstation Design," Technical Report 22100.07B, Man–Machine Integration Division, Naval Air Development Center, Warminster, PA.

Bruderlin, A., and Williams, L. (1995), "Motion Signal Procession," in *Proceedings of Computer Graphics,* Annual Conference Series, pp. 97–104.

Bubb, H. (1999), "Human Modeling in the Past and Future: The Lines of European Development," Keynote address at the 2nd SAE International Digital Human Modeling Conference, The Hague.

Burandt, U. (1978), *Ergonomie für Design und Entwicklung,* Schmidt, Cologne, p. 154.

Cassell, J., Vilhjalmsson, H. (1999), "Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous," *Autonomous Agents and Multi-Agent Systems* Vol. 2, pp. 45–64.

Cats-Baril, W., and Frymoyer, J. W. (1991), "The Economics of Spinal Disorders," in *The Adult Spine,* J. W. Frymoyer, T. B. Ducker, N. M. Hadler, J. P. Kostuik, J. N. Weinstein, and T. S. Whitecloud, Eds., Raven Press, New York, 85–105.

Chaffin, D. B. (2000), *Case Studies in Simulating People for Workspace Design,* SAE, Pittsburgh (forthcoming).

Chaffin, D. B., and Erig, M. (1991), "Three-Dimensional Biomechanical Static Strength Prediction Model Sensitivity to Postural and Anthropometric Inaccuracies," *IIE Transactions,* Vol. 23, No. 3, pp. 215–227.

Chaffin, D. B., Andersson, G. B. J., and Martin, B. J. (1999), *Occupational Biomechanics,* 3rd Ed., John Wiley & Sons, New York.

Chaffin, D. B., Faraway, J., Zhang, X., and Woolley, C. (2000), "Stature, Age, and Gender Effects on Reach Motion Postures," *Human Factors,* Vol. 42, No. 3, pp. 408–420.

Chao, E. Y. (1978), "Experimental Methods for Biomechanical Measurements of Joint Kinematics," in *CRC Handbook for Engineering in Medicine and Biology,* Vol. 1, B. N. Feinberg and D. G. Fleming, Eds., CRC Press, Cleveland, OH, pp. 385–411.

Chao, E. Y., and Rim, K. (1973), "Application of Optimization Principals in Determining the Applied Moments in Human Leg Joints During Gait," *Journal of Biomechanics,* Vol. 29, pp. 1393–1397.

Ciriello, V. M., and Snook, S. H. 1991, "The Design of Manual Handling Tasks: Revised Tables of Maximum Acceptable Weights and Forces," *Ergonomics,* Vol. 34, pp. 1197–1213.

Davis, R. B., Ounpuu S., Tyburski D., and Gage, J. R. (1991), "A Gait Analysis Data Collection and Reduction Technique," *Human Movement Science,* Vol. 10, pp. 575–587.

Dreyfuss, H. (1993), *The Measure of Man and Woman: Human Factors in Design,* Whitney Library of Design, New York.

Faraway, J. J. (1997), "Regression Analysis for a Functional Response," *Technometrics,* Vol. 39, No. 3, pp. 254–262.

Frymoyer, J. W., Pope, M. H., Clements, J. H., et al. (1983), "Risk Factors in Low Back Pain: An Epidemiologic Survey," *Journal of Bone and Joint Surgery,* Vol. 65A, pp. 213–216.

Funge, J., Tu, X., and Terzopoulos, D. (1999), "Cognitive Modeling: Knowledge, Reasoning and Planning for Intelligent Characters," in *SIGGRAPH Conference Proceedings* (Los Angeles, August).

Garg, A., Chaffin, D. B., and Herrin, G. D. (1978), "Prediction of Metabolic Rates for Manual Materials Handling Jobs," *American Industrial Hygiene Association Journal,* Vol. 39, No. 8, pp. 661–674.

Gordon, C. C., Bradtmiller, B., Churchill, T., Clauser, C. E., McConville, J. T., Tebbetts, I. O., and Walker, R. A. (1988), "1988 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics," Technical Report Natick/TR-89/044.

Grandjean, E. (1980), "Sitting Posture of Car Drivers from the Point of View of Ergonomics," in *Human Factors in Transport Research* (*Part 1*), E. Grandjean, Ed., Taylor & Francis, London, pp. 20–213.

Hodgins, J. K., and Pollard, N. S. (1997), "Adapting Simulated Behaviors for New Characters," *SIGGRAPH '97,* pp. 153–162.

Johnson, W. L., Rickel, J. W., and Lester, J. C. (2000), "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments," *International Journal of Artificial Intelligence in Education,* Vol. 11, pp. 47–78.

Karhu, O., Kansi, P., and Kuorina, I. (1977), "Correcting Working Postures in Industry: A Practical Method for Analysis," *Applied Ergonomics,* Vol. 8, pp. 199–201.

Krist, R. (1994), *Modellierung des Sitzkomforts: Eine experimentelle Studie,* Schuch, Weiden.

Laurig, W. (1973), "Suitability of Physiological Indicators of Strain for Assessment of Active Light Work," *Applied Ergonomics* (cited in Rohmert 1973b).

McAtamney, L., and Corlett, E. N. (1993), "RULA: A Survey Method for the Investigation of Work-Related Upper Limb Disorders," *Applied Ergonomics,* Vol. 24, pp. 91–99.

National Institute for Occupational Safety and Health (NIOSH) (1997), *Musculoskeletal Disorders (MSDs) and Workplace Factors: A Critical Review of Epidemiologic Evidence for Work Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back,* U.S. Department of Health and Human Services, Cincinnati.

NHANES (1994), "National Health and Nutrition Examination Survey, III: 1988–1994," U.S. Department of Health and Human Services, Center for Disease Control and Prevention, Vital and Health Statistics, Series 1, No. 32.

Nussbaum, M. A., Martin, B. J., and Chaffin, D. B. (1997), "A Neural Network Model for Simulation of Torso Muscle Coordination," *Journal of Biomechanics,* Vol. 30, pp. 251–258.

Pandy, M. G., and Zajac, F. E. (1991), "Optimal Muscular Coordination Strategies for Jumping," *Journal of Biomechanics*, Vol. 24, No. 1, pp. 1–10.

Pandya, A., Hasson, S., Aldridge, A., Maida, J., and Woolford, B. (1992), "The Validation of a Human Force Model to Predict Dynamic Forces Resulting from Multi-joint Motions," NASA Technical Report 3206.

Popovic, Z., and Witkins, A. (1999), "Physical Based Motion Transformation," *SIGGRAPH Conference Proceedings* (Los Angeles, August).

Porter, M. J., and Gyi, D. E. (1998), "Exploring the Optimum Posture for Driver Comfort," *International Journal of Vehicle Design,* Vol. 19, No. 3, pp. 255–266.

Raschke, U., Martin, B. J., and Chaffin, D. B. (1996), "Distributed Moment Histogram: A Neurophysiology Based Method of Agonist and Antagonist Trunk Muscle Activity Prediction," *Journal of Biomechanics,* Vol. 29, pp. 1587–1596.

Raschke, U., Schutte, L., and Volberg, O. (1998), "Control Strategies for Simulating Human Movement," SAE Technical Paper 981306.

Rebiffé, R. (1966), "An Ergonomic Study of the Arrangement of the Driving Position in Motor Cars," in *Proceedings of Journal of the Institute of Mechanical Engineers Symposium* (London).

Reed, M. P. (1998), "Statistical and Biomechanical Prediction of Automobile Driving Posture," Ph.D. Dissertation, Department of Industrial Engineering, University of Michigan.

Roebuck, J. A. (1995), *Anthropometric Methods: Designing to Fit the Human Body,* Human Factors and Ergonomics Society, Santa Monica, CA.

Rohmert, W. (1973a), "Problems in Determining Rest Allowances. Part 1: Use of Modern Methods to Evaluate Stress and Strain in Static Muscular Work," *Applied Ergonomics,* Vol. 4, No. 2, pp. 91–95.

Rohmert, W. (1973b), "Problems in Determining Rest Allowances. Part 2: Determining Rest Allowance in Different Human Tasks," *Applied Ergonomics,* Vol. 4, No. 2, pp. 158–162.

Ryan, P. W., and Springer, W. E. (1969), "Cockpit Geometry Evaluation, Phase 1," Final Report, Vol V. JANAIR Report 69105.

Scheepers, F., Parent, R. E., Carlson, W. E., and May, S. F. (1997), "Anatomy-Based Modeling of the Human Musculature," in *SIGGRAPH 97 Conference Proceedings* (Los Angeles, August), pp. 163–172.

Schultetus, W. (1987), "Montagegestaltung: Daten, Hinweise und Beispiele zur ergonomischen Arbeitsgestaltung," in *Praxis der Ergonomie*. TÜV Rheinland, Cologne, p. 122.

Seidl, A. (1993), "Das Menschmodell RAMSIS, Analyse, Synthese und Simulation dreidimensionaler Körperhaltungen des Menschen," Doctoral Dissertation, Department of Engineering Technical University of Munich.

Sjogaard, G., Savard, G., and Juel, C. (1988), "Muscle Blood Flow During Isometric Activity and Its Relation to Muscle Fatigue," *European Journal of Applied Physiology,* Vol. 57, No. 3, pp. 327–335.

Van Dieën, J. H. (1997), "Are Recruitment Patterns of the Trunk Musculature Compatible with a Synergy Based on the Maximization of Endurance?" *Journal of Biomechanics,* Vol. 30, pp. 1095–1100.

Van Dieën, J. H., and Vrielink, H. H. E (1994), "The Use of the Relation Between Relative Force and Endurance Time," *Ergonomics,* Vol. 37, No. 2, pp. 231–243.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks," *Ergonomics,* Vol. 36, No. 7, pp. 749–776.

Winters, J., and Woo, S. L. Y. (1990), *Multiple Muscle Systems, Biomechanics and Movement Organization,* J. Winters, and S. Woo, Eds., Springer, New York.

Witkins, A., and Zoran, P. (1995), "Motion Warping," in *Computer Graphics Proceedings,* Annual Conference Series, pp. 105–108.

Zehner, G. F., Meindl, R. S., and Hudson, J. A. (1993), "A Multivariate Anthropometric Method for Crew Station Design," Armstrong Laboratory Technical Report AL-TR-1993-0054.

ADDITIONAL READING

Consumer Product Safety Commission (CPSC), "Anthropometry of Infants, Children, and Youths to Age 18 for Product Safety Design," Technical report UM-HSRI-77-17, Prepared for the U.S.

Consumer Product Safety Commission by the Highway Safety Research Institute, University of Michigan, Ann Arbor, MI, 1977.

HQL, *Japanese Body Size Data 1992–1994,* Research Institute of Human Engineering for Quality Life, 1994.

Winter, D. A., *The Biomechanics and Motor Control of Human Gait: Normal, Elderly and Pathological,* University of Waterloo Press, Waterloo, ON, 1991.

CHAPTER 42

Human Factors Audit

COLIN G. DRURY
State University of New York at Buffalo

1. **THE NEED FOR AUDITING HUMAN FACTORS** **1131**
2. **DESIGN REQUIREMENTS FOR AUDIT SYSTEMS** **1132**
    - 2.1. Breadth, Depth, and Application Time 1132
    - 2.2. Use of Standards 1133
    - 2.3. Evaluation of an Audit System 1134
3. **AUDIT SYSTEM DESIGN** **1135**
    - 3.1. The Sampling Scheme 1135
    - 3.2. The Data-Collection Instrument 1136
        - 3.2.1. Checklists and Surveys 1137
        - 3.2.2. Other Data-Collection Methods 1145
    - 3.3. Data Analysis and Presentation 1145
4. **AUDIT SYSTEMS IN PRACTICE** **1146**
    - 4.1. Auditing a Decentralized Business 1146
    - 4.2. Error Reduction at a Colliery 1150
5. **FINAL THOUGHTS ON HUMAN FACTORS AUDITS** **1151**

REFERENCES **1152**

When we audit an entity, we perform an examination of it. Dictionaries typically emphasize official examinations of (financial) accounts, reflecting the accounting origin of the term. Accounting texts go further: for example, "testing and checking the records of an enterprise to be certain that acceptable policies and practices have been consistently followed" (Carson and Carlson 1977, p. 2). In the human factors field, the term is broadened to include nonfinancial entities, but it remains faithful to the concepts of checking, acceptable policies/practices, and consistency.

Human factors audits can be applied, as can human factors itself, to both products and processes. Both applications have much in common, as any process can be considered as a product of a design procedure, but this chapter emphasizes process audits because product evaluation is covered in detail in Chapter 49. Product usability audits have their own history (e.g., Malde 1992), which is best accessed through the product design and evaluation literature (e.g., McClelland 1990).

A second point needs to be made about the scope of this chapter: the role of checklists. As will be seen, checklists have assumed importance as techniques for conducting human factors audits. They can also be used alone as evaluation devices, in applications as diverse as VDT workplaces (Cakir et al. 1980), and risk factor assessment (Keyserling et al. 1992). Hence, the structure and use of checklists will be covered in some detail independently of their use as an auditing technique.

1. THE NEED FOR AUDITING HUMAN FACTORS

Human factors or ergonomics programs have become a permanent feature of many companies, with typical examples shown in Alexander and Pulat (1985). Like any other function, human factors/ergonomics needs tools to measure its effectiveness. Earlier, when human factors operated through individual projects, evaluation could take place on a project-by-project basis. Thus, the interventions to improve apparel-sewing workplaces described by Drury and Wick (1984) could be evaluated to show changes in productivity and reductions in cumulative trauma disorder causal factors. Similarly, Hasslequist (1981) showed productivity, quality, safety, and job satisfaction following human factors

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

interventions in a computer-component assembly line. In both cases, the objectives of the intervention were used to establish appropriate measures for the evaluation.

Ergonomics/human factors, however, is no longer confined to operating in a project mode. Increasingly, the establishment of a permanent function within an industry has meant that ergonomics is more closely related to the strategic objectives of the company. As Drury et al. (1989) have observed, this development requires measurement methodologies that also operate at the strategic level. For example, as a human factors group becomes more involved in strategic decisions about identifying and choosing the projects it performs, evaluation of the individual projects is less revealing. All projects performed could have a positive impact, but the group could still have achieved more with a more astute choice of projects. It could conceivably have had a more beneficial impact on the company's strategic objectives by stopping all projects for a period to concentrate on training the management, workforce, and engineering staff to make more use of ergonomics.

Such changes in the structure of the ergonomics/human factors profession indeed demand different evaluation methodologies. A powerful network of individuals, for example, who can, and do, call for human factors input in a timely manner can help an enterprise more than a number of individually successful project outcomes. Audit programs are one of the ways in which such evaluations can be made, allowing a company to focus its human factors resources most effectively. They can also be used in a prospective, rather than retrospective, manner to help quantify the needs of the company for ergonomics/human factors. Finally, they can be used to determine which divisions, plants, departments, or even product lines are in most need of ergonomics input.

2. DESIGN REQUIREMENTS FOR AUDIT SYSTEMS

Returning to the definition of an audit, the emphasis is on checking, acceptable policies, and consistency. The aim is to provide a fair representation of the business for use by third parties. A typical audit by a certified public accountant would comprise the following steps (adapted from Koli 1994):

1. *Diagnostic investigation:* description of the business and highlighting of areas requiring increased care and high risk
2. *Test for transaction:* trace samples of transactions grouped by major area and evaluate
3. *Test of balances:* analyze content
4. *Formation of opinion:* communicate judgment in an audit report

Such a procedure can also form a logical basis for human factors audits. The first step chooses the areas of study, the second samples the system, the third analyzes these samples, and the final step produces an audit report. These define the broad issues in human factors audit design:

1. *How to sample the system:* how many samples and how these are distributed across the system
2. *What to sample:* specific factors to be measured, from biomechanical to organizational
3. *How to evaluate the sample:* what standards, good practices, or ergonomic principles to use for comparison
4. *How to communicate the results:* techniques for summarizing the findings, how far separate findings can be combined

A suitable audit system needs to address all of these issues (see Section 3), but some overriding design requirements must first be specified.

2.1. Breadth, Depth, and Application Time

Ideally, an audit system would be broad enough to cover any task in any industry, would provide highly detailed analysis and recommendations, and would be applied rapidly. Unfortunately, the three variables of breadth, depth, and application time are likely to trade off in a practical system. Thus a thermal audit (Parsons 1992) sacrifices breadth to provide considerable depth based on the heat balance equation but requires measurement of seven variables. Some can be obtained rapidly (air temperature, relative humidity), but some take longer (clothing insulation value, metabolic rate). Conversely, structured interviews with participants in an ergonomics program (Drury 1990a) can be broad and rapid but quite deficient in depth.

At the level of audit instruments such as questionnaires or checklists, there are comprehensive surveys such as the position analysis questionnaire (McCormick 1979), the Arbeitswissenschaftliche Erhebungsverfahren zur Tätikgkeitsanalyse (AET) (Rohmert and Landau 1989), which takes two to three hours to complete, or the simpler work analysis checklist (Pulat 1992). Alternatively, there are simple single-page checklists such as the ergonomics—working position—sitting Checklist (SHARE 1990), which can be completed in a few minutes.

Analysis and reporting can range in depth from merely tabulating the number of ergonomic standards violated to expert systems that provide prescriptive interventions (Ayoub and Mital 1989).

Most methodologies fall between the various extremes given above, but the goal of an audit system with an optimum trade-off between breadth, depth and time is probably not realizable. A better practical course would be to select several instruments and use them together to provide the specific breadth and depth required for a particular application.

2.2 Use of Standards

The human factors/ergonomics profession has many standards and good practice recommendations. These differ by country (ANSI, BSI, DIN), although commonality is increasing through joint standards such as those of the International Standards Organization (ISO). Some standards are quantitative, such as heights for school furniture (BSI 1965), sizes of characters or a VDT screen (ANSI/HFS-100), and occupational exposure to noise. Other standards are more general in nature, particularly those that involve management actions to prevent or alleviate problems, such as the OSHA guidelines for meat-packing plants (OSHA 1990). Generally, standards are more likely to exist for simple tasks and environmental stressors and are hardly to be expected for the complex cognitive activities with which human factors predictions increasingly deal. Where standards exist, they can represent unequivocal elements of audit procedures, as a workplace which does not meet these standards is in a position of legal violation. A human factors program that tolerates such legal exposure should clearly be held accountable in any audit.

However, merely meeting legal requirements is an insufficient test of the quality of ergonomics/human factors efforts. Many legal requirements are arbitrary or outdated, such as weight limits for manual materials handling in some countries. Additionally, other aspects of a job with high ergonomic importance may not be covered by standards, such presence of multiple stressors, work in restricted spaces resulting in awkward postures, or highly repetitive upper extremity motions. Finally, there are many human factors good practices that are not the subject of legal standards. Examples are the NIOSH lifting equation (Waters et al. 1993), the Illuminating Engineering Society (IES) codes (1993), and the zones of thermal comfort defined by ASHRAE (1989) or Fanger (1970). In some cases, standards are available in a different jurisdiction from that being audited. As an example, the military standard MIL-1472D (DOD 1989) provides detailed standards for control and display design that are equally appropriate to process controls in manufacturing industry but have no legal weight there.

Standards, in the legal sense, are a particularly reactive phenomenon. It may take many years (any many injuries and accidents) before a standard is found necessary and agreed upon. The NIOSH lifting equation referenced above addresses a back injury problem that is far from new, yet it still has no legal force. Standards for upper extremity cumulative trauma disorder prevention have lagged disease incidence by many years. Perhaps because of busy legislative agendas, we cannot expect rapid legal reaction, unless a highly visible major disaster occurs. Human factors problems are both chronic and acute, so that legislation based on acute problems as the sole basis for auditing is unlikely ever to be effective.

Despite the lack of legislation covering many human factors concerns, standards and other instantiations of good practice do have a place in ergonomics audits. Where they exist, they can be incorporated into an audit system without becoming the only criterion. Thus, noise levels in the United States have a legal limit for hearing protection purposes of 90 dBA. But at levels far below this, noise can disrupt communications (Jones and Broadbent 1987) and distract from task performance. An audit procedure can assess the noise on multiple criteria, that is, on hearing protection and on communication interruptions, with the former criterion used on all jobs and the latter only where verbal communication is an issue.

If standards and other good practices are used in a human factors audit, they provide a quantitative basis for decision making. Measurement reliability can be high and validity self-evident for legal standards. However, it is good practice in auditing to record only the measurement used, not its relationship to the standard, which can be established later. This removes any temptation by the analyst to bend the measurement to reach a predetermined conclusion. Illumination measurements, for example, can vary considerably over a workspace, so that an audit question:

Work Surface Illumination >750 Lux ☐ yes ☐ no

could be legitimately answered either way for some workspaces by choice of sampling point. Such temptation can be removed, for example, by an audit question.

Illumination at four points on workstation:

☐ ☐ ☐ ☐ Lux

Later analysis can establish whether, for example, the mean exceeds 750 Lux or whether any of the four points fall below this level.

It is also possible to provide later analyses that combine the effects of several simple checklist responses, as in Parsons's (1992) thermal audit, where no single measure would exceed good practice even though the overall result would be cumulative heat stress.

2.3. Evaluation of an Audit System

For a methodology to be of value, it must demonstrate validity, reliability, sensitivity, and usability. Most texts that cover measurement theory treat these aspects in detail (e.g., Kerlinger 1964). Shorter treatments are found in human factors methodology texts (e.g., Drury 1990b; Osburn 1987).

Validity is the extent to which a methodology measures the phenomenon of interest. Does our ergonomics audit program indeed measure the quality of ergonomics in the plant? It is possible to measure validity in a number of ways, but ultimately all are open to argument. For example, if we do not know the true value of the quality of ergonomics in a plant, how can we validate our ergonomics audit program? Broadly, there are three ways in which validation can be tested.

Content validity is perhaps the simplest but least convincing measure. If each of the items of our measurement device displays the correct content, then validity is established. Theoretically, if we could list all of the possible measures of a phenomenon, content validity would describe how well our measurement device samples these possible measures. In practice it is assessed by having experts in the field judge each item for how well its content represents the phenomenon studied. Thus, the heat balance equation would be judged by most thermal physiologists to have a content that well represents the thermal load on an operator. Not all aspects are as easily validated!

Concurrent (or *prediction*) *validity* has the most immediate practical impact. It measures empirically how well the output of the measurement device correlates with the phenomenon of interest. Of course, we must have an independent measure of the phenomenon of interest, which raises difficulties. To continue our example, if we used the heat balance equation to assess the thermal load on operators, then there should be a high correlation between this and other measures of the effects of thermal load—perhaps measures such as frequency of temperature complaints or heat disorders: heat stroke, hyperthermia, hypothermia, and so on. In practice, however, measuring such correlations would be contaminated by, for example, propensity to report temperature problems or individual acclimatization to heat. Overall outputs from a human factors audit (if such overall outputs have any useful meaning) should correlate with other measures of ergonomic inadequacy, such as injuries, turnover, quality measures, or productivity. Alternatively, we can ask how well the audit findings agree with independent assessments of qualified human factors engineers (Keyserling et al. 1992; Koli et al. 1993) and thus validate against one interpretation of current good practice.

Finally, there is *construct validity,* which is concerned with inferences made from scores, evaluated by considering all empirical evidence and models. Thus, a model may predict that one of the variables being measured should have a particular relationship to another variable not in the measurement device. Confirming this relationship empirically would help validate the particular construct underlying our measured variable. Note that different parts of an overall measurement device can have their construct validity tested in different ways. Thus, in a board human factors audit, the thermal load could differentiate between groups of operators who do and do not suffer from thermal complaints. In the same audit, a measure of difficulty in a target aiming task could be validated against Fitts's law. Other ways to assess construct validity are those that analyze clusters or factors within a group of measures. Different workplaces audited on a variety of measures and the scores, which are then subjected to factor analysis, should show an interpretable, logical structure in the factors derived. This method has been used on large databases for job evaluation-oriented systems such as McCormick's position analysis questionnaire (PAQ) (McCormick 1979).

Reliability refers to how well a measurement device can repeat a measurement on the same sample unit. Classically, if a measurement X is assumed to be composed of a true value X_t and a random measurement error X_e, then

$$X = X_t + X_e$$

For uncorrelated X_t and X_e, taking variances gives:

$$\text{Variance}\ (X) = \text{variance}\ (X_t) + \text{variance}\ (\mathrm{X}_e)$$

or

$$V(X) = V(X_t) + V(X_e)$$

We can define the reliability of the measurement as the fraction of measurement variance accounted for by true measurement variance:

$$\text{Reliability} = \frac{V(X_t)}{V(X_t) + V(X_e)}$$

Typically, reliability is measured by correlating the scores obtained through repeated measurements. In an audit instrument, this is often done by having two (or more) auditors use the instrument on the same set of workplaces. The square of the correlation coefficient between the scores (either overall scores, or separately for each logical construct) is then the reliability. Thus, PAQ was found to have an overall reliability of 0.79, tested using 62 jobs and two trained analysts (McCormick 1979).

Sensitivity defines how well a measurement device differentiates between different entities. Does an audit system for human–computer interaction find a difference between software generally acknowledged to be "good" and "bad"? If not, perhaps the audit system lacks sensitivity, although of course there may truly be no difference between the systems except what blind prejudice creates. Sensitivity can be adversely affected by poor reliability, which increases the variability in a measurement relative to a fixed difference between entities, that is, gives a poor signal-to-noise ratio. Low sensitivity can also come from a floor or ceiling effect. These arise where almost all of the measurements cluster at a high or low limit. For example, if an audit question on the visual environment was:

Does illumination exceed 10 lux? ☐ yes ☐ no

then almost all workplaces could answer "yes" (although the author has found a number that could not meet even this low criterion). Conversely, a floor effect would be a very high threshold for illluminance. Sensitivity can arise too when validity is in question. Thus, heart rate is a valid indicator of heat stress but not of cold stress. Hence, exposure to different degrees of cold stress would be only insensitively measured by heart rate.

Usability refers to the auditor's ease of use of the audit system. Good human factors principles should be followed, such as document design guidelines in constructing checklists (Patel et al. 1993; Wright and Barnard 1975). If the instrument does not have good usability, it will be used less often and may even show reduced reliability due to auditors' errors.

3. AUDIT SYSTEM DESIGN

As outlined in Section 2, the audit system must choose a sample, measure that sample, evaluate it, and communicate the results. In this section we approach these issues systematically.

An audit system is not just a checklist; it is a methodology that often includes the technique of a checklist. The distinction needs to be made between methodology and techniques. Over three decades ago, Easterby (1967) used Bainbridge and Beishon's (1964) definitions:

Methodology: a principle for defining the necessary procedures
Technique: a means to execute a procedural step.

Easterby notes that a technique may be applicable in more than one methodology.

3.1. The Sampling Scheme

In any sampling, we must define the unit of sampling, the sampling frame, and the sample choice technique. For a human factors audit the unit of sampling is not as self-evident as it appears. From a job-evaluation viewpoint (e.g., McCormick 1979), the natural unit is the job that is composed of a number of tasks. From a medical viewpoint the unit would be the individual. Human factors studies focus on the task/operator/machine/environment (TOME) system (Drury 1992) or equivalently the software/hardware/environment/liveware (SHEL) system (ICAO 1989). Thus, from a strictly human factors viewpoint, the specific combination of TOME can become the sampling unit for an audit program.

Unfortunately, this simple view does not cover all of the situations for which an audit program may be needed. While it works well for the rather repetitive tasks performed at a single workplace, typical of much manufacturing and service industry, it cannot suffice when these conditions do not hold. One relaxation is to remove the stipulation of a particular incumbent, allowing for jobs that require frequent rotation of tasks. This means that the results for one task will depend upon the incumbent chosen, or that several tasks will need to be combined if an individual operator is of interest. A second relaxation is that the same operator may move to different workplaces, thus changing environment as well as task. This is typical of maintenance activities, where a mechanic may perform any one of a repertoire of hundreds of tasks, rarely repeating the same task. Here the rational sampling unit is the task, which is observed for a particular operator at a particular machine in a particular environment. Examples of audits of repetitive tasks (Mir 1982; Drury 1990a) and maintenance tasks (Chervak and Drury 1995) are given below to illustrate these different approaches.

Definition of the sampling frame, once the sampling unit is settled, is more straightforward. Whether the frame covers a department, a plant, a division, or a whole company, enumeration of all sampling units is at least theoretically possible. All workplaces or jobs or individuals can in principle be listed, although in practice the list may never be up to date in an agile industry where change is the normal state of affairs. Individuals can be listed from personnel records, tasks from work orders or planning documents, and workplaces from plant layout plans. A greater challenge, perhaps, is to decide whether indeed the whole plant really is the focus of the audit. Do we include office jobs or just production? What about managers, chargehands, part-time janitors, and so on? A good human factors program would see all of these tasks or people as worthy of study, but in practice they may have had different levels of ergonomic effort expended upon them. Should some tasks or groups be excluded from the audit merely because most participants agree that they have few pressing human factors problems? These are issues that need to be decided explicitly before the audit sampling begins.

Choice of the sample from the sampling frame is well covered in sociology texts. Within human factors it typically arises in the context of survey design (Sinclair 1990). To make statistical inferences from the sample to the population (specifically to the sampling frame), our sampling procedure must allow the laws of probability to be applied. The most often-used sampling methods are:

Random sampling: Each unit within the sampling frame is equally likely to be chosen for the sample. This is the simplest and most robust method, but it may not be the most efficient. Where subgroups of interest (strata) exist and these subgroups are not equally represented in the sampling frame, one collects unnecessary information on the most populous subgroups and insufficient information on the least populous. This is because our ability to estimate a population statistic from a sample depends upon the absolute sample size and not, in most practical cases, on the population size. As a corollary, if subgroups are of no interest, then random sampling loses nothing in efficiency.

Stratified random sampling: Each unit within a particular stratum of the sampling frame is equally likely to be chosen for the sample. With stratified random sampling we can make valid inferences about each of the strata. By weighting the statistics to reflect the size of the strata within the sampling frame, we can also obtain population inferences. This is often the preferred auditing sampling method as, for example, we would wish to distinguish between different classes of tasks in our audits: production, warehouse, office, management, maintenance, security, and so on. In this way our audit interpretation could give more useful information concerning where ergonomics is being used appropriately.

Cluster sampling: Clusters of units within the sampling frame are selected, followed by random or nonrandom selection within clusters. Examples of clusters would be the selection of particular production lines within a plant (Drury 1990a) or selection of representative plants within a company or division. The difference between cluster and stratified sampling is that in cluster sampling only a subset of possible units within the sampling frame is selected, whereas in stratified sampling all of the sampling frame is used because each unit must belong to one stratum. Because clusters are not randomly selected, the overall sample results will not reflect population values, so that statistical inference is not possible. If units are chosen randomly within each cluster, then statistical inference within each cluster is possible. For example, if three production lines are chosen as clusters, and workplaces sampled randomly within each, the clusters can be regarded as fixed levels of a factor and the data subjected to analysis of variance to determine whether there are significant differences between levels of that factor. What is sacrificed in cluster sampling is the ability to make *population* statements. Continuing this example, we could state that the lighting in line A is better than in lines B or C but still not be able to make statistically valid statements about the plant as a whole.

3.2. The Data-Collection Instrument

So far we have assumed that the instrument used to collect the data from the sample is based upon measured data where appropriate. While this is true of many audit instruments, this is not the only way to collect audit data. Interviews with participants (Drury 1990a), interviews and group meetings to locate potential errors (Fox 1992), and archival data such as injury of quality records (Mir 1982) have been used. All have potential uses with, as remarked earlier, a judicious range of methods often providing the appropriate composite audit system.

One consideration on audit technique design and use is the extent of computer involvement. Computers are now inexpensive, portable, and powerful and can thus be used to assist data collection, data verification, data reduction, and data analysis (Drury 1990a). With the advent of more intelligent interfaces, checklist questions can be answered from mouse-clicks on buttons or selection from menus, as well as the more usual keyboard entry. Data verification can take place at entry time by checking for out-of-limits data or odd data such as the ratio of luminance to illuminance, implying a reflectivity greater than 100%. In addition, branching in checklists can be made easier, with only valid follow-on questions highlighted. The checklist user's manual can be built into the checklist software using context-sensitive help facilities, as in the EEAM checklist (Chervak and Drury 1995). Computers can, of course, be used for data reduction (e.g., finding the insulation value of clothing from a clothing inventory), data analysis, and results presentation.

With the case for computer use made, some cautions are in order. Computers are still bulkier than simple pencil-and-paper checklists. Computer reliability is not perfect, so inadvertent data loss is still a real possibility. Finally, software and hardware date much more rapidly than hard copy, so results safely stored on the latest media may be unreadable 10 years later. How many of us can still read punched cards or eight-inch floppy disks? In contrast, hard-copy records are still available from before the start of the computer era.

3.2.1. Checklists and Surveys

For many practitioners the proof of the effectiveness of an ergonomics effort lies in the ergonomic quality of the TOME systems it produces. A plant or office with appropriate human–machine function allocation, well-designed workplaces, comfortable environment, adequate placement/training, and inherently satisfying jobs almost by definition has been well served by human factors. Such a facility may not have human factors specialists, just good designers of environment, training, organization, and so on working independently, but this would generally be a rare occurrence. Thus, a checklist to measure such inherently ergonomic qualities has great appeal as part of an audit system.

Such checklists are almost as old as the discipline. Burger and deJong (1964) list four earlier checklists for ergonomic job analysis before going on to develop their own, which was commissioned by the International Ergonomics Association in 1961 and is usually known as the IEA checklist. It was based in part on one developed at the Philips Health Centre by G. J. Fortuin and provided in detail in Burger and deJong's paper.

Checklists have their limitations, though. The cogent arguments put forward by Easterby (1967) provide a good early summary of these limitations, and most are still valid today. Checklists are only of use as an aid to designers of systems at the earliest stages of the process. By concentrating on simple questions, often requiring yes/no answers, some checklists may reduce human factors to a simple stimulus–response system rather than encouraging conceptual thinking. Easterby quotes Miller (1967): "I still find that many people who should know better seem to expect magic from analytic and descriptive procedures. They expect that formats can be filled in by dunces and lead to inspired insights. . . . We should find opportunity to exorcise this nonsense" (Easterby 1967, p. 554)

Easterby finds that checklists can have a helpful structure but often have vague questions, make nonspecified assumptions, and lack quantitative detail. Checklists are seen as appropriate for some parts of ergonomics analysis (as opposed to synthesis) and even more appropriate to aid operators (not ergonomists) in following procedural steps. This latter use has been well covered by Degani and Wiener (1990) and will not be further presented here.

Clearly, we should be careful, even 30 years on, to heed these warnings. Many checklists are developed, and many of these published, that contain design elements fully justifying such criticisms.

A checklist, like any other questionnaire, needs to have both a helpful overall structure and well-constructed questions. It should also be proven reliable, valid, sensitive, and usable, although precious few meet all of these criteria. In the remainder of this section, a selection of checklists will be presented as typical of (reasonably) good practice. Emphasis will be on objective, structure, and question design.

3.2.1.1. The IEA Checklist The IEA checklist (Burger and de Jong 1964) was designed for ergonomic job analysis over a wide range of jobs. It uses the concept of functional load to give a logical framework relating the physical load, perceptual load, and mental load to the worker, the environment, and the working methods/tools/machines. Within each cell (or subcell, e.g., physical load could be static or dynamic), the load was assessed on different criteria such as force, time, distance, occupational, medical, and psychological criteria. Table 1 shows the structure and typical questions. Dirken (1969) modified the IEA checklist to improve the questions and methods of recording. He found that it could be applied in a median time of 60 minutes per workstation. No data are given on evaluation of the IEA checklist, but its structure has been so influential that it included here for more than historical interest.

3.2.1.2. Position Analysis Questionnaire The PAQ is a structured job analysis questionnaire using worker-oriented elements (187 of them) to characterize the human behaviors involved in jobs (McCormick et al. 1969). The PAQ is structured into six divisions, with the first three representing the classic experimental psychology approach (information input, mental process, work output) and

TABLE 1 IEA Checklist: Structure and Typical Questions

A: Structure of the Checklist		A	B	C
Load 1. Mean 2. Peaks Intensity, Frequency, Duration		Worker	Environment	Working method, tools, machines
I. Physical load	1. Dynamic 2. Static			
II. Perceptual load	1. Perception 2. Selection, decision 3. Control of movement			
III. Mental load	1. Individual 2. Group			
B: Typical Question				
I B. Physical load/environment		2.1. Physiological Criteria		
	1. Climate: high and low temperatures 1. Are these extreme enough to affect comfort or efficiency? 2. If so, is there any remedy? 3. To what extent is working capacity adversely affected? 4. Do personnel have to be specially selected for work in this particular environment?			

the other three a broader sociotechnical view (relationships with other persons, job context, other job characteristics). Table 2 shows these major divisions, examples of job elements in each and the rating scales employed for response (McCormick 1979).

Construct validity was tested by factor analyses of databases containing 3700 and 2200 jobs, which established 45 factors. Thirty-two of these fit neatly into the original six-division framework, with the remaining 13 being classified as "overall dimensions." Further proof of construct validity was based on 76 human attributes derived from the PAQ, rated by industrial psychologists and the ratings subjected to principal components analysis to develop dimensions "which had reasonably similar attribute profiles" (McCormick 1979, p. 204). Interreliability, as noted above, was 0.79, based on another sample of 62 jobs.

The PAQ covers many of the elements of concern to human factors engineers and has indeed much influenced subsequent instruments such as AET. With good reliability and useful (though perhaps dated), construct validity, it is still a viable instrument if the natural unit of sampling is the job. The exclusive reliance on rating scales applied by the analyst goes rather against current practice of comparison of measurements against standards or good practices.

3.2.1.3. AET (*Arbeit the Arbeitswissenschaftliche Erhebungsverfahren zur Tätikgkeitsanalyse*) The AET, published in German (Landau and Rohmert 1981) and later in English (Rohmert and Landau 1983), is the job-analysis subsystem of a comprehensive system of work studies. It covers "the analysis of individual components of man-at-work systems as well as the description and scaling of their interdependencies" (Rohmert and Landau 1983, pp. 9–10). Like all good techniques, it starts from a model of the system (REFA 1971, referenced in Wagner 1989), to which is added Rohmert's stress/strain concept. This latter sees strain as being caused by the intensity and duration of stresses impinging upon the operator's individual characteristics. It is seen as useful in the analysis of requirements and work design, organization in industry, personnel management, and vocational counseling and research.

AET itself was developed over many years, using PAQ as an initial starting point. Table 3 shows the structure of the survey instrument with typical questions and rating scales. Note the similarity between AET's job demands analysis and the first three categories of the PAQ and the scales used in AET and PAQ (Table 2).

Measurements of validity and reliability of AET are discussed by H. Luczak in an appendix to Landau and Rohment, although no numerical values are given. Cluster analysis of 99 AET records produced groupings which supported the AET constructs. Seeber et al. (1989) used AET along with

TABLE 2. PAQ: Structure and Typical Questions

A: Structure of the Checklist

Division	Definition	Examples of Questions
1. Information input	Where and how does the worker get the information he uses in performing his job?	1. Use of written materials 2. Near-visual differentiation
2. Mental processes	What reasoning, decision making, planning, and information processing activities are involved in performing the job?	1. Level of reasoning in problem solving 2. Coding/decoding
3. Work output	What physical activities does the worker perform and what tools or devices does he use?	1. Use of keyboard devices 2. Assembling/unassembling
4. Relationships with other persons	What relationships with other people are required in performing the job?	1. Instructing 2. Contacts with public or customers
5. Job context	In what physical or social contexts is the work performed?	1. High temperature 2. Interpersonal; conflict situations
6. Other job characteristics	What activities, conditions, or characteristics other than those described above are relevant to the job?	1. Specified work pace 2. Amount of job structure

B: Scales used to rate elements

Types of scale		Scale values	
Identification	Type of Rating	Rating	Definition
U	Extend to Use	N	Does not apply
I	Importance of the job	1	Very minor
T	Amount of Time	2	Low
P	Possibility of Occurrence	3	Average
A	Applicability (yes/no only)	4	High
S	Special code	5	Extreme

two other work-analysis methods on 170 workplaces. They found that AET provided the most differentiating aspects (suggesting sensitivity). They also measured postural complaints and showed that only the AET groupings for 152 female workers found significant differences between complaint levels, thus helping establish construct validity.

AET, like PAQ before it, has been used on many thousands of jobs, mainly in Europe. A sizable database is maintained that can be used for both norming of new jobs analyzed and analysis to test research hypotheses. It remain a most useful instrument for work analysis.

3.2.1.4. Ergonomics Audit Program (*Mir 1982; Drury 1990a*) This program was developed at the request of a multinational corporation to be able to audit its various divisions and plants as ergonomics programs were being instituted. The system developed was a methodology of which the workplace survey was one technique. Overall, the methodology used archival data or outcome measures (injury reports, personnel records, productivity) and critical incidents to rank order departments within a plant. A cluster sampling of these departments gives either the ones with highest need (if the aim is to focus ergonomic effort) or a sample representative of the plant (if the objective is an audit). The workplace survey is then performed on the sampled departments.

The workplace survey was designed based on ergonomic aspects derived from a task/operator/machine/environment model of the person at work. Each aspect formed a section of the audit, and sections could be omitted if there were clearly not relevant, for example, manual materials-handling aspects for data-entry clerks. Questions within each section were based on standards, guidelines, and models, such as the NIOSH (1981) lifting equation, *ASHRAE Handbook of Fundamentals* for thermal aspects, and Givoni and Goldman's (1972) model for predicting heart rate. Table 4 shows the major sections and typical questions.

TABLE 3 AET: Structure and Typical Questions

A: Structure of the Checklist

Part	Major Division	Section
A: Work systems analysis	1. Work objects	1.1. Material work objects 1.2 Energy as work object 1.3 Information as work object 1.4 Man, animals, plants as work objects
	2. Equipment	2.1 Working equipment 2.2 Other equipment
	3. Work environment	3.1 Physical environment 3.2 Organizational and social environment 3.3 Principles and methods of remuneration
B: Task analysis	1. Tasks relating to material work objects 2. Tasks relating to abstract work objects 3. Man-related tasks 4. Number and repetitiveness of tasks	
C: Job demand analysis	1. Demands on perception	1.1 Mode of perception 1.2 Absolute/relative evaluation of perceived information 1.3 Accuracy of perception
	2. Demands for decision	2.1 Complexity of decisions 2.2 Pressure of time 2.3 Required knowledge
	3. Demands for response/activity	3.1 Body postures 3.2 Static work 3.3 Heavy muscular work 3.4 Light muscular work, active light work 3.5 Strenuousness and frequency of moves

B: Types of scale		Typical Scale values	
Code	Type of Rating	Duration Value	Definition
A	Does this apply?	0	Very infrequent
F	Frequency	1	Less than 10% of shift time
S	Significance	2	Less than 30% of shift time
D	Duration	3	30% to 60% of shift time
		4	More than 60% of shift time
		5	Almost continuously during whole shift

Data were entered into the computer program and a rule-based logic evaluated each section to provide messages to the user in the form of either a "section shows no ergonomic problems" message:

```
MESSAGE
Results from analysis of auditory aspects:
       Everything OK in this section
```

or discrepancies from a single input:

```
MESSAGE
Seats should be padded, covered with non-slip materials and have front
edge rounded
```

or discrepancies based on the integration of several inputs:

TABLE 4 Workplace Survey: Structure and Typical Questions

Section	Major Classification	Examples of Questions
1. Visual aspects		Nature of task Measure illuminance at task midfield outer field
2. Auditory aspects		Noise level, dBA Main source of noise
3. Thermal aspects		Strong radiant sources present? Wet bulb temperature (Clothing inventory)
4. Instruments, controls, displays	Standing vs. Seated Displays Labeling Coding Scales, dials, counters Control/display relationships Controls	Are controls mounted between 30 in. and 70 in. Signals for crucial visual checks Are trade names deleted? Color codes same for control & display? All numbers upright on fixed scales? Grouping by sequence or subsystem? Emergency button diameter $>$ 0.75 in.?
5. Design of workplaces	Desks Chairs Posture	Seat to underside of desk $>$ 6.7 in.? Height easily adjustable 15–21 in.? Upper arms vertical?
6. Manual materials handling	(NIOSH Lifting Guide, 1981)	Task, H, V, D, F
7. Energy expenditure		Cycle time Object weight Type of work
8. Assembly/repetitive aspects		Seated, standing, or both? If heavy work, is bench 6–16 in. below elbow height?
9. Inspection aspects		Number of fault types? Training time until unsupervised?

```
MESSAGE
The total metabolic workload is 174 watts
Intrinsic clothing insulation is 0.56 clo
Initial rectal temperature is predicted to be 36.0°C
Final rectal temperature is predicted to be 37.1°C
```

Counts of discrepancies were used to evaluate departments by ergonomics aspect, while the messages were used to alert company personnel to potential design changes. This latter use of the output as a training device for nonergonomic personnel was seen as desirable in a multinational company rapidly expanding its ergonomics program.

Reliability and validity have not been assessed, although the checklist has been used in a number of industries (Drury 1990a). The Workplace Survey has been included here because, despite its lack of measured reliability and validity, it shows the relationship between audit as methodology and checklist as technique.

3.2.1.5. ERGO, EEAM, and ERNAP (*Koli et al. 1993; Chervak and Drury 1995*) These checklists are both part of complete audit systems for different aspects of civil aircraft hangar activities. They were developed for the Federal Aviation Administration to provide tools for assessing human factors in aircraft inspection (ERGO) and maintenance (EEAM) activities, respectively. Inspection and maintenance activities are nonrepetitive in nature, controlled by task cards issued to technicians at the start of each shift. Thus, the sampling unit is the task card, not the workplace, which is highly variable between task cards. Their structure was based on extensive task analyses of inspection and maintenance tasks, which led to generic function descriptions of both types of work (Drury et al. 1990). Both systems have sampling schemes and checklists. Both are computer based with initial data collection on either hard copy or direct into a portable computer. Recently, both have been combined into a single program (ERNAP) distributed by the FAA's Office of Aviation Medicine. The structure of ERNAP and typical questions are given in Table 5.

TABLE 5 ERNAP Structure and Typical Questions

Audit Phase	Major Classification	Examples of Questions
I. Premaintenance	Documentation	Is feedforward information on faults given?
	Communication	Is shift change documented?
	Visual characteristics	If fluorescent bulbs are used, does flicker exist?
	Electric/pneumatic equipment	Do push buttons prevent slipping of fingers?
	Access equipment	Do ladders have nonskid surfaces on landings?
II. Maintenance	Documentation (M)	Does inspector sign off workcard after each task?
	Communication (M)	Explicit verbal instructions from supervisor?
	Task lighting	Light levels in four zones during task, fc.
	Thermal issues	Wet bulb temperature in hanger bay, °C
	Operator perception	Satisfied with summer thermal environment?
	Auditory issues	Noise levels at five times during task, dBA
	Electrical and pneumatic	Are controls easily differentiated by touch?
	Access equipment (M)	Is correct access equipment available?
	Hand tools	Does the tool handle end in the palm?
	Force measurements	What force is being applied, kg?
	Manual Materials handling	Does task require pushing or pulling forces?
	Vibration	What is total duration of exposure on this shift?
	Repetitive motion	Does the task require flexion of the wrist?
	Access	How often was access equipment repositioned?
	Posture	How often were following postures adopted?
	Safety	Is inspection area adequately cleaned for inspect?
	Hazardous material	Were hazardous materials signed out and in?
III. Postmaintenance	Buy back	Are discrepancy worksheets readable?

As in Mir's Ergonomics Audit Program, the ERNAP, the checklist is again modular, and the software allows formation of data files, selection of required modules, analysis after data entry is completed, and printing of audit reports. Similarly, the ERGO, EEAM, and ERNAP instruments use quantitative or Yes/No questions comparing the entered value with standards and good practice guides. Each takes about 30 minutes per task. Output is in the form of an audit report for each workplace, similar to the messages given by Mir's Workplace Survey, but in narrative form. Output in this form was chosen for compatibility with existing performance and compliance audits used by the aviation maintenance community.

Reliability of a first version of ERGO was measured by comparing the output of two auditors on three tasks. Significant differences were found at $P < 0.05$ on all three tasks, showing a lack of interrater reliability. Analysis of these differences showed them to be largely due to errors on questions requiring auditor judgment. When such questions were replaced with more quantitative questions, the two auditors had no significant disagreements on a later test. Validity was measured using concurrent validation against six Ph.D. human factors engineers who were asked to list all ergonomic issues on a power plant inspection task. The checklist found more ergonomic issues than the human factors engineers. Only a small number of issues were raised by the engineers that were missed by the checklist. For the EEAM checklist, again an initial version was tested for reliability with two auditors, and it only achieved the same outcome for 85% of the questions. A modified version was tested and the reliability was considered satisfactory with 93% agreement. Validity was again tested against four human factors engineers, this time the checklist found significantly more ergonomic issues than the engineers without missing any issues they raised.

The ERNAP audits have been included here to provide examples of a checklist embedded in an audit system where the workplace is *not* the sampling unit. They show that non-repetitive tasks can be audited in a valid and reliable manner. In addition, they demonstrate how domain-specific audits can be designed to take advantage of human factors analyses already made in the domain.

3.2.1.6. Upper-Extremity Checklist (*Keyserling et al. 1993*) As its name suggests, this checklist is narrowly focused on biomechanical stresses to the upper extremities that could lead to cumulative trauma disorders (CTDs). It does not claim to be a full-spectrum analysis tool, but it is included here as a good example of a special-purpose checklist that has been carefully constructed and validated. The checklist (Table 6) was designed for use by management and labor to fulfill a requirement in the OSHA guidelines for meat-packing plants. The aim is to screen jobs rapidly for harmful exposures rather than to provide a diagnostic tool. Questions were designed based upon the biomechanical literature, structured into six sections. Scoring was based on simple presence or absence of a condition, or on a three-level duration score. As shown in Table 6, the two or three levels were scored as o, ·, or * depending upon the stress rating built into the questionnaire. These symbols represented insignificant, moderate, or substantial exposures. A total score could be obtained by summing moderate and substantial exposures.

The upper extremity checklist was designed to be biased towards false positives, that is, to be very sensitive. It was validated against detailed analyses of 51 jobs by an ergonomics expert. Each section (except the first, which only recorded dominant hand) was considered as giving a positive screening if at least one * rating was recorded. Across the various sections, there was reasonable agreement between checklist users and the expert analysis, with the checklist, being generally more sensitive, as was its aim. The original reference shows the findings of the checklist when applied to 335 manufacturing and warehouse jobs.

As a special-purpose technique in an area of high current visibility for human factors, the upper extremity checklist has proven validity, can be used by those with minimal ergonomics training for screening jobs, and takes only a few minutes per workstation. The same team has also developed

TABLE 6 Upper Extremity Checklist: Structure, Questions, and Scoring

A: Structure of the Checklist

Major Section	Examples of Questions
Worker information	Which hand is dominant?
Repetitiveness	Repetitive use of the hands and wrists? If "yes" then: Is cycle < 30 sec? Repeated for > 50% cycle?
Mechanical stress	Do hard or sharp objects put pressure localized pressure on: back or side of fingers? Palm or base of hand . . .
Force	Lift, carry, push or pull objects > 4.5 kg? If gloves worn, do they hinder gripping? . . .
Posture	Is pinch grip used? Is there wrist deviation? . . .
Tools, hand-held objects and equipment	Is vibration transmitted to the operator's hand? Does cold exhaust air blow on the hand or wrist? . . .

B. Scoring scheme

Question	Scoring		
Is there wrist deviation?	No o	Some √	> 33% cycle *

C. Overall evaluation

Total Score	Number of √ + Number of *

and validated a legs, trunk, and neck job screening procedure along similar lines (Keyserling et al. 1992).

3.2.1.7. Ergonomic Checkpoints The Workplace Improvement in Small Enterprises (WISE) methodology (Kogi 1994) was developed by the International Ergonomics Association (IEA) and the International Labour Office (ILO) to provide cost-effective solutions for smaller organizations. It consists of a training program and a checklist of potential low-cost improvements. This checklist, called ergonomics checkpoints, can be used both as an aid to discovery of solutions and as an audit tool for workplaces within an enterprise.

The 128-point checklist has now been published (Kogi and Kuorinka 1995). It covers the nine areas shown in Table 7. Each item is a statement rather than a question and is called a checkpoint. For each checkpoint there are four sections, also shown in Table 7. There is no scoring system as such; rather, each checkpoint becomes a point of evaluation of each workplace for which it is appropriate. Note that each checkpoint also covers why that improvement is important, and a description of the core issues underlying it. Both of these help the move from rule-based reasoning to knowledge-based reasoning as nonergonomists continue to use the checklist. A similar idea was embodied in the Mir (1982) ergonomic checklist.

3.2.1.8. Other Checklists The above sample of successful audit checklists has been presented in some detail to provide the reader with their philosophy, structure, and sample questions. Rather then continue in the same vein, other interesting checklists are outlined in Table 8. Each entry shows the domain, the types of issues addressed, the size or time taken in use, and whether validity and reliability have been measured. Most textbooks now provide checklists, and a few of these are cited. No claim is made that Table 8 is comprehensive. Rather, it is rather a sampling with references so that readers can find a suitable match to their needs. The first nine entries in the table are conveniently colocated in Landau and Rohmert (1989). Many of their reliability and validity studies are reported in this publication. The next entries are results of the Commission of European Communities fifth ECSC program, reported in Berchem-Simon (1993). Others are from texts and original references. The author has not personally used all of these checklists and thus cannot specifically endorse them. Also, omission of a checklist from this table implies nothing about its usefulness.

TABLE 7 Ergonomic Checkpoints: Structure, Typical Checkpoints, and Checkpoint Structure

A: Structure of the Checklist

Major Section	Typical Checkpoints
Materials handling	• Clear and mark transport ways.
Handtools	• Provide handholds, grips, or good holding points for all packages and containers.
Productive machine safety	• Use jigs and fixtures to make machine operations stable, safe, and efficient.
Improving workstation design	• Adjust working height for each worker at elbow level or slightly below it.
Lighting	• Provide local lights for precision or inspection work.
Premises	• Ensure safe wiring connections for equipment and lights.
Control of hazards	• Use feeding and ejection devices to keep the hands away from dangerous parts of machinery.
Welfare facilities	• Provide and maintain good changing, washing, and sanitary facilities to keep good hygiene and tidiness.
Work organization	• Inform workers frequently about the results of their work.

B. Structure of each checkpoint

WHY?	Reasons why improvments are important.
HOW?	Description of several actions each of which can contribute to improvement.
SOME MORE HINTS	Additional points which are useful for attaining the improvement.
POINTS TO REMEMBER	Brief description of the core element of the checkpoint.

From Kogi, private communication, November 13, 1995.

TABLE 8 A Selection of Published Checklists

Name	Authors	Coverage	Reliability	Validity
TBS	Hacker et al. 1983	Mainly mental work		vs. AET
VERA	Volpert et al. 1983	Mainly mental work		vs. AET
RNUR	RNUR, 1976	Mainly physical work		
LEST	Guèlaud. 1975	Mainly physical work		
AVISEM	AVISEM. 1977	Mainly physical work		
GESIM	GESIM. 1988	Mainly physical work		
RHIA	Leitner et al. 1987	Task hindrances, stress	0.53–0.79	vs. many
MAS	Groth. 1989	Open structure, derived from AET		vs. AET
JL and HA	Mattila and Kivi. 1989	Mental, physical work, hazards	0.87–0.95	
	Bolijn 1993	Physical work checklist for women	tested	
	Panter 1993	Checklist for load handling		
	Portillo Sosa 1993	Checklist for VDT standards		
Work Analy.	Pulat 1992	Mental and physical work		
Thermal Aud.	Parsons 1992	Thermal audit from heat balance		content
WAS	Yoshida and Ogawa, 1991	Workplace and environment	tested	vs. expert
Ergonomics	Occupational Health and Safety Authority 1990	Short workplace checklists		
	Cakir et al. 1980	VDT checklist		

First nine from Landau and Rohmert 1989; next three from Berchem-Simon 1993.

3.2.2. Other Data-Collection Methods

Not all data come from checklists and questionnaires. We can audit a human factors program using outcome measures alone (e.g., Chapter 47). However, outcome measures such as injuries, quality, and productivity are nonspecific to human factors: many other external variables can affect them. An obvious example is changes in the reporting threshold for injuries, which can lead to sudden apparent increases and decreases in the safety of a department or plant. Additionally, injuries are (or should be) extremely rare events. Thus, to obtain enough data to perform meaningful statistical analysis may require aggregation over many disparate locations and/or time periods. In ergonomics audits, such outcome measures are perhaps best left for long-term validation or for use in selecting cluster samples.

Besides outcome measures, interviews represent a possible data-collection method. Whether directed or not (e.g., Sinclair 1990) they can produce critical incidents, human factors examples, or networks of communication (e.g., Drury 1990a), which have value as part of an audit procedure. Interviews are routinely used as part of design audit procedures in large-scale operations such as nuclear power plants (Kirwan 1989) or naval systems (Malone et al. 1988).

A novel interview-based audit system was proposed by Fox (1992) based on methods developed in British Coal (reported in Simpson 1994). Here an error-based approach was taken, using interviews and archival records to obtain a sampling of actual and possible errors. These were then classified using Reason's (1990) active/latent failure scheme and orthogonally by Rasmussen's (1987) skill-, rule-, knowledge-based framework. Each active error is thus a conjunction of skill/mistake/violation with skill/rule/knowledge. Within each conjunction, performance-shaping factors can be deduced and sources of management intervention listed. This methodology has been used in a number of mining-related studies: examples will be presented in Section 4.

3.3. Data Analysis and Presentation

Human factors as a discipline covers wide range of topics, from workbench height to function allocation in automated systems. An audit program can only hope to abstract and present a part of this range. With our consideration of sampling systems and data collection devices we have seen different ways in which an unbiased abstraction can be aided. At this stage the data consist of large numbers of responses to large numbers of checklist items, or detailed interview findings. How can, or should, these data be treated for best interpretation?

Here there are two opposing viewpoints: one is that the data are best summarized across sample units, but not across topics. This is typically the way the human factors professional community treats the data, giving summaries in published papers of the distribution of responses to individual items on the checklist. In this way, findings can be more explicit, for example that the lighting is an area that needs ergonomics effort, or that the seating is generally poor. Adding together lighting and seating discrepancies is seen as perhaps obscuring the findings rather than assisting in their interpretation.

The opposite viewpoint, in many ways, is taken by the business community. For some, an overall figure of merit is a natural outcome of a human factors audit. With such a figure in hand, the relative needs of different divisions, plants, or departments can be assessed in terms of ergonomic and engineering effort required. Thus, resources can be distributed rationally from a management level. This view is heard from those who work for manufacturing and service industries, who ask after an audit "How did we do?" and expect a very brief answer. The proliferation of the spreadsheet, with its ability to sum and average rows and columns of data, has encouraged people to do just that with audit results. Repeated audits fit naturally into this view because they can become the basis for monthly, quarterly, or annual graphs of ergonomic performance.

Neither view alone is entirely defensible. Of course, summing lighting and seating needs produces a result that is logically indefensible and that does not help diagnosis. But equally, decisions must be made concerning optimum use of limited resources. The human factors auditor, having chosen an unbiased sampling scheme and collected data on (presumably) the correct issues, is perhaps in an excellent position to assist in such management decisions. But so too are other stakeholders, primarily the workforce.

Audits, however, are not the only use of some of the data-collection tools. For example, the Keyserling et al. (1993) upper extremity checklist was developed specifically as a screening tool. Its objective was to find which jobs/workplaces are in need of detailed ergonomic study. In such cases, summing across issues for a total score has an operational meaning, that is, that a particular workplace needs ergonomic help.

Where interpretation is made at a deeper level than just a single number, a variety of presentation devices have been used. These must show scores (percent of workplaces, distribution of sound pressure levels, etc.) separately but so as to highlight broader patterns. Much is now known about separate vs. integrated displays and emergent features (e.g., Wickens 1992, pp. 121–122), but the traditional profiles and spider web charts are still the most usual presentation forms. Thus, Wagner (1989) shows the AVISEM profile for a steel industry job before and after automation. The nine different issues (rating factors) are connected by lines to show emergent shapes for the old and the new jobs. Landau and Rohmert's (1981) original book on AET shows many other examples of profiles. Klimer et al. (1989) present a spider web diagram to show how three work structures influenced ten issues from the AET analysis. Mattila and Kivi (1989) present their data on the job load and hazard analysis system applied to the building industry in the form of a table. For six occupations, the rating on five different loads/hazards is presented as symbols of different sizes within the cells of the table.

There is little that is novel in the presentation of audit results: practitioners tend to use the standard tabular or graphical tools. But audit results are inherently multidimensional, so some thought is needed if the reader is to be helped towards an informed comprehension of the audit's outcome.

4. AUDIT SYSTEMS IN PRACTICE

Almost any of the audit programs and checklists referenced in previous sections give examples of their use in practice. Only two examples will be given here, as others are readily accessible. These examples were chosen because they represent quite different approaches to auditing.

4.1. Auditing a Decentralized Business

From 1992 to 1996, a major U.S.-based apparel manufacturer had run an ergonomics program aimed primarily at the reduction of workforce injuries in backs and upper extremities. As detailed in Drury et al. (1999), the company during that time was made up of nine divisions and employed about 45,000 workers. Of particular interest was the fact that the divisions enjoyed great autonomy, with only a small corporate headquarters with a single executive responsible for all risk-management activities. The company had grown through mergers and acquisitions, meaning that different divisions had different degrees of vertical integration. Hence, core functions such as sewing, pressing, and distribution were common to most divisions, while some also included weaving, dyeing, and embroidery. In addition, the products and fabrics presented quite different ergonomic challenges, from delicate undergarments to heavy jeans to knitted garments and even luggage.

The ergonomics program was similarly diverse. It started with a corporate launch by the highest-level executives and was rolled out to the divisions and then to individual plants. The pace of change was widely variable. All divisions were given a standard set of workplace analysis and modification tools (based on Drury and Wick 1984) but were encouraged to develop their own solutions to problems in a way appropriate to their specific needs.

Evaluation took place continuously, with regular meetings between representatives of plants and divisions to present results of before-and-after workplace studies. However, there was a need for a broader audit of the whole corporation aimed at understanding how much had been achieved for the multimillion-dollar investment, where the program was strong or weak, and what program needs were emerging for the future. A team of auditors visited all nine divisions, and a total of 12 plants spread across eight divisions, during 1995. This was three years after the initial corporate launch and about two years after the start of shop-floor implementation.

A three-part audit methodology was used. First, a workplace survey was developed based on elements of the program itself, supplemented by direct comparisons to ergonomics standards and good practices. Table 9 shows this 50-item survey form, with data added for the percentage of "yes" answers where the responses were not measures or scale values. The workplace survey was given at a total of 157 workplaces across the 12 plants. Second, a user survey (Table 10) was used in an interview format with 66 consumers of ergonomics, typically plant managers, production managers, human resource managers, or their equivalent at the division level, usually vice presidents. Finally, a total of 27 providers of ergonomics services were given a similar provider survey (Table 11) interview. Providers were mainly engineers, with three human resources specialists and one line supervisor. From these three audit methods the corporation wished to provide a time snapshot of how effectively the current ergonomics programs was meeting their needs for reduction of injury costs. While the workplace survey measured how well ergonomics was being implemented at the workplace, the user and provider surveys provided data on the roles of the decision makers beyond the workplace.

Detailed audit results are provided in Drury et al. (1999), so only examples and overall conclusions are covered in this chapter. Workplaces showed some evidence of good ergonomic practice, with generally satisfactory thermal, visual, and auditory environments. There were some significant differences ($p < 0.05$) between workplace types rather than between divisions or plants; for example, better lighting (> 700 lux) was associated with inspection and sewing. Also, higher thermal load was associated with laundries and machine load/unload. Overall, 83% of workplaces met the ASHRAE (1990) summer comfort zone criteria. As seen in Table 12, the main ergonomics problem areas were in poor posture and manual materials handling. Where operators were seated (only 33% of all workplaces) seating was relatively good. In fact, many in the workforce had been supplied with well-designed chairs as part of the ergonomics program.

To obtain a broad perspective, the three general factors at the end of Table 9 were analyzed. Apart from cycle time (W48), the questions related to workers having seen the corporate ergonomics video (W49) and having experienced a workplace or methods change (W50). Both should have received a "yes" response if the ergonomics program were reaching the whole workforce. In fact, both showed highly significant differences between plants ($X_8^2 = 92.0$, $p < 0.001$, and $X_8^2 = 22.2$, $p < 0.02$, respectively). Some of these differences were due to two divisions lagging in ergonomics implementation, but even beyond this were large between-plant differences. Overall, 62% of the workforce had seen the ergonomics video, a reasonable value but one with wide variance between plants and divisions. Also, 38% of workplaces had experienced some change, usually ergonomics-related, a respectable figure after only two to three years of the program.

From the user and provider surveys an enhanced picture emerged. Again, there was variability between divisions and plants, but 94% of the users defined ergonomics as fitting the job to the operator rather than training or medical management of injuries. Most users had requested an ergonomic intervention within the past two months, but other "users" had never in fact used ergonomics.

The solutions employed ranged widely, with a predominance of job aids such as chairs or standing pads. Other frequent categories were policy changes (e.g., rest breaks, rotation, box weight reduction) and workplace adjustment to the individual operator. There were few uses of personal aids (e.g. splints) or referrals to MDs as ergonomic solutions. Changes to the workplace clearly predominated over changes to the individual, although a strong medical management program was in place when required. When questioned about ergonomics results, all mentioned safety (or workplace comfort or ease of use), but some also mentioned others. Cost or productivity benefits were the next most common response, with a few additional ones relating to employee relations, absence/turnover, or job satisfaction. Significantly, only one respondent mentioned quality.

The major user concern at the plant level was time devoted to ergonomics by providers. At the corporate level, the need was seen for more rapid job-analysis methods and corporate policies, such as on back belts or "good" chairs. Overall, 94% of users made positive comments about the ergonomics program.

Ergonomics providers were almost always trained in the corporate or division training seminars, usually near the start of the program. Providers' chief concern was for the amount of time and resources they could spend on ergonomics activities. Typically, ergonomics was only one job responsibility among many. Hence, broad programs, such as new chairs or back belts, were supported enthusiastically because they gave the maximum perceived impact for the time devoted. Other solutions presented included job aids, workplace redesign (e.g., moving from seated to standing jobs for long-seam sewing), automation, rest breaks, job rotation, packaging changes, and medical man-

TABLE 9 Ergonomics Audit: Workplace Survey with Overall Data

Number	Division	Plant	Job Type

1. Postural aspects

	Yes	No	Factor
W1	68%		Frequent extreme motions of back, neck, shoulders, wrists
W2	66%		Elbows raised or unsupported more than 50% of time
W3	22%		Upper limbs contact nonrounded edges
W4	73%		Gripping with fingers
W5	36%		Knee/foot controls

1.1 Seated

	Yes	No	Factor
W6	12%		Leg clearance restricted
W7	21%		Feet unsupported/legs slope down
W8	17%		Chair/table restricts thighs
W9	22%		Back unsupported
W10	37%		Chair height not adjustable easily

1.2 Standing

	Yes	No	Factor
W11	3%		Control requires weight on one foot more than 50% time
W12	37%		Standing surface hard
W13	92%		Work surface height not adjustable easily

1.3 Hand tools

	Yes	No	Factor
W14	77%		Tools require hand/wrist bending
W15	9%		Tools vibrate
W16	63%		Restricted to one-handed use
W17	39%		Tool handle ends in palm
W18	20%		Tool handle has nonrounded edges
W19	56%		Tool uses only 2 or 3 fingers
W20	9%		Requires continuous or high force
W21	41%		Tool held continuously in one hand

2. Vibration

	Yes	No	Factor
W22	14%		Vibration reaches body from any source

3. Manual materials handling

	Yes	No	Factor
W23	40%		More than 5 moves per minute
W24	36%		Loads unbalanced
W25	14%		Lift above head
W26	28%		Lift off floor
W27	83%		Reach with arms
W28	78%		Twisting
W29	60%		Bending trunk
W30	3%		Floor wet or slippery
W31	0%		Floor in poor condition
W32	17%		Area obstructs task
W33	4%		Protective clothing unavailable
W34	2%		Handles used

TABLE 9 (*Continued*)

Number			Division	Plant	Job Type
4. Visual aspects					
	Yes	No	Factor		
W35			Task nature: 1 = rough, 2 = moderate, 3 = fine, 4 = very fine		
W36			Glare/reflection: 0 = none, 1 = noticeable, 2 = severe		
W37			Colour contrast: 0 = none, 1 = noticeable, 2 = severe		
W38			Luminance contrast: 0 = none, 1 = noticeable, 2 = severe		
W39			Task illuminance, foot candles		
W40	69%		Luminance: Task > Midfield > Outerfield = yes		
5. Thermal aspects					
			Factor		
W41			Dry bulb temperature, °F		
W42			Relative humidity, %		
W43			Air speed: 1 = just perceptible, 2 = noticeable, 3 = severe		
W44			Metabolic cost		
W45			Clothing, clo value		
6. Auditory aspects					
			Factor		
W46			Maximum sound pressure level, dBA		
W47			Noise sources 1 = m/c, 2 = other m/c, 3 = general, 4 = other		
7. General factors					
	Yes	No	Factor		
W48			Primary cycle time, sec		
W49	62%		Seen ergonomics video		
W50	38%		Any ergonomics changes to workplace or methods		

agement. Specific needs were seen in the area of corporate or supplier help in obtaining standard equipment solutions and of more division-specific training. As with users, the practitioners enjoyed their ergonomics activity and thought it worthwhile.

Recommendations arising from this audit were that the program was reasonably effective at that time but had some long-term needs. The corporation saw itself as an industry leader and wanted to move beyond a relatively superficial level of ergonomics application. To do this would require more time resources for job analysis and change implementation. Corporate help could also be provided in developing more rapid analysis methods, standardized video-based training programs, and more standardized solutions to recurring ergonomics problems. Many of these changes have since been implemented.

On another level, the audit was a useful reminder to the company of the fact that it had incurred most of the up-front costs of a corporate ergonomics program, and was now beginning to reap the benefits. Indeed, by 1996, corporate injury costs and rates had decreased by about 20% per year after

TABLE 10 Ergonomics Audit: User Survey

Number	Division	Plant	Job Type
U1. What is ergonomics?			
U2. Who do you call to do ergonomics?			
U3. When did you last ask them to do ergonomics?			
U4. Describe what they did?			
U5. Who else should we talk to about ergonomics?			
U6. General comments on ergonomics.			

TABLE 11 Ergonomics Audit: Provider Survey

Number	Division	Plant	Job Type

P1. What do you do?
P2. How do you get contacted to do ergonomics?
P3. When were you last asked to do ergonomics?
P4. Describe what you did.
P5. How long have you been doing ergonomics?
P6. How were you trained in ergonomics?
P7. What percent of your time is spent on ergonomics?
P8. Where do you go for more detailed ergonomics help?
P9. What ergonomics implementation problems have you had?
P10. How well are you regarded by management?
P11. How well are you regarded by workforce?
P12. General comments on ergonomics.

peaking in 1993. Clearly, the ergonomics program was not the only intervention during this period, but it was seen by management as the major contributor to improvement. Even on the narrow basis of cost savings, the ergonomics program was a success for the corporation.

4.2. Error Reduction at a Colliery

In a two-year project, reported by Simpson (1994) and Fox (1992), the human error audit described in Section 3.2 was applied to two colliery haulage systems. The results of the first study will be presented here. In both systems, data collection focused on potential errors and the performance-shaping factors (PSFs) that can influence these errors. Data was collected by ''observation, discussion and measurement within the framework of the broader man-machine systems and checklist of PSFs,'' taking some 30–40 shifts at each site. The whole haulage system from surface operations to delivery at the coal face was covered.

The first study found 40 active failures (i.e., direct error precursors) and nine latent failures (i.e., dormant states predisposing the system to later errors). Four broad classes of active failures were:

1. Errors associated with locomaintenance (7 errors), e.g., fitting incorrect thermal cut-offs
2. Errors associated with locooperation (10 errors), e.g., locos not returned to service bay for 24-hour check.
3. Errors associated with loads and load security (7 errors); e.g., failure to use spacer wagons between overhanging loads
4. Errors associated with the design/operation of the haulage route (10 errors), e.g., continued use despite potentially unsafe track
5. Plus a small miscellaneous category

The latent failures were (Fox 1992):

1. Quality assurance in supplying companies
2. Supplies ordering procedures within the colliery
3. Locomotive design
4. Surface make-up of supplies
5. Lack of equipment at specific points
6. Training
7. Attitudes to safety
8. The safety inspection/reporting/action procedures

As an example from 3, Locomotive design, the control positions were not consistent across the locomotives fleet, despite all originating from the same manufacturer.

Using the slip/mistake/violation categorization, each potential error could be classified so that the preferred source of action (intervention) could be specified.

This audit led to the formation of two teams, one to tackle locomotive design issues and the other for safety reporting and action. As a result of team activities, many ergonomic actions were implemented. These included management actions to ensure a uniform wagon fleet, autonomous inspection/repair teams for tracks, and multifunctional teams for safety initiatives.

TABLE 12 Responses to Ergonomics User

Question and Issue	Corporate Mgt	Corporate Staff	Plant Mgt	Plant Staff
1. What is Ergonomics?				
1.1 Fitting job to operator	1	6	10	5
1.2 Fitting operator to job	0	6	0	0
2. Who do you call on to get ergonomics work done?				
2.1 Plant ergonomics people	0	3	3	2
2.2 Division ergonomics people	0	4	5	2
2.3 Personnel department	3	0	0	0
2.4 Engineering department	1	8	6	11
2.5 We do it ourselves	0	2	1	0
2.6 College interns	0	0	4	2
2.7 Vendors	0	0	0	1
2.8 Everyone	0	1	0	0
2.9 Operators	0	1	0	0
2.10 University faculty	0	0	1	0
2.11 Safety	0	1	0	0
3. When did you last ask them for help?				
3.1 Never	0	4	2	0
3.2 Sometimes/infrequently	2	0	1	0
3.3 One year or more ago	0	1	4	0
3.4 One month or so ago	0	0	2	0
3.5 less than 1 month ago	1	0	3	4
5. Who else should we talk to about ergonomics?				
5.1 Engineers	0	0	3	2
5.2 Operators	1	1	2	0
5.3 Everyone	0	0	2	0
6. General Ergonomics Comments				
6.1 Ergonomics Concerns				
6.11 Workplace design for safety/ease/stress/fatigue	2	5	13	5
6.12 Workplace design for cost savings/productivity	1	0	2	1
6.13 Workplace design for worker satisfaction	1	1	0	1
6.14 Environment design	2	1	3	0
6.15 The problem of finishing early	0	0	1	1
6.16 The Seniority/bumping problem	0	3	1	0
6.2 Ergonomics program concerns				
6.21 Level of reporting of ergonomics	0	1	7	0
6.22 Communication/who does erognomics	7	1	4	0
6.23 Stability/staffing of ergonomics	0	0	10	4
6.24 General evaluation of ergonomics				
Positive	1	3	3	4
Negative	4	10	10	3
6.25 Lack of financial support for ergonomics	0	0	1	0
6.26 Lack of priority for ergonomics	2	2	1	4
6.27 Lack of awareness of ergonomics	2	1	6	1

The outcome was that the accident rate dropped from 35.40 per 100,000 person-shifts to 8.03 in one year. This brought the colliery from worst in the regional group of 15 collieries to best in the group, and indeed in the United Kingdom. In addition, personnel indicators, such as industrial relations climate and absence rates, improved.

5. FINAL THOUGHTS ON HUMAN FACTORS AUDITS

An audit system is a specialized methodology for evaluating the ergonomic status of an organization at a point in time. In the form presented here, it follows auditing practices in the accounting field, and indeed in such other fields as safety. Data is collected, typically with a checklist, analyzed, and presented to the organization for action. In the final analysis, it is the action that is important to human factors engineers, as the colliery example above shows. Such actions could be taken using other methodologies, such as active redesign by job incumbents (Wilson 1994); audits are only one method of tackling the problems of manufacturing and service industries. But as Drury (1991) points out, industry's moves towards quality are making it more measurement driven. Audits fit naturally into modern management practice as measurement, feedback, and benchmarking systems for the human factors function.

REFERENCES

Alexander, D. C., and Pulat, B. M. (1985), *Industrial Ergonomics: A Practitioner's Guide,* Industrial Engineering and Management Press, Atlanta.

American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) (1989), "Physiological Principles, Comfort and Health," in *Fundamentals Handbook,* Atlanta.

AVISEM (1977), Techniques d'amélioration des conditions de travail dans l'industrie, Editions Hommes et Techniques, Suresnes, France.

Ayoub, M. M., and Mital, A. (1989), *Manual Materials Handling,* Taylor & Francis, London.

Bainbridge, L., and Beishon, R. J. (1964), "The Place of Checklists in ergonomic Job Analysis," in *Proceedings of the 2nd I.E.A. Congress* (Dortmund), Ergonomics Congress Proceedings Supplement.

Berchem-Simon, O., Ed. (1993), *Ergonomics Action in the Steel Industry*, EUR 14832 EN, Commission of the European Communities, Luxembourg.

Bolijn, A. J. (1993), "Research into the Employability of Women in Production and Maintenance Jobs in Steelworks," in *Ergonomics Action in the Steel Industry,* O. Berchem-Simon, Ed., EUR 14832 EN, Commission of the European Communities, Luxembourg, 201–208.

British Standards Institution (1965), "Office Desks, Tables and Seating," British Standard 3893, London.

Burger, G. C. E., and de Jong, J. R. (1964), "Evaluation of Work and Working Environment in Ergonomic Terms," *Aspects of Ergonomic Job Analysis,* 185–201.

Carson, A. B., and Carlson, A. E. (1977), *Secretarial Accounting,* 10th Ed. South Western, Cincinnati.

Cakir, A., Hart, D. M., and Stewart, T. F. M. (1980), *Visual Display Terminals,* John Wiley & Sons, New York, pp. 144–152, 159–190, App. I.

Chervak, S., and Drury, C. G. (1995), "Simplified English Validation," in *Human Factors in Aviation Maintenance—Phase 6 Progess Report,* DOT/FAA/AM-95/xx, Federal Aviation Administration/ Office of Aviation Medicine, National Technical Information Service, Springfield, VA.

Degani, A., and Wiener, E. L. (1990), "Human Factors of Flight-Deck Checklists: The Normal Checklist," NASA Contractor Report 177548, Ames Research Center, CA.

Department of Defense (1989), "Human Engineering Design Criteria for Military Systems, Equipment and Facilities," MIL-STD-1472D, Washington, DC.

Dirken, J. M. (1969), "An Ergonomics Checklist Analysis of Printing Machines," *ILO,* Geneva, Vol. 2, pp. 903–913.

Drury, C. G. (1990a), "The Ergonomics Audit," in *Contemporary Ergonomics,* E. J. Lovesey, Ed., Taylor & Francis, London, pp. 400–405.

Drury, C. G. (1990b), "Computerized Data Collection in Ergonomics," in *Evaluation of Human Work,* J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London, pp. 200–214.

Drury, C. G. (1991), "Errors in Aviation Maintenance: Taxonomy and Control," in *Proceedings of the 35th Annual Meeting of the Human Factors Society* (San Francisco), pp. 42–46.

Drury, C. G. (1992), "Inspection Performance," in *Handbook of Industrial Engineering,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 2282–2314.

Drury, C. G., and Wick, J. (1984), "Ergonomic Applications in the Shoe Industry," in *Proceedings of the International Conference on Occupational Ergonomics,* Vol. 1, pp. 489–483.

Drury, C. G., Kleiner, B. M., and Zahorjan, J. (1989), "How Can Manufacturing Human Factors Help Save a Company: Intervention at High and Low Levels," in *Proceedings of the Human Factors Society 33rd Annual Meeting,* Denver, pp. 687–689.

Drury, C. G., Prabhu, P., and Gramopadhye, A. (1990), "Task Analysis of Aircraft Inspection Activities: Methods and Findings," in *Proceedings of the Human Factors Society 34th Annual Conference,* (Santa Monica, CA), pp. 1181–1185.

Drury, C. G., Broderick, R. L., Weidman, C. H., and Mozrall, J. R. (1999), "A Corporate-Wide Ergonomics Programme: Implementation and Evaluation," *Ergonomics,* Vol. 42, No. 1, pp. 208–228.

Easterby, R. S. (1967), "Ergonomics Checklists: An Appraisal," *Ergonomics,* Vol. 10, No. 5, pp. 548–556.

Fanger, P. O. (1970), *Thermal Comfort, Analyses and Applications in Environmental Engineering,* Danish Technical Press, Copenhagen.

Fox, J. G. (1992), ‘‘The Ergonomics Audit as an Everyday Factor in Safe and Efficient Working,’’ *Progress in Coal, Steel and Related Social Research,* pp. 10–14.

Givoni, B., and Goldman, R. F. (1972), ‘‘Predicting Rectal Temperature Response to Work, Environment, and Clothing,’’ *Journal of Applied Physiology,* Vol. 32, No. 6, pp. 812–822.

Groth, K. M. (1989), ‘‘The Modular Work Analysis System (MAS),’’ in *Recent Developments in Job Analysis, Proceedings of the International Symposium on Job Analysis,* (University of Hohenheim, March 14–15), Taylor & Francis, New York, pp. 253–261.

Groupement des Entreprises Sidérurgiques et Minières (GESIM) (1988), Connaissance du poste de travail, II conditions de l'activitC, GESIM, Metz.

Guélaud, F., Beauchesne, M.-N., Gautrat, J. and Roustang, G. (1975), *Pour une analyse des conditions de travail ouvrier dans l'entreprise,* 3rd Ed., Armand Colin, Paris.

Hacker, W., Iwanowa, A., and Richter, P. (1983), *Tätigkeitsbewertungssystem,* Psychodiagnostisches Zentrum, Berlin.

Hasslequist, R. J. (1981), ‘‘Increasing Manufacturing Productivity Using Human Factors Principles,’’ in *Proceedings of the Human Factors Society 25th Annual Conference,* (Santa Monica, CA), pp. 204–206.

Illuminating Engineering Society (1993), *Lighting Handbook, Reference and Application,* 8th Ed., The Illuminating Engineering Society of North America, New York.

International Civil Aviation Organization (ICAO) (1989), *Human Factors Digest No. 1: Fundamental Human Factors Concepts,* Circular 216-AN/131, Montreal.

International Organization for Standardization (ISO) (1987), *Assessment of Noise-Exposure During Work for Hearing Conservation Purposes,* ISO, Geneva.

Jones, D. M., and Broadbent, D. E. (1987), ‘‘Noise,’’ in *Handbook of Human Factors Engineering,* G. Salvendy, Ed., John Wiley & Sons, New York.

Kerlinger, F. N. (1964), *Foundations of Behavioral Research,* Holt, Rinehart & Winston, New York.

Keyserling, W. M., Stetson, D. S., Silverstein, B. A., and Brouwer, M. L. (1993), ‘‘A Checklist for Evaluating Ergonomic Risk Factors Associated with Upper Extremity Cumulative Trauma Disorders, *Ergonomics,* Vol. 36, No. 7, pp. 807–831.

Keyserling, W. M., Brouwer, M., and Silverstein, B. A. (1992), ‘‘A Checklist for Evaluation Ergonomic Risk Factors Resulting from Awkward Postures of the Legs, Truck and Neck,’’ *International Journal of Industrial Ergonomics,* Vol. 9, No. 4, pp. 283–301.

Kirwan, B. (1989), ‘‘A Human Factors and Human Reliability Programme for the Design of a Large UK Nuclear Chemical Plant,’’ in *Proceedings of the Human Factors Society 33rd Annual Meeting—1989* (Denver), pp. 1009–1013.

Klimer, F., Kylian, H., Schmidt, K.-H., and Rutenfranz, J. (1989), ‘‘Work Analysis and Load Components in an Automobile Plant after the Implementation of New Technologies,’’ in *Recent Developments in Job Analysis,* K. Landau and W. Rohmert, Eds., Taylor & Francis, New York, pp. 331–340.

Kogi, K. (1994), ‘‘Introduction to WISE (Work Improvement in Small Enterprises) Methodology and Workplace Improvements Achieved by the Methodology in Asia,’’ in *Proceedings of the 12th Triennial Congress of the International Ergonomics Association,* Vol. 5, Human Factors Association of Canada, Toronto, pp. 141–143.

Kogi, K., and Kuorinka, I., Eds. (1995), *Ergonomic Checkpoints,* ILO, Switzerland.

Koli, S. T. (1994), ‘‘Ergonomic Audit for Non-Repetitive Task,’’ M.S. Thesis. State University of New York at Buffalo.

Koli, S., Drury, C. G., Cuneo, J. and Lofgren, J. (1993), ‘‘Ergonomic Audit for Visual Inspection of Aircraft,’’ in *Human Factors in Aviation Maintenance—Phase Four, Progress Report, DOT/FAA/AM-93/xx,* National Technical Information Service, Springfield, VA.

Landau, K., and Rohmert, W. (1981), *Fallbeispiele zurArbeitsanalyse,* Hans Huber, Bern.

Landau, K., and Rohmert, W., Eds. (1989), *Recent Developments in Job Analysis: Proceedings of the International Symposium on Job Analysis* (University of Hohenheim, March 14–15), Taylor & Francis, New York.

Leitner, K., and Greiner, B. (1989), ‘‘Assessment of Job Stress: The RHIA Instrument,’’ in *Recent Developments in Job Analysis: Proceedings of the International Symposium on Job Analysis,* K. Landau and W. Rohmert, Eds. (University of Hohenheim, March 14–15), Taylor & Francis, New York.

Malde, B. (1992), "What Price Usability Audits? The Introduction of Electronic Mail into a User Organization," *Behaviour and Information Technology,* Vol. 11, No. 6, pp. 345–353.

Malone, T. B., Baker, C. C., and Permenter, K. E. (1988), "Human Engineering in the Naval Sea Systems Command," in *Proceedings of the Human Factors Society—32nd Annual Meeting—1988* (Anaheim, CA), Vol. 2, pp. 1104–1107.

Mattila, M., and Kivi, P. (1989), "Job Load and Hazard Analysis: A Method for Hazard Screening and Evaluation," in *Recent Developments in Job Analysis: Proceedings of the International Symposium on Job Analysis,* K. Landau and W. Rohmert, Eds. (University of Hohenheim, March 14–15), Taylor & Francis, New York, pp. 179–186.

McClelland, I. (1990), "Product Assessment and User Trials," in *Evaluation of Human Work,* J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, New York, pp. 218–247.

McCormick, E. J. (1979), *Job Analysis: Methods and Applications,* AMACOM, New York.

McCormick, W. T., Mecham, R. C., and Jeanneret, P. R. (1969), *The Development and Background of the Position Analysis Questionnaire,* Occupational Research Center, Purdue University, West Lafayette, IN.

Mir, A. H. (1982), "Development of Ergonomic Audit System and Training Scheme," M.S. Thesis, State University of New York at Buffalo.

Muller-Schwenn, H. B. (1985), "Product Design for Transportation," in *Ergonomics International 85,* pp. 643–645.

National Institute for Occupational Safety and Health (NIOSH) (1981), *Work Practices Guide for Manual Lifting*, DHEW-NIOSH publication 81-122, Cincinnati.

Occupational Health and Safety Authority (1990), *Inspecting the Workplace,* Share Information Booklet, Occupational Health and Safety Authority, Melbourne.

Occupational Safety and Health Administration (OSHA), (1990), *Ergonomics Program Management Guidelines for Meatpacking Plants*, Publication No. OSHA-3121, U.S. Department of Labor, Washington, DC.

Osburn, H. G. (1987), "Personnel Selection," in *Handbook of Human Factors,* G. Salvendy Ed., John Wiley & Sons, New York, pp. 911–933.

Panter, W. (1993), "Biomechanical Damage Risk in the Handling of Working Materials and Tools: Analysis, Possible Approaches and Model Schemes," in *Ergonomics Action in the Steel Industry,* O. Berchem-Simon, Ed., EUR 14832 EN, Commission of the European Communities, Luxembourg.

Parsons, K. C. (1992), "The Thermal Audit: A Fundamental Stage in the Ergonomics Assessment of Thermal Environment," in *Contemporary Ergonomics 1992,* E. J. Lovesey, Ed., Taylor & Francis, London, pp. 85–90.

Patel, S., Drury, C. G., and Prabhu, P. (1993), "Design and Usability Evaluation of Work Control Documentation," in *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (Seattle), pp. 1156–1160.

Portillo Sosa, J. (1993), "Design of a Computer Programme for the Detection and Treatment of Ergonomic Factors at Workplaces in the Steel Industry," in *Ergonomics Action in the Steel Industry,* O. Berchem-Simon, Ed., EUR 14832 EN, Commission of the European Communities, Luxembourg, pp. 421–427.

Pulat, B. M. (1992), *Fundamentals of Industrial Ergonomics,* Prentice Hall, Englewood Cliffs, NJ.

Putz-Anderson, V. (1988), *Cumulative Trauma Disorders: A Manual for Musculo-Skeletal Diseases of the Upper Limbs,* Taylor & Francis, London.

Rasmussen, J. (1987), "Reasons, Causes and Human Error," in *New Technology and Human Error,* J. Rasmussen, K. Duncan, and J. Leplat, Eds., John Wiley & Sons, New York, pp. 293–301.

Reason, J. (1990), *Human Error,* Cambridge University Press, New York.

Régie Nationale des Usines Renault (RNUR) (1976), *Les profils de ostes, Méthode d'analyse des conditions de travail,* Collection Hommes et Savoir, Masson, Sirtès, Paris.

Rohmert, W., and Landau, K. (1983), *A New Technique for Job Analysis,* Taylor & Francis, London.

Rohmert, W., and Landau, K. (1989), "Introduction to Job Analysis," in *A New Technique for Job Analysis, Part 1,* Taylor & Francis, London, pp. 7–22.

Seeber, A., Schmidt, K.-H., Kierswelter, E., and Rutenfranz, J. (1989), "On the Application of AET, TBS and VERA to Discriminate between Work Demands at Repetitive Short Cycle Tasks, in *Recent Developments in Job Analysis,* K. Landau and W. Rohmert, Eds., Taylor & Francis, New York, pp. 25–32.

Simpson, G. C. (1994), "Ergonomic Aspects in Improvement of Safe and Efficient Work in Shafts," in *Ergonomics Action in Mining,* EUR 14831, Commission of the European Communities, Luxembourg, pp. 245–256.

Sinclair, M. A. (1990), "Subjective Assessment," in *Evaluation of Human Work,* J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London, pp. 58–88.

Volpert, W., Oesterreich, R., Gablenz-Kolakovic, S., Krogoll, T., and Resch, M. (1983), *Verfahren zur Ermittlung von Regulationserfordernissen in der Arbeitstätigkeit* (*VERA*), TÜV Rheinland, Cologne.

Wagner, R. (1989), "Standard Methods Used in French-Speaking Countries for Workplace Analysis," in *Recent Developments in Job Analysis,* K. Landau and W. Rohmert, Eds., Taylor & Francis, New York, pp. 33–42.

Wagner, R. (1993), "Ergonomic Study of a Flexible Machining Cell Involving Advanced Technology," in *Ergonomics Action in the Steel Industry,* EUR 14832, Commission of the European Communities, Luxembourg, pp. 157–170.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks, *Rapid Communications, Ergonomics,* Vol. 36, No. 7, pp. 748–776.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance,* 2nd Ed., HarperCollins, New York.

Wilson, J. R. (1994), "A Starting Point for Human-Centered Systems," in *Proceedings of the 12th Triennial Congress of the International Ergonomics Association* (Toronto), Canada, Vol. 6, No. 1, pp. 141–143.

Wright, P., and Barnard, P. (1975), "Just Fill in This Form—A Review for Designers," *Applied Ergonomics,* Vol. 6, pp. 213–220.

Yoshida, H., and Ogawa, K. (1991), "Workplace Assessment Quideline—Checking Your Workplace," in *Advances in Industrial Ergonomics and Safety III,* W. Karwowski and J. W. Yates, Eds., Taylor & Francis, London, pp. 23–28.

CHAPTER 43

Design for Occupational Health and Safety

Michael J. Smith
Pascale Carayon
Ben-Tzion Karsh
University of Wisconsin-Madison

1. **INTRODUCTION** **1157**
2. **INTERDISCIPLINARY NATURE OF OCCUPATIONAL SAFETY AND HEALTH** **1157**
3. **A PUBLIC HEALTH MODEL FOR OCCUPATIONAL SAFETY AND HEALTH PROTECTION** **1157**
4. **A BALANCE MODEL OF OCCUPATIONAL SAFETY AND HEALTH PERFORMANCE** **1159**
   - 4.1. The Person 1159
   - 4.2. Technology and Materials 1160
   - 4.3. Task Factors 1160
   - 4.4. The Work Environment 1161
   - 4.5. Organizational Structure 1161
5. **SAFETY AND HEALTH ORGANIZATIONS, AGENCIES, LAWS, AND REGULATIONS** **1162**
   - 5.1. The Occupational Safety and Health Administration 1162
   - 5.2. The National Institute for Occupational Safety and Health 1163
   - 5.3. State Agencies 1164
   - 5.4. The Centers for Disease Control and Prevention 1164
   - 5.5. The Bureau of Labor Statistics 1164
   - 5.6. The Environmental Protection Agency 1164
   - 5.7. Other Agencies and Groups 1164
   - 5.8. Safety and Ergonomics Program Standards 1165
6. **DEFINING OCCUPATIONAL INJURIES AND DISEASES** **1168**
7. **WORKPLACE HAZARDS** **1168**
8. **MEASURING HAZARD POTENTIAL AND SAFETY PERFORMANCE** **1171**
   - 8.1. Inspection Programs 1171
   - 8.2. Illness and Injury Statistics 1173
   - 8.3. Incident Reporting 1174
9. **CONTROLLING WORKPLACE HAZARDS** **1175**
   - 9.1. Engineering Controls 1175
   - 9.2. Human Factors Controls 1176
     - 9.2.1. Informing 1176
     - 9.2.2. Promoting Safe and Healthful Behavior 1177
     - 9.2.3. Workplace and Job Design 1177
   - 9.3. Organizational Design 1179
     - 9.3.1. Safety Training 1180
     - 9.3.2. Hazard Reduction through Improved Work Practices 1181
10. **SAFETY PROGRAMS** **1183**
11. **PARTICIPATIVE APPROACHES TO RESPOND TO THE EMERGING HAZARDS OF NEW TECHNOLOGIES** **1184**
    - 11.1. Quality Improvement 1184
    - 11.2. International Organization for Standardization 1185
    - 11.3. Social Democracy 1186

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

11.4. Hazard Survey 1186
11.5. Employee/Management Ergonomics Committee 1187
12. CONCLUSIONS 1187
REFERENCES 1188
ADDITIONAL READING 1190
APPENDIX: USEFUL WEB INFORMATION SOURCES 1190

1. INTRODUCTION

Each year in the United States, thousands of employees are killed on the job, many times that number die of work-related diseases, and millions suffer a work-related injury or health disorder (BLS 1998a, 1999). According to the International Labour Organization (ILO 1998), about 250 million workers worldwide are injured annually on the job, 160 million suffer from occupational diseases, and approximately 335,0000 die each year from occupational injuries. In the United States, the occupational injury and illness incidence rates per 100 full-time workers have been generally decreasing since 1973, but as of 1998, the illness and injury rate was still 6.7 per 100 employees and the injury rate was 6.2 (BLS 1998a). There were a total of 5.9 million occupational injuries and illnesses in 1998 in private industry (BLS 1998a). These figure represented the sixth year in a row of declining injury rates. Overall lost workday rates have steadily declined from 1990 to 1998, but cases with days of restricted work activity have increased. There were just over 60000 occupational fatalities in the private sector in 1998. These work-related deaths and injuries have enormous costs. In the United States alone, it was estimated that in 1992 the direct costs (e.g., medical, property damage) totaled $65 billion and the indirect costs (e.g., lost earnings, workplace training and restaffing, time delays) totaled $106 billion (Leigh et al. 1997). Of the U.S. dollar figures presented, approximately $230 million of the direct costs and $3.46 billion of the indirect costs were related to fatal occupational injuries. Nonfatal injuries accounted for $48.9 billion in direct costs and $92.7 billion in indirect costs (the rest was cost due to death and morbidity from occupational illnesses). These estimates assumed 6500 occupational fatalities and 13.2 million nonfatal injuries.

The workplace continues to undergo rapid change with the introduction of new technologies and processes. Many new processes, such as genetic engineering and biotechnology, introduce new hazards that are challenging, particularly since we do not know much about their potential risks. Will current hazard-control methods be effective in dealing with these new hazards? Our challenge is to protect workers from harm while taking advantage of the benefits of this new technology. To achieve this, we must be ready to develop new safety and health methods to deal with new technology.

This chapter will examine the causation and prevention of occupational diseases and injuries, with an emphasis on recognizing and evaluating hazards, determining disease/injury potential, and defining effective intervention strategies. Due to the huge amount of pertinent information on each of these topic, it cannot be all inclusive. Rather, it will provide direction for establishing effective detection and control methods. Additional resources are provided in the Appendix for more detailed information about the subjects covered.

2. INTERDISCIPLINARY NATURE OF OCCUPATIONAL SAFETY AND HEALTH

Occupational health and safety has its roots in several disciplines, including such diverse fields as engineering, toxicology, epidemiology, medicine, sociology, psychology, and economics. Essentially, occupational health and safety is a multidisciplinary endeavor requiring knowledge from diverse sources to deal with the interacting factors of people, technology, the work environment, and the organization of work activities. Any successful approach for the prevention of injuries and health disorders must recognize the need to deal with these diverse factors using the best available tools from various disciplines and to organize a systematic and balanced effort. Large companies have many resources that can be called upon, but small companies do not have such resources and may need to contact local, state, and federal agencies for information, advice, and consultation.

3. A PUBLIC HEALTH MODEL FOR OCCUPATIONAL SAFETY AND HEALTH PROTECTION

Figure 1 illustrates a general public health approach for improving the safety and health of the workforce (HHS 1989). It begins with surveillance. We have to know what the hazards are and their safety and health consequences before we can establish priorities on where to apply our limited resources and develop intervention strategies. At the national level, there are statistics on occupational

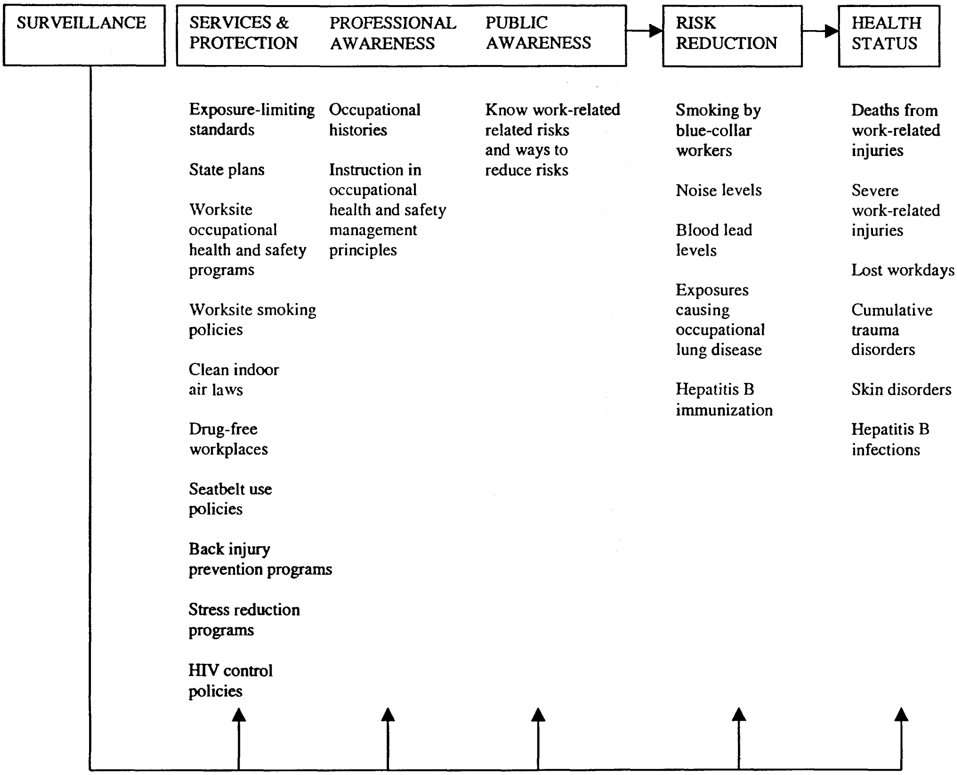

SURVEILLANCE
SERVICES & PROTECTION
PROFESSIONAL AWARENESS
PUBLIC AWARENESS
RISK REDUCTION
HEALTH STATUS
Exposure-limiting standards
State plans
Worksite occupational health and safety programs
Worksite smoking policies
Clean indoor air laws
Drug-free workplaces
Seatbelt use policies
Back injury prevention programs
Stress reduction programs
HIV control policies
Occupational histories
Instruction in occupational health and safety management principles
Know work-related related risks and ways to reduce risks
Smoking by blue-collar workers
Noise levels
Blood lead levels
Exposures causing occupational lung disease
Hepatitis B immunization
Deaths from work-related injuries
Severe work-related injuries
Lost workdays
Cumulative trauma disorders
Skin disorders
Hepatitis B infections

Figure 1 A Public Health Model for Improving the Safety and Health of the Workforce. (*Source:* HHS 1989)

injuries and illnesses from various sources including the U.S. Bureau of Labor Statistics, U.S. National Center for Health Statistics, and National Safety Council. These statistics provide an indication of the most hazardous jobs and industries. In addition, each state has workers' compensation statistics that can provide information about local conditions. This same kind of surveillance can be done by each company at the plant level to identify hazardous jobs and operations. Plant level exposure, hazard, and injury/illness records can be examined periodically to establish trends and determine plant hot spots that need immediate attention.

The second level of the model defines specific services and protection measures to prevent the occurrence of hazards and injuries/illnesses. It also includes services for quick and effective treatment if injury or illness should occur. For example, the safety staff keep track of the state and federal requirements regarding workplace exposure standards. The safety staff then establishes a process for enforcing the standards through inspections and correction activities. Additional plant safety and health programs deal with employee safety and health training (Cohen and Colligan 1998), emergency medical treatment facilities, and arrangements with local clinics. The basic thrust of these multifaceted approaches is to reduce or eliminate adverse workplace exposures and their consequences and provide effective treatment when injuries and illnesses occur.

At the next level of the model is the need to heighten the awareness of the professionals in the workplace who have to make the decisions that affect safety, such as the manufacturing engineers, accountants, operations managers, and supervisors. In addition, workers need to be informed so that they know about the risks and consequences of occupational exposures. There is substantial workplace experience to attest that managers and employees do not always agree on the hazardous nature of workplace exposures. It is vital that those persons who can have a direct impact on plant exposures, such as managers and employees, have the necessary information in an easily understandable and useful form to be able to make informed choices that can lead to reductions in adverse exposures. Providing the appropriate information is the first basic step for good decision making. However, it does not ensure good decisions. Knowledgeable and trained professionals are also needed to provide proper advice on how to interpret and use the information.

The next stage in the model is the reduction of risk by having known adverse agents of injury or illness controlled or removed. The reduction in risk leads to the final stage, which is an improvement in health and safety status of the workforce. This general model leads to a more specific approach that can be applied at specific workplaces.

4. A BALANCE MODEL OF OCCUPATIONAL SAFETY AND HEALTH PERFORMANCE

An important consideration in conceptualizing an approach to occupational health and safety is an understanding of the many personal and workplace factors that interact to cause exposures and accidents. Any strategy to control these exposures and accidents should consider a range of factors and their influences on each other. A model of human workplace interaction is presented in Figure 2. Each element of this model can produce hazardous exposures, for instance a work environment with chemical exposures. These elements also interact to produce hazardous exposures. Examples of these interactions are when high-workload tasks are performed in environments with chemical exposures creating greater fatigue, or more inhalation of the chemicals. Another example is when the person uses machinery and tools that have hazardous characteristics and there is high work pressure to complete the task quickly. Then, the potential for an acute injury increases. Each single factor of the balance model has specific characteristics that can influence exposures to hazards and accident potential or disease risk. At the same time, each interacts with the others to increase exposures and risks or reduce exposures and risks. The model indicates that the person is exposed to loads and hazards that create acute or chronic strains. These strains can lead directly to injury or illness, or they may increase accident potential and/or disease risk.

4.1. The Person

A wide range of individual attributes can affect exposure and accident potential. These include intellectual capabilities and aptitudes, perceptual–motor abilities, physical capabilities such as strength and endurance, current health status, susceptibilities to disease, and personality. A person's education, knowledge, and aptitude affect his or her ability to recognize hazards. They also influence how much a person will learn from training about hazards and safety. An important aspect of injury prevention is to have knowledgeable employees who can determine the potential danger of an exposure and respond appropriately. This requires some previous experience with a hazard and/or training about the nature of the hazard, its injury potential, and ways to control it. Employees must have the ability to learn and retain the information that they are given in training classes. There is a fundamental need for employees to have adequate background and education to be able to apply their intelligence and acquire new knowledge through training. Of specific importance are reading and language skills

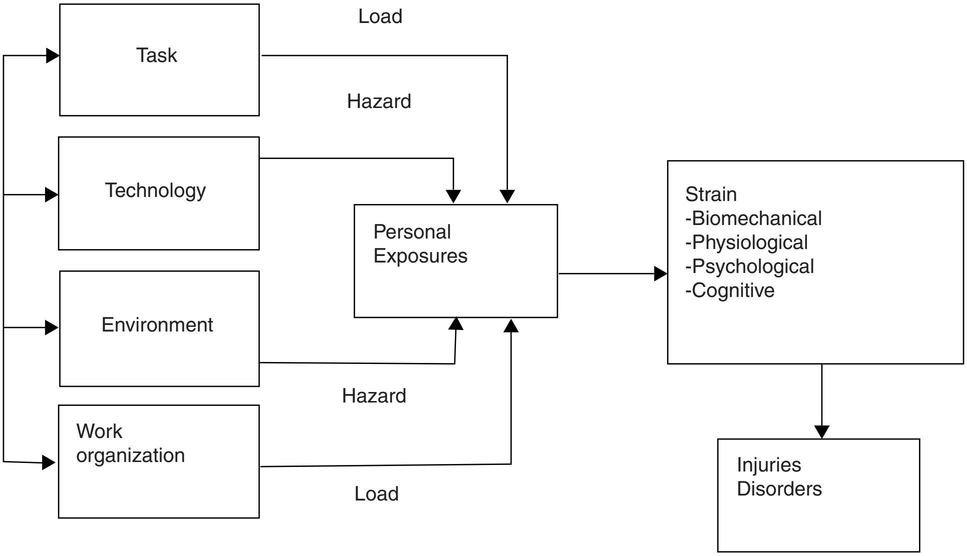


Figure 2 Model of the Work System Risk Process. (Adapted from Smith and Sainfort 1989; Smith et al. 1999)

so that employees can be trained and instructed properly. When employees lack sufficient intelligence or knowledge-acquisition skills, much greater emphasis must be placed on engineering controls.

Physiological considerations such as strength, endurance, and susceptibilities to fatigue, stress, or disease are also of importance. Some jobs demand high energy expenditure and strength requirements. For these, employees must have adequate physical resources to do the work safely.

Another example deals with a concern about women being exposed to reproductive hazards in select industries, for instance in lead processing or synthetic hormone production. Biological sensitivity to substances may increase the risk of an adverse health outcome. Where adequate protection can be provided, there is no logical reason to exclude employees on the basis of gender or biological sensitivity. However, with certain exposures the courts in the United States have ruled that biologically sensitive employees may be barred from jobs in which these biologically adverse exposures cannot be adequately controlled or guarded against.

An attribute related to physical capacity is the perceptual/motor skills of an individual, such as eye–hand coordination. These skills vary widely among individuals and may have more health and safety significance than strength or endurance because they come into play in the moment-by-moment conduct of work tasks. While strength may influence the ability to perform a specific component of a task, perceptual/motor skills are involved in all aspects of manual tasks. Thus, perceptual/motor skills affect the quality with which a task is carried out as well as the probability of a mistake that could cause an exposure or accident.

An individual attribute that should also be considered is personality. For many years it was believed that an employee's personality was the most significant factor in accident causation and that certain workers were more ''accident prone'' than other workers. There is some evidence that ''affectivity'' is related to occupational injuries (Iverson and Erwin 1997). However, in an earlier review of the accident proneness literature (Century Research Corp. 1973), it was determined that individual characteristics such as personality, age, sex, and intellectual capabilities were not significant determinants of accident potential or causation. Rather, situational considerations such as the hazard level of the job tasks and daily personal problems were more important in determining accident risk. There is some evidence that individuals are at greater or lesser risk at different times in their working careers due to these situational considerations. Such situational considerations may account for findings that younger and older employees have higher than average injury rates (Laflamme 1997; Laflamme and Menckel 1995).

It is critical that a proper fit be achieved among employees and other elements of the model. This can occur with proper hazard orientation, training, skill enhancement, ergonomic improvements, and proper engineering of the tasks, technology, and environment.

4.2. Technology and Materials

As with personal attributes, characteristics of the machinery, tools, technology, and materials used by the worker can influence the potential for an exposure or accident. One consideration is the extent to which machinery and tools influence the use of the most appropriate and effective perceptual/motor skills and energy resources. The relationship between the controls of a machine and the action of that machine dictates the level of perceptual/motor skill necessary to perform a task. The action of the controls and the subsequent reaction of the machinery must be compatible with basic human perceptual/motor patterns. If not, significant interference with performance can occur which may lead to improper responses that can cause accidents. In addition, the adequacy of feedback about the action of the machine affects the performance efficiency that can be achieved and the potential for an operational error.

The hazard characteristics of materials will affect exposure and risk. More hazardous materials inherently have a greater probability of adverse health outcomes upon exposure. Sometimes employees will be more careful when using materials that they know have a high hazard potential. But this can only be true when they are knowledgeable about the material's hazard level. If a material is very hazardous, often less-hazardous materials available can be substituted. The same is true for hazardous work processes. Proper substitution can decrease the risk of injury or illness, but care must be taken to ensure that the material or process being substituted is really safer and that it mixes well with the entire product formulation or production/assembly process.

4.3. Task Factors

The demands of a work activity and the way in which it is conducted can influence the probability of an exposure or accident. In addition, the influence of the work activity on employee attention, satisfaction, and motivation can affect behavior patterns that increase exposure and accident risk. Work task considerations can be broken into the physical requirements, mental requirements, and psychological considerations. The physical requirements influence the amount of energy expenditure necessary to carry out a task. Excessive physical requirements can lead to fatigue, both physiological and mental, which can reduce worker capabilities to recognize and respond to workplace hazards.

Typically, relatively high workloads can be tolerated for short periods of time. However, with longer exposure to heavy workloads and multiple exposures to shorter-duration heavy workloads, fatigue accumulates and worker capacity is diminished.

Other task considerations dealing with the content of the task that are related to the physical requirements include the pace or rate of work, the amount of repetition in task activities, and work pressure due to production demands. Task activities that are highly repetitive and paced by machinery rather than by the employee tend to be stressful. Such conditions also diminish an employee's attention to hazards and the capability to respond to a hazard due to boredom. These conditions may produce cumulative trauma disorders to the musculoskeletal system when the task activity cycle time is short and constant. Tasks with relatively low workload and energy expenditure can be very hazardous due to the high frequency of muscle and joint motions and boredom, which leads to employee inattention to hazards.

Psychological task content considerations, such as satisfaction with job tasks, the amount of control over the work process, participation in decision making, the ability to use knowledge and skills, the amount of esteem associated with the job, and the ability to identify with the end products of the task activity, can influence employee attention and motivation. They can also cause job stress, which can affect employee ability to attend to, recognize, and respond to hazards as well as the motivation needed to be concerned with their health and safety considerations. Negative influences can bring about emotional disturbances that limit the employee's capabilities to respond. Task considerations are a central aspect in reducing worker fatigue and stress and enhancing worker motivation for positive health and safety behavior. Tasks must be designed to fit the workforce capabilities and needs and be compatible with the other elements of the model.

4.4. The Work Environment

The work environment exposes employees to materials, chemicals, and physical agents that can cause harm or injury if the exposure exceeds safe limits. Such exposures vary widely from industry to industry, from job to job, and from task to task. Hazard exposure in the work environment influences the probability of an injury or illness, and the extent of exposure often determines the seriousness. Differences in hazard potential are a central basis for determining the rates companies pay for workers' compensation insurance. The central concept is one of relative risk. The greater the number of hazards, the more serious their potential to inflict injury or illness, then the greater the probability of an accident. The greater the probability of a serious accident, the higher the insurance premium. The hazard potential of different environmental factors can be evaluated using various federal, state, and local codes and standards for worker protection and limits established by scientific groups.

Environmental conditions can also hamper the ability of employees to use their senses (poor lighting, excessive noise) and reduce an employee's abilities to respond or react to a hazardous situation. Moderate levels of noise, high levels of heat, and the presence of dust/fumes or gases/vapors have been linked to higher risk of occupational fatalities (Barreto et al. 1997). The environment should be compatible with worker perceptual/motor, energy expenditure, and motivational needs to encourage hazard recognition, precautions, and the desire to do tasks in the proper way.

4.5. Organizational Structure

A company's health and safety performance can be influenced by management policies and procedures, the way that work tasks are organized into plant-wide activities, the style of employee supervision, the motivational climate in the plant, the amount of socialization, interaction between employees, the amount of social support employees receive, and management attitude towards safety. The last point, management attitude, has often been cited as the most critical element in a successful safety program (Cohen, 1977; Smith et al. 1978; Cleveland et al. 1979). If the individuals who manage an organization have a disregard for safety considerations, then employees tend not be very motivated to work safely. Conversely, if the management attitude is that safety considerations are paramount, that is, even more important than production goals, then managers, supervisors, and employees will put great diligence into health and safety efforts.

There are other organizational considerations important in safety performance that are related to management atmosphere and attitudes. For instance, a management structure that provides for frequent employee interaction, positive supervisor relations, and frequent social support leads to an organizational climate that is conducive to cooperative efforts in hazard recognition and control. Such a structure also allows for the motivational climate necessary to encourage appropriate safety behavior. Supervisor and coworker social support have been shown to reduce the risk of occupational injuries (Iverson and Erwin 1997).

A consistent factor in accident causation is work pressure for greater production, or faster output, or to correct problems quickly to continue production or reduce customer waiting time. This work pressure can be exacerbated by technology malfunctions, insufficient staffing, and improper work standards. Management emphasis on reducing costs, enhancing profits, and increasing stock price often stretch the limits of the capabilities of the workforce and technology. When breakdowns occur or operations are not running normally, employees tend to take risks to keep production online or get it back online quickly. It is during these nonnormal operations that many accidents occur.

Management must provide adequate resources to meet production goals and accommodate nonnormal operations. Management must also establish policies to ensure that employees and supervisors do not take unnecessary risks to ensure production.

5. SAFETY AND HEALTH ORGANIZATIONS, AGENCIES, LAWS, AND REGULATIONS

History has shown that ensuring the safety and health of the workforce cannot be left solely to the discretion of owners and employers. There are those who take advantage of their position and power and do not provide adequate safeguards. Today this is true for only a small percentage of employers but unfortunately, even well-intentioned employers sometimes expose the workforce to hazardous conditions through ignorance of the risks. Factory safety laws were enacted in Europe in past centuries to deal with employer abuses. The basic concepts were adopted in the United States when some of the states took action in the form of factory laws and regulations, both for worker safety and for compensation in case of injury. Over the years these laws were strengthened and broadened until 1969 and 1970, when the U.S. Congress created two federal laws regulating safety and health for those employers engaged in interstate commerce in coal mining and all other private industry. In 1977, federal legislation dealing with other forms of mining was added. However, public institutions such as state and local governments and universities were not covered by the federal laws in the United States and still are not covered by federal safety legislation.

The Occupational Safety and Health Act of 1970 (OSHAct) remains the primary federal vehicle for ensuring workplace safety and health in the United States. This law requires that employers provide a place of employment free from recognized hazards to employee safety or health. The critical word is ‘‘recognized’’ because today’s workplaces have many new materials and processes for which hazard knowledge is absent. This places a large responsibility on the employer to keep abreast of new knowledge and information about workplace hazards for their operations. The OSHAct established three agencies to deal with workplace safety and health. These were the Occupational Safety and Health Administration (OSHA), the National Institute for Occupational Safety and Health (NIOSH), and the Occupational Safety and Health Review Commission.

5.1. The Occupational Safety and Health Administration

OSHA, located within the U.S. Department of Labor, has the responsibility for establishing federal workplace safety and health standards and enforcing them. Over the last three decades, OSHA has developed hundreds of standards that have been published in the code of federal regulations (CFR) Section 29 CFR, subsections 1900–1928, which cover General Industry (1910), Longshoring (1918), Construction (1926), and Agriculture (1928) (http://www.osha-slc.gov/OshStd_toc/OSHA_Std_toc.html). This code is revised periodically and new standards are added continually. Current and proposed rules and standards include Process Safety Management of Highly Hazardous Chemicals (1910.119), Personal Protective Equipment (1910.132 to 1910.139), the Proposed Safety and Health Program Rule, and the Final Ergonomic Program Standard. It is important for employers to keep up with these new and revised standards. One way is to keep in frequent contact with your area office of OSHA and request that they send you updates. Another way is to subscribe to one or more of the many newsletters for occupational safety and health that provide current information and updates. A third way is to access the OSHA web page (http://www.osha.gov), which provides information about current activities, regulations, and updates.

Workplaces are required by law to be in compliance with the federal safety and health standards. Failure to do so can be the basis for federal fines in the first instance, meaning that when a violation is found, the inspector will propose a fine and an abatement period. Under some of the former state safety programs, an identified hazard did not bring a fine if it was controlled within the abatement period. Under the federal process, fines will be assessed immediately after a hazard is identified. These fines can be substantial, and the threat of such fines is felt to be an incentive for compliance with the federal standards. Many employers feel that this first-instance fine is punitive and takes resources away from abatement efforts. However, the logic of first-instance fines is to motivate employers to be proactive in looking for and correcting workplace hazards. If an employer is found to be in violation of a standard and disagrees with the inspector about this, the employer has the right to a review at higher OSHA levels. If no recourse is obtained from OSHA, then the employer has the right for review by the Occupational Safety and Health Review Commission. The employer also has the right to subsequent legal recourse through the federal courts.

Current estimates of the numbers of OSHA inspectors and workplaces indicate that the average workplace can expect an OSHA inspection about every 9 to 10 years. To apply its inspection resources in the most effective manner, OSHA has adopted a strategy to concentrate on the most hazardous industries. Thus, these industries can expect an inspection much more often. OSHA concentrates on select high-risk industries, such as construction, and select injuries, such as cumulative trauma, which is widespread in meat processing and assembly jobs. OSHA has contracted with several state workers' compensation agencies to obtain data on the employers in these states with the greatest injury frequency and severity and uses these data to select employers for inspection. Inspections can also be triggered by a complaint from an employee or employee representative. When the inspector arrives at your workplace, you will be shown an official credential indicating that the inspector is from OSHA. It is a good policy to cooperate with the OSHA inspector. However, at the initial contact by the inspector you have the right, by ruling of the U.S. Supreme Court, to require a search warrant for entry into your workplace. Most employers provide immediate access to their facilities because requiring a search warrant may create an antagonistic situation once the warrant is obtained by the inspector.

When an inspection is triggered by a complaint, the inspector will typically only examine the hazard(s) and work areas defined in the complaint. If it is a general inspection, usually the entire workplace is examined. The first step in the inspection is a review of your injury and illness log (OSHA 300). The inspector will be looking for jobs or areas of the plant that have many injuries and illnesses. After this, the inspector will conduct a walk-through of the facility. You have the right to accompany the inspector on the inspection and point out trade secrets that you want to be kept confidential. OSHA will meet your request unless there is good reason to believe that your trade secret is not really a secret. A representative of your employees is also allowed to accompany the inspector. During the inspection the inspector will talk to employees about their working conditions and any safety and health problems they are encountering. The inspector will take notes on what is observed and heard. Some inspections are complex and require special measurements of the equipment, air contamination, noise level, or other aspects of hazard exposure. Depending on the size of the plant or store location and the number of hazards, an inspection can take from one day up to several months. At the end of the inspection, the inspector will have a closing conference with management and the employee representative to discuss what was observed and possible violations of the standards. If standards were violated, some days after the inspection you will receive a formal citation indicating the standards violated and proposed penalties via registered mail. At that time you can contest any of the proposed violations and/or penalties. These will be reviewed with you by the OSHA area director and some agreement may be reached at that time. If not, you can contest to higher levels in OSHA, and then to the Occupational Safety and Health Review Commission. The final recourse is the federal courts.

OSHA also provides professional training and short courses to safety and health professionals, employee representatives, and employees at the OSHA Training Institute and through contracts with universities, colleges, technical schools, and unions. See the OSHA website for information on training (http://www.osha.gov).

5.2. The National Institute for Occupational Safety and Health

NIOSH is a subunit of the Centers for Disease Control and Prevention (see website http://www.cdc.gov), which conducts research into the causes and cures of occupationally caused injuries and illnesses. It also has a training function that promotes the development of occupational safety and health professionals. NIOSH research covers a broad range of topics, from the toxicology of chemicals, metals, and biological agents to the causes of accidents to the psychological stress aspects of the workplace, to name a few. The results of this research are published in the scientific literature and public reports disseminated by NIOSH and the U.S. Government Printing Office. A listing of these reports is available on the NIOSH website (http://www.cdc.gov/niosh). Oftentimes enough research knowledge is accumulated to develop a recommendation for a federal safety and health standard. At that time a criteria or technical document is developed that defines the scope of the hazard(s), the evidence for adverse safety or health effects, and recommendations for exposure limits and control mechanisms. These criteria documents and technical reports are sent to OSHA for consideration as the basis for a safety and health federal standard. They also often appear in court litigation as the state-of-the-art knowledge about the hazards of products in product liability suits, even though they may not have been adopted as standards. Information on these criteria documents and NIOSH research reports is available from the U.S. Superintendent of Documents in Washington, DC, and from the NIOSH website.

NIOSH also conducts health hazard investigations in situations where employees become ill from unknown causes or where they have been exposed to agents for which only scarce knowledge is available and special expertise is needed. These investigations are triggered by a request for technical assistance from OSHA, a state health agency, a state safety agency, or a health hazard evaluation request from a company, union, or employee. NIOSH has a right of entry into workplaces for these investigations but can be required to obtain a search warrant by the company. These investigations are often much more complex than OSHA investigations and can entail extensive environmental measurement, employee physical examinations and interviews, examinations of company records, and discussions with management. Reports are developed based on the evaluation, which include a determination of hazards and proposed methods of control. These reports may be sent to OSHA for compliance action (inspection and/or citation).

5.3. State Agencies

The state(s) in which your operation(s) are located may also have jurisdiction for enforcing occupational safety and health standards, and conducting investigations based on an agreement with OSHA. In many states the state health agency and/or labor department has agreements with OSHA to provide consultative services to employers. You can find out by contacting these agencies directly. In several states the safety and health agencies enforce safety and health rules and standards for local and state government employees, or other employees not covered by OSHA. See website http://www.cdc.gov/niosh/statosh.html.

5.4. Centers for Disease Control and Prevention

The purpose of the Centers for Disease Control and Prevention (CDC) is to promote health and quality of life by preventing and controlling disease, injury, and disability. The CDC provides limited information on occupational safety and health. For example, their web page has information about accident causes and prevention, back belts, cancer—occupational exposure, effects of workplace hazards on male reproductive health, latex allergies, needle stick, occupational injuries, teen workers, and violence in the workplace (see website http://www.cdc.gov). The Center for Health Statistics is located within CDC and provides basic health statistics on the U.S. population. This information is used to identify potential occupational health risks by occupational health researchers (see website http://www.cdc.gov/nchs).

5.5. The Bureau of Labor Statistics

The Bureau of Labor Statistics (BLS) is the principal fact-finding agency for the federal government in the broad field of labor economics and statistics (see website http://stats.bls.gov). It collects, processes, analyzes, and disseminates essential statistical data. Among the data are occupational safety and health data, including annual reports, by industry, of rates of injuries, illnesses, and fatalities (http://stats.bls.gov/oshhome.htm).

5.6. The Environmental Protection Agency

The Environmental Protection Agency (EPA) was established as an independent agency in 1970 with the purpose of protecting the air, water, and land (see website http://www.epa.gov). To this end, the EPA engages in a variety of research, monitoring, standard setting, and enforcement activities. The Agency administers 10 comprehensive environmental protection laws: the Clean Air Act (CAA); the Clean Water Act (CWA); the Safe Drinking Water Act (SDWA); the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA, or Superfund); the Resource Conservation and Recovery Act (RCRA); the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA); the Toxic Substances Control Act (TSCA); the Marine Protection, Research, and Sanctuaries Act (MPRSA); Uranium Mill Tailings Radiation Control Act (UMTRCA); and the Pollution Prevention Act (PPA). The EPA's Strategic Plan for the 21st century includes 10 goals, among which is one dealing with preventing pollution and reducing risk in communities, homes, workplaces, and ecosystems. For the purposes of this chapter, we will focus on the issues related to workplaces. According to the EPA's plan (EPA 1999), over 75,000 chemicals are in commerce today, with an estimated 2,000 new chemicals and 40 genetically engineered microorganisms introduced each year. Among those are potentially toxic chemicals that may present risks to workers, such as persistent, bioaccumulative and toxic chemicals (PBTs). Reducing PBTs should lead to safer manufacturing processes and eliminate some occupational exposures. Strategies to deal with such chemicals include better management of, training about, and reduced use of pesticides; better programs to deal with the chemical industry; industrial pollution prevention; better building construction to promote quality indoor air; and industrial waste minimization.

5.7. Other Agencies and Groups

The American National Standards Institute (ANSI) was founded in 1918 and has been the administrator and coordinator of the United States private sector voluntary standardization system (see website http://www.ansi.org). ANSI does not itself develop standards but rather facilitates development by establishing consensus among qualified groups. ANSI has developed a number of occupational health- and safety-related standards, including standards related to information management for occupational safety and health (ANSI Z16.2-1995), preparation of hazardous industrial chemical material safety data sheets (ANSI Z400.1-1998), construction safety and health audit programs (ANSI A10.39-1996), and human factors engineering of visual display terminal workstations (ANSI/HFS 100-1988).

The mission of the American Society for Testing and Materials (ASTM) is to be the foremost developer and provider of voluntary consensus standards, related technical information, and services having internationally recognized quality and applicability that (1) promote public health and safety, (2) contribute to the reliability of materials, products, systems and services, and (3) facilitate national, regional, and international commerce. See website http://www.astm.org.

The International Labour Organization (ILO) was created in 1919 and is a United Nations (UN) specialized agency that seeks the promotion of social justice and internationally recognized human and labor rights (see webpage http://www.ilo.org). The ILO formulates minimum standards of basic labor rights, such as freedom of association, the right to organize, and collective bargaining. It also provides technical assistance in areas such as vocational training and vocational rehabilitation, employment policy, working conditions, and occupational safety and health. There is a branch concerned with occupational safety and health (see website http://www.ilo.org/public/english/protection/safework/intro.htm) that focuses on reducing the number and seriousness of occupational accidents and diseases, adapting the working environment, equipment, and work processes to the physical and mental capacity of the worker, enhancing the physical, mental, and social well being of workers in all occupations, encouraging national policies and programs of member states, and providing appropriate assistance. To achieve those aims, the ILO works with government and nongovernment agencies to design and implement policies and programs to improve working conditions. The ILO is currently working on a global program for occupational safety and health.

The constitution of the World Health Organization (WHO) was approved in 1946 (see website http://www.who.org). Its goal is good health for all people. To this end, the WHO directs international health activity, promotes technical cooperation, helps governments strengthen their own health services, provides technical assistance, conducts research, and establishes international standards for biological and pharmaceutical products. They also provide information on global occupational safety and health issues (see website http://www.who.org/peh/Occupational_health/occindex.html), such as biological agents, noise, radiation, chemicals, occupational carcinogens, and allergenic agents. WHO established the international statistical classification of diseases and related health problems in occupational health. The organization has a global strategy on occupational safety and health that includes 10 priority areas:

1. Strengthening international and national policies for health at work and developing necessary policy tools
2. Developing healthy work environments
3. Developing healthy work practices and promoting health at work
4. Strengthening occupational health services
5. Establishing support services for occupational health
6. Developing occupational health standards
7. Developing human resources for occupational health
8. Establishing registration and data systems, developing information services, and raising public awareness
9. Strengthening research
10. Developing collaboration in occupational health with other services

The European Agency for Safety and Health at Work (see website http://europe.osha.eu.int/) was established in 1996 and is based in Bilbao, Spain. The Agency has put forth numerous directives for employers in its member states to follow. These include general safety requirement directives, directives regarding temporary workers, pregnant workers, and young people and directives on manual handling, work equipment, and safety signs. The agency also conducts information campaigns, such as the recently launched campaign aimed at reducing the number of work-related back injuries and other musculoskeletal disorders. Other information about occupational health and safety in the European Union can be found using HASTE (see website http://www.occuphealth.fi/e/eu/haste/), the European Union health and safety database, which lists databases from member states.

5.8. Safety and Ergonomics Program Standards

The purpose of the OSHA Proposed Safety and Health Program Rule (see website http://www.osha-slc.gov/SLTC/safetyhealth/nshp.html) is to reduce the number of job-related fatalities, illnesses, and injuries by requiring employers to establish a workplace safety and health program to ensure compliance with OSHA standards and the General Duty Clause of the OSHAct. All employers covered under the OSHAct are covered by this rule (except construction and agriculture), and the rule applies to all hazards covered by the General Duty Clause and OSHA standards. Five elements make up the program:

1. Management leadership and employee participation
2. Hazard identification and assessment
3. Hazard prevention and control
4. Information and training
5. Evaluation of program effectiveness

Employers that already have safety and health programs with these five elements can continue using their existing programs if they are effective. Late in 2000, OSHA announced its Final Ergonomic Program Standard (see website http://www.osha-slc.gov/ergonomics-standard/index.html). The proposed Standard specifies employer's obligations to control musculoskeletal disorder (MSD) hazards and provide MSD medical management for injured employees. The Proposed Ergonomic Standard uses a program approach—that is, the proposal specifies the type of a program to set up to combat MSD, as opposed to specifying the minimum or maximum hazard levels. According to the proposed ergonomic standard, an ergonomic program consists of the following six program elements:

1. Management leadership and employee participation
2. Hazard information and reporting
3. Job hazard analysis and control
4. Training
5. MSD management
6. Program evaluation

These are similar to the elements in the proposed safety and health program rule. The proposed ergonomics standard covers workers in general industry, though construction, maritime, and agriculture operations may be covered in future rulemaking. The proposal specifically covers manufacturing jobs, manual material-handling jobs, and other jobs in general industry where MSDs occur. If an employer has an OSHA recordable MSD, the employer is required to analyze the job and control any identified MSD hazards. Public hearings were ongoing regarding the proposed ergonomic standard as of March–April of 2000 and written comments were being taken. Some states have proposed ergonomic standards to control MSDs.

The California Ergonomic Standard (see website http://www.dir.ca.gov/Title8/5110.html) went into effect on July 3, 1997. The standard targets jobs where a repetitive motion injury (RMI) has occurred and the injury can be determined to be work related and at a repetitive job. The injury must have been diagnosed by a physician. The three main elements of the California standard are work site evaluation, control of exposures that have caused RMIs, and employee training. The exact language of the standard has been undergoing review in the California judicial system.

The purpose of the Washington State Proposed Ergonomics Program Rule (see website http://www.lni.wa.gov/wisha) is to reduce employee exposure to workplace hazards that can cause or aggravate work-related musculoskeletal disorders (WMSDs). There are no requirements for medical management in the proposed rule. The proposal covers employers with caution zone jobs, which the proposed rule defines based on the presence of any one or more of a number physical job factors. For example, a caution zone job exists if the job requires "working with the neck, back or wrist(s) bent more than 30 degrees for more than 2 hours total per workday" (WAC 296-62-05105). The proposed standard specifies the type of ergonomic awareness education that must be provided to employees who work in caution zone jobs. The standard also states that if caution zone jobs have WMSD hazards, employers must reduce the WMSD hazards identified. Several tools are suggested that can be used to analyze the caution zone jobs for WMSD hazards, and thresholds are provided that indicate sufficient hazard reduction. The rule also states that employees should be involved in analyzing caution zone jobs and in controlling the hazards identified. This proposed rule was under review in 2000.

The proposed North Carolina Ergonomic Standard (see website http://www.dol.state.nc.us/news/ergostd.htm) was first announced in November 1998. Like the proposed national standard and California's standard, North Carolina's is a program standard without physical hazard thresholds. The proposal states employers shall provide ergonomic training within 90 days of employment and no less than every three years thereafter. It also specifies the nature of that training, which includes information on ergonomic physical hazards, MSDs that can arise from ergonomic hazards, workplace ways to control ergonomic hazards, and the importance of reporting symptoms. Under the proposed standard, if an MSD is thought to be causally related to work, the employer has to implement

TABLE 1 Descriptions of NIOSH Top Eight Occupational Disease and Injury Categories

Disease or Injury	Description
Allergic and irritant dermatitis	Allergic and irritant dermatitis (contact dermatitis) is overwhelmingly the most important cause of occupational skin diseases, which account for 15% to 20% of all reported occupational diseases. There is virtually no occupation or industry without potential exposure to the many diverse agents that cause allergic and irritant dermatitis.
Asthma and chronic obstructive pulmonary disease	Occupationally related airway diseases, including asthma and chronic obstructive pulmonary disease (COPD), have emerged as having substantial public health importance. Nearly 30% of COPD and adult asthma may be attributable to occupational exposure. Occupational asthma is now the most frequent occupational respiratory disease diagnosis. More than 20 million U.S. workers are exposed to substances that can cause airway diseases.
Fertility and pregnancy abnormalities	While more than 1000 workplace chemicals have shown reproductive effects in animals, most have not been studied in humans. In addition, most of the 4 million other chemical mixtures in commercial use remain untested. Physical and biological agents in the workplace that may affect fertility and pregnancy outcomes are practically unstudied.
Hearing loss	Occupational hearing loss may result from an acute traumatic injury, but it is far more likely to develop gradually as a result of chronic exposure to ototraumatic (damaging to the ear or hearing process) agents. Noise is the most important occupational cause of hearing loss, but solvents, metals, asphyxiants, and heat may also play a role. Exposure to noise combined with other agents can result in hearing losses greater than those resulting from exposure to noise or other agents alone.
Infectious disease	Health care workers are at risk of tuberculosis (TB), hepatitis B and C viruses, and the human immunodeficiency virus (HIV). Social service workers, corrections personnel, and other occupational groups who work regularly with populations having increased rates of TB may also face increased risk. Laboratory workers are at risk of exposure to infectious diseases when working with infective material.
Low-back disorders	Back pain is one of the most common and significant musculoskeletal problems in the world. In 1993, back disorders accounted for 27% of all nonfatal occupational injuries and illnesses involving days away from work in the United States. The economic costs of low back disorders are staggering. According to NIOSH (1996b), a recent study showed the average cost of a workers' compensation claim for a low-back disorder was $8,300, which was more than twice the average cost of $4075 for all compensable claims combined. Estimates of the total cost of low-back pain to society in 1990 were between $50 billion and $100 billion per year, with a significant share (about $11 billion) borne by the workers' compensation system. Moreover, as many as 30% of American workers are employed in jobs that routinely require them to perform activities that may increase risk of developing low-back disorders.
Musculoskeletal disorders of the upper extremities	Musculoskeletal disorders of the upper extremities (such as carpal tunnel syndrome and rotator cuff tendinitis) due to work factors are common and occur in nearly all sectors of our economy. More than $2 billion in workers' compensation costs are spent annually on these work-related problems. Musculoskeletal disorders of the neck and upper extremities due to work factors affect employees in every type of workplace and include such diverse workers as food processors, automobile and electronics assemblers, carpenters, office data-entry workers, grocery store cashiers, and garment workers. The highest rates of these disorders occur in the industries with a substantial amount of repetitive, forceful work. Musculoskeletal disorders affect the soft tissues of the neck, shoulder, elbow, hand, wrist, and fingers.

TABLE 1 **(*Continued*)**

Disease or Injury	Description
Traumatic injuries	During the period 1980 through 1992, more than 77,000 workers died as a result of work-related injuries. This means that an average of 16 workers die every day from injuries suffered at work. The leading causes of occupational injury fatalities over this 13-year period were motor vehicles, machines, homicides, falls, electrocutions, and falling objects. In four industries—mining, construction, transportation, and agriculture—occupational injury fatality rates were notably and consistently higher than all other industries. In 1994, 6.3 million workers suffered job-related injuries that resulted in lost work time, medical treatment other than first aid, loss of consciousness, restriction of work or motion, or transfer to another job. The leading causes of nonfatal occupational injuries involving time away from work in 1993 were overexertion, contact with objects or equipment, and falls to the same level.

engineering controls, work practice controls, and/or administrative controls to reduce the impact of the ergonomic hazards. Employee participation in the ergonomic program is encouraged.

6. DEFINING OCCUPATIONAL INJURIES AND DISEASES

Early in the 1980s, NIOSH defined the 10 most serious occupational disease and injury areas (CDC 1983). These were occupational lung diseases, musculoskeletal injuries, occupational cancers, acute trauma, cardiovascular diseases, disorders of reproduction, neurotoxic disorders, noise-induced hearing loss, dermatologic conditions, and psychological disorders. In April of 1996, NIOSH developed the National Occupational Research Agenda (NORA) (NIOSH 1996a), which identified 21 priority research areas to target and coordinate occupational safety and health research. Eight of the 21 target areas focus on occupational diseases and injuries. This was an update of the list from the early 1980s defined above. The new list identifies allergic and irritant dermatitis, asthma and chronic obstructive pulmonary disease, fertility and pregnancy abnormalities, hearing loss, infectious disease, low-back disorders, musculoskeletal disorders of the upper extremities, and traumatic injuries (NIOSH 1996b) as serious problems. More detail on each disease or condition is provided in Table 1. Table 2 provides brief descriptions of various types of occupational diseases and injuries not included in Table 1 that were highlighted previously by CDC and NIOSH.

7. WORKPLACE HAZARDS

A list of all currently recognized and potential workplace hazards would be larger than this entire Handbook. The best places to start accumulating hazard information pertinent to your operations are the OSHA standards, NIOSH criteria documents, and government reports and publications. These are available on the websites listed in this chapter and from the U.S. Superintendent of Documents in Washington, DC. The websites have a great deal of useful information. Other excellent sources of information include the National Safety Council Safety manuals, NIOSH (1984), and Best's Loss Control Guide. Other federal, state, and local agencies can also provide some aspects of occupational health and safety hazard information. At the federal level these include the Environmental Protection Agency (EPA), the National Institute for Environmental Health Sciences (NIEHS), and the Centers for Disease Control and Prevention (CDC).

It is important to comprehend the breadth and nature of occupational hazard exposures. To do this we can classify workplace hazard sources into broad categories that help us to understand their nature and potential controls. These are:

1. Physical agents such as noise and heat
2. Powered mechanical agents such as machinery and tools
3. Nonpowered mechanical agents such as hammers, axes, and knives
4. Liquid chemical agents such as benzene and toluene
5. Powdered materials such as pesticides, asbestos, sand, and coal dust
6. Gaseous or vaporous chemical agents such as nitrous oxide, carbon monoxide, and anhydrous ammonia

TABLE 2 Descriptions of Various Occupational Disorders and Diseases Not Included in Table 1

Disease or Injury	Description
Occupational lung disease	The latent period for many lung diseases can be several years. For instance, for silicosis it may be as long as 15 years and for asbestos-related diseases as long as 30 years. The lung is a primary target for disease related to toxic exposures because it is often the first point of exposure through breathing. Many chemicals and dusts are ingested through breathing. The six most severe occupational lung disorders are asbestosis, byssinosis, silicosis, coal workers' pneumoconiosis, lung cancer, and occupational asthma.
Asbestosis	This disease produces scarring of the lung tissue, which causes progressive shortness of breath. The disease continues to progress even after exposures end, and there is no specific treatment. The latent period is 10–20 years. The agent of exposure is asbestos, and insulation and shipyard workers are those most affected.
Byssinosis	This disease produces chest tightness, cough, and airway obstruction. Symptoms can be acute (reversible) or chronic. The agents of exposure are dusts of cotton, flax, and hemp, and textile workers are those most affected.
Silicosis	This is a progressive disease that produces nodular fibrosis, which inhibits breathing. The agent of exposure is free crystalline silica, and miners, foundry workers, abrasive blasting workers, and workers in stone, clay, and glass manufacture are most affected.
Coal Miners' pneumoconiosis	This disease produces fibrosis and emphysema. The agent of exposure is coal dust. The prevalence of this disorder among currently employed coal miners has been estimated at almost 5%.
Lung cancer	This disease has many symptoms and multiple pathology. There are several agents of exposure, including chromates, arsenic, asbestos, chloroethers, ionizing radiation, nickel, and polynuclear aromatic hydrocarbon compounds.
Occupational cancers	There is some debate on the significance of occupational exposures in the overall rate of cancer ranging from 4 to 20% due to occupation, yet there is good agreement that such occupational exposures can produce cancer. There are many types of cancer that are related to workplace exposures, including hemangiosarcoma of the liver; mesothelioma; malignant neoplasm of the nasal cavities, bone, larynx, scrotum, bladder, kidney, and other urinary organs; and lymphoid leukemia and erythroleukemia.
Traumatic injuries Amputations	The vast majority of amputations occur to the fingers. The agents of exposure include powered hand tools and powered machines. Many industries have this type of injury, as do many occupations. Machine operators are the single most injured occupation for amputations.
Fractures	Falls and blows from objects are the primary agents that cause fractures. The major sources of these injuries include floors, the ground, and metal items. This suggests falling down or being struck by an object as the primary reasons for fractures. Truck drivers, general laborers, and construction laborers were the occupations having the most fractures.
Eye loss	It was estimated that in 1982 there were over 900,000 occupational eye injuries. Most were due to particles in the eye such as pieces of metal, wood, or glass, but a smaller number were caused by chemicals in the eye. A number of occupations are affected, including those in woodworking, metalwork, construction, and agriculture.

TABLE 2 **(*Continued*)**

Disease or Injury	Description
Traumatic injuries	
Lacerations	Over 2,000,000 lacerations occur each year with the majority to the fingers, followed by the arms, legs, and head/neck. Lacerations occur primarily from being struck or stuck by an object or from striking against an object. The major agents of exposure include knives, sharp metal objects, saws, glass items, nails, and machines.
Cardiovascular disease	These disorders include hypertensive disease, ischemic heart disease, other forms of heart disease, and cerebrovascular disease. As with cancer, the specific contribution of occupational factors to the causation of CVD has been debated, but there is agreement that some workplace factors contribute to or cause CVD (see Smith and Sainfort 1990 for a detailed discussion of psychosocial factors and their contribution). Four main occupational sources of CVD causation are agents that affect cardiopulmonary capacity, chemicals, noise, and psychosocial stress.
Cardiopulmonary capacity reducers	Agents such as dusts, mists, heavy metals, silica, and other trace elements make the lungs work much harder than normal and can induce congestive heart failure. Metal such as beryllium, antimony, lead, and cobalt as well as silica and asbestos can produce heart disorders.
Chemicals	Some chemicals act to sensitize the heart muscle and the smooth muscle of the blood vessels, while others reduce the oxygen-carrying capacity of the blood. These include nitroglycerin, carbon monoxide, carbon disulfide, and halogenated hydrocarbons.
Noise	Studies have shown that noise can produce transient increases in blood pressure that may lead to CVD. This may be due to psychological factors related to stress reactions.
Psychosocial stress	Research from longitudinal studies of cardiovascular fitness has demonstrated a relationship between perceived job satisfaction and stress and cardiovascular illness. Epidemiological studies have shown that particular jobs with specific characteristics such as high demands and low control have a higher incidence of coronary heart disease. Organizational demands and relations, job task demands, social relationships at work, work schedule, work content features, discretionary control and participation, and physical working conditions have all been shown to influence the level of job stress (Smith 1986; Kalimo et al. 1997).
Neurotoxic disorders	Neurotoxic disorders are produced by damage to the central nervous system, damage to the peripheral nervous system, and intoxication. These cause deficits in attention, reasoning, thinking, remembering, and making judgments. They may also cause peripheral neuropathy, neuroses and psychoses, personality changes, aberrant behavior, or reduced reaction time and motor skill. One of the first workplace-related neurological disorders identified was lead poisoning, which produced palsy in workers exposed to lead dust. NIOSH publishes a list of the materials known to have neurotoxic effects.
Psychological disorders	Psychological disorders related to working conditions include sleep disturbances, mood disturbances, reduced motivation to work or recreate, somatic and psychosomatic complaints, neuroses, psychoses, and dysfunctional coping behavior. The effects of stress on an individual are influenced by the nature of the exposures and the individual's physical and psychological characteristics and coping behaviors that may accentuate or mitigate the exposure.

7. Heavy metals such as lead and mercury
8. Biological agents such as bacteria and viruses
9. Genetically engineered agents
10. Other hazards, such as wet working surfaces, unguarded floor openings, job stress, and the unsafe behavior of others

These hazards enter the body through various avenues, including inhalation into the lungs and nose, absorption through the skin and other membranes, ingestion into the throat and stomach, traumatic contact with various body surfaces and organs, and, in the case of job stress, through the cognitive mental processes. Descriptions of many of these hazards and definitions of adverse exposure levels are contained in NIOSH (1977).

Traditional hazards such as unexpected energy release and chemicals are still major concerns in the workplace. The use of lasers, robots, microwaves, x-rays, and imaging devices will become more common and will make many of the traditional problems of controlling energy release and limiting worker access to hazardous machine components even more challenging. These technologies will be even more problematic because of the complex nature of the mechanisms of energy release and the increased power of the forces involved. For instance, using x-rays for lithographic etching of computer chips could produce exposures that are substantially higher than with conventional diagnostic x-rays. The safety precautions for this type of instrument have to be much better than current standards for diagnostic equipment.

In addition to these emerging hazards, other new hazards will appear. Some will be the exotic products of genetic engineering and biotechnology, while others will be the products of our ability to harness the laws of physics and chemistry with advanced engineering designs. The future will see commercial uses of plasma gas generators for tool hardening, electron accelerators for generating tremendous power for x-ray lithography in microchip production and fusion power generation. These will become everyday tools used by thousands of workers, many of whom will not be well educated or knowledgeable about the tremendous power of the technology they will be working with.

While these physical and biological hazards will become more prevalent and dangerous than they are today, there will also be more physical and psychological work demands that can lead to psychological stress problems. Currently, the two fastest-rising workers' compensation claim areas in the United States are cumulative musculoskeletal trauma and psychological distress. The rise in these problems can generally be related to two factors: first, greater media, worker, and employer awareness and knowledge about how the workplace can contribute to such problems; and second, huge increases in workplace automation that create conditions that produce these disorders. Dealing with these stress-induced problems may be even more difficult than dealing with the biological, chemical, and physical hazards.

8. MEASURING HAZARD POTENTIAL AND SAFETY PERFORMANCE

To control occupational hazards and related illness and injuries, it is necessary to define their nature and predict when and where they will occur. This requires that some system of hazard detection be developed that can define the frequency of the hazard, its seriousness, and its amenability to control. Traditionally, two parallel systems of information have been used to attain this purpose. One is hazard identification, such as plant inspections, fault-free analysis, and employee hazard-reporting programs, which have been used to define the nature and frequency of company hazards. In this case, action is taken before an injury or illness occurs. The second system is after the fact in that it uses employee injury and company loss-control information to define problem spots based on the extent of injuries and costs to the organization. When pre- and postinjury systems have been integrated, they have been successful in predicting high-risk plant areas or working conditions where remedial programs can be established for hazard control.

8.1. Inspection Programs

Hazard identification prior to the occurrence of an occupational injury is a major goal of a hazard inspection program. In the United States, such programs have been formalized in terms of federal and state regulations that require employers to monitor and abate recognized occupational health and safety hazards. These recognized hazards are defined in the federal and state regulations that provide explicit standards of unsafe exposures. The standards can be the basis for establishing an in-plant inspection program because they specify the explicit subject matter to be investigated and corrected.

Research has shown that inspections are most effective in identifying permanent fixed physical and environmental hazards that do not vary over time. Inspections are not very effective in identifying transient physical and environmental hazards or improper workplace behaviors because these hazards may not be present when the inspection is taking place (Smith et al. 1971). A major benefit from inspections, beyond hazard recognition, is the positive motivational influence on employees. Inspections demonstrate management interest in the health and safety of employees and a commitment to a safe working environment. To capitalize on this positive motivational influence, an inspection should not be a punitive or confrontational process of placing blame. Indicating the good aspects of a work area and not just the hazards is important in this respect. It is also important to have employees participate in hazard inspections because this increases hazard-recognition skills and increases motivation for safe behavior.

The first step in an inspection program is to develop a checklist that identifies all potential hazards. A good starting point is the state and federal standards. Many insurance companies have developed general checklists of OSHA standards that can be tailored to a particular plant. These are a good source when drawing up the checklist. A systematic inspection procedure is preferred. This requires that the inspectors know what to look for and where to look for it and have the proper tools to conduct an effective assessment. It is important that the checklist be tailored to each work area after an analysis of that work area's needs has been undertaken. This analysis should determine the factors to be inspected: (1) the machinery, tools, and materials, (2) chemicals, gases, vapors, and biological agents, and (3) environmental conditions. The analysis should also determine (1) the frequency of inspections necessary to detect and control hazards, (2) the individuals who should conduct and/or participate in the inspections, and (3) the instrumentation needed to make measurements of the hazard(s).

The hazards that require inspection can be determined by (1) their potential to cause an injury or illness, (2) the potential seriousness of the injuries or illnesses, (3) the number of people exposed to the hazard, (4) the number of injuries and illnesses at a workplace related to a specific hazard, and (5) hazardous conditions defined by federal, state, and local regulations. The frequency of inspections should be based on the nature of the hazards being evaluated. For instance, once a serious fixed physical hazard has been identified and controlled, it is no longer a hazard. It will only have to be reinspected periodically to be sure the situation is still no longer hazardous. Random spot checking is another method that can indicate whether the hazard control remains effective. Other types of hazards that are intermittent will require more frequent inspection to assure proper hazard abatement. In most cases, monthly inspections are warranted, and in some cases daily inspections are reasonable.

Inspections should take place when and where the highest probability of a hazard exists, while reinspection can occur on an incidental basis to ensure that hazard control is effectively maintained. Inspections should be conducted when work processes are operating, and on a recurring basis at regular intervals. According to the National Safety Council (1974), a general inspection of an entire premises should be conducted at least once a year, except for those work areas scheduled for more frequent inspections because of their high hazard level. Because housekeeping is an important aspect of hazard control, inspection of all work areas should be conducted at least weekly for cleanliness, clutter, and traffic flow. The National Safety Council (1974) indicated that a general inspection should cover the following:

1. Plant grounds
2. Building and related structures
3. Towers, platforms, or other additions
4. Transportation access equipment and routes
5. Work areas
6. Machinery
7. Tools
8. Materials handling
9. Housekeeping
10. Electrical installations and wiring
11. Floor loading
12. Stairs and stairways
13. Elevators
14. Roofs and chimneys

We would add to this:

15. Chemicals, biological agents, radiation, etc.
16. Ergonomic stressors
17. Psychosocial stressors

Intermittent inspections are the most common type and are made at irregular intervals, usually on an ad hoc basis. Such inspections are unannounced and are often limited to a specific work area or process. Their purpose is to keep first-line supervisors and workers alert to safety considerations and hazard detection. Such inspections do not always require a checklist. Systematic informal inspections made by first-line supervisors on a daily basis in their work area can be effective in identifying intermittent hazards and also keeping employees aware of good safety practices. Continuous inspections occur when employees are aware of safety considerations and detect and report hazards as they occur. Maintenance staff can also play a role in defining hazardous conditions while carrying out their duties of machinery repair.

As indicated above, all employees in an organization can become involved in inspecting for hazards, some formally, some informally. Technical, periodic health and safety inspections should be conducted by the plant medical, industrial hygiene, and safety staff. These persons have special expertise to define and evaluate hazards. That expertise can be supplemented by outside experts from insurance companies, government safety and health agencies, and private consultants. Conducting a formal inspection requires some planning and structure. First it must be determined what, where, and when to inspect. That decision will be based on hazard and illness/injury potential. A determination must be made whether to give prior warning to the employees in the area to be inspected. If hazards are primarily behavioral in nature, the prior warning may reduce the effectiveness of the inspection.

When conducting the inspection, the department supervisor should be asked to identify hot spots or special problem areas. A checklist can be used to identify each hazard and its nature, exact location, potential to cause serious damage, and possible control measures. During the walk-through, employee input should be solicited. Photographs and videotapes of hazards are effective in documenting the nature and potential seriousness of hazards. Once the inspection is completed, a report should be prepared that specifies pertinent information about the nature of the hazards, illness and injury potential, and abatement recommendations. This report needs to be detailed and provide step-by-step instructions for instituting hazard-control procedures in a timely manner. First, all potential hazards and their contributors should be listed. Second, the hazard identification analysis should provide solutions to deal with the hazards. The methods used to conduct the hazard identification should produce accurate estimates for the risks of harm to employees. Finally, resources should be allocated for abating the hazards and should be prioritized by those safety improvements that should yield the best results, that is, the best safety performance.

It is not sufficient simply to write up the results; they should be shared with all parties concerned in a face-to-face meeting. This meeting will give the results greater significance and serve as the basis for further interaction and possible modification of recommendations. Such meetings will enhance employee understanding and allow for in-depth discussions of the findings and recommendations. This makes the entire inspection process more relevant to supervisors and employees and facilitates the favorable acceptance of the results and any subsequent recommendations.

The quality of a hazard-identification system can be evaluated by answering the following four questions (Suokas 1993):

1. How well has the analysis identified hazards and their contributors?
2. How effectively has the analysis produced potential solutions needed in the system?
3. How accurately has the analysis estimated the risks of the system?
4. What is the cost-effectiveness of the hazard identification analysis?

8.2. Illness and Injury Statistics

There are four main uses of injury statistics: (1) to identify high-risk jobs or work areas, (2) to evaluate company health and safety performance, (3) to evaluate the effectiveness of hazard-abatement approaches, and (4) to identify factors related to illness and injury causation. An illness and injury-reporting and analysis system requires that detailed information must be collected about the characteristics of illness and injuries and their frequency and severity. The Occupational Safety and Health Act (1970) established illness and injury reporting and recording requirements that are mandatory for all employers, with certain exclusions such as small establishments and government agencies. Regulations have been developed to define how employers are to adhere to these requirements (BLS 1978).

The OSHAct requirements specify that any illness or injury to an employee that causes time lost from the job, treatment beyond first aid, transfer to another job, loss of consciousness, or an occupational illness must be recorded on a daily log of injuries and illnesses, the OSHA 300 form (previously the 200 form). This log identifies the injured person, the date and time of the injury, the department or plant location where the injury occurred, and a brief description of the occurrence of the injury, highlighting salient facts such as the chemical, physical agent, or machinery involved and the nature of the injury. An injury should be recorded on the day that it occurs, but this is not always possible with MSDs and other cumulative trauma injuries. The number of days that the person is absent from the job is also recorded upon the employee's return to work. In addition to the daily log, a more detailed form is filled out for each injury that occurs. This form provides a more detailed description of the nature of the injury, the extent of damage to the employee, the factors that could be related to the cause of the injury, such as the source or agent that produced the injury, and events surrounding the injury occurrence. A workers' compensation form can be substituted for the OSHA 301 form (previously the 101 form) because equivalent information is gathered on these forms.

The OSHA Act injury and illness system specifies a procedure for calculating the frequency of occurrence of occupational injuries and illnesses and an index of their severity. These can be used by companies to monitor their health and safety performance. National data by major industrial categories are compiled by the U.S. Bureau of Labor Statistics annually and can serve as a basis of comparison of individual company performance within an industry. Thus, a company can determine whether its injury rate is better or worse than that of other companies in its industry. This industry-wide injury information is available on the OSHA website (http://www.osha.gov).

The OSHA system uses the following formula in determining company annual injury and illness incidence. The total number of recordable injuries is multiplied by 200,000 and then divided by the number of hours worked by the company employees. This gives an injury frequency per 100 person hours of work (injury incidence). These measures can be compared to an industry average.

$$\text{Incidence} = \frac{\text{number of recordable injuries and illnesses} \times 200{,}000}{\text{number of hours worked by company employees}}$$

where The number of recordable injuries and illnesses is taken from the OSHA 300 daily log of injuries.
The number of hours worked by employees is taken from payroll records and reports prepared for the Department of Labor or the Social Security Administration.

It is also possible to determine the severity of company injuries. Two methods are typically used. In the first, the total number of days lost due to injuries is compiled from the OSHA 300 daily log and divided by the total number of injuries recorded on the OSHA 300 daily log. This gives an average number of days lost per injury. In the second, the total number of days lost is multiplied by 200,000 and then divided by the number of hours worked by the company employees. This gives a severity index per 100 person hours of work. These measures can also be compared to an industry average.

Injury incidence and severity information can be used by a company to monitor its injury and illness performance over the years to examine improvement and the effectiveness of health and safety interventions. Such information provides the basis for making corrections in the company's approach to health and safety and can serve as the basis of rewarding managers and workers for good performance. However, it must be understood that injury statistics give only a crude indicator of safety company performance and an even cruder indicator of individual manager or worker performance. This information can be used to compare company safety performance with the industry average.

Because injuries are rare events, they do not always reflect the sum total of daily performance of company employees and managers. Thus, while they are an accurate measure of overall company safety performance, they are an insensitive measure at the individual and departmental levels. Some experts feel that more basic information has to be collected to provide the basis for directing health and safety efforts. One proposed measure is to use first-aid reports from industrial clinics. These provide information on more frequent events than the injuries required to be reported by the OSHAct. It is thought that these occurrences can provide insights into patterns of hazards and/or behaviors that may lead to the more serious injuries and that their greater number provides a larger statistical base for determining accident potential.

8.3. Incident Reporting

Another approach is for a company to keep track of all accidents whether an illness or injury is involved or not. Thus, property damage accidents without illness or injury would be recorded, as would near accidents and incidents that almost produced damage or injury. The proponents of this system feel that a large database can be established for determining accident-causation factors. As with the first-aid reports, the large size of the database is the most salient feature of this approach. A major difficulty in both systems is the lack of uniformity of recording and reporting the events of interest. The method of recording is much more diffuse because the nature of the events will differ substantially from illnesses or injuries, making their description in a systematic or comparative way difficult.

This critique is aimed not at condemning these approaches but at indicating how difficult they are to define and implement. These systems provide a larger base of incidents than the limited occurrences in injury-recording systems. The main problem is in organizing them into a meaningful pattern. A more fruitful approach than looking at these after-the-fact occurrences may be to look at the conditions that can precipitate injuries, that is, hazards. They can provide a large body of information for a statistical base and can also be organized into meaningful patterns.

9. CONTROLLING WORKPLACE HAZARDS

With the workplace hazards identified and defined, the next logical step is to eliminate or control them. Historically, there have been two predominant concepts about hazard controls. The first concept is to think of hazard control as a hierarchy of methods of control. In the hierarchy, the best control method is to eliminate the hazard through redesign or substitution. If elimination or substitution cannot be achieved, then the next-best approach is to block employee access to the hazard. Finally, if blocking cannot be achieved, then a last approach would be to warn the employees of the hazard and train them how to avoid the hazard.

A second way to conceptualize hazard control is in terms of the type of control: engineering controls, human factors controls, and organizational controls. Engineering controls include modifying the technology, workstation, tools, environment, or other physical aspects of work to eliminate, remove, substitute, or block access to the hazard.

Human factors controls deal with fitting the work activity to the employee's capabilities. Organizational controls involve things such as improving work procedures and practices, providing training, rotating employees to reduce the amount of exposure, and providing rest breaks designed to reduce the impact of hazards. All of these types of controls are not mutually exclusive and should be used together to achieve hazard reductions.

9.1. Engineering Controls

It would seem that the simplest way to deal with a hazard would be to get rid of it. This can be accomplished by redesigning a product, tool, machine, process, or environment or through substitution of a nonhazardous or less hazardous material or machine. For example, the loading of a mechanical punch press can be accomplished by placing a part directly onto the die with the employee's hand, which puts the hand directly into the point of operation. If the press should inadvertently cycle, the employee could injure his or her hand. To eliminate this hazard, a fixture can be designed so that the employee can place the part onto the fixture and then slide the fixture with the part into the point of operation. Thus, the fixture and not the employee's hand goes into the point of operation. This redesign removes the hand from the hazardous area of the machine. Likewise, a barrier guard could be put over the point of operation so that the employee's hand could not fit into the danger zone. This will be discussed in the next paragraph. Another example is substituting a less hazardous chemical for a more hazardous chemical, thereby reducing the extent of risk or the level of exposure.

The second class of engineering interventions is blocking employee access to the hazard. This can be achieved by putting up a barrier that keeps the employee from entering a hazardous area. The best example of this is fencing off an area such as high-voltage transformers. With this type of intervention, the hazard remains but access to the hazard is limited. However, often the hazardous area must be accessed for maintenance or other reasons. In this case, there are often secondary hazard controls to protect those who cross the barrier. For example, robots usually have a barrier around them to keep employees outside of their arc of swing so that they do not inadvertently come into contact with the robot's arm. But when the robot has to be programmed or maintained, an employee has to go across the barrier to access the robot. A secondary control is to have the robot automatically shut down when the barrier is breached. This is a form of interlock that keeps the hazard inactive while employees are present in the danger zone. In the case of many hazards, such as the high-voltage transformer, it may not be possible to have a secondary hazard control. Then we must rely on the knowledge, skills, and good sense of the employee and/or the person breaching the barrier. These human factor controls will be discussed below.

Containment is a form of a barrier guard that is used primarily with very dangerous chemical and physical hazards. An example is the ionizing radiation from a nuclear reactor. This radiation at the core of the reactor is restrained from leaving the reactor by lead-lined walls, but if leakage should occur through the walls, a back-up barrier contains the leakage. In the case of a closed system, the employee never comes in contact with the source (such as the reactor core) of the hazard. The system is designed through automation to protect the employee from the hazard source. Many chemical plants use the concept of a closed system of containment. The only time an employee would contact these specific deadly hazards would be in the case of a disaster in which the containment devices failed.

Another form of barrier control that looks something like a secondary hazard control is a guard, which is used to cover moving parts that are accessible by the employees—for example, inrunning nip points on a machine. Such guards are most often fixed and cannot be removed except for maintenance. Sometimes the guard needs to be moved to access the product. For example, when a power press is activated, there is a hazard at the point of operation. When the ram is activated, guards are engaged that prohibit an employee's contact with the die. When the ram is at rest, the guard can be lifted to access the product. If the guard is lifted, an interlock prohibits the ram from being activated. In this situation, there is a barrier to keep the employee from the area of the hazard only when the hazard is present. The guard allows access to the area of the hazard for loading, unloading, and other job operations that can be carried out without activation. But when the machine energy is activated, the guard moves into place to block the employee from access to the site of the action. In the case of the robot, the hazard area is quite large and a perimeter barrier is used; but in the case of a mechanical press, the hazard area is limited to the point of operation, which is quite small.

Yet another engineering control that is important for dealing with workplace hazards is the active removal of the hazard before it contacts the employee during the work process. An example is a local scavenger ventilation system that sucks the fumes produced by an operation such as spot welding or laser surgery away from the employees. This exhausts the fumes into the air outside of the plant (surgery room) and away from the employees. The ventilation systems must comply with federal, state, and local regulations in design and in the level of emissions into the environment. Thus, the fumes may need to be scrubbed clean by a filter before being released into the open environment. A related ventilation approach is to dilute the extent of employee exposure to airborne contaminants by bringing in more fresh air from outside the plant on a regular basis. The fresh air dilutes the concentration of the contaminant to which the employee is exposed to a level that is below the threshold of dangerous exposure. The effectiveness of this approach is verified by measuring the ambient air level of contamination and employee exposure levels on a regular basis. When new materials or chemicals are introduced into the work process or when other new airborne exposures are introduced into the plant, the adequacy of the ventilation dilution approach to provide safe levels of exposure(s) must be reverified. (See Hagopian and Bastress 1976 for recommendations for ventilation guidelines.)

When guarding or removal systems (e.g., saw guards, scavenger and area ventilation) cannot provide adequate employee protection, then personal protective equipment (PPE) must be worn by the employees (safety glasses, respirator). Because it relies on compliance by the employees, this is not a preferred method of control. A cardinal rule of safety and health engineering is that the primary method of controlling hazards is through engineering controls. Human factors controls are to be used primarily when engineering controls are not practical, feasible, solely effective in hazard control, or cost effective. It is recognized that human factor controls are often necessary as adjuncts (supplements) to engineering controls and in many instances are the only feasible and effective controls.

9.2. Human Factors Controls

In the traditional scheme of hazard control, there are two elements of human factors considerations for controlling hazards: warning and training. These can also be conceptualized as informing employees about hazards and promoting safe and healthful employee behavior. Cohen and Colligan (1998) conducted a literature review of safety training effectiveness studies and found that occupational safety and health training was effective in reducing employee hazard risks and injuries.

9.2.1. Informing

Informing employees about workplace hazards has three aspects: the right to know, warnings, and instructions. Regarding the right to know, federal safety and health regulations and many state and local regulations (ordinances) specify that an employer has the obligation to inform employees of hazardous workplace exposures to chemicals, materials, or physical agents that are known to cause harm. The local requirements of reporting vary and employers must be aware of the reporting requirements in the areas where they have facilities. Generally, an employer must provide information on the name of the hazard, its potential health effects, exposure levels that produce adverse health effects, and the typical kinds of exposures encountered in the plant. In addition, if employees are exposed to a toxic agent, information about first aid and treatment should be available. For each chemical or material or physical agent classified as toxic by OSHA, employers are required to maintain a standard data sheet that provides detailed information on its toxicity, control measures, and standard operating procedures (SOPS) for using the product. A list of hazardous chemicals, materials, and physical agents is available from your local OSHA office or the OSHA website (http://www.osha.gov). These standard data sheets (some are referred to as material safety data sheets [MSDS]) must be supplied to purchasers by the manufacturer who sells the product. These data sheets must be shared by employers with employees who are exposed to the specific hazardous products, and must be available at the plant (location) as an information resource in case of an exposure or emergency. The motivation behind the right-to-know concept is that employees have a basic right to knowledge about their workplace exposures and that informed employees will make better choices and use better judgment when they know they are working with hazardous materials.

Warnings are used to convey the message of extreme danger. They are designed to catch the attention of the employee, inform the employee of a hazard, and instruct him or her in how to avoid the hazard. The OSHA regulations require that workplace warnings meet the appropriate ANSI standards, including Z35.1-1972 specifications for accident prevention signs; Z35.4-1973 specifications for informational signs complementary to ANSI Z35.1-1972; and ANSI Z53.1-1971 safety color code for marking physical hazards. These ANSI standards were revised in 1991 as Z535.1-535.4. Warnings are primarily visual but can also be auditory, as in the case of a fire alarm. Warnings use sensory techniques that capture the attention of the employee. For instance, the use of the color red has a cultural identification with danger. The use of loud, discontinuous noise is culturally associated with emergency situations and can serve as a warning. After catching attention, the warning must provide information about the nature of the hazard. What is the hazard and what will it do to you? This provides the employee with an opportunity to assess the risk of ignoring the warning. Finally, the warning should provide some information about specific actions to take to avoid the hazard, such as ''Stay clear of the boom'' or ''Stand back 50 feet from the crane'' or ''Stay away from this area.''

Developing good warnings requires following the ANSI standards, using the results of current scientific studies and good judgment. Lehto and Miller (1986) wrote a book on warnings, and Lehto and Papastavrou (1993) define critical issues in the use and application of warnings. Laughery and Wogalter (1997) define the human factors aspects of warnings and risk perception and considerations for designing warnings. Peters (1997) discusses the critical aspects of technical communications that need to be considered from both human factors and legal perspectives. Considerations such as the level of employee's word comprehension, the placement of the warning, environmental distortions, wording of instructions, and employee sensory overload, just to name a few, must be taken into account for proper warning design and use. Even when good warnings are designed, their ability to influence employee behavior varies widely. Even so, the regulations require their use and they do provide the employee an opportunity to make a choice. Warnings should never be used in place of engineering controls. Warnings always serve as an adjunct to other means of hazard control.

Instructions provide direction to employees that will help them to avoid or deal more effectively with hazards. They are the behavioral model that can be followed to ensure safety. The basis of good instructions is the job analysis, which provides detailed information on the job tasks, environment, tools, and materials used. The job analysis will identify high-risk situations. Based on verification of the information in the job analysis, a set of instructions on how to avoid hazardous situations can be developed. The implementation of such instructions as employee behavior will be covered in the next section under training and safe behavior improvement.

9.2.2. Promoting Safe and Healthful Behavior

There are four basic human factors approaches that can be used in concert to influence employee behavior to control workplace hazards:

1. Applying methods of workplace and job design to provide working situations that capitalize on worker skills
2. Designing organizational structures that encourage healthy and safe working behavior
3. Training workers in the recognition of hazards and proper work behavior(s) for dealing with these hazards
4. Improving worker health and safety behavior through work practices improvement

Each of these approaches is based on certain principles that can enhance effective safety performance.

9.2.3. Workplace and Job Design

The sensory environment in which job tasks are carried out influences worker perceptual capabilities to detect hazards and respond to them. Being able to see or smell a hazard is an important prerequisite in dealing with it; therefore, workplaces have to provide a proper workplace sensory environment for hazard detection. This means proper illumination and noise control and adequate ventilation.

There is some evidence that appropriate illumination levels can produce significant reductions in accident rate (McCormick 1976). The environment can influence a worker's ability to perceive visual and auditory warnings such as signs or signals. To ensure the effectiveness of warnings, they should be highlighted. For visual signals, use the colors defined in the ANSI standard (Z535.1, ANSI 1991) and heightened brightness. For auditory signals, use changes in loudness, frequency, pitch, and phasing.

Work environments that are typically very loud and do not afford normal conversation can limit the extent of information exchange and may even increase the risk of occupational injury (Barreto et al. 1997). In such environments, visual signs are a preferred method for providing safety information. However, in most situations of extreme danger, an auditory warning signal is preferred because it attracts attention more quickly and thus provides for a quicker worker response. In general, warning signals should quickly attract attention, be easy to interpret, and provide information about the nature of the hazard.

Proper machinery layout, use, and design should be a part of good safety. Work areas should be designed to allow for traffic flow in a structured manner in terms of the type of traffic, the volume of traffic, and the direction of flow. The traffic flow process should support the natural progression of product manufacture and/or assembly. This should eliminate unnecessary traffic and minimize the complexity and volume of traffic. There should be clearly delineated paths for traffic to use and signs giving directions on appropriate traffic patterns and flow.

Work areas should be designed to provide workers with room to move about in performing tasks without having to assume awkward postures or come into inadvertent contact with machinery. Task-analysis procedures can determine the most economical and safest product-movement patterns and should serve as the primary basis for determining layout of machinery, work areas, traffic flow, and storage for each workstation.

Equipment must conform to principles of proper engineering design so that the controls that activate the machine, the displays that provide feedback of machine action, and the safeguards to protect workers from the action of the machine are compliant with worker skills and expectations. The action of the machine must be compliant with the action of the controls in temporal, spatial, and force characteristics.

The layout of controls on a machine is very important for proper machinery operation, especially in an emergency. In general, controls can be arranged on the basis of (1) their sequence of use, (2) common functions, (3) frequency of use, and (4) relative importance. Any arrangement should take into consideration (1) the ease of access, (2) the ease of discrimination, and (3) safety considerations such as accidental activation. The use of a sequence arrangement of controls is often preferred because it ensures smooth, continuous movements throughout the work operation. Generally, to enhance spatial compliance, the pattern of use of controls should sequence from left to right and from top to bottom. Sometimes controls are more effective when they are grouped by common functions. Often controls are clustered by common functions that can be used in sequence so that a combination of approaches is used.

To prevent unintentional activation of controls, the following steps can be taken: (1) recess the control, (2) isolate the control to an area on the control panel where it will be hard to trip unintentionally, (3) provide protective coverings over the control, (4) provide lock-out of the control so that it cannot be tripped unless unlocked, (5) increase the force necessary to trip the control so that extra effort is necessary and/or (6) require a specific sequence of control actions such that one unintentional action does not activate the machinery.

A major deficiency in machinery design is the lack of adequate feedback to the machine operator about the machine action, especially at the point of operation. Such feedback is often difficult to provide because there typically are no sensors at the point of operation (or other areas) to determine when such action has taken place. However, an operator should have some information about the results of the actuation of controls to be able to perform effectively. Operators may commit unsafe behaviors to gain some feedback about the machine's performance as the machine is operating that may put them in contact with the point of operation. To avoid this, machinery design should include feedback of operation. The more closely this feedback reflects the timing and action of the machinery, the greater the amount of control that can be exercised by the operator. The feedback should be displayed in a convenient location for the operator at a distance that allows for easy readability.

Work task design is a consideration for controlling safety hazards. Tasks that cause employees to become fatigued or stressed can contribute to exposures and accidents. Task design has to be based on considerations that will enhance employer attention and motivation. Thus, work tasks should be meaningful in terms of the breadth of content of the work that will eliminate boredom and enhance the employee's mental state. Work tasks should be under the control of the employees, and machine-paced operations should be avoided. Tasks should not be repetitious. This last requirement is sometimes hard to achieve. When work tasks have to be repeated often, providing the employee with some control over the pacing of the task reduces stress associated with such repetition. Because boredom is also a consideration in repetitious tasks, employee attention can be enhanced by providing frequent breaks from the repetitious activity to do alternative tasks or take a rest. Alternative tasks enlarge the job and enhance the breadth of work content and employee skills.

The question of the most appropriate work schedule is a difficult matter. There is evidence that rotating-shift systems produce more occupational injuries than fixed-shift schedules (Smith et al. 1982). This implies that fixed schedules are more advantageous for injury control. However, for many younger workers (without seniority and thus often relegated to afternoon and night shifts) this may produce psychosocial problems related to family responsibilities and entertainment needs, and therefore lead to stress. Because stress can increase illness and injury potential, the gain from the fixed-shift systems may be negated by stress. This suggests that one fruitful approach may be to go to fixed shifts with volunteers working the nonday schedules. Such an approach provides enhanced biological conditions and fewer psychosocial problems.

Overtime work should be avoided because of fatigue and stress considerations. It is preferable to have a second shift of workers than to overtax the physical and psychological capabilities of employees. Since a second shift may not be economically feasible, some considerations need to be given for determining appropriate amounts of overtime. This is a judgmental determination since there is inadequate research evidence on which to base a definitive answer. It is reasonable that job tasks that create high levels of physical fatigue and/or psychological stress should not be performed more than 10 hours in one day and 50 hours in one week. Jobs that are less fatiguing and stressful can probably be safely performed for up to 12 hours per day. There is some evidence that working more than 50 hours per week can increase the risk of coronary heart disease (Breslow and Buell 1960; Russek and Zohman 1958), and therefore working beyond 50 hours per week for extended periods should be avoided.

9.3. Organizational Design

Organizational policies and practices can have a profound influence on a company's health and safety record and the safety performance of its employees. To promote health and safety, organizational policies and practices should demonstrate that safety is an important organizational objective. The first step in this process is to establish a written organizational policy statement on health and safety. This should be followed up with written procedures to implement the policy. Such a formalized structure is the foundation on which all health and safety activities in the company are built. It provides the legitimate basis for undertaking health- and safety-related actions and curtails the frequent arguments among various levels of management about what constitutes acceptable activities. Such a policy statement also alerts employees to the importance of health and safety.

For employees, the policy statement is the declaration of an intent to achieve a goal. However, employees are skeptical of bureaucratic policies and look for more solid evidence of management commitment. Thus, the timing and sequence of health- and safety-related decisions demonstrate how the policy will be implemented and the importance of health and safety considerations. A health and safety policy with no follow-through is worthless and in fact may be damaging to employee morale by showing employees a lack of management commitment. This can backfire and can lead to poor employee safety attitudes and behaviors. Thus, an employer has to put the ''money where the mouth is'' to demonstrate commitment. If not, a policy is an empty promise.

Since physical conditions are the most obvious health and safety hazards, it is important that they be dealt with quickly to demonstrate management commitment. Relations with local, state, and federal health and safety agencies reflect on management commitment to health and safety. Companies that have written safety policies and guidelines with adequate follow-through but are constantly at odds with government health and safety officials are sending a confusing message to their employees. It is important to have a good public image and good public relations with government agencies, even through there may be specific instances of disagreement and even hostility. This positive public image will enhance employee attitudes and send a consistent message to employees about the importance of health and safety.

In this regard, organizations must ensure an adequate flow of information in the organization. The flow must be bidirectional, that is, upward as well as downward. One approach for dealing with safety communications is to establish communication networks. These are formal structures to ensure that information gets to the people who need to know the message(s) in a timely way. These networks are designed to control the amount of information flow to guard against information overload, misinformation, or a lack of needed information. Such networks have to be tailored to the specific needs of an organization. They are vital for hazard awareness and general health and safety information. For instance, in a multishift plant, information on a critical hazardous condition can be passed from shift to shift so that workers can be alerted to the hazard. Without a communication network, this vital information may not get to all affected employees and an avoidable exposure or accident could occur.

Organizational decision making is an important motivational tool for enhancing employee health and safety performance. Decisions about work task organization, work methods, and assignments should be delegated to the lowest level in the organization at which they can be logically made; that is, they should be made at the point of action. This approach has a number of benefits. For example, this level in the organization has the greatest knowledge of the work processes and operations and of their associated hazards. Such knowledge can lead to better decisions about hazard control. Diverse input to decision making from lower levels up to higher levels makes for better decisions as there are more options to work with. Additionally, this spreading of responsibility by having people participate in the inputs to decision making promotes employee and line supervisor consideration of safety and health issues. Such participation has been shown to be a motivator and to enhance job satisfaction (French 1963; Korunka et al. 1993; Lawler 1986). It also gives employees greater control over their work tasks and a greater acceptance of the decisions concerning hazard control due to the shared responsibility. All of this leads to decreased stress and increased compliance with safe behavior(s).

Organizations have an obligation to increase company health and safety by using modern personnel practices. These include appropriate selection and placement approaches, skills training, promotion practices, compensation packages, and employee-assistance programs. For safety purposes, the matching of employee skills and needs to job task requirements is an important consideration. It is inappropriate to place employees at job tasks for which they lack the proper skills or capacity. This will increase illness and injury risk and job stress. Selection procedures must be established to obtain a properly skilled workforce. When a skilled worker is not available, training must be undertaken to increase skill levels to the proper level before a task is undertaken. This assumes that the employer has carried out a job task analysis and knows the job skills that are required. It also assumes that the employer has devised a way to test for the required skills. Once these two conditions have been met, the employer can improve the fit between employee skills and job task requirements through proper selection, placement, and training. Many union contracts require that employees with seniority be given first consideration for promotions. Such consideration is in keeping with this approach as long as the worker has the appropriate skills to do the job task or the aptitude to be trained to attain the necessary skills. If a person does not have the necessary knowledge and skills or cannot be adequately trained, there is good reason to exclude that individual from a job regardless of seniority.

9.3.1. Safety Training

Training workers to improve their skills and recognize hazardous conditions is a primary means for reducing exposures and accidents. Cohen and Colligan (1998) found that safety and health training was effective in reducing employee risk. Training can be defined as a systematic acquisition of knowledge, concepts, or skills that can lead to improved performance or behavior. Eckstrand (1964) defined seven basic steps in training: (1) defining the training objectives, (2) developing criteria measures for evaluating the training process and outcomes, (3) developing or deriving the content and materials to be learned, (4) designing the techniques to be used to teach the content, (5) integrating the learners and the training program to achieve learning, (6) evaluating the extent of learning, and (7) modifying the training process to improve learner comprehension and retention of the content. These steps provide the foundation for the application of basic guidelines that can be used for designing the training content, and integrating the content and the learner.

In defining training objectives, two levels can be established: global and specific. The global objectives are the end goals that are to be met by the training program. For instance, a global objective might be the reduction of eye injuries by 50%. The specific objectives are those that are particular to each segment of the training program, including the achievements to be reached by the completion of each segment. A specific objective might be the ability to recognize eye-injury hazards by all employees by the end of the hazard-education segment. A basis for defining training objectives is the assessment of company safety problem areas. This can be done using hazard-identification methods such as injury statistics, inspections, and hazard surveys. Problems should be identified, ranked in importance, and then used to define objectives.

To determine the success of the training process, criteria for evaluation need to be established. Hazard-identification measures can be used to determine overall effectiveness. Thus, global objectives can be verified by determining a reduction in injury incidence (such as eye injuries) or the elimination of a substantial number of eye hazards. However, it is necessary to have more sensitive measures of evaluation that can be used during the course of training to assess the effectiveness of specific aspects of the training program. This helps to determine the need to redirect specific training segments if they prove to be ineffective. Specific objectives can be examined through the use of evaluation tools. For instance, to evaluate the ability of workers to recognize eye hazards, a written or oral examination can be used. Hazards that are not recognized can be emphasized in subsequent training and retraining.

The content of the training program should be developed based on the learners' knowledge level, current skills, and aptitudes. The training content should be flexible enough to allow for individual differences in aptitudes, skills, and knowledge, as well as for individualized rates of learning. The training content should allow all learners to achieve a minimally acceptable level of health and safety knowledge and competence by the end of training. The specifics of the content deal with the skills to be learned and the hazards to be recognized and controlled.

There are various techniques that can be used to train workers. Traditionally, on-the-job training (OJT) has been emphasized to teach workers job skills and health and safety considerations. The effectiveness of such training will be influenced by the skill of the supervisor or lead worker in imparting knowledge and technique as well as his or her motivation to successfully train the worker. First-line supervisors and lead workers are not educated to be trainers and may lack the skills and motivation to do the best job. Therefore, OJT has not always been successful as the sole safety training method. Since the purpose of a safety training program is to impart knowledge and teach skills, it is important to provide both classroom experiences to gain knowledge and OJT to attain skills.

Classroom training is used to teach concepts and improve knowledge and should be carried out in small groups (not to exceed 15 employees). A small group allows for the type of instructor–student interaction needed to monitor class progress, provide proper motivation and determine each learner's comprehension level. Classroom training should be given in an area free of distractions to allow learners to concentrate on the subject matter. Training sessions should not exceed 30 minutes, after which workers can return to their regular duties. There should be liberal use of visual aids to increase comprehension and make the training more concrete and identifiable to the learners. In addition, written materials should be provided that can be taken from the training session for study or reference away from the classroom.

For OJT the major emphasis should be on enhancing skills through observation and practice. Key workers with exceptional skills can be used as role models and mentors. Learners can observe these key workers and pick up tips from them. They then can practice what they have learned under the direction of the key workers to increase their skill, obtain feedback on their technique, and be motivated to improve.

Once the learner and the training program have been integrated, it will be necessary to evaluate the extent of learning. This can be done by testing learner knowledge and skills. Such testing should be done frequently throughout the training process to provide the learners with performance feedback and allow for program redirection as needed. Knowledge is best tested by written examinations that test acquisition of facts and concepts. Pictorial examinations (using pictures or slides of working conditions) can be used to determine hazard recognition ability. Oral questioning on a frequent basis can provide the instructor with feedback on the class comprehension of materials being presented, but should not be used for individual learner evaluation, since some learners may not be highly verbal and could be demotivated by being asked to recite. Skills testing should take place in the work area under conditions that control hazard exposures. Skills can be observed during practice sessions to determine progress under low-stress conditions.

The final stage in a training program, the success of the program having been determined, is to make modifications to improve the learning process. Such modifications should be done on a continuous basis as feedback on learner performance is acquired. In addition, at the end of the program it is necessary to determine whether the company objectives have been met. If so, should the objectives be modified? The answers to these questions can lead to modifications in the training program.

9.3.2. Hazard Reduction through Improved Work Practices

A large number of the hazards in the workplace are produced by the interaction between employees and their tools and environment. These hazards cannot be completely controlled through hazard inspection and machine guarding. They can be controlled by increasing employee recognition of the hazards and by proper worker behavior. Such behavior may be an evasive action when a hazard occurs, or it may be the use of safe work procedures to ensure that hazards will not occur. There are very few hazard-control efforts that are not in some way dependent on employee behavior. Making employees aware of hazards is meaningless if they do not choose to do something about them. For example, when controlling chemical exposures, personal protective equipment is useless if it is not worn. Likewise, an inspection system is useless if hazards are not reported or not corrected when reported. Thus, taking positive action (behavior) is central to hazard control. It is often true that there are no ideal engineering control methods to deal with a certain hazard. In such a case, it is usually necessary to use proper work practices to avoid hazardous exposure when engineering controls are not feasible. Likewise, even when engineering control will work successfully it is necessary to have employees use good work practices to get the engineering controls to work properly.

Conard (1983) has defined work practices as employee behaviors that can be simple or complex and that are related to reducing a hazardous situation in occupational activities. A series of steps can be used in developing and implementing work practices for eliminating occupational hazards:

1. The definition of hazardous work practices
2. The definition of new work practices to reduce the hazards
3. Training employees in the desired work practices
4. Testing the new work practices in the job setting
5. Installing the new work practices using motivators
6. Monitoring the effectiveness of the new work practices
7. Redefining the new work practices
8. Maintaining proper employee habits regarding work practices

In defining hazardous work practices, there are a number of sources of information that should be examined. Injury and accident reports such as the OSHA 301 Form provide information about the circumstances surrounding an injury. Often employee or management behaviors that contributed to the injury can be identified. Employees are a good source of information about workplace hazards. They can be asked to identify critical behaviors that may be important as hazard sources or hazard controls. First-line supervisors are also a good source of information because they are constantly observing employee behavior. All of these sources should be examined; however, the most important source of information is in directly observing employees at work.

There are a number of considerations when observing employee work behaviors. First, observation must be an organized proposition. Before undertaking observations, it is useful to interview employees and first-line supervisors and examine injury records to develop a checklist of significant behaviors to be observed. This should include hazardous behaviors as well as those that are used to enhance engineering control or directly control hazards. The checklist should identify the job task being observed, the types of behaviors being examined, their frequency of occurrence, and a time frame of their occurrence. The observations should be made at random times so that employees do not change their natural modes of behavior when observed. The time of observation should be long enough for a complete cycle of behaviors associated with a work task(s) of interest to be examined. Two or three repetitions of this cycle should be examined to determine consistency in behavior with an employee and among employees. Random times of recording behavior are most effective in obtaining accurate indications of typical behavior. The recorded behaviors can be analyzed by the frequency and pattern of their occurrence as well as their significance for hazard control. Hot spots can be identified. All behaviors need to be grouped into categories in regard to hazard control efforts and then prioritized.

The next step is to define the proper work practices that need to be instilled to control the hazardous procedures observed. Sometimes the observations provide the basis for the good procedures that you want to implement. Often, however, new procedures need to be developed. There are four classes of work practices that should be considered: (1) hazard recognition and reporting, (2) housekeeping, (3) doing work tasks safely, and (4) emergency procedures. The recognition of workplace hazards requires that the employee be cognizant of hazardous conditions through training and education and that employees actively watch for these conditions. Knowledge is useless unless it is applied. These work practices ensure the application of knowledge and the reporting of observed hazards to fellow workers and supervisors. Housekeeping is a significant consideration for two reasons. A clean working environment makes it easier to observe hazards. It is also a more motivating situation that enhances the use of other work practices.

The most critical set of work practices deals with carrying out work tasks safely through correct skill use and hazard-avoidance behaviors. This is where the action is between the employee and the environment, and it must receive emphasis in instilling proper work practices. Situations occur that are extremely hazardous and require the employee to get out of the work area or stay clear of the work area. These work practices are often life-saving procedures that need special consideration because they are used only under highly stressful conditions, such as emergencies.

Each of these areas needs to have work practices spelled out. These should be statements of the desired behaviors specified in concise, easily understandable language. Statements should typically be one sentence long and should never exceed three sentences. Details should be excluded unless they are critical to the proper application of the work practice. The desired work practices having been specified, employees should be given classroom and on-the-job training to teach them the work practices. Training approaches discussed earlier should be applied. This includes classroom training as well as an opportunity for employees to test the work practices in the work setting.

To ensure the sustained use of the learned work practices, it is important to motivate workers through the use of incentives. There are many types of incentives, including money, tokens, privileges, social rewards, recognition, feedback, participation, and any other factors that motivate employees, such as enriched job tasks. Positive incentives should be used to develop consistent work practice patterns.

Research has demonstrated that the use of financial rewards in the form of increased hourly wage can have a beneficial effect on employee safety behaviors and reduced hazard exposure. One study (Smith et al. 1983; Hopkins et al. 1986) evaluated the use of behavioral approaches for promoting employee use of safe work practices to reduce their exposure to styrene. The study was conducted in three plants and had three components: (1) the development and validation of safe work practices for working with styrene in reinforced fiberglass operations, (2) the development and implementation of an employee training program for learning the safe work practices, and (3) the development and testing of a motivational technique for enhancing continued employee use of the safe work practices. Forty-three work practices were extracted from information obtained from a literature search, walk-through plant survey, interviews with employees and plant managers, and input from recognized experts in industrial safety and hygiene. The work practices were pilot tested for their efficacy in reducing styrene exposures. A majority of the work practices were found to be ineffective in reducing styrene exposures, and only those that were effective were incorporated into a worker training program.

The worker training program consisted of classroom instruction followed up with on-the-job application of the material learned in class. Nine videotapes were made to demonstrate the use of safe work practices. Basic information about each work practice and its usefulness was presented, followed by a demonstration of how to perform the work practice. Employees observed one videotape for 15 minutes per week for nine weeks. After each showing, a discussion session was held, followed up by on-the-job application of the work practice given by the research training instructor. Once training was completed, each employee was included in the motivational program. This program was based on a financial reward of $10 per week for using the safe work practices. Observations of employee behavior were made by researchers four times daily on a random basis. These observations served as the basis for an employee's receipt of the financial reward.

The effectiveness of the training and motivational programs was measured by examining the changes in employee behavior from before the programs to the end of the study. Approximately 35% of the safe work practices were observed prior to training for the 41 employees studied in the three plants. At the end of the study, approximately 95% of the safe work practices were observed. The real significance of this increased use of safe work practices lies in the effectiveness in reducing employee exposures to styrene. The results indicated a reduction in styrene exposure from before training to the end of the study of 36%, 80%, and 65% for each plant respectively. In a follow-up evaluation a few years after the behavioral management program was discontinued, it was found that approximately 90% of the safe work practices were still being used by the employees even in the absence of rewards. The study results demonstrated the effectiveness of behavioral techniques for increasing worker use of safe work practices, as well as the effectiveness of such usage in reducing employee exposures to workplace hazards.

10. SAFETY PROGRAMS

The preceding materials provide the basis for developing an effective company hazard-control program. However, there are a number of other elements to consider in developing a safety program or upgrading your current program. These include organizational policies, managing various elements of the program, motivational practices, hazard-control procedures, dealing with employees, accident investigations, and injury recording. Aspects of each of these have already been discussed, and in this section they are integrated into an effective safety program. There has been considerable research into the necessary elements for a successful safety program (see Cohen 1977; Smith et al. 1978; Cleveland et al. 1979; Zimolong 1997) and how these elements should be applied. One primary factor emerges from every study on this subject. A safety program will not be successful unless there is a commitment to the program by top management. This dictates that there be a written organizational policy statement on the importance of safety and the general procedures the corporation intends to use to meet this policy. Having such a policy is just the first step toward effective management commitment.

Smith et al. (1978) have shown that it takes more than a written policy to ensure successful safety performance. It takes involvement on the part of all levels of management in the safety program. From the top managers it means that they must get out onto the shop floor often and talk to employees about plant conditions and safety problems. This can be on a scheduled basis, but it seems to be more effective on an informal basis. For middle managers there is a need to participate in safety program activities such as monthly hazard awareness meetings or weekly toolbox meetings. This does not necessitate active presentations by these managers, but it does mean active participation in group discussions and answering worker questions. These activities bring the upper and middle managers in touch with potential hazard sources and educates them to shop floor problems. It also demonstrates to employees that management cares about their safety and health.

Another aspect of management commitment is the level of resources that are made available for safety programming. Cohen (1977), in reviewing successful program research, found that organizational investment in full-time safety staff was a key feature to good plant safety performance. The effectiveness of safety and health staff was greater the higher they were in the management structure. The National Safety Council (1974) has suggested that plants with less than 500 employees and a low to moderate hazard level can have an effective program with a part-time safety professional. Larger plants or those with more hazards need more safety staff.

Along with funds for staffing, successful programs also make funds available for hazard abatement in a timely fashion. Thus, segregated funds are budgeted to be drawn upon when needed. This gives the safety program flexibility in meeting emergencies when funds may be hard to get quickly from operating departments. An interesting fact about companies with successful safety programs is that they are typically efficient in their resource utilization, planning, budgeting, quality control, and other aspects of general operations and include safety programming and budgeting as just another component of their overall management program. They do not single safety out or make it special; instead, they integrate it into their operations to make it a natural part of daily work activities.

Organizational motivational practices will influence employee safety behavior. Research has demonstrated that organizations that exercise humanistic management approaches have better safety performance (Cohen 1977; Smith et al. 1978; Cleveland et al. 1979). These approaches are sensitive to employee needs and thus encourage employee involvement. Such involvement leads to greater awareness and higher motivation levels conducive to proper employee behavior. Organizations that use punitive motivational techniques for influencing safety behavior have poorer safety records than those using positive approaches. An important motivational factor is encouraging communication between various levels of the organization (employees, supervisors, managers). Such communication increases participation in safety and builds employee and management commitment to safety goals and objectives. Often informal communication is a more potent motivator and provides more meaningful information for hazard control.

An interesting research finding is that general promotional programs aimed at enhancing employee awareness and motivation, such as annual safety awards dinners and annual safety contests, are not very effective in influencing worker behavior or company safety performance (Smith et al. 1978). The major reason is that their relationship in time and subject matter (content) to actual plant hazards and safety considerations is so abstract that workers cannot translate the rewards to specific actions that need to be taken. It is hard to explain why these programs are so popular in industry despite being so ineffective. Their major selling points are that they are easy to implement and highly visible, whereas more meaningful approaches take more effort.

Another important consideration in employee motivation and improved safety behavior is training. Two general types of safety training are of central importance: skills training and training in hazard awareness. Training is a key component to any safety program because it is important to employee knowledge of workplace hazards and proper work practices and provides the skills necessary to use the knowledge and the work practices. Both formal and informal training seem to be effective in enhancing employee safety performance (Cohen and Colligan 1998). Formal training programs provide the knowledge and skills for safe work practices, while informal training by first-line supervisors and fellow employees maintains and sharpens learned skills.

All safety programs should have a formalized approach to hazard control. This often includes an inspection system to define workplace hazards, accident investigations, record keeping, a preventive maintenance program, a machine guarding program, review of new purchases to ensure compliance with safety guidelines, and housekeeping requirements. All contribute to a safety climate that demonstrates to workers that safety is important. However, the effectiveness of specific aspects of such a formalized hazard-control approach have been questioned (Cohen 1977; Smith et al. 1978). For instance, formalized inspection programs have been shown to deal with only a small percentage of workplace hazards (Smith et al. 1971). In fact, Cohen (1977) indicated that more frequent informal inspections may be more effective than more formalized approaches. However, the significance of formalized hazard-control programs is that they establish the groundwork for other programs such as work practice improvement and training. In essence, they are the foundation for other safety approaches. They are also a source of positive motivation by demonstrating management interest in employees by providing a clean workplace free of physical hazards. Smith et al. (1978) have demonstrated that sound environmental conditions are a significant contribution to company safety performance and employee motivation.

11. PARTICIPATIVE APPROACHES TO RESPOND TO THE EMERGING HAZARDS OF NEW TECHNOLOGIES

Hazard control for new technologies requires a process that will be dynamic enough to be able to deal with the increasing rate of hazards caused by technological change. Research on successful safety program performance in plants with high hazard potential has shown a number of factors that contribute to success (Cohen 1977; Smith et al. 1978). These are having a formal, structured program so people know where to go for help, management commitment and involvement in the program, good communications between supervisors and workers, and worker involvement in the safety and health activities. These considerations are important because they provide a framework for cooperation between management and employees in identifying and controlling hazards. These factors parallel the basic underlying principles of quality management, social democracy, hazard surveys, ergonomic committees, and other employee-involvement approaches that will be discussed below—that is, developing management and employee cooperation, participation, and honest exchange of ideas about problems in a controlled format.

11.1. Quality Improvement

The development of total quality management (TQM) approaches may produce some positive results with regard to occupational safety and health (Zink 1999). Power and Fallon (1999) have proposed TQM as a framework for integration of health and safety activities with other functions. They argue that the practice of safety management should include the following TQM principles: management commitment to occupational safety and health objectives, plans and policies; development of a health and safety culture; employee involvement in safety activities, such as risk assessment and training of new employees; measurement and monitoring of health and safety performance; and continuous improvement. Smith (1999) has proposed a model for integrating ergonomics, safety, and quality based on behavioral cybernetics. From a behavioral cybernetics perspective, participatory ergonomics and safety and quality management are effective because they enable workers to control sensory feedback from job-related decisions or working conditions that affect them and in turn to generate sensory feedback for the control and benefit of other workers. Worker involvement in decision making, worker control over the production process, and job enrichment enhance the overall level of worker self-control. Use of workers as resource specialists and emphasis on skill development can benefit the integration of ergonomics, safety management, and quality management of the organization.

11.2. International Organization for Standardization

The International Organization for Standardization (ISO) has been developing technical standards over many sectors of business, industry, and technology since 1947. With the exception of ISO 9000 and ISO 14000, the vast majority of ISO standards are highly specific. They are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics. The goal of these standards is to ensure that materials, products, processes, and services are fit for their purpose. Then in 1987 came ISO 9000, followed nearly 10 years later by ISO 14000. These two standards are very different from the majority of ISO's highly specific standards.

ISO 9000 and ISO 14000 are known as generic management system standards. Management system standards provide the organization with a model to follow in setting up and operating the management system. ISO 9000 is concerned with quality management, whereas ISO 14000 is concerned with environmental management. Quality management regards what the organization does to ensure that its products conform to the customer's requirements. Environmental management regards what the organization does to minimize harmful effects on the environment caused by its activities. Both ISO 9000 and ISO 14000 concern the way an organization goes about its work, and not directly the result of this work. That is, ISO 9000 and ISO 14000 concern processes, and not products, at least directly. Both standards provide requirements for what the organization must do to manage processes influencing quality (ISO 9000) or the processes influencing the impact on the environment (ISO 14000).

The ISO 9000 family of standards currently contains over 20 standards and documents. By the end of the year 2000, a new version of the ISO 9000 quality management standards was issued. The three primary standards in the Year 2000 ISO 9000 are:

- ISO 9000: Quality management systems—Fundamentals and vocabulary
- ISO 9001: Quality management systems—Requirements
- ISO 9004: Quality management systems—Guidance for performance improvement.

The new ISO 9000 family of quality management standards is being developed to achieve a coherent terminology with the ISO 14000 family and other management standards, including possibly an OSH management standard.

In 1996, the ISO held an international conference in Geneva to test stakeholder views on developing a standard on occupational health and safety. Given the limited support from the main stakeholders for the ISO to develop international standards in this field, ISO decided that no further action should be taken. Recently, the ISO has reopened the issue of whether to develop management system standards to help organizations meet their responsibilities. The British Standards Institution (BSI) has submitted a proposal to the ISO for creation of a new ISO technical committee on OHS management standards. The BSI has proposed to transform BS 8800, the British "noncertifiable" OHS management system guidelines, into an ISO standard. The ISO is looking into the issue of whether or not to develop an occupational health and safety management standard.

Many companies have invested considerable resources in order to obtain certification of their quality management systems according to the ISO 9000 standards. From a safety point of view, one may wonder whether the implementation of ISO 9000 management systems can encourage the development of safer and healthier work environments. In Sweden, Karltun et al. (1998) examined the influences on working conditions, following the implementation of ISO 9000 quality systems in six small and medium-sized companies. Improvements to the physical work environment triggered by the ISO implementation process were very few. There were improvements in housekeeping and production methods. Other positive aspects present in some of the companies included job enrichment and a better of understanding of employees' role and importance to production. However, the implementation of ISO 9000 was accompanied by increased physical strain, stress, and feelings of lower appreciation. According to Karltun and his colleagues (1998), improved working conditions could be triggered by the implementation of ISO quality management standards if additional goals, such as improved working conditions, are considered by top management and if a participative implementation process is used. Others have argued that quality management systems and environmental management systems can be designed to address occupational health and safety (Wettberg 1999;

Martin 1999). A study by Eklund (1995) showed a relationship between ergonomics and quality in assembly work. Tasks that had ergonomic problems (e.g., physical and psychological demands) were also the tasks that had quality deficiencies. The evidence for the integration between quality management and occupational safety and health is weak. However, there is reason to believe that improved health and safety can be achieved in the context of the implementation of ISO 9000 management standards.

11.3. Social Democracy

One framework for addressing the health and safety issues of new technology is the social democratic approach practiced in Norway and Sweden (Emery and Thorsrud 1969; Gardell 1977). This approach is based on the concept that workers have a right to participate in decisions about their working conditions and how their jobs are undertaken. In Sweden, there are two central federal laws that establish the background for health and safety. One, similar to U.S. Occupational Safety and Health Act, established agencies to develop and enforce standards as well as to conduct research. The second is the Law of Codetermination, which legislates the right of worker representatives to participate in decision making on all aspects of work. This law is effective because over 90% of the Swedish blue- and white-collar workforce belong to a labor union and the unions take the lead in representing the interests of the employees in matters pertaining to working conditions, including health and safety. The Scandinavian approach puts more emphasis on the quality of working life in achieving worker health and well being. Thus, there is emphasis on ensuring that job design and technology implementation do not produce physical and psychological stress. This produces discussion and action when safety and health problems are first reported.

11.4. Hazard Survey

Organizational and job-design experts have long proposed that employee involvement in work enhances motivation and produces production and product quality benefits (Lawler 1986). Smith and Beringer (1986) and Zimolong (1997) have recommended that employees be involved in safety programming and hazard recognition to promote safety motivation and awareness. An effective example of using this concept in health and safety is the hazard survey program (Smith 1973; Smith and Beringer 1986). Smith et al. (1971) showed that most occupational hazards were either transient or due to improper organizational or individual behavior. Such hazards are not likely to be observed during formal inspections by safety staff or compliance inspections by state or federal inspectors. The theory proposes that the way to keep on top of these transient and behavioral hazards is to have them identified on a continuous basis by the employees as they occur through employee participation.

One approach that gets employee involvement is the hazard survey. While inspection and illness/injury analysis systems can be expected to uncover a number of workplace hazards, they cannot define all of the hazards. Many hazards are dynamic and occur only infrequently. Thus, they may not be seen during an inspection or may not be reported as a causal factor in an illness or injury. To deal with hazards that involve dynamically changing working conditions and/or worker behaviors requires a continuously operating hazard-identification system. The hazard survey is a cooperative program between employees and managers to identify and control hazards. Since the employee is in direct contact with hazards on a daily basis, it is logical to use employees' knowledge of hazards in their identification. The information gathered from employees can serve as the basis of a continuous hazard identification system that can be used by management to control dynamic workplace hazards.

A central concept of this approach is that hazards exist in many forms as fixed physical conditions, as changing physical conditions, as worker behaviors, and as an operational interaction that causes a mismatch between worker behavior and physical conditions (Smith 1973). This concept defines worker behavior as a critical component in the recognition and control of all of these hazards. Involving workers in hazard recognition sensitizes them to their work environment and acts as a motivator to use safe work behaviors. Such behaviors include using safe work procedures to reduce hazard potential, using compensatory behaviors when exposed to a known hazard, or using avoidance behaviors to keep away from known hazards. The hazard survey program also establishes communication between supervisors and employees about hazards.

The first step in a hazard survey program is to formalize the lines of communication. A primary purpose of this communication network is to get critical hazard information to decision makers as quickly as possible so that action can be taken to avert an exposure or accident. Traditional communication routes in most companies do not allow for quick communication between workers and decision makers, and thus serious hazards may not be corrected before an exposure or accident occurs. Each company has an established organizational structure that can be used to set up a formalized communication network. For instance, most companies are broken into departments or work units. These can serve as the primary segments within which workers report hazards. These hazards can be dealt with at the departmental level or communicated to higher-level decision makers for action.

Once primary communication units are established, a process to communicate hazard information has to be established. This requires structure and rules. The structure of the program should be simple so that information can flow quickly and accurately. It is important to designate a company decision maker who has the responsibility and authority to respond to serious hazards through the expenditure of company resources. Sometimes the hazards can be dealt with immediately at the departmental level by the supervisor. This is often true when large expenditures are not involved. Each department may decide to select someone in that department to serve as the primary communication source between the workers in the department and the supervisor. Hazards are reported directly to this person, who then reports them to the supervisor or, in the case of a serious hazard, immediately to the company decision maker.

It is best to have a formal procedure for recording employee-identified hazards. This can be easily accomplished by a hazard form that provides a written record of the hazard, its location, and other pertinent information, such as the number of employees exposed and possible hazard-control measures. These forms can be distributed to each employee and be available from the department committee member. Employees may wish to express their views about the existence of potential hazards anonymously on the forms. Employees should report near-miss accidents, property damage incidents, and potential injury-producing hazards. It is essential in a program such as this that employees be given anonymity if desired and that they be assured that no action will be taken against them for their participation (even if they report silly hazards).

Once hazards have been examined, rated, and ranked by the committee, some plan of action for their control should be developed either by the committee or by company management. The results of the hazards review by the committee and the action to be taken should be fed back to employees. Experience using this type of program indicates that for every 100 hazards reported, 1 is very serious, needing immediate action, 24 require attention quickly to avert a potential accident, 50 require some minor action to improve the quality of working conditions but do not concern a serious hazard, and 25 concern gripes and hassles of employees that are not related to safety hazards.

Employees are expected to fulfill the following duties in this program:

1. Examine the workplace to determine whether there are hazards
2. Report hazards on the form and return it to the department committee member or supervisor
3. Make an effort to find out what has happened to the hazard(s) reported
4. Continue to report hazards as they are observed

This program will provide continuous monitoring of safety hazards using employee input. This participation should stimulate employee awareness of safety and motivate the employees to work more safely. The continued use of the program should encourage the employees to have a vested interest in their safety and that of their fellow employees. This sense of involvement can carry over into improvement in individual work habits.

11.5. Employee/Management Ergonomics Committee

Another employee-involvement approach that could be successful in addressing some of the emerging issues of new technology is the joint union/management ergonomic committee (Hagglund 1981). This approach starts with a joint training course for union stewards and line managers about the hazards of chronic trauma and possible ergonomic interventions to resolve these problems. The course covers how to recognize ergonomic hazards, how to measure the hazard potential, and how to develop dialogue and cooperation between labor and management. This training is led by a facilitator (typically a university staff person), and is conducted at the company during work hours. Employees and supervisors are given time from their jobs to participate, which demonstrates the importance of the program. One main purpose of the training is to generate discussion between line managers/supervisors and union representatives about specific hazards and worker perceptions. This give and take develops an understanding of the other person's perspective and concerns. It often generates good solutions, especially toward the end of the course, when an understanding of the course technical material is integrated within the specific context of the plant.

After the training, an ergonomics committee composed of top management, select line management, and select union stewards is established that meets on a regular basis to discuss ergonomic problems and potential solutions. Employees with ergonomic problems can report them to a member of this committee, which typically tends to be a union steward. Semiannual retraining is given to the ergonomics committee on emerging issues that are generated by the kinds of problems being reported at the company. This approach has been extremely successful in reducing the extent of chronic trauma in electronic assembly plants in Wisconsin.

12. CONCLUSIONS

Designing for successful occupational health and safety performance requires a systematic approach. This includes understanding that the workplace is a system where changes in one element lead to influences on the other system components. It also means that efforts to make improvements must be multifaceted and address all of the elements of the work system. Health and safety improvements begin with an understanding of the hazards, the evaluation of injury and illness experience, the development of interventions, the implementation of improvements, follow-up to evaluate the results of improvements, and continuous efforts of evaluation and improvement. Good programming starts at the top of the company and includes all levels of the organizational structure. Employee input and involvement are critical for success. Often there is a need for technical expertise when dealing with complex or new hazards. In the end, having everyone in the company alert to health and safety issues should lead to improved health and safety performance.

REFERENCES

American National Standards Institute (ANSI), ‘‘Safety Color Code,’’ Z535.1-1991, ANSI, New York.

Barreto, S. M., Swerdlow, A. J., Smith, P. G., and Higgins, C. D. (1997), ‘‘A Nested Case-Control Study of Fatal Work Related Injuries among Brazilian Steelworkers,’’ *Occupational and Environmental Medicine,* Vol. 54, pp. 599–604.

Breslow, L., and Buell, P. (1960), ‘‘Mortality from Coronary Heart Disease and Physical Activity of Work in California,’’ *Journal of Chronic Diseases,* Vol. 11, pp. 615–626.

Bureau of Labor Statistics (BLS) (1978), ‘‘Recordkeeping Requirements under the Occupational Safety and Health Act of 1970,’’ U.S. Department of Labor, Washington, DC.

Bureau of Labor Statistics (1998a), http://www.osha.gov/oshstats/bltable.html.

Bureau of Labor Statistics (BLS) (1998b), ‘‘Occupational Injuries and Illnesses in US Industry, 1997,’’ U.S. Department of Labor, Washington, DC.

Bureau of Labor Statistics (BLS) (1999), ‘‘Fatal Workplace Injuries in 1997,’’ U.S. Department of Labor, Washington DC.

Century Research Corp. (1973), *Are Some People Accident Prone?* Century Research Corp., Arlington, VA.

Centers for Disease Control (CDC) (1983), ‘‘Leading Work-Related Diseases and Injuries—United States; Musculoskeletal Injuries,’’ *Morbidity and Mortality Weekly Report,* Vol. 32, pp. 189–191.

Cleveland, R. J., Cohen, H. H., Smith, M. J., and Cohen, A. (1979), ‘‘Safety Program Practices in Record-Holding Plants,’’ DHEW (NIOSH) Publication No. 79-136, U.S. GPO, Washington, DC.

Cohen, A. (1977), ‘‘Factors in Successful Occupational Safety Programs,’’ *Journal of Safety Research,* Vol. 9, No. 4, pp. 168–178.

Cohen, A., and Colligan, M. J. (1998), *Assessing Occupational Safety and Health Training: A Literature Review,* National Institute for Occupational Safety and Health, Cincinnati.

Conard, R. (1983), *Employee Work Practices,* U.S. Department of Health and Human Services, National Institute for Occupational Safety and Health, Cincinnati.

Eckstrand, G. (1964), ‘‘Current Status of the Technology of Training,’’ AMRL Doc. Tech. Rpt. 64-86, U.S. Department of Defense, Washington, DC.

Eklund, J. A. E. (1995), ‘‘Relationships between ergonomics and quality in assembly work,’’ *Applied Ergonomics,* Vol. 26, No. 1, pp. 15–20.

Emery, F. E., and Thorsrud, E. (1969), *The Form and Content of Industrial Democracy,* Tavistock Institute, London.

Environmental Protection Agency (EPA) (1999), http://www.epa.gov/ocfo/plan/plan.htm.

French, J. (1963), ‘‘The Social Environment and Mental Health,’’ *Journal of Social Issues,* Vol. 19.

Gardell, B. (1977), ‘‘Autonomy and Participation at Work,’’ *Human Relations,* Vol. 30, pp. 515–533.

Hagglund, G. (1981), ‘‘Approaches to Safety and Health Hazard Abatement,’’ *Labor Studies Journal,* Vol. 6.

Hagopian, J. H., and Bastress, E. K. (1976), *Recommended Industrial Ventilation Guidelines,* U.S. Government Printing Office, Washington, DC.

Hopkins, B. L., Conard, R. J., and Smith, M. J. (1986), ‘‘Effective and Reliable Behavioral Control Technology,’’ *American Industrial Hygiene Association Journal,* Vol. 47, pp. 785–791.

International Labour Organization (ILO) (1998), Occupational injury and illness statistics, http://www.ilo.org/public/english/protection/safework/indexold.htm.

Iverson, R. D., and Erwin, P. J. (1997), ‘‘Predicting Occupational Injury: The Role of Affectivity,’’ *Journal of Occupational and Organizational Psychology,* Vol. 70, No. 2, pp. 113–128.

Kalimo, R., Lindstrom, K., and Smith, M. J. (1997), ‘‘Psychosocial Approach in Occupational Health,’’ *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1059–1084.

Karltun, J., Axelsson, J., and Eklund, J. (1998), "Working Conditions and Effects of ISO 9000 in Six Furniture-Making Companies: Implementation and Processes," *Applied Ergonomics,* Vol. 29, No. 4, pp. 225–232.

Korunka, C., Weiss, A., and Karetta, B. (1993), "Effects of New Technologies with Special Regard for the Implementation Process per se," *Journal of Organizational Behavior,* Vol. 14, pp. 331–348.

Laflamme, L. (1997), "Age-Related Injuries among Male and Female Assembly Workers: A Study in the Swedish Automobile Industry," *Industrial Relations,* Vol. 52, No. 3, pp. 608–618.

Laflamme, L., and Menckel, E. (1995), "Aging and Occupational Accidents: A Review of the Literature of the Last Three Decades," *Safety Science,* Vol. 21, No. 2, pp. 145.

Laughery, K. R., and Wogalter, M. S. (1997), "Warnings and Risk Perception" in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1174–1197.

Lawler, E. E. (1986), *High Involvement Management,* Jossey-Bass, San Francisco.

Lehto, M. R., and Miller, J. M. (1986), *Warnings,* Fuller Technical, Ann Arbor, MI.

Lehto, M. R., and Papastavrou, J. D. (1993), "Models of the Warning Process: Important Implications towards Effectiveness," *Safety Science,* Vol. 16, pp. 569–595.

Leigh, J. P., Markowitz, S. B., Fahs, M., Shin, C., and Landrigan, P. J. (1997), "Occupational Injuries and Illness in the United States," *Archives of Internal Medicine,* Vol. 157, pp. 1557–1568.

Martin, H. (1999), "Aspects of the Integration of Quality Management and Occupational Health and Safety Systems in German Enterprises," in *Proceedings of the International Conference on TQM and Human Factors: Towards Successful Integration,* Vol. 1, J. Axelsson, B. Bergman, and J. Eklund, Eds., Centre for Studies of Humans, Technology and Organization, Linkoping, Sweden, pp. 427–432.

McCormick, E. (1976), *Human Factors Guide to Engineering,* 4th Ed., McGraw-Hill, New York.

National Safety Council (1974), *Accident Prevention Manual for Industrial Operations,* 7th Ed., National Safety Council, Chicago.

National Institute for Occupational Safety and Health (NIOSH) (1977), *Occupational Diseases: A Guide to Their Recognition,* U.S. Government Printing Office, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1984), *The Industrial Environment: Its Evaluation and Control,* 3rd Ed., U.S. Government Printing Office, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1996a), National Occupational Research Agenda, http://www.cdc.gov/niosh/norhmpg.html.

National Institute for Occupational Safety and Health (NIOSH) (1996b), National Occupational Research Agenda—Disease and Injury Priority Areas, http://www.cdc.gov/niosh/diseas.html.

Peters, G. A. (1997), "Technical Communication: Assessment of How Technical Information Is Communicated," *Technology, Law and Insurance,* Vol. 2, pp. 187–190.

Power, F. P., and Fallon, E. F. (1999), "Integrating Occupational Health and Safety Activities with Total Quality Management," in *Proceedings of the International Conference on TQM and Human Factors: Towards Successful Integration,* Vol. 1, J. Axelsson, B. Bergman, and J. Eklund, Eds., Centre for Studies of Humans, Technology and Organization, Linkoping, Sweden, pp. 445–450.

Russek, H., and Zohman, B. (1958), "Relative Significance of Heredity, Diet, and Occupational Stress in CHF of Young Adults," *American Journal of Medical Science,* Vol. 235, pp. 266–275.

Smith, K. U. (1973), "Performance Safety Codes and Standards for Industry: The Cybernetic Basis of the Systems Approach to Accident Prevention," in *Selected Readings in Safety,* J. T. Widner, Ed., Academy Press, Macon, GA.

Smith, M. J. (1986), "Occupational Stress," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 844–860.

Smith, M. J., and Beringer, D. B. (1986), "Human Factors in Occupational Injury Evaluation and Control," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York.

Smith M. J., and Sainfort, P. C. (1989), "A Balance Theory of Job Design and for Stress Reduction," *International Journal of Industrial Ergonomics,* Vol. 4, pp. 67–79.

Smith, M. J., and Carayon, P. (1990), "Demands of the Work Environment: Stress and CVD Implications," working paper, University of Wisconsin-Madison, Department of Industrial Engineering.

Smith, M. J., Bauman, R. D., Kaplan, R. P., Cleveland, R., Derks, S., Sydow, M., and Coleman, P. J. (1971), *Inspection Effectiveness,* Occupational Safety and Health Administration, Washington, DC.

Smith, M. J., Cohen, H. H., Cohen, A., and Cleveland, R. (1978), "Characteristics of Successful Safety Programs," *Journal of Safety Research,* Vol. 10, pp. 5–15.

Smith, M. J., Colligan, M., and Tasto, D. (1982), "Shift Work Health Effects for Food Processing Workers," *Ergonomics,* Vol. 25, pp. 133–144.

Smith, M., Anger, W. K., Hopkins, B., and Conrad, R. (1983), "Behavioral-Psychological Approaches for Controlling Employee Chemical Exposures," in *Proceedings of the Tenth World Congress on the Prevention of Occupational Accidents and Diseases,* International Social Security Association, Geneva.

Smith, M. J., Karsh, B.-T., and Moro, F. B. (1999), "A Review of Research on Interventions to Control Musculoskeletal Disorders," in J. Suokas and V. Rouhiainen, Eds., *Work-Related Musculoskeletal Disorders,* National Academy Press, Washington, DC, pp. 200–229.

Smith, T. J. (1999), "Synergism of Ergonomics, Safety and Quality: A Behavioral Cybernetic Analysis," *International Journal of Occupational Safety and Ergonomics,* Vol. 5, No. 2, pp. 247–278.

United States Department of Health and Human Services (HHS) (1989), *Promoting Health/Preventing Disease: Year 2000 Objectives for the Nation,* HHS, Washington, DC.

Wettberg, W. (1999), "Health and Environmental Conservation: Integration into a Quality Management System of a Building Serve Contract Company," in *Proceedings of the International Conference on TQM and Human Factors—Towards Successful Integration,* Vol. 2, J. Axelsson, B. Bergman, and J. Eklund, Eds., Centre for Studies of Humans, Technology and Organization, Linkoping, Sweden, pp. 108–113.

Zimolong, B. (1997), "Occupational Risk Management," in *Handbook of Human Factors and Ergonomics,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 989–1020.

Zink, K. (1999), "Human Factors and Business Excellence," in *Proceedings of the International Conference on TQM and Human Factors—Towards Successful Integration,* Vol. 1, J. Axelsson, B. Bergman, and J. Eklund, Eds., Centre for Studies of Humans, Technology and Organization, Linkoping, Sweden, pp. 9–27.

ADDITIONAL READING

American National Standards Institute (ANSI), "Method of Recording Basic Facts Relating to the Nature and Occurrence of Work Injuries," Z16.2-1969, ANSI, New York, 1969.

American National Standards Institute (ANSI), "Safety Color Code For Marking Physical Hazards," Z53.1-1971, ANSI, New York, 1971.

American National Standards Institute (ANSI), "Specifications for Accident Prevention Signs," Z35.1-1972, ANSI, New York, 1972.

American National Standards Institute (ANSI), "Specifications for Informational Signs Complementary to ANSI Z35.1-1972, Accident Prevention Signs," Z35.4-1973, ANSI, New York, 1973.

American National Standards Institute (ANSI), "Method of Recording and Measuring Work Injury Experience," Z16.1-1973, ANSI, New York, 1973.

American National Standards Institute (ANSI), "Office Lighting," A132.1-1973, ANSI, New York, 1973.

American National Standards Institute (ANSI), "Product Safety Signs and Labels," Z535.4-1991, ANSI, New York, 1991.

American National Standards Institute (ANSI), "Environmental and Facility Safety Signs," Z535.2-1991, ANSI, New York, 1991.

American National Standards Institute (ANSI), "Criteria for Safety Symbols," Z535.3-1991, ANSI, New York, 1991.

Centers for Disease Control (CDC), "Noise Induced Loss of Hearing," *Morbidity and Mortality Weekly Report,* Vol. 35, 1986, pp. 185–188.

Cooper, C. L., and Marshall, J., "Occupational Sources of Stress: A Review of the Literature Relating to Coronary Heart Disease and Mental Ill Health," *Journal of Occupational Psychology,* Vol. 49, 1976, pp. 11–28.

Gyllenhammar, P. G., *People at Work,* Addison-Wesley, Reading, MA, 1977.

Suokas, J., "Quality of Safety Analysis," in *Quality Management of Safety and Risk Analysis,* Elsevier Science, Amsterdam, pp. 25–43, 1993.

APPENDIX
Useful Web Information Sources

American Association of Occupational Health Nurses, http://www.aaohn.org
American Board of Industrial Hygiene, http://www.abih.org

American College of Occupational and Environmental Medicine, http://www.acoem.org
American Council of Government Industrial Hygienists, http://www.acgih.org
American Industrial Hygiene Association, http://www.aiha.org
American Psychological Association, http://www.apa.org
Bureau of Labor Statistics, http://www.bls.gov
Centers for Disease Control and Prevention, http://www.cdc.gov
Department of Justice, http://www.usdoj.gov
Department of Labor, http://www.dol.gov
National Institute for Environmental Health Sciences, http://www.niehs.gov
National Institute for Occupational Safety and Health, http://www.niosh.gov
National Safety Council, http://www.nsc.org
Occupational Safety and Health Administration, http://www.osha.gov

CHAPTER 44

Human–Computer Interaction

Kay M. Stanney
University of Central Florida
Michael J. Smith
University of Wisconsin-Madison
Pascale Carayon
University of Wisconsin-Madison
Gavriel Salvendy
Purdue University

1. OVERVIEW **1193**

2. ERGONOMICS **1194**

- 2.1. Components of the Work System 1194
- 2.2. Critical Ergonomics Issues in Human–Computer Interaction 1195
- 2.3. Ergonomics of Computer Interfaces 1195
  - 2.3.1. The Screen and Viewing 1195
  - 2.3.2. Screen Character Features 1196
  - 2.3.3. Viewing Distance 1197
  - 2.3.4. Screen Flicker and Image Stability 1197
  - 2.3.5. Screen Swivel and Tilt 1197
- 2.4. The Visual Environment 1198
  - 2.4.1. Lighting 1198
  - 2.4.2. Illumination 1198
  - 2.4.3. Luminance 1199
  - 2.4.4. Glare 1199
- 2.5. The Auditory Environment 1200
  - 2.5.1. Noise 1200
  - 2.5.2. Heating, Ventilating, and Air Conditioning (HVAC) 1200
- 2.6. Computer Interfaces 1201
  - 2.6.1. The Keyboard 1201
  - 2.6.2. Accessories 1202
  - 2.6.3. The Mouse 1202
- 2.7. The Workstation 1202
  - 2.7.1. Working Surfaces 1202
  - 2.7.2. The Chair 1204
  - 2.7.3. Other Workstation Considerations 1204
- 2.8. Work Practices 1205
  - 2.8.1. Work Breaks 1205

3. COGNITIVE DESIGN **1205**

- 3.1. Overview 1205
- 3.2. Requirements Definition 1206
- 3.3. Contextual Task Analysis 1206
  - 3.3.1. Background Information 1206
  - 3.3.2. Characterizing Users 1207
  - 3.3.3. Collecting and Analyzing Data 1208
  - 3.3.4. Constructing Models of Work Practices 1210
  - 3.3.5. Task Allocation 1210
- 3.4. Competitive Analysis and Usability Goal Setting 1212
- 3.5. User Interface Design 1212
  - 3.5.1. Initial Design Definition 1212
  - 3.5.2. Detailed Design 1213
  - 3.5.3. Prototyping 1216
- 3.6. Usability Evaluation of Human–Computer Interaction 1216

4. SOCIAL, ORGANIZATIONAL, AND MANAGEMENT FACTORS **1217**

- 4.1. Social Environment 1217
- 4.2. Organizational Factors 1222
- 4.3. Management Factors 1225
- 4.4. An International Perspective 1228

5. ITERATIVE DESIGN **1228**

REFERENCES **1230**

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. OVERVIEW

The utilities of information technology are spreading into all walks of life, from the use of self-standing personal computers and networking to Internet and intranet. This technology has allowed for tremendous growth in Web-based collaboration and commerce and has expanded into information appliances (e.g., pagers, cellular phones, two-way radios) and other consumer products. It is important that these interactive systems be designed so that they are easy to learn and easy to operate, with minimal errors and health consequences and maximal speed and satisfaction. Yet it can be challenging to achieve an effective design that meets these criteria.

The design of interactive systems has evolved through several paradigm shifts. Initially, designers focused on functionality. The more a system could do, the better the system was deemed to be. This resulted in system designs whose functionality often could not be readily accessed or utilized, or tended to physically stress users (Norman 1988). For example, how many homes have you walked into where the VCR is flashing 12:00? This example shows that even devices that should be simple to configure can be designed in such a manner that users cannot readily comprehend their use. Further, the occurrence of repetitive strain injuries rose as users interacted with systems that engendered significant physical stress. The development of such systems led to a shift in design focus from functionality to usability. Usability engineering (Nielsen 1993) focuses on developing interactive systems that are ergonomically suitable for the users they support (Grandjean 1979; Smith 1984), as well as cognitively appropriate (Vicente 1999). This approach aims to ensure the ease of learning, ease of use, subjective satisfaction, and physical comfort of interactive systems. While these design goals are appropriate and have the potential to engender systems that are effective and efficient to use, system designers have found that this focus on usability does not always lead to the most user-acceptable system designs. In recent years, environmental concerns (i.e., social, organizational, and management factors) have led to design practices that incorporate a greater emphasis on studying and understanding the semantics of work environments (Vicente 1999), often through ethnographic approaches (Nardi 1997; Takahashi 1998). Through participant-observation practices, efforts are made to understand more completely the tasks, work practices, artifacts, and environment that the system will become a part of (Stanney et al. 1997). This is often achieved by designers immersing themselves in the target work environment, thereby becoming accustomed to and familiar with the various factors of interactive system design. These factors include users' capabilities and limitations (both cognitive and physical), organizational factors (e.g., management and social issues), task requirements, and environmental conditions that the work environment supports (see Figure 1). Through the familiarity gained by this involvement, designers can develop systems that are more uniquely suited to target users and the organizations for which they work.

This chapter provides guidelines and data on how to achieve these objectives through the effective design of human–computer interaction, which takes into account the human's physical, cognitive, and social abilities and limitations in reference to interacting with computers and/or computer based appliances. In doing so, it relies on the available standards, practices, and research findings. Much of it is guided by currently available technology but may also be applicable as technology changes and new applications evolve.

The overall thrust of the chapter is that good physical design of the workplace will minimize the probabilities of the occurrence of health consequences; good cognitive design will maximize the utility of interactive systems; and good social and organizational design will effectively integrate these systems into existing work domains. In general, it is suggested that human–computer interaction will be optimized when the following are observed:

- The system design is ergonomically suited to the user.
- Interactive design matches the mental models of users.
- Only information needed for decision making is presented.
- Information of a similar nature is chunked together.
- The interface is adaptive to individual differences due to innate, acquired, or circumstantial reasons.
- The system design supports existing work practices and related artifacts.

Interactive system design is thus about many interfaces; it considers how users relate to each other, how they physically and cognitively interact with systems, how they inhabit their organizations, and how these interfaces can best be supported by mediating technologies. Focusing on each of these areas highlights the need for a multidisciplinary interactive system design team. As Duffy and Salvendy (1999) have documented, in teams that consist of design and manufacturing engineers, marketing specialists, and a team leader, even when they have common goals, each member retrieves and uses different information and has a different mental model that focuses on unique aspects in achieving the same design objectives.

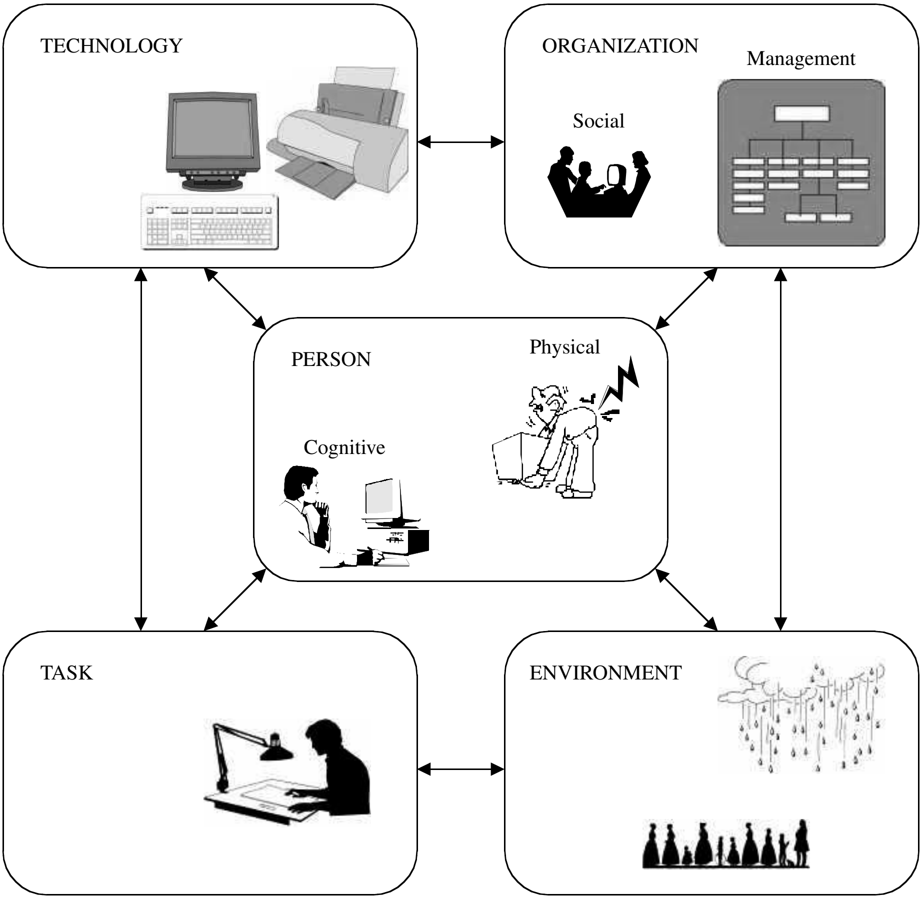


Figure 1 Model of the Work System. (Adapted from Smith and Sainfort 1989)

The following sections will focus on different aspects of interactive system design, including ergonomics, cognitive design, and social, organizational, and management factors.

2. ERGONOMICS

Ergonomics is the science of fitting the environment and activities to the capabilities, dimensions, and needs of people. Ergonomic knowledge and principles are applied to adapt working conditions to the physical, psychological, and social nature of the person. The goal of ergonomics is to improve performance while at the same time enhancing comfort, health, and safety. In particular, the efficiency of human–computer interaction, as well as the comfort, health, and safety of users, can be improved by applying ergonomic principles (Grandjean 1979; Smith 1984). However, no simple recommendations can be followed that will enhance all of these aspects simultaneously. Compromise is sometimes necessary to achieve a set of balanced objectives while ensuring user health and safety (Smith and Sainfort 1989; Smith and Cohen 1997). While no one set of rules can specify all of the necessary combinations of proper working conditions, the use of ergonomic principles and concepts can help in making the right choices.

2.1. Components of the Work System

From an ergonomic point of view, the different components of the work system (e.g., the environment, technology, work tasks, work organization, and people) interact dynamically with each other and function as a total system (see Figure 1). Since changing any one component of the system influences the other aspects of the system, the objective of ergonomics is to *optimize the whole system* rather than maximize just one component. In an ergonomic approach, the person is the central focus and the other factors of the work system are designed to help the person be effective, motivated, and comfortable. The consideration of physical, physiological, psychological, and social needs of the person is necessary to ensure the best possible workplace design for productive and healthy human–computer interaction. Table 1 shows ergonomic recommendations for fixed desktop video display terminal (VDT) use that improve the human interface characteristics. Ergonomic conditions for laptop computer use should conform as closely as possible to the recommendations presented in Table 1.

2.2. Critical Ergonomics Issues in Human–Computer Interaction

A major feature of the ergonomics approach is that *the job task characteristics will define the ergonomic interventions* and the priorities managers should establish for workplace design requirements. The following discussion of critical areas—the technology, sensory environment, thermal environment, workstation design, and work practices—will highlight the major factors that engineers and managers should be aware of in order to optimize human–computer interaction and protect user health. Specific recommendations and guidelines will be derived from these discussions, but please be advised that the recommendations made throughout this chapter may have to be modified to account for differences in technology, personal, situational, or organizational needs at your facility, as well as improved knowledge about human–computer interaction. It cannot be overstated that these considerations represent recommendations and guidelines and not fixed specifications or standards. The realization that any one modification in any single part of the work system will affect the whole system and particularly the person (see Figure 1) is essential for properly applying the following recommendations and specifications.

2.3. Ergonomics of Computer Interfaces

Today, the primary display interfaces in human–computer interaction are the video display with a cathode ray tube and the flat panel screen. In the early 1980s, the US Centers for Disease Control (CDC 1980) and the U.S. National Academy of Sciences defined important design considerations for the use of cathode ray tubes (NAS 1983). The Japan Ergonomics Society (JES) established a Committee for Flat Panel Display Ergonomics in 1996, which proposed ergonomic guidelines for use of products with flat panels, such as liquid crystal displays (LCDs) (JES 1996). These Japanese guidelines were subsequently reviewed by the Committee on Human–Computer Interaction of the International Ergonomics Association (IEA). The JES guidelines addressed the following issues: (1) light-related environmental factors, (2) device use and posture factors, (3) environmental factors, (4) job design factors, and (5) individual user factors. These guidelines will be discussed in appropriate sections of this chapter.

The use of CRTs and flat panel displays has been accompanied by user complaints of visual fatigue, eye soreness, general visual discomfort, and various musculoskeletal complaints and discomfort with prolonged use (Grandjean 1979; Smith et al. 1981; NIOSH 1981; NAS 1983; Smith 1984; JES 1996). Guidelines for providing the proper design of the VDT and the environment in which it is used have been proposed by the Centers for Disease Control (CDC 1980) and the Human Factors and Ergonomics Society (ANSI 1988), and for the laptop and palm computer, by the Japan Ergonomics Society (JES 1996). The following sections deal with the visual environment for using desktop computers, but the discussion can be extrapolated to other types of computer use.

The major interfaces of employees with computers are the screen (CRT, flat panel), the keyboard, and the mouse. Other interfaces are being used more and more, such as voice input, pointers, hand-actuated motion devices, and apparatuses for virtual environment immersion.

2.3.1. The Screen and Viewing

Poor screen images, fluctuating and flickering screen luminances, and screen glare cause user visual discomfort and fatigue (Grandjean 1979; NAS 1983). There are a range of issues concerning readability and screen reflections. One is the adequacy of contrast between the characters and screen background. Screens with glass surfaces have a tendency to pick up glare sources in the environment and reflect them. This can diminish the contrast of images on the screen. To reduce environmental glare, the luminance ratio within the user's near field of vision should be approximately 1:3, and within the far field approximately 1:10 (NIOSH 1981). For luminance on the screen itself, the character-to-screen background luminance contrast ratio should be at least 7:1 (NIOSH 1981). To give the best readability for each operator, it is important to provide VDTs with adjustments for character contrast and brightness. These adjustments should have controls that are obvious to observe and manipulate and easily accessible from normal working position (e.g., located at the front of the screen) (NIOSH 1981).

TABLE 1 Ergonomic Recommendations for the VDT Technology, Work Environment, and Workstation

Ergonomic Consideration	Recommendation
1. Viewing screen	
a. Character/screen contrast	7:1 minimum
b. Screen character size	height = 20–22 min of visual arc width = 70–80% of height
c. Viewing distance	Usually 50 cm or less, but up to 70 cm is acceptable
d. Line refresh rate	70 hz minimum
e. Eye viewing angle from horizon	10–40° (from top to bottom gaze)
2. Illumination	
a. No hardcopy	300 lux minimum
b. With normal hard copy	500 lux
c. With poor hard copy	700 lux
d. Environmental luminance contrast	
• Near objects	1:3
• Far objects	1:10
e. Reflectance from surfaces	
• Working surface	40–60%
• Floor	30%
• Ceiling	80–90%
• Walls	40–60%
3. HVAC	
a. Temperature—winter	20–24°C (68–75°F)
b. Temperature—summer	23–27°C (73–81°F)
c. Humidity	50–60%
d. Airflow	0.15–0.25 m/sec
4. Keyboard	
a. Slope	0–15° preferred, 0–25° acceptable
b. Key top area	200 mm^2
c. Key top horizontal width	12 mm (minimum)
d. Horizontal key spacing	18–19 mm
e. Vertical key spacing	18–20 mm
f. Key force	0.25N–1.5N (0.5–0.6N preferred)
5. Workstation	
a. Leg clearance	51 cm minimum (61 cm preferred minimum)
b. Leg depth	38 cm minimum
c. Leg depth with leg extension	59 cm minimum
d. Work surface height—nonadjustable	70 cm
e. Work surface height—adjustable for one surface	70–80 cm
f. Work surface height—adjustable for two surfaces	Keyboard surface 59–71 cm Screen surface 70–80 cm
6. Chair	
a. Seat pan width	45 cm minimum
b. Seat pan depth	38–43 cm
c. Seat front tilt	5° forward to 7° backward
d. Seat back inclination	110–130°
e. Seat pan height adjustment range	38–52 cm
f. Backrest inclination	Up to 130°
g. Backrest height	45–51 cm above seat pan surface

2.3.2. Screen Character Features

Good character design can help improve image quality, which is a major factor for reducing eyestrain and visual fatigue. The proper size of a character is dependent on the task and the display parameters (brightness, contrast, glare treatment, etc.) and the viewing distance. Character size that is too small can make reading difficult and cause the visual focusing mechanism to overwork. This produces eyestrain and visual fatigue (NAS 1983). Character heights should preferably be at least 20–22 min of visual arc, while character width should be 70–80% of the character height (Smith 1984; ANSI 1988). This approximately translates into a minimum lowercase character height of 3.5 mm with a width of 2.5 mm at a normal viewing distance of 50 cm.

Good character design and proper horizontal and vertical spacing of characters can help improve image quality. To ensure adequate discrimination between characters and good screen readability, the character spacing should be in the range of 20–50% of the character width. The interline spacing should be 50–100% of the character height (Smith 1984; ANSI 1988).

The design of the characters influences their readability. Some characters are hard to decipher, such as lowercase *g*, which looks like numeral *8*. A good font design minimizes character confusion and enhances the speed at which characters can be distinguished and read. Two excellent fonts are Huddleston and Lincoln-Mitre (NAS 1983). Most computers have a large number of fonts to select from. Computer users should choose a font that is large enough to be easy for them to read.

2.3.3. Viewing Distance

Experts have traditionally recommended a viewing distance between the screen and the operator's eye of 45–50 cm but no more than 70 cm (Grandjean 1979; Smith 1984). However, experience in field studies has shown that users may adopt a viewing distance greater than 70 cm and are still able to work efficiently and not develop visual problems. Thus, viewing distance should be determined in context with other considerations. It will vary depending on the task requirements, CRT screen characteristics, and individual's visual capabilities. For instance, with poor screen or hard copy quality, it may be necessary to reduce viewing distances for easier character recognition. Typically, the viewing distance should be 50 cm or less due to the small size of characters on the VDT screen. LNCs are often used in situations where the computer is placed on any convenient surface, for example a table at the airport waiting room. Thus, the viewing distance is defined by the available surface, not a fixed workstation. When the surface is farther away from the eyes, the font size used should be larger.

Proper viewing distance will be affected by the condition of visual capacity and by the wearing of spectacles/lenses. Persons with myopia (near-sightedness) may find that they want to move the screen closer to their eyes; while persons with presbyopia (far-sightedness) or bifocal lenses may want the screen farther away from their eyes. Many computer users who wear spectacles have a special pair of spectacles with lenses that are matched to their particular visual defect and a comfortable viewing distance to the screen. Eyecare specialist can have special spectacles made to meet computer users' screen use needs.

2.3.4. Screen Flicker and Image Stability

The stability of the screen image is another characteristic that contributes to CRT and LCD quality. Ideally, the display should be completely free of perceptible movements such as flicker or jitter (NIOSH 1981). CRT screens are refreshed a number of times each second so that the characters on the screen appear to be solid images. When this refresh rate is too low, users perceive screen flicker. LCDs have less difficulty with flicker and image stability than CRT displays. The perceptibility of screen flicker depends on illumination, screen brightness, polarity, contrast, and individual sensitivity. For instance, as we get older and our visual acuity diminishes, so too does our ability to detect flicker. A screen with a dark background and light characters has less flicker than screens with dark lettering on a light background. However, light characters on a dark background show more glare. In practice, flicker should not be observable, and to achieve this a screen refresh rate of at least 70 cycles per second needs to be achieved for each line on the CRT screen (NAS 1983; ANSI 1988). With such a refresh rate, flicker should not be a problem for either screen polarity (light on dark or dark on light). It is a good idea to test a screen for image stability. Turn the lights down, increase the screen brightness/contrast settings, and fill the screen with letters. Flickering of the entire screen or jitter of individual characters should not be perceptible, even when viewed peripherally.

2.3.5. Screen Swivel and Tilt

Reorientation of the screen around its vertical and horizontal axes can reduce screen reflections and glare. Reflections can be reduced by simply tilting the display slightly back or down or to the left or right, depending on the angle of the source of glare. These adjustments are easiest if the screen can be tilted about its vertical and horizontal axes. If the screen cannot be tilted, it should be approximately vertical to help eliminate overhead reflections, thus improving legibility and posture.

The perception of screen reflection is influenced by the tilt of the screen up or down and back and forth and by the computer user's line of sight toward the screen. If the screen is tilted toward sources of glare and these are in the computer user's line of sight to the screen, the screen images will have poorer clarity and reflections can produce disability glare (see Section 2.4.4). In fact, the line of sight can be a critical factor in visual and musculoskeletal discomfort symptoms. When the line of sight can observe glare or reflections, then eyestrain often occurs. For musculoskeletal considerations, experts agree that the line of sight should never exceed the straight-ahead horizontal gaze, and in fact it is best to provide a downward gaze of about 10–20° from the horizontal when viewing the top of the screen and about 40° when viewing the bottom edge of the screen (NIOSH 1981; NAS 1983; Smith 1984; ANSI 1988). This will help reduce neck and shoulder fatigue and pain. These gaze considerations are much harder to obtain when using LNCs because of the smaller screen size and workstation features (eg., airport waiting room table).

2.4. The Visual Environment

2.4.1. Lighting

Lighting is an important aspect of the visual environment that influences readability and glare on the screen and viewing in the general environment. There are four types of general workplace illumination of interest to the computer user's environment:

1. *Direct radiants:* The majority of office lighting is direct radiants. These can be incandescent lights, which are most common in homes, or fluorescent lighting, which is more prevalent in workplaces and stores. Direct radiants direct 90% or more of their light toward the object(s) to be illuminated in the form of a cone of light. They have a tendency to produce glare.
2. *Indirect lighting:* This approach uses reflected light to illuminate work areas. Indirect lighting directs 90% or more of the light onto the ceiling and walls, which reflect it back into the room. Indirect lighting has the advantage of reducing glare, but supplemental lighting is often necessary, which can be a source of glare.
3. *Mixed direct radiants and indirect lighting:* In this approach, part of the light (about 40%) radiates in all directions while the rest is thrown directly or indirectly onto the ceiling and walls.
4. *Opalescent globes:* These lights give illumination equally in all directions. Because they are bright, they often cause glare.

Modern light sources used in these four general approaches to workplace illumination are typically of two kinds: electric filament lamps and fluorescent tubes. Following are the advantages and drawbacks of these two light sources:

1. *Filament lamps:* The light from filament lamps is relatively rich in red and yellow rays. It changes the apparent colors of objects and so is unsuitable when correct assessment of color is essential. Filament lamps have the further drawback of emitting heat. On the other hand, employees like their warm glow, which is associated with evening light and a cozy atmosphere.
2. *Fluorescent tubes:* Fluorescent lighting is produced by passing electricity through a gas. Fluorescent tubes usually have a low luminance and thus are less of a source of glare. They also have the ability to match their lighting spectrum to daylight, which many employees find preferable. They may also be matched to other spectrums of light that can fit office decor or employee preferences. Standard-spectrum fluorescent tubes are often perceived as a cold, pale light and may create an unfriendly atmosphere. Fluorescent tubes may produce flicker, especially when they become old or defective.

2.4.2. Illumination

The intensity of illumination or the illuminance being measured is the amount of light falling on a surface. It is a measure of the quantity of light with which a given surface is illuminated and is measured in lux. In practice, this level depends on both the direction of flow of the light and the spatial position of the surface being illuminated in relation to the light flow. Illuminance is measured in both the horizontal and vertical planes. At computer workplaces, both the horizontal and vertical illuminances are important. A document lying on a desk is illuminated by the horizontal illuminance, whereas the computer screen is illuminated by the vertical illuminance. In an office that is illuminated from overhead luminaires, the ratio between the horizontal and vertical illuminances is usually between 0.3 and 0.5. So if the illuminance in a room is said to be 500 lux, the horizontal illuminance is 500 lux while the vertical illuminance is between 150 and 250 lux (0.3 and 0.5 of the horizontal illuminance).

The illumination required for a particular task is determined by the visual requirements of the task and the visual ability of the employees concerned. In general, an illuminance in the range of 300–700 lux measured on the horizontal working surface (not the computer screen) is normally preferable (CDC 1980; NAS 1983). The JES (1996) recommends office lighting levels ranging from 300–1,000 lux for flat panel displays. Higher illumination levels are necessary to read hard copy and lower illumination levels are better for work that just uses the computer screen. Thus, a job in which hard copy and a computer screen are both used should have a general work area illumination level of about 500–700 lux, while a job that only requires reading the computer screen should have a general work area illumination of 300–500 lux. Conflicts can arise when both hardcopy and computer screens are used by different employees who have differing job task requirements or differing visual capabilities and are working in the same area. As a compromise, room lighting can be set at the recommended lower (300 lux) or intermediate level (500 lux) and additional task lighting can be provided as needed. Task lighting refers to localized lighting at the workstation to replace or supplement ambient lighting systems used for more generalized lighting of the workplace. Task lighting is handy for illuminating hardcopy when the room lighting is set at a low level, which can hinder document visibility. Such additional lighting must be carefully shielded and properly placed to avoid glare and reflections on the computer screens and other adjacent working surfaces of other employees. Furthermore, task lighting should not be too bright in comparison to the general work area lighting since looking between these two different light levels may produce eyestrain.

2.4.3. Luminance

Luminance is a measure of the brightness of a surface, that is, the amount of light leaving the surface of an object, either reflected by the surface (as from a wall or ceiling), emitted by the surface (as from the CRT or LCD characters), or transmitted (as light from the sun that passes through translucent curtains). Luminance is expressed in units of candelas per square meter. High-intensity luminance sources (such as windows) in the peripheral field of view should be avoided. In addition, the balance among the luminance levels within the computer user's field of view should be maintained. The ratio of the luminance of a given surface or object to another surface or object in the central field of vision should be around 3:1, while the luminance ratio in the peripheral field of vision can be as high as 10:1 (NAS 1983).

2.4.4. Glare

Large differences in luminance or high-luminance lighting sources can cause glare. Glare can be classified with respect to its effects (disability glare vs. discomfort glare) or the source of glare (direct glare vs. reflected glare). Glare that results in an impairment of vision (e.g., reduction of visual acuity) is called disability glare, while discomfort glare is experienced as a source of discomfort to the viewer but does not necessarily interfere with visual performance. With regard to the source, direct glare is caused by light sources in the field of view of the computer user, while reflected glare is caused by reflections from illuminated, polished, or glossy surfaces or by large luminance differences in the visual environment. In general, glare is likely to increase with the luminance, size, and proximity of the lighting source to the line of sight.

Direct and reflected glare can be limited through one or more of the following techniques:

1. Controlling the light from windows: This can be accomplished by closing drapes, shades, and/or blinds over windows or awnings on the outside, especially during sunlight conditions.
2. Controlling the view of luminaires:
   - **(a)** By proper positioning of CRT screen with regard to windows and overhead lighting to reduce direct or reflected glare and images. To accomplish this, place VDTs parallel to windows and luminaires and between luminaires rather than underneath them.
   - **(b)** Using screen hoods to block luminaires from view.
   - **(c)** Recessing light fixtures.
   - **(d)** Using light-focusing diffusers.
3. Controlling glare at the screen surface by:
   - **(a)** Adding antiglare filters on the VDT screen.
   - **(b)** Proper adjustment up or down/left or right of the screen.
4. Controlling the lighting sources using:
   - **(a)** Appropriate glare shields or covers on the lamps.
   - **(b)** Properly installed indirect lighting systems.

Glare can also be caused by reflections from surfaces, such as working surfaces, walls, or the floor covering. These surfaces do not emit light themselves but can reflect it. The ratio of the amount of light reflected by a surface (luminance) to the amount of light striking the surface (illuminance) is called reflectance. Reflectance is unitless. The reflectance of the working surface and the office machines should be on the order of 40–60% (ANSI 1988). That is, they should not reflect more than 60% of the illuminance striking their surface. This can be accomplished if surfaces have a matte finish.

Generally, floor coverings should have a reflectance of about 30%, ceilings, of 80–90%, and walls, 40–60%. Reflectance should increase from the floor to the ceiling. Although the control of surface reflections is important, especially with regard to glare control, it should not be at the expense of a pleasant working environment where employees feel comfortable. Walls and ceilings should not be painted dark colors just to reduce light reflectance, nor should windows be completely covered or bricked up to keep out sunlight. Other, more reasonable luminance control approaches can give positive benefits while maintaining a psychologically pleasing work environment.

2.5. The Auditory Environment

2.5.1. Noise

A major advantage of computer technology over the typewriter is less noise at the workstation. However, it is not unusual for computer users to complain of bothersome office noise, particularly from office conversation. Noise levels commonly encountered in offices are below established limits that could cause damage to hearing (i.e., below 85 dBA). The JES (1996) proposed that the noise level should not exceed 55 dBA. The expectations of office employees are for quiet work areas because their tasks often require concentration. Annoying noise can disrupt their ability to concentrate and may produce stress.

Actually, there are many sources of annoyance noise in computer operations. Fans in computers, printers, and other accessories, which are used to maintain a favorable internal device temperature, are a source of noise. Office ventilation fans can also be a source of annoyance noise. The computers themselves may be a source of noise (e.g., the click of keys or the high-pitched squeal of the CRT). The peripheral equipment associated with computers, such as printers, can be a source of noise. Problems of noise may be exacerbated in open-plan offices, in which noise is harder for the individual employee to control than in enclosed offices.

Acoustical control can rely upon ceiling, floor and wall, furniture, and equipment materials that absorb sound rather than reflect it. Ceilings that scatter, absorb, and minimize the reflection of sound waves are desirable to promote speech privacy and reduce general office noise levels. The most common means of blocking a sound path is to build a wall between the source and the receiver. Walls are not only sound barriers but are also a place to mount sound-absorbent materials. In open-plan offices, free-standing acoustical panels can be used to reduce the ambient noise level and also to separate an individual from the noise source. Full effectiveness of acoustical panels is achieved in concert with the sound-absorbent materials and finishes applied to the walls, ceiling, floor, and other surfaces. For instance, carpets not only cover the floor but also serve to reduce noise. This is achieved in two ways: (1) carpets absorb the incident sound energy and (2) gliding and shuffling movements on carpets produce less noise than on bare floors. Furniture and draperies are also important for noise reduction.

Acoustical control can also be achieved by proper space planning. For instance, workstations that are positioned too closely do not provide suitable speech privacy and can be a source of disturbing conversational noise. As a general rule, a minimum of 8–10 ft between employees, separated by acoustical panels or partitions, will provide normal speech privacy.

2.5.2. Heating, Ventilating, and Air Conditioning (HVAC)

Temperature, humidity, air flow, and air exchanges are important parameters for employees' performance and comfort.

It is unlikely that offices will produce excessive temperatures that could be physically harmful to employees. However, thermal comfort is an important consideration in employee satisfaction that can influence performance. Satisfaction is based not on the ability to tolerate extremes but on what makes an individual happy. Many studies have shown that most office employees are not satisfied with their thermal comfort. The definition of a comfortable temperature is usually a matter of personal preference. Opinions as to what is a comfortable temperature vary within an individual from time to time and certainly among individuals. Seasonal variations of ambient temperature influence perceptions of thermal comfort. Office employees sitting close to a window may experience the temperature as being too cold or hot, depending on the outside weather. It is virtually impossible to generate one room temperature in which all employees are equally well satisfied over a long period of time.

As a general rule, it is recommended that the temperature be maintained in the range of 20–24°C (68–75°F) in winter and 23–27°C (73–81°F) in summer (NIOSH 1981; Smith 1984). The JES (1996) recommends office temperatures of 20–23°C in winter and 24–27°C in summer.

Air flows across a person's neck, head, shoulders, arms, ankles, and knees should be kept low (below 0.15 m/sec in winter and below 0.25 m/sec in summer). It is important that ventilation not produce currents of air that blow directly on employees. This is best handled by proper placement of the workstation.

Relative humidity is an important component of office climate and influences an employee's comfort and well being. Air that is too dry leads to drying out of the mucous membranes of the eyes, nose, and throat. Individuals who wear contact lenses may be made especially uncomfortable by dry air. In instances where intense, continuous near-vision work at the computer is required, very dry air has been shown to irritate the eyes. As a general rule, it is recommended that the relative humidity in office environments be at least 50% and less than 60% (NIOSH 1981; Smith 1984). The JES (1996) recommends humidity levels of 50–60%. Air that is too wet enhances the growth of unhealthy organisms (molds, fungus, bacteria) that can cause disease (legionnaires', allergies).

2.6. Computer Interfaces

Computer interfaces are the means by which users provide instructions to the computer. There are a wide variety of devices for interfacing, including keyboards, mice, trackballs, joy sticks, touch panels, light pens, pointers, tablets, and hand gloves. Any mechanical or electronic device that can be tied to a human motion can serve as a computer interface. The most common interfaces in use today are the keyboard and the mouse. The keyboard will be used as an example to illustrate how to achieve proper human–computer interfaces.

2.6.1. The Keyboard

In terms of computer interface design, a number of keyboard features can influence an employee's comfort, health, and performance. The keyboard should be detachable and movable, thus providing flexibility for independent positioning of the keyboard and screen. This is a major problem with LNCs because the keyboard is built into the top of the computer case for portability and convenience. It is possible to attach a separate, detachable keyboard to the LNC, and this should be done when the LNC is used at a fixed workstation in an office or at home. Clearly, it would be difficult to have a separate keyboard when travelling and the LNC portability feature is paramount.

The keyboard should be stable to ensure that it does not slide on the tabletop. This is a problem when an LNC is held in the user's lap or some other unstable surface. In order to help achieve a favorable user arm height positioning, the keyboard should be as thin as possible. The slope or angle of the keyboard should be between 0° and 15°, measured from the horizontal. LNCs are limited in keyboard angle because the keyboard is often flat (0°). However, some LNCs have added feet to the computer case to provide an opportunity to increase the keyboard angle. Adjustability of keyboard angle is recommended. While the ANSI standard (ANSI 1988) suggests 0–25°, we feel angles over 15° are not necessary for most activities.

The shape of the key tops must satisfy several ergonomic requirements, such as minimizing reflections, aiding the accurate location of the operator's finger, providing a suitable surface for the key legends, preventing the accumulation of dust, and being neither sharp nor uncomfortable when depressed. For instance, the surface of the key tops, as well as the keyboard itself, should have a matte finish. The key tops should be approximately 200 mm (ANSI 1988) with a minimum horizontal width of 12 mm. The spacing between the key centers should be about 18–19 mm horizontally and 18–20 mm vertically (ANSI 1988). There should be slight protrusions on select keys on the home row to provide tactile information about finger position on the keyboard.

The force to depress the key should ideally be between 0.5 N and 0.6 N (ANSI 1988). However, ranges from 0.25–1.5 N have been deemed acceptable (ANSI 1988). The HFES/ANSI-100 standard is currently being revised, and this recommendation may change soon. Some experts feel that the keying forces should be as low as feasible without interfering with motor coordination. Research has shown that light-touch keys require less operator force in depressing the key (Rempel and Gerson, 1991; Armstrong et al. 1994; Gerard et al. 1996). The light-touch force keyboards vary between 0.25–0.40 N.

Feedback from typing is important for beginning typists because it can indicate to the operator that the keystroke has been successfully completed. There are two main types of keyboard feedback: tactile and auditory. Tactile feedback can be provided by a collapsing spring that increases in tension as the key is depressed or by a snap-action mechanism when key actuation occurs. Auditory feedback (e.g., "click" or "beep") can indicate that the key has been actuated. Of course, there is also visual feedback on the computer screen. For experienced typists, the feedback is not useful, as their fingers are moving in a ballistic way that is too fast for the feedback to be useful for modifying finger action (Guggenbuhl and Krueger 1990, 1991; Rempel and Gerson 1991; Rempel et al. 1992).

The keyboard layout can be the same as that of a conventional typewriter, that is, the QWERTY design, or some other proven style, such as the DVORAK layout. However, it can be very difficult for operators to switch between keyboards with different layouts. Traditional keyboard layout has straight rows and staggered columns. Some authors have proposed curving the rows to provide a better fit for the hand to reduce biomechanical loading on the fingers (Kroemer 1972). However, there is no research evidence that such a design provides advantages for operator's performance or health.

Punnett and Bergqvist (1997) have proposed that keyboard design characteristics can lead to upper-extremity musculoskeletal disorders. There is controversy about this contention by Punnett and Bergqvist because there are many factors involved in computer typing jobs independent of the keyboard characteristics that may contribute to musculoskeletal disorders. Some ergonomists have designed alternative keyboards in attempts to reduce the potential risk factors for musculoskeletal disorders (Kroemer 1972; Nakaseko et al. 1985; Ilg 1987). NIOSH (1997) produced a publication that describes various alternative keyboards. Studies have been undertaken to evaluate some of these alternative keyboards (Swanson et al. 1997; Smith et al. 1998). The research results indicated some improvement in hand/wrist posture from using the alternative keyboards, but no decrease in musculoskeletal discomfort.

2.6.2. Accessories

The use of a wrist rest when keying can help to minimize extension (backward bending) of the hand. A wrist rest should have a fairly broad surface (approximately 5 cm) with a rounded front edge to prevent cutting pressures on the wrist and hands. Padding further minimizes skin compression and irritation. Height adjustability is important so that the wrist rest can be set to a preferred level in concert with the keyboard height and slope. Some experts are concerned that resting the wrist on a wrist rest during keying could cause an increase in intercarpal canal pressure. They prefer that wrist rests be used only when the user is not keying for the purpose of resting the hands and wrist. Thus, they believe users need to be instructed (trained) about when and how to use a wrist rest. Arm holders are also available to provide support for the hands, wrists, and arms while keyboarding. However, these may also put pressure on structures that may produce nerve compression. As with a wrist rest, some experts feel these devices are best used only during rest from keying.

2.6.3. The Mouse

The most often-used computer pointing device is the mouse. While there are other pointing devices, such as the joystick, touch panel, trackball, and light pen, the mouse is still the most universally used of these devices. An excellent discussion of these pointing devices can be found in Bullinger et al. (1977). The mouse provides for an integration of both movement of the cursor and action on computer screen objects, simultaneously. Many mice have multiple buttons to allow for several actions to occur in sequence. The ease of motion patterns and multiple-function buttons give the mouse an advantage over other pointing devices. However, a disadvantage of the mouse is the need for tabletop space to achieve the movement function. Trankle and Deutschmann (1991) conducted a study to determine which factors influenced the speed of properly positioning a cursor with a mouse. The results indicated that the most important factors were the target size and the distance traveled. Also of lesser importance was the display size arc. The control/response ratio or the sensitivity of the control to movement was not found to be important. Recently, studies have indicated that operators have reported musculoskeletal discomfort due to mouse use (Karlqvist et al. 1994; Armstrong et al. 1995; Hagberg 1995; Fogelman and Brogmus 1995; Wells et al. 1997).

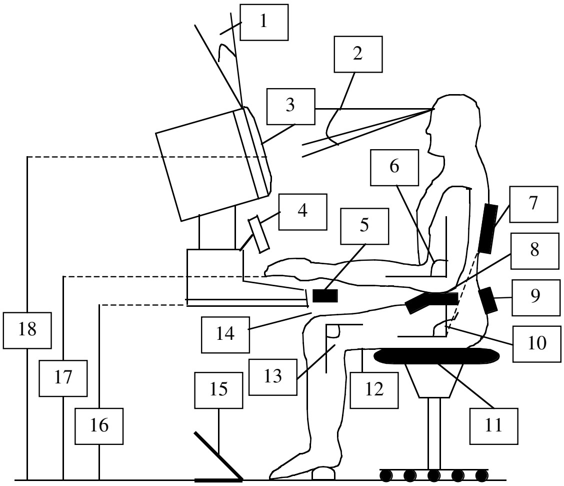
2.7. The Workstation

Workstation design is a major element in ergonomic strategies for improving user comfort and particularly for reducing musculoskeletal problems. Figure 2 illustrates the relationships among the working surface, VDT, chair, documents, and various parts of the body. Of course, this is for a fixed workstation at the office or home. Use of LNCs often occurs away from fixed workstations where it is difficult to meet the requirements described below. However, efforts should be made to meet these requirements as much as possible, even when using LNCs.

The task requirements will determine critical layout characteristics of the workstation. The relative importance of the screen, keyboard, and hard copy (i.e., source documents) depends primarily on the task, and this defines the design considerations necessary to improve operator performance, comfort, and health. Data-entry jobs, for example, are typically hard copy oriented. The operator spends little time looking at the screen, and tasks are characterized by high rates of keying. For this type of task it is logical for the layout to emphasize the keyboard, mouse, and hard copy, because these are the primary tools used in the task, while the screen is of lesser importance. On the other hand, data-acquisition operators spend most of their time looking at the screen and seldom use hard copy. For this type of task, the screen and the keyboard layout should be emphasized.

2.7.1. Working Surfaces

The size of the work surface is dependent on the task(s), documents, and technology. The primary working surface (e.g., supporting the keyboard, display, and documents) should be sufficient to: (1) permit the screen to be moved forward or backward to a comfortable viewing distance for a range



1. Screen tilt angle
2. Visual angle between the horizontal and the center of the display
3. Eye-screen distance
4. Document holder and source document
5. Wrist wrest
6. Elbow angle
7. Backrest
8. Elbow rest
9. Lumbar support
10. Seat back angle (from horizontal)
11. Seat pan angle (from horizontal)
12. Clearance between leg and seat
13. Knee angle
14. Clearance between leg and table
15. Footrest
16. Table height
17. Home row (middle row height)
18. Screen height to center of screen

Figure 2 Definitions of VDT Workstation Terminology. (Adapted from Helander 1982)

of users, (2) allow a detachable keyboard to be placed in several locations, and (3) permit source documents to be properly positioned for easy viewing. Additional working surfaces (i.e., secondary working surfaces) may be required in order to store, lay out, read, and/or write on documents or materials. Often users have more than one computer, so a second computer is placed on a secondary working surface. In such a situation, workstations are configured so that multiple pieces of equipment and source materials can be equally accessible to the user. In this case, additional working surfaces are necessary to support these additional tools and are arranged to allow easy movement while seated from one surface to another.

The tabletop should be as thin as possible to provide clearance for the user's thighs and knees. Moreover, it is important to provide unobstructed room under the working surface for the feet and legs so that users can easily shift their posture. Knee space height and width and leg depth are the three key factors for the design of clearance space under working surfaces (see Figure 2). Recommendations for minimum width for leg clearance is 51 cm, while the preferred minimum width is 61 cm (ANSI, 1988). The minimum depth under the work surface from the operator edge of the work surface should be 38 cm for clearance at the knee level and 60 cm at the toe level (ANSI 1988). A good workstation design accounts for individual body sizes and often exceeds minimum clearances to allow for free postural movement.

Table height has been shown to be an important contributor to computer user musculoskeletal problems. In particular, tables that are too high cause the keyboard to be too high for many operators. The standard desk height of 30 in. (76 cm) is often too high for most people to attain the proper arm angle when using the keyboard. This puts undue pressure on the hands, wrists, arms, shoulders, and neck. It is desirable for table heights to vary with the trunk height of the operator. Height-adjustable tables are effective for this. Adjustable multisurface tables enable good posture by allowing the keyboard and display to be independently adjusted to appropriate keying and viewing heights for each individual and each task. Tables that cannot be adjusted easily are not appropriate when used by several individuals of differing sizes. If adjustable tables are used, ease of adjustment is essential. Adjustments should be easy to make and operators should be instructed (trained) about how to adjust the workstation to be comfortable and safe.

Specifications for the height of working surfaces vary by whether the table is adjustable or fixed in height and depending on a single working surface or multiple working surfaces. Remember that adjustable-height working surfaces are strongly recommended. However, if the working surface height is not adjustable, the proper height for a nonadjustable working surface is about 70 cm (floor to top of surface) (ANSI 1988). Adjustable tables allow vertical adjustments of the keyboard and display. Some allow for independent adjustment of the keyboard and display. For single adjustable working surfaces, the working surface height adjustment should be 70–80 cm. For independently adjustable working surfaces for the keyboard and screen, the appropriate height range for the keyboard surface is 59–71 cm, and 70–80 cm for the screen (ANSI 1988).

2.7.2. The Chair

Poorly designed chairs can contribute to computer user discomfort. Chair adjustability in terms of height, seat angle, lumbar support, and armrest height and angle reduces the pressure and loading on the musculoskeleture of the back, legs, shoulders, neck, and arms. In addition, how the chair supports the movement of the user (the chair's action) helps to maintain proper seated posture and encourages good movement patterns. A chair that provides swivel action encourages movement, while backward tilting increases the number of postures that can be assumed. The chair height should be adjustable so that the feet can rest firmly on the floor with minimal pressure beneath the thighs. The minimum range of adjustment for seat height should be 38–52 cm (NAS 1983; Smith 1984; ANSI 1988). Modern chairs also provide an action that supports the back (spine) when seated. Examples of such chairs are the Leap by Steelcase, Inc. and the Aeron by Herman Miller.

To enable shorter users to sit with their feet on the floor without compressing their thighs, it may be necessary to add a footrest. A well-designed footrest has the following features: (1) it is inclined upward slightly (about 5–15°), (2) it has a nonskid surface, (3) it is heavy enough that it does not slide easily across the floor, (4) it is large enough for the feet to be firmly planted, and (5) it is portable.

The seat pan is where the user's buttocks sits on the chair. It is the part that directly supports the weight of the buttocks. The seat pan should be wide enough to permit operators to make slight shifts in posture from side to side. This not only helps to avoid static postures but also accommodates a large range of individual buttock sizes with a few seat pan widths. The minimum seat pan width should be 45 cm and the minimum depth 38–43 cm (ANSI 1988). The front edge of the chair should be well rounded downward to reduce pressure on the underside of the thighs, which can affect blood flow to the legs and feet. The seat needs to be padded to the proper firmness that ensures an even distribution of pressure on the thighs and buttocks. A properly padded seat should compress about one-half to one inch when a person sits on it.

Some experts feel that the seat front should be elevated slightly (up to 7°), while others feel it should be lowered slightly (about 5°) (ANSI 1988). There is little agreement among the experts about which is correct (Grandjean 1979, 1984). Many chair manufacturers provide adjustment of the front angle so the user can have the preferred tilt angle, either forward or backward.

The tension for leaning backward and the backward tilt angle of the backrest should be adjustable. Inclination of chair backrest is important for users to be able to lean forward or back in a comfortable manner while maintaining a correct relationship between the seat pan angle and the backrest inclination. A back seat inclination of about 110° is considered as the best position by many experts (Grandjean 1984). However, studies have shown that operators may incline backward as much as 125°. Backrests that tilt to allow an inclination of up to 125–130° are a good idea. The advantage of having an independent tilt angle adjustment is that the backrest tilt will then have little or no effect on the front seat height. This also allows operators to shift postures easily and often.

Chairs with full backrests that provide lower back (lumbar) support and upper back (lower shoulder) support are preferred. This allows employees to lean backward or forward, adopting a relaxed posture and resting the back muscles. A full backrest with a height around 45–51 cm is recommended (ANSI 1988). However, some of the newer chair designs do not have the bottom of the backrest go all the way to the seat pan. This is acceptable as long as the lumbar back is properly supported. To prevent back strain with such chairs, it is recommended that they have midback (lumbar) support since the lumbar region is one of the most highly loaded parts of the spine.

For most computer workstations, chairs with rolling castors (or wheels) are desirable. They are easy to move and facilitate the postural adjustment of users, particularly when the operator has to access equipment or materials that are on secondary working surfaces. Chairs should have a five-star base for tipping stability (ANSI 1988).

Another important chair feature is armrests. Pros and cons for the use of armrests at computer workstations have been advanced. On the one hand, some chair armrests can present problems of restricted arm movement, interference with keyboard operation, pinching of fingers between the armrest and table, restriction of chair movement such as under the work table, irritation of the arm or elbows, and adoption of awkward postures.

On the other hand, well-designed armrests or elbow rests can provide support for resting the arms to prevent or reduce fatigue, especially during breaks from typing. Properly designed armrests can overcome the problems mentioned because they can be raised, lowered, and angled to fit the user's needs. Removable armrests are an advantage because they provide greater flexibility for individual user preference, especially for users who develop discomfort and pain from the pressure of the armrest on their arms.

2.7.3. Other Workstation Considerations

An important component of the workstation that can help reduce musculoskeletal loading is a document holder. When properly designed and proportioned, document holders reduce awkward inclinations, as well as frequent movements up and down and back and forth of the head and neck. They permit source documents to be placed in a central location at approximately the same viewing distance and height as the computer screen. This eliminates needless head and neck movements and reduces eyestrain. In practice, some flexibility about the location, adjustment, and position of the document holder should be maintained to accommodate both task requirements and operator preferences. The document holder should have a matte finish so that it does not produce reflections or a glare source.

Privacy requirements include both visual and acoustical control of the workplace. Visual control prevents physical intrusions and distractions, contributes to protecting confidential/private conversations, and prevents the individual from feeling constantly watched. Acoustical control prevents distracting and unwanted noise—from machine or conversation—and permits speech privacy. While certain acoustical methods and materials such as free-standing panels are used to control general office noise level, they can also be used for privacy. In open-office designs they can provide workstation privacy. Generally, noise control at a computer workstation can be achieved through the following methods:

- Use of vertical barriers, such as acoustical screens or panels.
- Selection of floor, ceiling, wall, and workstation materials and finishes according to their power to control noise.
- Placement of workstations to enhance individual privacy.
- Locating workstations away from areas likely to generate noise (e.g., printer rooms, areas with heavy traffic).

Each of these methods can be used individually or combined to account for the specific visual and acoustical requirements of the task or individual employee needs. Planning for privacy should not be made at the expense of visual interest or spatial clarity. For instance, providing wide visual views can prevent the individual from feeling isolated. Thus, a balance between privacy and openness enhances user comfort, work effectiveness, and office communications. Involving the employee in decisions of privacy can help in deciding the compromises between privacy and openness.

2.8. Work Practices

Good ergonomic design of computer workstations has the potential to reduce visual and musculoskeletal complaints and disorders as well as increase employee performance. However, regardless of how well a workstation is designed, if operators must adopt static postures for a long time, they can still have performance, comfort, and health problems. Thus, designing tasks that induce employee movement in addition to work breaks can contribute to comfort and help relieve employees' fatigue.

2.8.1. Work Breaks

As a minimum, a 15-minute break from working should be taken after 2 hours of continuous computer work (CDC 1980; NIOSH 1981). Breaks should be more frequent as visual, muscular, and mental loads are high and as users complain of visual and musculoskeletal discomfort and psychological stress. With such intense, high-workload tasks, a work break of 10 minutes should be taken after 1 hour of continuous computer work. More frequent breaks for alternative work that does not pose demands similar to the primary computer work can be taken after 30 minutes of continuous computer work. Rest breaks provide an opportunity for recovery from local visual, musculoskeletal, and mental fatigue, to break from monotonous activities, or to engage in activities that provide variety in sensory, motor, and cognitive requirements.

While ergonomics considers users' physiological interface with interactive systems, cognitive design focuses on the psychological interface between users and computers. This will be addressed in the next section.

3. COGNITIVE DESIGN

3.1. Overview

Cognitive design, also referred to as cognitive engineering, is a multidisciplinary approach to system design that considers the analysis, design, and evaluation of interactive systems (Vicente 1999). Cognitive design involves developing systems through an understanding of human capabilities and limitations. It focuses on how humans process information and aims to identify users' mental models, such that supporting metaphors and analogies can be identified and designed into systems (Eberts 1994). The general goal of cognitive design is thus to design interactive systems that are predictable (i.e., respond to the way users perceive, think, and act). Through the application of this approach, human–computer interaction has evolved into a relatively standard set of interaction techniques, including typing, pointing, and clicking. This set of ‘‘standard’’ interaction techniques is evolving, with a transition from graphical user interfaces to perceptual user interfaces that seek to more naturally interact with users through multimodal and multimedia interaction (Turk and Robertson 2000). In either case, however, these interfaces are characterized by interaction techniques that try to match user capabilities and limitations to the interface design.

Cognitive design efforts are guided by the requirements definition, user profile development, tasks analysis, task allocation, and usability goal setting that result from an intrinsic understanding gained from the target work environment. Although these activities are listed and presented in this order, they are conducted iteratively throughout the system development life cycle.

3.2. Requirements Definition

Requirements definition involves the specification of the necessary goals, functions, and objectives to be met by the system design (Eberts 1994; Rouse 1991). The intent of the requirements definition is to specify what a system should be capable of doing and the functions that must be available to users to achieve stated goals. Karat and Dayton (1995) suggest that developing a careful understanding of system requirements leads to more effective initial designs that require less redesign. Ethnographic evaluation can be used to develop a requirements definition that is necessary and complete to support the target domain (Nardi 1997).

Goals specify the desired system characteristics (Rouse 1991). These are generally qualitatively stated (e.g., automate functions, maximize use, accommodate user types) and can be met in a number of ways. Functions define what the system should be capable of doing without specifying the specifics of how the functions should be achieved. Objectives are the activities that the system must be able to accomplish in support of the specified functions. Note that the system requirements, as stated in terms of goals, functions, and objectives, can be achieved by a number of design alternatives. Thus, the requirements definition specifies what the system should be able to accomplish without specifying how this should be realized. It can be used to guide the overall design effort to ensure the desired end is achieved. Once a set of functional and feature requirements has been scoped out, an understanding of the current work environment is needed in order to design systems that effectively support these requirements.

3.3. Contextual Task Analysis

The objective of contextual task analysis is to achieve a user-centered model of current work practices (Mayhew 1999). It is important to determine how users currently carry out their tasks, which individuals they interact with, what tools support the accomplishment of their job goals, and the resulting products of their efforts. Formerly this was often achieved by observing a user or set of users in a laboratory setting and having them provide verbal protocols as they conducted task activities in the form of use cases (Hackos and Redish 1998; Karat 1988; Mayhew 1999; Vermeeren 1999). This approach, however, fails to take into consideration the influences of the actual work setting. Through an understanding of the work environment, designers can leverage current practices that are effective while designing out those that are ineffective. The results of a contextual task analysis include work environment and task analyses, from which mental models can be identified and user scenarios and task-organization models (e.g., use sequences, use flow diagrams, use workflows, and use hierarchies) can be derived (Mayhew 1999). These models and scenarios can then help guide the design of the system. As depicted in Figure 3, contextual task analysis consists of three main steps.

Effective interactive system design thus comes from a basis in direct observation of users in their work environments rather than assumptions about the users or observations of their activities in contrived laboratory settings (Hackos and Redish 1998). Yet contextual tasks analysis is sometimes overlooked because developers assume they know users or that their user base is too diverse, expensive, or time consuming to get to know. In most cases, however, observation of a small set of diverse users can provide critical insights that lead to more effective and acceptable system designs. For usability evaluations, Nielsen (1993) found that the greatest payoff occurs with just three users.

3.3.1. Background Information

It is important when planning a task analysis to first become familiar with the work environment. If analysts do not understand work practices, tools, and jargon prior to commencing a task analysis, they can easily get confused and become unable to follow the task flow. Further, if the first time users see analysts they have clipboard and pen in hand, users are likely to resist being observed or change their behaviors during observation. Analysts should develop a rapport with users by spending time with them, participating in their task activities when possible, and listening to their needs and concerns. Once users are familiar and comfortable with analysts and analysts are likewise versed on work practices, data collection can commence. During this familiarization, analysts can also capture data to characterize users.

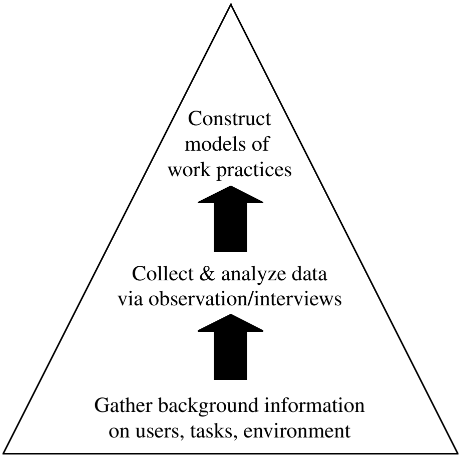

Construct
models of
work practices
Collect & analyze data
via observation/interviews
Gather background information
on users, tasks, environment

Figure 3 The Steps of Contextual Task Analysis.

3.3.2. Characterizing Users

It is ultimately the users who will determine whether a system is adopted into their lives. Designs that frustrate, stress, or annoy users are not likely to be embraced. Based on the requirements definition, the objective of designers should be to develop a system that can meet specified user goals, functions, and objectives. This can be accomplished through an early and continual focus on the target user population (Gould et al. 1997). It is inconceivable that design efforts would bring products to market without thoroughly determining who the user is. Yet developers, as they expedite system development to rush products to market, are often reluctant to characterize users. In doing so, they may fail to recognize the amount of time they spend speculating upon what users might need, like, or want in a product (Nielsen 1993). Ascertaining this information directly by querying representative users can be both more efficient and more accurate.

Information about users should provide insights into differences in their computer experience, domain knowledge, and amount of training on similar systems (Wixon and Wilson 1997). The results can be summarized in a narrative format that provides a user profile of each intended user group (e.g., primary users, secondary users, technicians and support personnel). No system design, however, will meet the requirements of all types of users. Thus, it is essential to identify, define, and characterize target users. Separate user profiles should be developed for each target user group. The user profiles can then feed directly into the task analysis by identifying the user groups for which tasks must be characterized (Mayhew 1999).

Mayhew (1999) presents a step-by-step process for developing user profiles. First, a determination of user categories is made by identifying the intended user groups for the target system. When developing a system for an organization, this information may come directly from preexisting job categories. Where those do not exist, marketing organizations often have target user populations identified for a given system or product. Next, the relevant user characteristics must be identified. User profiles should be specified in terms of psychological (e.g., attitudes, motivation), knowledge and experience (e.g., educational background, years on job), job and task (e.g., frequency of use), and physical (e.g., stature, visual impairments) characteristics (Mayhew 1999; Nielsen 1993; Wixon and Wilson 1997). While many of these user attributes can be obtained via user profile questionnaires or interviews, psychological characteristics may be best identified via ethnographic evaluation, where a sense of the work environment temperament can be obtained. Once this information is obtained, a summary of the key characteristics for each target user group can be developed, highlighting their implications to the system design. By understanding these characteristics, developers can better anticipate such issues as learning difficulties and specify appropriate levels of interface complexity. System design requirements involve an assessment of the required levels of such factors as ease of learning, ease of use, level of satisfaction, and workload for each target user group (see Table 2)

Individual differences within a user population should also be acknowledged (Egan 1988; Hackos and Redish 1998). While users differ along many dimensions, key areas of user differences have been identified that significantly influence their experience with interactive systems. Users may differ

TABLE 2 Example of a User Profile

Characteristic	Questionnaire Response	System Design Requirement
Attitude	Negative	System should be subjectively pleasing
Motivation	Generally low	Usefulness of system should be readily apparent
Education level	High school	Simplicity important; training requirements should be minimal
Computer experience	Low	High ease of learning required
Frequency of computer use	Discretionary	High ease of use required; system workload should be minimized
Typing skills	Poor	Minimize typing; use icons and visual displays
Gender	Mostly males	Consider color blindness
Age	Average = 42.5 (s.d. = 3.6)	Text and symbol size should be readily legible

in such attributes as personality, physical or cognitive capacity, motivation, cultural background, education, and training. Users also change over time (e.g., transitioning from novice to expert). By acknowledging these differences, developers can make informed decisions on whether or not to support them in their system designs. For example, marketing could determine which group of individuals it would be most profitable to target with a given system design.

3.3.3. Collecting and Analyzing Data

Contextual task analysis focuses on the behavioral aspects of a task, resulting in an understanding of the general structure and flow of task activities (Mayhew 1999; Nielsen 1993; Wixon and Wilson 1997). This analysis identifies the major tasks and their frequency of occurrence. This can be compared to cognitive task analysis, which identifies the low-level perceptual, cognitive, and motor actions required during task performance (Card et al. 1983; Corbett et al. 1997). Beyond providing an understanding of tasks and workflow patterns, the contextual task analysis also identifies the primary objects or artifacts that support the task, information needs (both inputs and outputs), workarounds that have been adopted, and exceptions to normal work activities. The result of this analysis is a task flow diagram with supporting narrative depicting user-centered task activities, including task goals; information needed to achieve these goals; information generated from achieving these goals; and task organization (i.e., subtasks and interdependencies).

Task analysis thus aims to structure the flow of task activities into a sequential list of functional elements, conditions of transition from one element to the next, required supporting tools and artifacts, and resulting products (Sheridan 1997a). There are both formal and informal techniques for task analysis (see Table 3). Such an analysis can be driven by formal models such as TAKD (task analysis for knowledge description; see Diaper 1989) or GOMS (goals, operators, methods, and selection rules; see Card et al. 1983) or through informal techniques such as interviews, observation and shadowing, surveys, and retrospectives and diaries (Jeffries 1997). With all of these methods, typically

TABLE 3 Task-Analysis Techniques for Interactive System Design

Design Objective	Task-Analysis Technique
Detailed description of task	TAKD, GOMS, interviews
Detailed description of task (when difficult to verbalize task knowledge)	Observation, shadowing
Task description for tasks with significant performance variation; determine specific task characteristics (e.g., frequency)	Surveys, observation, shadowing
Clarify task areas	Surveys, observation, shadowing, retrospectives and diaries

a domain expert is somehow queried about their task knowledge. It may be beneficial to query a range of users, from novice to expert, to identify differences in their task practices. In either case, it is important to select individuals that can readily verbalize how a task is carried out to serve as informants (Ebert 1994).

When a very detailed task analysis is required, formal techniques such as TAKD (Diaper 1989; Kirwan and Ainsworth 1992) or GOMS (Card et al. 1983) can be used to delineate task activities (see Chapter 39), TAKD uses knowledge-representation grammars (i.e., sets of statements used to described system interaction) to represent task-knowledge in a task-descriptive hierarchy. This technique is useful for characterizing complex tasks that lack fine-detail cognitive activities (Eberts 1994). GOMS is a predictive modeling technique that has been used to characterize how humans interact with computers. Through a GOMS analysis, task goals are identified, along with the operators (i.e., perceptual, cognitive, or motor acts) and methods (i.e., series of operators) to achieve those goals and the selection rules used to elect between alternative methods. The benefit of TAKD and GOMS is that they provide an in-depth understanding of task characteristics, which can be used to quantify the benefits in terms of consistency (TAKD) or performance time gains (GOMS) of one design vs. another (see Gray et al. 1993; McLeod and Sherwood-Jones 1993 for examples of the effective use of GOMS in design). This deep knowledge, however, comes at a great cost in terms of time to conduct the analysis. Thus, it is important to determine the level of task analysis required for informed design. While formal techniques such as GOMS can lead to very detailed analyses (i.e., at the perceive, think, act level), often such detail is not required for effective design. Jeffries (1997) suggests that one can loosely determine the right level of detail by determining when further decomposition of the task would not reveal any “interesting” new subtasks that would enlighten the design. If detailed task knowledge is not deemed requisite, informal task-analysis techniques should be adopted.

Interviews are the most common informal technique to gather task information (Jeffries 1997; Kirwan and Ainsworth 1992; Meister 1985). In this technique, informants are asked to verbalize their strategies, rationale, and knowledge used to accomplish task goals and subgoals (Ericsson and Simon 1980). As each informant’s mental model of the tasks they verbalize is likely to differ, it is advantageous to interview at least two to three informants to identify the common flow of task activities. Placing the informant in the context of the task domain and having him or her verbalize while conducting tasks affords more complete task descriptions while providing insights on the environment the task is performed within. It can sometimes be difficult for informants to verbalize their task performance because much of it may be automatized (Eberts 1994). When conducting interviews, it is important to use appropriate sampling techniques (i.e., sample at the right time with enough individuals), avoid leading questions, and follow up with appropriate probe questions (Nardi 1997). While the interviewer should generally abstain from interfering with task performance, it is sometimes necessary to probe for more detail when it appears that steps or subgoals are not being communicated. Eberts (1994) suggests that the human information-processing model can be used to structure verbal protocols and determine what information is needed and what is likely being left out.

Observation during task activity or shadowing workers throughout their daily work activities are time-consuming task-analysis techniques, but they can prove useful when it is difficult for informants to verbalize their task knowledge (Jeffries 1997). These techniques can also provide information about the environment in which tasks are performed, such as tacit behaviors, social interactions, and physical demands, which are difficult to capture with other techniques (Kirwan and Ainsworth 1992).

While observation and shadowing can be used to develop task descriptions, surveys are particularly useful task-analysis tools when there is significant variation in the manner in which tasks are performed or when it is important to determine specific task characteristics, such as frequency (Jeffries 1997; Nielsen 1993). Surveys can also be used as a follow-on to further clarify task areas described via an interview. Focused observation, shadowing, retrospectives, and diaries are also useful for clarifying task areas. With retrospectives and diaries, an informant is asked to provide a retrospective soon after completing a task or to document his or her activities after several task events, the latter being a diary.

Whether formal or informal techniques are used, the objective of the task analysis is to identify the goals of users and determine the techniques they use to accomplish these goals. Norman (1988) provides a general model of the stages users go through when accomplishing goals (see Figure 4). Stanton (1998) suggests that there are three main ways in which this process can go awry: by users forgetting a required action, executing an errant action, or misperceiving or misinterpreting the current state of the system. In observing users of a vending machine, Verhoef (1988) indeed found that these types of errors occur during system interaction. In this case study, users of the vending machine failed to perceive information presented by the machine, performed actions in the wrong order, and misinterpreted tasks when they were not clearly explained or when incomplete information was provided.

By understanding the stages of goal accomplishment and the related errors that can occur, developers can more effectively design interactive systems. For example, by knowing that users perceive and interpret the system state once an action has been executed, designers can understand why it is

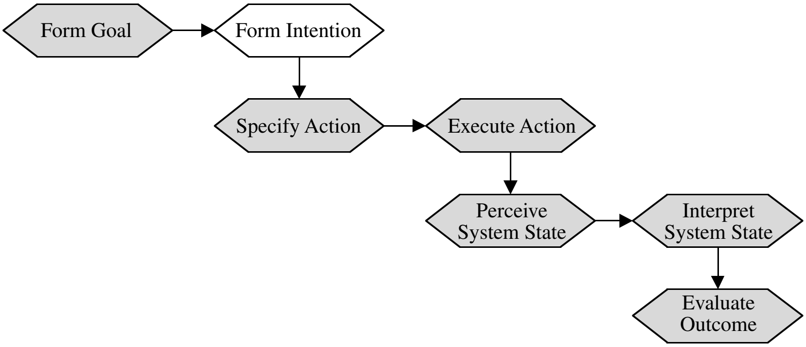


Figure 4 Stages of Goal Accomplishment.

essential to provide feedback (Nielsen 1993) to the executed action. Knowing that users often specify alternative methods to achieve a goal or change methods during the course of goal seeking, designers can aim to support diverse approaches. Further, recognizing that users commit errors emphasizes the need for undo functionality.

The results from a task analysis provide insights into the optimal structuring of task activities and the key attributes of the work environment that will directly affect interactive system design (Mayhew 1999). The analysis enumerates the tasks that users may want to accomplish to achieve stated goals through the preparation of a task list or task inventory (Hackos and Redish 1998; Jeffries 1997). A model of task activities, including how users currently think about, discuss, and perform their work, can then be devised based on the task analysis. To develop task-flow models, it is important to consider the timing, frequency, criticality, difficulty, and responsible individual of each task on the list. In seeking to conduct the analysis at the appropriate level of detail, it may be beneficial initially to limit the task list and associated model to the primary 10–20 tasks that users perform. Once developed, task models can be used to determine the functionality necessary to support in the system design. Further, once the task models and desired functionality are characterized, use scenarios (i.e., concrete task instances, with related contextual [i.e., situational] elements and stated resolutions) can be developed that can be used to drive both the system design and evaluation (Jeffries 1997).

3.3.4. Constructing Models of Work Practices

While results from the task analysis provide task-flow models, they also can provide insights on the manner in which individuals model these process flows (i.e., mental models). Mental models synthesize several steps of a process into an organized unit (Allen 1997). An individual may model several aspects of a given process, such as the capabilities of a tool or machine, expectations of coworkers, or understandings of support processes (Fischer 1991). These models allow individuals to predict how a process will respond to a given input, explain a process event, or diagnose the reasons for a malfunction. Mental models are often incomplete and inaccurate, however, so understandings based on these models can be erroneous.

As developers design systems, they will develop user models of target user groups (Allen 1997). These models should be relevant (i.e., able to make predictions as users would), accurate, adaptable to changes in user behavior, and generalizable. Proficient user modeling can assist developers in designing systems that interact effectively with users. Developers must recognize that users will both come to the system interaction with preconceived mental models of the process being automated and develop models of the automated system interaction. They must thus seek to identify how users represent their existing knowledge about a process and how this knowledge fits together in learning and performance so that they can design systems that engender the development of an accurate mental model of the system interaction (Carroll and Olson 1988). By understanding how users model processes, developers can determine how users currently think and act, how these behaviors can be supported by the interactive system design when advantageous, and how they can be modified and improved upon via system automation.

3.3.5. Task Allocation

In moving toward a system design, once tasks have been analyzed and associated mental models characterized, designers can use this knowledge to address the relationship between the human and the interactive system. Task allocation is a process of assigning the various tasks identified via the task analysis to agents (i.e., users), instruments (e.g., interactive systems) or support resources (e.g., training, manuals, cheat sheets). It defines the extent of user involvement vs. computer automation in system interaction (Kirwan and Ainsworth 1992). In some system-development efforts, formal task allocation will be conducted; in others, it is a less explicit yet inherent part of the design process.

While there are many systematic techniques for conducting task analysis (see Kirwan and Ainsworth 1992), the same is not true of task allocation (Sheridan 1997a). Further, task allocation is complicated by the fact that seldom are tasks or subtasks truly independent, and thus their interdependence must be effectively designed into the human–system interaction. Rather than a deductive assignment of tasks to human or computer, task allocation thus becomes a consideration of the multitude of design alternatives that can support these interdependencies. Sheridan (1997a,b) delineates a number of task allocation considerations that can assist in narrowing the design space (see Table 4).

In allocating tasks, one must also consider what will be assigned to support resources. If the system is not a walk-up-and-use system but one that will require learning, then designers must identify what knowledge is appropriate to allocate to support resources (e.g., training courses, manuals, online help).

Training computer users in their new job requirements and how the technology works has often been a neglected element in office automation. Many times the extent of operator training is limited to reading the manual and learning by trial and error. In some cases, operators may have classes that go over the material in the manual and give hands-on practice with the new equipment for limited periods of time. The problem with these approaches is that there is usually insufficient time for users to develop the skills and confidence to adequately use the new technology. It is thus essential to determine what online resources will be required to support effective system interaction.

Becoming proficient in hardware and software use takes longer than just the training course time. Often several days, weeks, or even months of daily use are needed to become an expert depending on the difficulty of the application and the skill of the individual. Appropriate support resources should be designed into the system to assist in developing this proficiency. Also, it is important to remember that each individual learns at his or her own pace and therefore some differences in proficiency will be seen among individuals. When new technology is introduced, training should tie in skills from the former methods of doing tasks to facilitate the transfer of knowledge. Sometimes new skills clash with those formerly learned, and then more time for training and practice is necessary to achieve good results. If increased performance or labor savings are expected with the new technology, it is prudent not to expect results too quickly. Rather, it is wise to develop the users' skills completely if the most positive results are to be achieved.

TABLE 4 Considerations in the Task-Allocation Process

Considerations in Task Allocation	Design Issue
Check task-analysis constraints	Strict task requirements can complicate or make infeasible appropriate task allocation
Identify obvious allocations	Highly repetitive tasks are generally appropriate for automation; dealing with the unexpected or cognitively complex tasks are generally appropriate for humans
Identify expected allocations	Users' mental models may uncover expected allocation schemes
Identify the extremes	Bound the design space by assessing total computer automation vs. total human manual control solutions
Consider points between the extremes	Sheridan (1997a,b) offers a 10-point scale of allocation between the extremes that assists in assessing intermediate solutions
Consider level of specificity required by allocation	Strict assignments are ineffectual; a general principle is to leave the big picture to the human and the details to the computer
Consider sequential vs. parallel processing	Will the computer and user trade outputs of their processing or will they concurrently collaborate in task performance?
Consider the range of criteria that can be used to judge appropriate allocation	While many criteria affect overall system interaction, a small number of criteria are generally important for an individual task

3.4. Competitive Analysis and Usability Goal Setting

Once the users and tasks have been characterized, it is sometimes beneficial to conduct a competitive analysis (Nielsen 1993). Identifying the strengths and weaknesses of competitive products or existing systems allows means to leverage strengths and resolve identified weaknesses.

After users have been characterized, a task analysis performed, and, if necessary, a competitive analysis conducted, the next step in interactive system design is usability goal setting (Hackos and Redish 1998; Mayhew 1999; Nielsen 1993; Wixon and Wilson 1993). Usability objectives generally focus around effectiveness (i.e., the extent to which tasks can be achieved), intuitiveness (i.e., how learnable and memorable the system is), and subjective perception (i.e., how comfortable and satisfied users are with the system) (Eberts 1994; Nielsen 1993; Shneiderman 1992; Wixon and Wilson 1997). Setting such objectives will ensure that the usability attributes evaluated are those that are important for meeting task goals; that these attributes are translated into operational measures; that the attributes are generally holistic, relating to overall system/task performance; and that the attributes relate to specific usability objectives.

Because usability is assessed via a multitude of potentially conflicting measures, often equal weights cannot be given to every usability criterion. For example, to gain subjective satisfaction, one might have to sacrifice task efficiency. Developers should specify usability criteria of interest and provide operational goals for each metric. These metrics can be expressed as absolute goals (i.e., in terms of an absolute quantification) or as relative goals (i.e., in comparison to a benchmark system or process). Such metrics provide system developers with concrete goals to meet and a means to measure usability. This information is generally documented in the form of a usability attribute table and usability specification matrix (see Mayhew 1999).

3.5. User Interface Design

While design ideas evolve throughout the information-gathering stages, formal design of the interactive system commences once relevant information has been obtained. The checklist in Table 5 can be used to determine whether the critical information items that support the design process have been addressed. Readied with information, interactive system design generally begins with an initial definition of the design and evolves into a detailed design, from which iterative cycles of evaluation and improvement transpire (Martel 1998).

3.5.1. Initial Design Definition

Where should one commence the actual design of a new interactive system? Often designers look to existing products within their own product lines or competitors' products. This is a sound practice because it maintains consistency with existing successful products. This approach may be limiting, however, leading to evolutionary designs that lack design innovations. Where can designers obtain the ideas to fuel truly innovative designs that uniquely meet the needs of their users? Ethnographic evaluations can lead to many innovative design concepts that would never be realized in isolation of the work environment (Mountford 1990). The effort devoted to the early characterization of users and tasks, particularly when conducted in the context of the work environment, often is rewarded in terms of the generation of innovative design ideas. Mountford (1990) has provided a number of techniques to assist in eliciting design ideas based on the objects, artifacts, and other information gathered during the contextual task analysis (see Table 6).

To generate a multitude of design ideas, it is beneficial to use a parallel design strategy (Nielsen 1993), where more than one designer sets out in isolation to generate design concepts. A low level

TABLE 5 Checklist of Information Items

Identified necessary goals, functions, and objectives to be met by system design
Became familiar with practices, tools, and vernacular of work environment
Characterized user profiles in terms of psychological characteristics, knowledge and experience, job and task characteristics, and physical attributes
Acknowledged individual differences within target user population
Developed user models that are relevant, accurate, adaptable to changes in user behavior, and generalizable
Developed a user-centered model of current work practices via task analysis
Defined extent of user involvement vs. computer automation in system interaction, as well as required support resources (e.g., manuals)
Conducted a competitive analysis
Set usability goals

TABLE 6 Design Idea Generation Checklist

Are there new uses for objects and artifacts identified in the task analysis?
Could objects and artifacts be adapted to be like something else? How would this change the organizational structure of the system interaction?
Could objects and artifacts be modified to serve a new purpose?
Can tools or other features be added to objects and artifacts?
Can interaction be streamlined by subtracting from objects and artifacts?
Are there alternative metaphors that would be more appropriate for the task domain being automated? Can the basic layout of a metaphor be modified or somehow rearranged?
Can a design scheme be reversed or transposed for alternative interaction approaches?
Are there large, encompassing metaphors that could be used to characterize more of the task domain?

of effort (e.g., a few hours to a few days) is generally devoted to this idea-generation stage. Storyboarding via paper prototypes is often used at this stage because it is easy to generate and modify and is cost effective (Martel 1998). Storyboards provide a series of pictures representing how an interface may look.

3.5.2. Detailed Design

Once design alternatives have been storyboarded, the best aspects of each design can be identified and integrated into a detailed design concept. The detailed design can be realized through the use of several techniques, including specification of nouns and verbs that represent interface objects and actions, as well as the use of metaphors (Hackos and Redish 1998). The metaphors can be further refined via use scenarios, use sequences, use flow diagrams, use workflows, and use hierarchies. Storyboards and rough interface sketches can support each stage in the evolution of the detailed design.

3.5.2.1. Objects and Actions Workplace artifacts, identified via the contextual task analysis, become the objects in the interface design (Hackos and Redish 1998). Nouns in the task flow also become interface objects, while verbs become interface actions. Continuing the use of paper prototyping, the artifacts, nouns, and verbs from the task flows and related models can each be specified on a sticky note and posted to the working storyboard. Desired attributes for each object or action can be delineated on the notes. The objects and actions should be categorized and any redundancies eliminated. The narratives and categories can generate ideas on how interface objects should look, how interface actions should feel, and how these might be structurally organized around specified categories.

3.5.2.2. Metaphors Designers often try to ease the complexity of system interaction by grounding interface actions and objects and related tasks and goals in a familiar framework known as a metaphor (Neale and Carroll 1997). A metaphor is a conceptual set of familiar terms and associations (Erickson 1990). If designed into a user interface, it can be used to incite users to relate what they already know about the metaphoric concept to the system interaction, thereby enhancing the learnability of the system (Carroll and Thomas 1982).

The purpose of developing interface metaphors is to provide users with a useful orienting framework to guide their system interaction. The metaphor provides insights into the spatial properties of the user interface and the manner in which they are derived and maintained by interaction objects and actions (Carroll and Mack 1985). It stimulates systematic system interaction that may lead to greater understanding of these spatial properties. Through this understanding, users should be able to tie together a configural representation (or mental model) of the system to guide their interactions (Kay 1990).

An effective metaphor will both orient and situate users within the system interaction. It will aid without attracting attention or energy away from the automated task process (Norman 1990). Providing a metaphor should help focus users to critical cues and away from irrelevant distractions. The metaphor should also help to differentiate the environment and enhance visual access (Kim and Hirtle 1995). Parunak (1989) accomplished this in a hypertext environment by providing between-path mechanisms (e.g., backtracking capability and guided tours), annotation capabilities that allow users to designate locations that can be accessed directly (e.g., bookmarks in hypertext), and links and filtering techniques that simplify a given topology.

It is important to note that a metaphor does not have to be a literal similarity (Ortony 1979) to be effective. In fact, Gentner (1983) and Gentner and Clement (1988) suggest that people seek to identify relational rather than object attribute comparisons in comprehending metaphors. Based on

Gentner's structure-mapping theory, the aptness of a metaphor should increase with the degree to which its interpretation is relational. Thus, when interpreting a metaphor, people should tend to extend relational rather than object attribute information from the base to the target. The learning efficacy of a metaphor, however, is based on more than the mapping of relational information between two objects (Carroll and Mack 1985). Indeed, it is imperative to consider the open-endedness of metaphors and leverage the utility of not only correspondence, but also noncorrespondence in generating appropriate mental models during learning. Nevertheless, the structure-mapping theory can assist in providing a framework for explaining and designing metaphors for enhancing the design of interactive systems.

Neale and Carroll (1997) have provided a five-stage process from which design metaphors can be conceived (see Table 7). Through the use of this process, developers can generate coherent, well-structured metaphoric designs.

3.5.2.3. Use Scenarios, Use Sequences, Use Flow Diagrams, Use Workflows, and Use Hierarchies Once a metaphoric design has been defined, its validity and applicability to task goals and subgoals can be identified via use scenarios, use sequences, use flow diagrams, use workflows, and use hierarchies (see Figure 5) (Hackos and Redish 1998). Use scenarios are narrative descriptions of how the goals and subgoals identified via the contextual task analysis will be realized via the interface design. Beyond the main flow of task activities, they should address task exceptions, individual differences, and anticipated user errors. In developing use scenarios, it can be helpful to reference task allocation schemes (see Section 3.3.5). These schemes can help to define what will be achieved by users via the interface, what will be automated, and what will be rendered to support resources in the use scenarios.

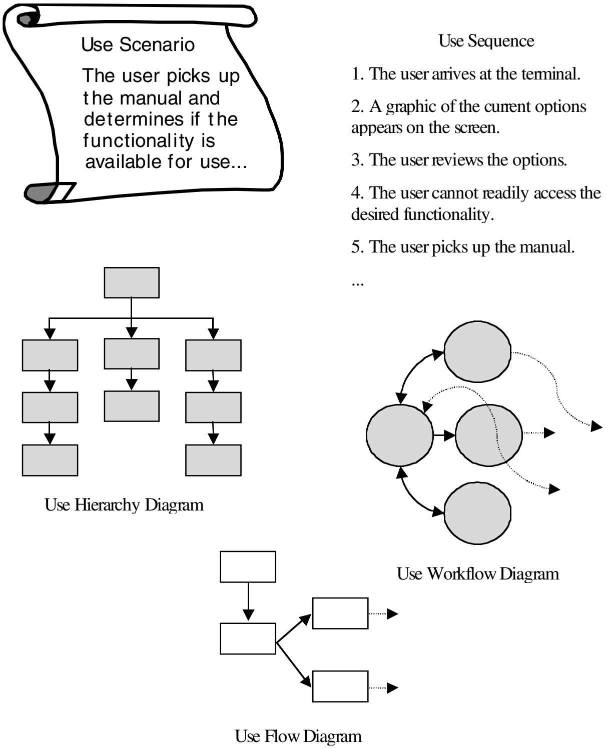
If metaphoric designs are robust, they should be able to withstand the interactions demanded by a variety of use scenarios with only modest modifications required. Design concepts to address required modifications should evolve from the types of interactions envisioned by the scenarios. Once the running of use scenarios fails to generate any required design modifications, their use can be terminated.

If parts of a use scenario are difficult for users to achieve or designers to conceptualize, use sequences can be used. Use sequences delineate the sequence of steps required for a scenario subsection being focused upon. They specify the actions and decisions required of the user and the interactive system, the objects needed to achieve task goals, and the required outputs of the system interaction. Task workarounds and exceptions can be addressed with use sequences to determine if the design should support these activities. Providing detailed sequence specifications highlights steps that are not appropriately supported by the design and thus require redesign.

When there are several use sequences supported by a design, it can be helpful to develop use flow diagrams for a defined subsection of the task activities. These diagrams delineate the alternative paths and related intersections (i.e., decision points) users encounter during system interaction. The representative entities that users encounter throughout the use flow diagram become the required objects and actions for the interface design.

TABLE 7 Stages of Metaphoric Design Generation

Stage	Design Outcome
Identify system functionality	Required functions, features, and system capabilities are identified (see Section 3.2)
Generate metaphoric concepts	Artifacts and other objects in the environment identified via the contextual task analysis (see Section 3.3) can assist in generating design concepts (see Table 6)
Identify metaphor–interface matches	Identify what users do (goals and subgoals), the methods they use to accomplish these objectives (actions and objects), and map to the physical elements available in the metaphor; use cases can be used for this stage
Identify metaphor–interface mismatches	Identify where the metaphor has no analogous function for desired goals and subgoals
Determine how to manage metaphor–interface mismatches	Determine where composite metaphors or support resources are needed (e.g., online help, agent assistance) so problems related to mismatches can be averted



The user picks up the manual and determines if the functionality is available for use...

1. The user arrives at the terminal.
2. A graphic of the current options appears on the screen.
3. The user reviews the options.
4. The user cannot readily access the desired functionality.
5. The user picks up the manual.

...

Figure 5 Design-Generation Techniques.

When interactive systems are intended to yield savings in the required level of information exchange, use workflows can be used. These flows provide a visualization of the movement of users or information objects throughout the work environment. They can clearly denote if a design concept will improve the information flow. Designers can first develop use workflows for the existing system interaction and then eliminate, combine, resequence, and simplify steps to streamline the flow of information.

Use hierarchies can be used to visualize the allocation of tasks among workers. By using sticky notes to represent each node in the hierarchy, these representations can be used to demonstrate the before- and after-task allocations. The benefits of the new task allocation engendered by the interactive system design should be readily perceived in hierarchical flow changes.

3.5.2.4. Design Support Developers can look to standards and guidelines to direct their design efforts. Standards focus on advising the look of an interface, while guidelines address the usability of the interface (Nielsen 1993). Following standards and guidelines can lead to systems that are easy to learn and use due to a standardized look and feel (Buie 1999). Developers must be careful, however, not to follow these sources of design support blindly. An interactive system can be designed strictly according to standards and guidelines yet fail to physically fit users, support their goals and tasks, and integrate effectively into their environment (Hackos and Redish 1998).

Guidelines aim at providing sets of practical guidance for developers (Brown 1988; Hackos and Redish 1998; Marcus 1997; Mayhew 1992). They evolve from the results of experiments, theory-based predictions of human performance, cognitive psychology and ergonomic design principles, and experience. Several different levels of guidelines are available to assist system development efforts, including general guidelines applicable to all interactive systems, as well as category-specific (i.e., voice vs. touch screen interfaces) and product-specific guidelines (Nielsen 1993).

Standards are statements (i.e., requirements or recommendations) about interface objects and actions (Buie 1999). They address the physical, cognitive, and affective nature of computer interaction. They are written in general and flexible terms because they must be applicable to a wide variety of applications and target user groups. International (e.g., ISO 9241), national (e.g., ANSI, BSI), military and government (e.g., MIL-STD 1472D), and commercial (e.g., Common User Access by IBM) entities write them. Standards are the preferred approach in Europe. The European Community promotes voluntary technical harmonization through the use of standards (Rada and Ketchell 2000).

Buie (1999) has provided recommendations on how to use standards that could also apply to guidelines. These include selecting relevant standards; tailoring these select standards to apply to a given development effort; referring to and applying the standards as closely as possible in the interactive system design; revising and refining the select standards to accommodate new information and considerations that arise during development; and inspecting the final design to ensure the system design complies with the standards where feasible. Developing with standards and guidelines does not preclude the need for evaluation of the system. Developers will still need to evaluate their systems to ensure they adequately meet users' needs and capabilities.

3.5.2.5. Storyboards and Rough Interface Sketches The efforts devoted to the selection of a metaphor or composite metaphor and the development of use scenarios, use sequences, use flow diagrams, use workflows, and use hierarchies result in a plethora of design ideas. Designers can brainstorm over design concepts, generating storyboards of potential ideas for the detailed design (Vertelney and Booker 1990). Storyboards should be at the level of detail provided by use scenarios and workflow diagrams (Hackos and Redish 1998). The brainstorming should continue until a set of satisfactory storyboard design ideas has been achieved. The favored set of ideas can then be refined into a design concept via interface sketches. Sketches of screen designs and layouts are generally at the level of detail provided by use sequences. Cardboard mockups and Wizard of Oz techniques (Newell et al. 1990), the latter of which enacts functionality that is not readily available, can be used at this stage to assist in characterizing designs.

3.5.3. Prototyping

Prototypes of favored storyboard designs are developed. These are working models of the preferred designs (Hackos and Redish 1998; Vertelney and Booker 1990). They are generally developed with easy-to-use toolkits (e.g., Macromedia Director, Toolbook, SmallTalk, or Visual Basic) or simpler tools (e.g., hypercard scenarios, drawing programs, even paper or plastic mockups) rather than high-level programming languages. The simpler prototyping tools are easy to generate and modify and cost effective; however, they demonstrate little if anything in the way of functionality, may present concepts that cannot be implemented, and may require a "Wizard" to enact functionality. The toolkits provide prototypes that look and feel more like the final product and demonstrate the feasibility of desired functionality; however, they are more costly and time consuming to generate. Whether high- or low-end techniques are used, prototypes provide means to provide cost-effective, concrete design concepts that can be evaluated with target users (usually three to six users per iteration) and readily modified. They prevent developers from exhausting extensive resources in formal development of products that will not be adopted by users. Prototyping should be iterated until usability goals are met.

3.6. Usability Evaluation of Human–Computer Interaction

Usability evaluation focuses on gathering information about the usability of an interactive system so that this information can be used to focus redesign efforts via iterative design. While the ideal approach is to consider usability from the inception of the system development process, often it is considered in later development stages. In either case, as long as developers are committed to implementing modifications to rectify the most significant issues identified via usability-evaluation techniques, the efforts devoted to usability are generally advantageous. The benefits of usability evaluation include, but are not limited to, reduced system redesign costs, increased system productivity, enhanced user satisfaction, decreased user training, and decreased technical support (Nielsen 1993; Mayhew 1999).

There are several different usability-evaluation techniques. In some of these techniques, the information may come from users of the system (through the use of surveys, questionnaires, or specific measures from the actual use of the system), while in others, information may come from usability experts (using design walk-throughs and inspection methods). In still others, there may be no observations or user testing involved at all because the technique involves a theory-based (e.g., GOMS modeling) representation of the user (Card et al. 1983). Developers need to be able to select a method that meets their needs or combine or tailor methods to meet their usability objectives and situation. Usability-evaluation techniques have generally been classified as follows (Karat 1997; Preece 1993):

- Analytic/theory based (e.g., cognitive task analysis; GOMS)
- Expert evaluation (e.g., design walk-throughs; heuristic evaluations)
- Observational evaluation (e.g., direct observation; video; verbal protocols)
- Survey evaluation (e.g., questionnaires; structured interviews)

- Psychophysiological measures of subjective perception (e.g., EEGs; heart rate; blood pressure)
- Experimental evaluation (e.g., quantitative data; compare design alternatives)

There are advantages and disadvantages to each of these evaluative techniques (see Table 8) (Preece 1993, Karat 1997). Thus, a combination of methods is often used in practice. Typically, one would first perform an expert evaluation (e.g., heuristic evaluation) of a system to identify the most obvious usability problems. Then user testing could be conducted to identify remaining problems that were missed in the first stages of evaluation. In general, a number of factors need to be considered when selecting a usability-evaluation technique or a combination thereof (see Table 9) (Dix et al. 1993; Nielsen 1993; Preece 1993).

As technology has evolved, there has been a shift in the technoeconomic paradigm, allowing for more universal access of computer technology (Stephanidis and Salvendy 1998). Thus, individuals with diverse abilities, requirements, and preferences are now regularly utilizing interactive products. When designing for universal access, participation of diverse user groups in usability evaluation is essential. Vanderheiden (1997) has suggested a set of principles for universal design that focuses on the following: simple and intuitive use; equitable use; perceptible information; tolerance for error; accommodation of preferences and abilities; low physical effort; and space for approach and use. Following these principles should ensure effective design of interactive products for all user groups.

While consideration of ergonomic and cognitive factors can generate effective interactive system designs, if the design has not taken into consideration the environment in which the system will be used, it may still fail to be adopted. This will be addressed in the next section.

4. SOCIAL, ORGANIZATIONAL, AND MANAGEMENT FACTORS

Social, organizational, and management factors related to human–computer interaction may influence a range of outcomes at both the individual and organizational levels: stress, physical and mental health, safety, job satisfaction, motivation, and performance. Campion and Thayer (1985) showed that some of these outcomes may be conflicting. In a study of 121 blue-collar jobs, they found that enriched jobs led to higher job satisfaction but lower efficiency and reliability. The correlations between efficiency on one hand and job satisfaction and comfort on the other hand were negative. Another way of looking at all the outcomes has been proposed by Smith and Sainfort (1989). The objective of the proposed balance theory is to achieve an optimal balance among positive and negative aspects of the work system, including the person, task and organizational factors, technology, and physical environment. See Figure 1 for a model of the work system. The focus is not on a limited range of variables or aspects of the work system but on a holistic approach to the study and design of work systems. In this section we will focus on how social, organizational and management factors related to human–computer interaction influence both individual and organizational outcomes.

4.1. Social Environment

The introduction of computer technology into workplaces may change the social environment and social relationships. Interactive systems become a new element of the social environment, a new communication medium, and a new source of information. With new computer technologies there may be a shift from face-to-face interaction toward computer-mediated communication, or at least a change in how people communicate and interact. This shift may be most obvious with electronic mail and teleconferencing systems. A study of Eveland and Bikson (1988) on electronic mail shows that people connected to a network of microcomputers with electronic mail relied more on scheduled meetings than people with conventional office support, who relied more on unscheduled meetings. The impact on face-to-face interaction was not studied. A study of electronic mail by Rice and Case (1983) did not find any reduction in face-to-face interaction, but increased communications as a result of using electronic mail. Computers seem to be just another way of communicating with coworkers, subordinates, and supervisors (Rice and Case 1983). However, other studies have found that there was not only a change in quantity of communications (more or new information to and from more or new recipients), but also a change in quality of communications (Kiesler et al. (1984).

Aydin (1989) showed that the use of medical information systems for communicating physicians' medication orders from the nursing department to the pharmacy led to increased interdependence between the two departments. The change in work organization (increased interdependence) was accompanied by changes in the social environment: communication and cooperation between the two departments improved, leading to better working relationships.

Computer technologies also allow work to be performed at remote distances. Recently there has been an increase in home-based work due to technological advances in computer and network technologies. Telework or working at home is most common for clerical workers performing routine transactions and for autonomous professionals (e.g., writers, designers) (Sproull and Kiesler 1991). In general, home-based work increases social isolation from coworkers and supervisors. This not only reduces opportunities to socialize and make friends but also reduces chances for advancement

TABLE 8 Advantages and Disadvantages of Existing Usability Evaluation Techniques

Evaluation Method	Example Tools/Techniques	General Use	Advantages	Disadvantages
Analytic/theory-based	• Cognitive Task Analysis • GOMS	Used early in usability design life cycle for prediction of expert user performance.	Useful in making accurate design decisions early in the usability life cycle without the need for a prototype or costly user testing.	Narrow in focus; lack of specific diagnostic output to guide design; broad assumption on users' experience (expert) and cognitive processes; results may differ based on the evaluators' interpretation of the task.
Expert Evaluation	• Design walk-throughs • Heuristic evaluations • Process/system Checklists • Free play • Group evaluations	Used early in the design life cycle to identify *theoretical* problems that may pose actual *practical* usability problems.	Strongly diagnostic; can focus on entire system; high potential return in terms of number of usability issues identified; can assist in focusing observational evaluations.	Even the best evaluators can miss significant usability issues; results are subject to evaluator bias; does not capture real user behavior.
Observational evaluation	• Direct observation • Video • Verbal protocols • Computer logging • Think aloud techniques • Field evaluations • Ethnographic studies • Facilitated free play	Used in iterative design stage for problem identification.	Quickly highlights usability issues; verbal protocols provide significant insights; provides rich qualitative data.	Observation can affect user performance with the system; analysis of data can be time and resource consuming.

Survey evaluation	• questionnaires • Structured interviews • Ergonomics checklists • Focus groups	Used any time in the design life cycle to obtain information on users' preferences and perception of a system.	Provides insights into users' opinions and understanding of the system; can be diagnostic; rating scales can provide quantitative data; can gather data from large subject pools.	User experience important; possible user response bias (e.g., only dissatisfied users respond); response rates can be low; possible interviewer bias; analysis of data can be time and resource consuming; evaluator may not be using appropriate checklist to suit the situation.
Psychophysiological measures of satisfaction or workload	• EEGs • Heart rate • Blood pressure • Pupil dilation • Skin conductivity • Level of adrenaline in blood	Used any time in the design life cycle to obtain information on user satisfaction or workload.	Eliminate user bias by employing objective measures of user satisfaction and workload.	Invasive techniques are involved that are often intimidating and expensive for usability practitioners.
Experimental evaluation	• Quantitative measures • Alternative design comparisons • Free play • Facilitated free play	Used for competitive analysis in final testing of the system.	Powerful and prescriptive method; provides quantitative data; can provide a comparison of alternatives; reliability and validity generally good.	Experiment is generally time and resource consuming; focus can be narrow; tasks and evaluative environment can be contrived; results difficult to generalize.

TABLE 9 Factors to Consider in Selecting Usability-Evaluation Techniques

Purpose of the evaluation
Stage in the development life cycle in which the evaluation technique will be carried out
Required level of subjectivity or objectivity
Necessity or availability of test participants
Type of data that need to be collected
Information that will be provided
Required immediacy of the response
Level of interference implied
Resources that may be required or available

and promotion (OTA 1985). On the other hand, home-based work allows workers to spend more time with their family and friends, thus increasing social support from family and friends. Telework allows for increased control over work pace and variability of workload. It has been found, however, that electronic communication and telework have led to feelings of not being able to get away from work and to the augmentation (rather than substitution) of regular office hours (Sproull and Kiesler 1991; Phizacklea and Wolkowitz 1995). In addition, increased distractions and interruptions may disrupt work rhythm (OTA 1985). From a social point of view, home-based work has both negative and positive effects.

Another important social factor is intragroup relationships and relationships with coworkers. The role of computer technologies in influencing intragroup functioning is multidimensional. If workers spend a lot of time working in isolation at a computer workstation, they may have less opportunity for socialization. This may affect the group performance, especially if tasks are interdependent. On the other hand, intragroup relationships may be improved if workers gain access to better information and have adequate resources to use computers. The positive or negative effects may also vary across jobs and organizations. And they may depend on the characteristics of the interactive system (e.g., single- vs. multiple-user computer workstation).

Aronsson (1989) found that work group cohesion and possibilities for contacts with coworkers and supervisors had become worse among low-level jobs (e.g., secretary, data-entry operator, planner, office service) but had not been affected among medium- to high-level jobs. Changes in job design were related to changes in social relationships. The higher the change in intensity demands, the lower the work group cohesion and the possibilities for contacts with coworkers and supervisors. That is, increase in workload demands came along with worsening of the social environment. The negative effect was more pronounced among low-level jobs, presumably because higher-level job holders have more resources, such as knowledge, power, and access to information and can have a say in the implementation/design process as well as more control over their job.

Access to organizational resources and expertise is another important facet of the social environment for computer users. Technology can break down or malfunction, and users may need help to perform certain tasks or to learn new software. In these situations, access to organizational resources and expertise is critical for the end users, especially when they are highly dependent on the computer technology to perform their job or when they use the technology in their contact with customers. Danziger et al. (1993) have studied the factors that determine the quality of end-user computing services in a survey of 1869 employees in local governments. Three categories of factors were identified that might influence the quality of computing services: (1) the structure of service provision (e.g., centralization vs. decentralization), (2) the level of technological problems, and (3) the service orientation of computing service specialists. The results do not provide support for the argument that structural factors are most important; whether computing services are centralized or decentralized within an organization does not explain the perceived quality of computing services.

On the other hand, the results demonstrate the importance of the attitudes of the service providers. Computer specialists who are clearly user oriented, that is, who are communicative and responsive to user needs and are committed to improving existing applications and proposing appropriate new ones, seem best able to satisfy end users' criteria for higher quality computing services. Researchers emphasize the importance of a positive sociotechnical interface between end users and computing specialists, in addition to good operational performance (e.g., low incidence of technical problems).

The introduction of computers in workplaces can also change how workers interact with their supervisor and management. That change in social interaction may result in changes in social support. Sproull and Kiesler (1988) showed that electronic mail affected social relationships within organizations. Status equalization was observed in that messages from subordinates were no different than messages from supervisors. Thus, computer technologies can have positive effects on worker–management relationships because workers have easier access to their supervisors and/or feel less restrained from communicating with their supervisors. However, expectations of rapid service and faster work completion may impose more supervisory pressure on workers (Johansson and Aronsson 1984). This is a negative effect of computer technologies on worker–supervisor relationships. A study by Yang and Carayon (1995) showed that supervisor support was an important buffer against worker stress in both low and high job demands conditions. Two hundred sixty-two computer users of three organizations participated in the study. Supervisor social support was an important buffer against worker stress; however, coworker social support did not affect worker stress.

The social environment can be influenced by computer technologies in various ways: quality, quantity, and means of communications, social isolation, extended network of colleagues, work group cohesion, quality of social interaction among workers, coworkers and supervisors, and social support. Table 10 summarizes the potential effects of computer technologies on the social environment. There are several strategies or interventions that can be applied to counteract the negative influences of computer technology on the social environment and foster positive effects.

Computerized monitoring systems have an important impact on how supervisors interact with their employees. It is natural that when supervisors are suddenly provided with instantaneous, detailed information about individual employee performance, they feels a commitment, in fact an obligation, to use this information to improve the performance of the employees. This use of hard facts in interacting with employees often changes the style of supervision. It puts inordinate emphasis on hourly performance and creates a coercive interaction. This is a critical mistake in a high-technology environment where employee cooperation is essential.

Supervision has to be helpful and supportive if employee motivation is to be maintained and stress is to be avoided. This means that supervisors should not use individual performance data as a basis for interaction. The supervisor should be knowledgeable about the technology and serve as a resource when employees are having problems. If management wants employees to ask for help, the relationship with the supervisor has to be positive (not coercive) so that the employee feels confident enough to ask for help. If employees are constantly criticized, they will shun the supervisor and problem situations that can harm productivity will go unheeded.

Employees are a good source of information about productive ways to work. Their daily contact with the job gives them insight into methods, procedures, bottlenecks, and problems. Many times they modify their individual work methods or behavior to improve their products and rate of output. Often these are unique to the individual job or employee and could not be adopted as a standardized approach or method. If work systems are set up in a rigid way, this compensatory behavior cannot occur. Further, if adverse relationships exist between supervisors and employees, the employees are unlikely to offer their innovative ideas when developers are conducting a contextual task analysis (see Section 3.3). It is in the interest of the employer to allow employees to exercise at least a nominal level of control and decision making over their own task activity. Here again, the computer hardware and software have to be flexible so that individual approaches and input can be accommodated as long as set standards of productivity are met.

One approach for providing employee control is through employee involvement and participation in making decisions about interactive system design—for instance, by helping management select ergonomic furniture through comparative testing of various products and providing preference data, or being involved in the determination of task allocations for a new job, or voicing opinions about ways to improve the efficiency of their work unit. Participation is a strong motivator to action and a good way to gain employee commitment to a work standard or new technology. Thus, participation can be used as a means of improving the social environment and foster the efficient use of interactive

TABLE 10 Potential Effects of Computer Technologies on the Social Environment

Less face-to-face interaction
More computer-mediated communication
Change in the quality of communications (status equalization, pressure)
Increased or decreased interdependence between departments/work units
Increased or decreased cooperation between departments/work units
Increased or decreased opportunities for contacts with coworkers and supervisor
Increased quantity/quality of information
Increased or decreased work group cohesion
Home-based work: Isolation Reduced chances for advancement and promotion Increased/decreased social support

systems. But participation will only be effective as long as employees see tangible evidence that their input is being considered and used in a way that benefits them.

Employees who make positive contributions to the success of the organization should be rewarded for their efforts. Rewards can be administrative, social, or monetary. Administrative rewards can be such things as extra rest breaks, extended lunch periods, and special parking spaces. They identify the person as someone special and deserving. Another type of reward is social in that it provides special status to the individual. This is best exemplified by the receipt of praise from the supervisor for a job well done. This enhances personal self-esteem. If the praise is given in a group setting, it can enhance peer group esteem toward the individual. Monetary rewards can also be used, but these can be a double-edged sword because they may have to be removed during low-profit periods, and this can lead to employee resentment, thus negating the entire purpose of the reward system. Some organizations use incentive pay systems based on performance data provided by the computers. Computers can be used to keep track of worker performance continuously (Carayon 1993). That quantitative performance data can then be used to set up incentive pay systems that reward good performers. In general, incentive pay systems can lead to increase in output but at the expense of worker health (Levi 1972). Schleifer and Amick (1989) have shown how the use of a computer-based incentive system can lead to an increase in worker stress.

Different ways of improving the social environment in computerized workplaces thus include helpful and supportive managers and supervisors, increased control over one's job, employee involvement and participation, and rewards.

4.2. Organizational Factors

The way work is organized changes with the introduction of computer technologies, such as changes in workflow. Computer technologies obviously provide opportunities for reorganizing how work flows and have the potential of increasing efficiency. However, increased worker dependence on the computer is a potential problem, especially when the computer breaks down or slows down. It may affect not only performance but also stress. Organizational redesign may be one way of alleviating problems linked to dependence on the computer. Aronsson and Johansson (1987) showed that organizational rearrangement was necessary to decrease workers' dependence on the computer system by expanding their jobs with new tasks and allowing them to rotate between various tasks.

Given their technical capabilities, computers can be used to bring people closer and make them work in groups. The concept of computer-supported cooperative work is based on the expectation that the computer favors group work. Researchers in this area focus on all aspects of how large and small groups can work together in using computer technology (Greif 1988). They develop interactive systems that facilitate group work and study the social, organizational, and management impacts of computer-supported work groups. For instance, Grief and Sarin (1988) identified data-management requirements of computer group work.

New computer technologies allow work to be performed at a distance. This new work organization has some potential negative and positive effects for workers and management. Benefits for workers include increased control over work schedule and eliminating the commute to work (OTA 1985; Bailyn 1989). Constraints for workers include social isolation, increased direct and indirect costs (e.g., increased heating bill, no health insurance), lack of control over physical environment, and fewer opportunities for promotion (OTA 1985; Bailyn 1989). Benefits for employers include lowered costs (e.g., floor space, direct labor costs, and workers' benefits), more intensive use of computers (e.g., outside peak hours), increased flexibility (workers can be used when needed), and increased productivity; while problems include change in traditional management and supervision techniques and loss of control (OTA 1985).

Within organizations, the use of computer technologies has been linked to various positive and negative effects on job design (see, e.g., the case study of Buchanan and Boddy 1982). Increased workload, work pressure and demand for concentration, decreased job control, and variety are some of the negative effects (Smith et al. 1981; Buchanan and Boddy 1982; Johansson and Aronsson 1984). Increased feedback, control over one's job, and career opportunities are some of the positive effects (Buchanan and Boddy 1982). For some, such as professionals, the job-design effects of the use of computer technology may be all positive, while for others, such as clerical workers or data-entry operators, the effects may all be negative (Smith et al. 1981; Sauter et al. 1983; Johansson and Aronsson 1984).

The computer technology itself may have characteristics that can affect worker stress by inducing negative characteristics. For instance, technology characteristics such as breakdown and slowdown may increase perceived workload and work pressure and reduce the amount of control one has over work (Carayon-Sainfort 1992; Asakura and Fujigaki 1993). Carayon-Sainfort (1992) found that computer system performance was indirectly related to stress through its effect on perceived workload, work pressure and job control. Specifically, greater frequencies of computer problems were related to increases in perceived workload and work pressure as well as decreases in job control. These can have negative effects on an organization. Asakura and Fujigaki (1993) examined the direct and indirect effects of computerization on worker well being and health in a sample of 4400 office workers. The results of their study paralleled Carayon-Sainfort (1992).

A major complaint of office employees who have undergone computerization is that their workload has increased substantially. This is most true for clerical employees, who typically have an increased number of transactions to process when computers are introduced into the work routine. This increase in transactions means more keystrokes and more time at the workstation. These can lead to greater physical effort than before and possibly more visual and muscular discomfort. This discomfort reinforces the feeling of increased workload and adds to employee dissatisfaction with the workload.

Quite often the workload of computer users is established by the data-processing department in concert with other staff departments such as human resources and line managers. An important consideration is the cost of the computer equipment and related upkeep such as software and maintenance. The processing capability of the computer(s) is a second critical element in establishing the total capacity that can be achieved. The technology cost, the capability to process work, and the desired time frame to pay for the technology are factored together to establish a staffing pattern and the required workload for each employee. This approach is based on the capacity of the computer(s) coupled with investment recovery needs and does not necessarily meet the objective of good human resource utilization. Workload should not be based solely on technological capabilities or investment recovery needs but must include important considerations of human capabilities and needs. Factors such as attentional requirements, fatigue, and stress should be taken into account in establishing the workload. A workload that is too great will cause fatigue and stress that can diminish work quality without achieving desired quantity. A workload that is too low will produce boredom and stress and also reduce quality and economic benefits of computerization.

Workload problems are not concerned solely with the immediate level of effort necessary but also deal with the issue of work pressure. This is defined as an unrelenting backlog of work or workload that will never be completed. This situation is much more stressing than a temporary increase in workload to meet a specific crisis. It produces the feeling that things will never get better, only worse. Supervisors have an important role in dealing with work pressure by acting as a buffer between the demands of the employer and the daily activities of the employees. Work pressure is a perceptual problem. If the supervisor deals with daily workload in an orderly way and does not put pressure on the employee about a pile-up of work, then the employee's perception of pressure will be reduced and the employee will not suffer from work pressure stress.

Work pressure is also related to the rate of work, or work pace. A very fast work pace that requires all of the employee's resources and skills to keep up will produce work pressure and stress. This is exacerbated when this condition occurs often. An important job-design consideration is to allow the employee to control the pace of the work rather than having this controlled automatically by the computer. This will provide a pressure valve to deal with perceived work pressure.

A primary reason for acquiring new technology is to increase individual employee productivity and provide a competitive edge. Getting more work out of employees means that fewer are needed to do the same amount of work. Often employees feel that this increased output means that they are working harder even though the technology may actually make their work easier. Using scientific methods helps establish the fairness of new work standards.

Once work standards have been established, they can serve as one element in an employee-performance-evaluation scheme. An advantage of computer technology is the ability to have instantaneous information on individual employee performance in terms of the rate of output. This serves as *one* objective measure of how hard employees are working. But managers have to understand that this is just one element of employee performance and emphasis on quantity can have an adverse effect on the quality of work. Therefore, a balanced performance-evaluation system will include quality considerations as well. These are not as easy to obtain and are not as instantaneously available as are quantity measures. However, managers must resist the temptation to emphasize quantity measures just because they are readily available. A key consideration in any employee evaluation program is the issue of fairness, just as in workload determination.

Jobs in which people use computer technology may require high mental effort. Some types of computer-mediated tasks may increase information-processing requirements and place great demands on attention, decision making, and memory. Increased levels of cognitive demands due to computer technology have been shown to influence employee stress and health (Lindstrom and Leino 1989; Czaja and Sharit 1993; Yang 1994). Several types of cognitive demands can be generated from the use of computer technology: (1) a great amount of information given in a certain unit of time, (2) abstract information being presented on the screen, and (3) difficult and concurrent tasks being performed at the same time.

Cognitive demands can be increased when the system response time is poor and the nature of workflow is not transparent to the workers. In other words, unpredictable demands and interruptions of workflow caused by system breakdowns may be difficult to deal with because of the disruptive effect on the cognitive control process. Overall, cognitive demands are associated with job characteristics such as intensity of computer work, the type of communication, and high speed/functions of computers. The implementation of computer technology in work organizations can lead to greater demands on cognitive resources in terms of memory, attention, and decision making that may have a negative impact on worker health and work performance. If, however, computer systems have been designed with the cognitive capabilities and limitations of the user in mind (see Section 3), these issues should not occur.

There has been interest in the role of occupational psychosocial stress in the causation and aggravation of musculoskeletal disorders for computer users (Smith et al. 1981; Smith 1984; Bammer and Blignault 1988; Smith and Carayon 1995; Hagberg et al. 1995). It has been proposed that work organization factors define ergonomic risks to upper-extremity musculoskeletal problems by specifying the nature of the work activities (variety or repetition), the extent of loads, the exposure to loads, the number and duration of actions, ergonomic considerations such as workstation design, tool and equipment design, and environmental features (Smith and Carayon 1995; Carayon et al. 1999). These factors interact as a system to produce an overall load on the person (Smith and Sainfort 1989; Smith and Carayon 1995; Carayon et al. 1999), and this load may lead to an increased risk for upper extremity musculoskeletal problems (Smith and Carayon 1995; Carayon et al. 1999). There are psychobiological mechanisms that make a connection between psychological stress and musculoskeletal disorders plausible and likely (Smith and Carayon 1995; Carayon et al. 1999). At the organizational level, the policies and procedures of a company can affect the risk of musculoskeletal disorders through the design of jobs, the length of exposures to stressors, establishing work–rest cycles, defining the extent of work pressures and establishing the psychological climate regarding socialization, career, and job security (Smith et al. 1992; NIOSH 1992, 1993).

Smith et al. (1992), Theorell et al. (1991) and Faucett and Rempel (1994) have demonstrated that some of these organizational features can influence the level of self-reported upper-extremity musculoskeletal health complaints. In addition, the organization defines the nature of the task activities (work methods), employee training, availability of assistance, supervisory relations, and workstation design. All of these factors have been shown to influence the risk of upper-extremity musculoskeletal symptoms, in particular among computer users and office workers (Linton and Kamwendo 1989; Smith et al. 1992; Lim et al. 1989; Lim and Carayon. 1995; NIOSH 1990, 1992, 1993; Smith and Carayon 1995).

The amount of esteem and satisfaction an employee gets from work are tied directly to the content of the job. For many jobs, computerization brings about fragmentation and simplification that act to reduce the content of the job. Jobs need to provide an opportunity for skill use, mental stimulation, and adequate physical activity to keep muscles in tone. In addition, work has to be meaningful for the individual. It has to provide for identification with the product and the company. This provides the basis for pride in the job that is accomplished.

Computerization can provide an opportunity for employees to individualize their work. This lets them use their unique skills and abilities to achieve the required standards of output. It provides cognitive stimulation because each employee can develop a strategy to meet his or her goals. This requires that software be flexible enough to accept different types and order of input. Then it is the job of the software to transform the diverse input into the desired product. Usually computer programmers will resist such an approach because it is easier for them to program using standardized input strategies. However, such strategies build repetition and inflexibility into jobs that reduce job content and meaning.

Being able to carry out a complete work activity that has an identifiable end product is an important way to add meaningfulness to a job. When employees understand the fruits of their labor, it provides an element of identification and pride in achievement. This is in contrast to simplifying jobs into elemental tasks that are repeated over and over again. Such simplification removes meaning and job content and creates boredom, job stress, and product-quality problems. New computer systems should emphasize software that allows employees to use existing skills and knowledge to start out. These then can serve as the base for acquiring new skills and knowledge. Job activities should exercise employee mental skills and should also require a sufficient level of physical activity to keep the employee alert and in good muscle tone.

Table 11 summarizes the potential impacts of computer technologies on organizational factors. Overall, the decision about the use or design of interactive systems should include considerations for work load, work pressure, determination of work standards, job content (variety and skill use), and skill development. Computerization holds the promise of providing significant improvements in the quality of jobs, but it also can bring about organizational changes that reduce employee satisfaction and performance and increase stress. Designing interactive systems that meet both the aims of the organization and the needs of employees can be difficult. It requires attention to important aspects of work that contribute to employee self-esteem, satisfaction, motivation, and health and safety.

TABLE 11 Potential Effects of Computer Technologies on Organizational Factors

- Job design:
    - Increased/decreased workload and work pressure
    - Increased demand for concentration
    - Increased/decreased job control and autonomy
    - Increased/decreased variety
    - Increased feedback
- Increased work efficiency
- Computer malfunctions and breakdowns
- Computer-supported work group
- Home-based work
- Electronic monitoring of worker performance
- Incentive pay systems

4.3. Management Factors

Consideration of management factors in human–computer interaction is highly relevant in understanding the global effects of interactive systems (Clement and Gotlieb 1988). The introduction of computer technology is often accompanied by or responsible for changes in management structure. For instance, computer technologies can be used to increase workers' access to information. That move toward decentralization can lead to more decisions being made at lower levels. There has been a long debate about whether computer technology leads to centralization or decentralization of decision making (Attewell and Rule 1984; Blackler 1988). There is no clear answer to this debate: Variables such as organizational size, past experiences, management style, and work force skill level play a role in these structural effects (Attewell and Rule 1984). Furthermore, power may not be a simple zero-sum relationship. Various organizational actors may experience increased power and control opportunities after the implementation of computer technology. Information systems specialists increase their power because they have valuable expertise and knowledge, and workers may depend on them when a technical problem occurs or when they need additional training. Worker control may also increase when workers are given efficient technologies and are taught new computer skills.

The amount of information and the ease of access to information are important management factors affected by computer technologies. Electronic mail systems tend to change how information flows in organizations. Sproull and Kiesler (1988) found that electronic mail added new recipients to information being circulated and also added new information. However, one could ask about the usefulness and relevancy of the new information for organizational functioning. Information has been identified as a potent source of power (Crozier 1964). Computer technology that changes the type and amount of information available is likely to change the power distribution between various organizational actors, such as workers, supervisors, managers, computer experts, and unions. In addition, the introduction of computer technologies may create new sources of power and increase status differences between computer experts and nonexperts, between heavy computer users and light users.

Computer technologies can be used for increasing management control over production/service processes. Electronic monitoring of worker performance is an example of this effect. Computers are used to get detailed online data on worker performance to, for instance, improve work schedule and planning and increase control over worker performance. This may greatly enhance management capabilities and improve overall organizational effectiveness, but may induce stressful working conditions (Carayon 1993).

Smith et al. (1992) conducted a questionnaire survey study examining the differences in stress responses between employees who were electronically monitored while doing computer work and those who were not. Both groups performed the same jobs. The results of the surveys completed by 745 telecommunication employees showed that employees who had their performance electronically monitored perceived more stressful working conditions and more job boredom, psychological tension, anxiety, depression, anger, health complaints, and fatigue. Smith et al. (1992) suggest that the results might have been due to job-design changes associated with the monitoring.

In fact, when Carayon (1994) reanalyzed data from two job categories (255 service representatives and 266 clerks) from Smith et al. (1992), the results supported the proposition that electronic performance monitoring had an indirect effect on worker stress through its effects on job design. Carayon (1994) also reported on a second study to specifically examine whether or not electronic performance monitoring had a direct or indirect effect on worker stress. The results revealed that monitored employees reported more supervisor feedback and control over work pace and less job content than nonmonitored employees. There were no differences between the monitored and nonmonitored groups with regard to stress or health.

The process by which computer technologies are implemented is only one of the management factors that affect the effectiveness and acceptance of computer use. Management attitudes toward the implementation of computer technologies are very important insofar as they can affect overall job and organizational design and worker perceptions, attitudes, and beliefs regarding the new technologies (Crozier 1964; Smith 1984; Blackler 1988; Kahn and Cooper 1986). Several models that link the organizational effects of computer systems to the process used to implement those systems have been proposed (Blackler and Brown 1987; Robey 1987; Flynn 1989). They all emphasize the need to identify potential technical and social impacts, advocate general planning, and emphasize the support of workers and managers for successful implementation of new computer technology.

Carayon and Karsh (2000) examined the implementation of one type of computer technology, that is, imaging technology into two organizations in the Midwest. Results showed that imaging users in the organization that utilized end-user participation in the implementation of their imaging system rated their imaging systems better and reported higher job satisfaction than imaging users in the organization that did not incorporate end-user participation in the implementation of the system. Studies by Korunka and his colleagues (Korunka et al. 1993, 1995, 1996) have also demonstrated the benefits of end-user participation in technological change on quality of working life, stress, and health.

Management needs to also consider retraining issues when introducing new computer technology. Kearsley (1989) defined three general effects of computer technology: skill twist (change in required skills), deskilling (reduction in the level of skills required to do a job), and upskilling (increase in the level of required skills). Each of these effects has different implications for retraining. For instance, skill twist requires that workers be able and willing to learn new skills. Training or retraining are critical issues for the successful implementation and use of new computer technology. Even more critical is the need for continuous retraining because of rapid changes in hardware and software capabilities of computer technologies (Smith et al. 1981; Smith 1984; Kearsley 1989). Training can serve to enhance employee performance and add new skills. Such growth in skills and knowledge is an important aspect of good job design. No one can remain satisfied with the same job activities over years and years. Training is a way to assist employees in using new technology to its fullest extent and reduce the boredom of the same old job tasks. New technology by its nature will require changes in jobs, and training is an important approach not only for keeping employees current but also in building meaning and content into their jobs.

Computer technologies have the potential to affect both positively and negatively the following management factors: organizational structure (e.g., decentralization vs. centralization), power distribution, information flow, and management control over the production process. Management's strategies for implementing new computer technologies are another important management factor to take into account to achieve optimal use of these technologies. Table 12 summarizes the potential impacts of computer technologies on management factors. Some of the negative effects of computers on management factors can be counteracted. The rest of this section proposes various means of ensuring that computerization leads to higher performance and satisfaction and lower stress.

Monitoring employee performance is a vital concern of labor unions and employees. Computers provide greatly enhanced capability to track employee performance, and this will follow from such close monitoring. Monitoring of employee performance is an important process for management. It helps to know how productive your workforce is and where bottlenecks are occurring. It is vital management information that can be used by top management to realign resources and to make important management decisions. However, it is not a good practice to provide individual employee performance information to first-line supervisors; it can lead to a coercive supervision style. To

TABLE 12 Potential Effects of Computer Technologies on Management Factors

Decentralization vs. centralization of decision making
Flow and amount of information
Management control over work process
Implementation of technological change
Training
Electronic monitoring of worker performance
Job security
Career development

enhance individual performance, it is helpful to give periodic feedback directly to employees about their own performance. This can be done in a noncoercive way directly by the computer on a daily basis. This will help employees judge their performance and also assist in establishing a supervisory climate that is conducive to satisfied and productive employees.

While computerized monitoring systems can be particularly effective in providing employees with feedback, the misuse of such systems can be particularly counterproductive and cause stress. The following principles contribute to the effective use of computerized monitoring for performance enhancement and reduced stress:

- Supervisors should not be involved directly in the performance feedback system. Information on the performance that is given by the computerized monitoring system should be directly fed back to the operator.
- Computerized monitoring systems should give a comprehensive picture of the operator's performance (quantity *and* quality).
- Performance information should not be used for disciplinary purposes.
- Electronic monitoring should not be used for payment purposes such as payment by keystrokes (piece rate) or bonuses for exceeding goals.

Any kind of change in the workplace produces fears in employees. New technology brings with it changes in staff and the way work is done. The fear of the unknown, in this case the new technology, can be a potent stressor. This suggests that a good strategy in introducing new technology is to keep employees well informed of expected changes and how they will affect the workplace. There are many ways to achieve this. One is to provide informational memorandums and bulletins to employees at various stages of the process of decision making about the selection of technology and, during its implementation, on how things are going. These informational outputs have to be at frequent intervals (at least monthly) and need to be straightforward and forthright about the technology and its expected effects. A popular approach being proposed by many organizational design experts is to involve employees in the selection, design, and implementation of the new technology. The benefit of this participation is that employees are kept abreast of current information, employees may have some good ideas that can be beneficial to the design process, and participation in the process builds employee commitment to the use of the technology.

A large employee fear and potent stressor is concern over job loss due to improved efficiency produced by new technology. Many research studies have demonstrated that the anticipation of job loss and not really knowing if you will be one of the losers is much more stressful and more detrimental to employee health than knowing right away about future job loss. Telling those employees who will lose their jobs early provides them with an opportunity to search for a new job while they still have a job. This gives them a better chance to get a new job and more bargaining power regarding salary and other issues. Some employers do not want to let employees know too soon for fear of losing them at an inopportune time. By not being fair and honest to employees who are laid off, employers can adversely influence the attitudes and behavior of those employees who remain.

For those employees who are retained when new technology is acquired, there is the concern that the new technology will deskill their jobs and provide less opportunity to be promoted to a better job. Often the technology flattens the organizational structure, producing similar jobs with equivalent levels of skill use. Thus, there is little chance to be promoted except into a limited number of supervisory positions, which will be less plentiful with the new technology. If this scenario comes true, then employees will suffer from the ''blue-collar blues'' that have been prevalent in factory jobs. This impacts negatively on performance and stress.

Averting this situation requires a commitment from management to enhancing job design that builds skill use into jobs as well as developing career paths so that employees have something to look forward to besides 30 years at the same job. Career opportunities have to be tailored to the needs of the organization to meet production requirements. Personnel specialists, production managers, and employees have to work together to design work systems that give advancement opportunity while utilizing technology effectively and meeting production goals. One effective technique is to develop a number of specialist jobs that require unique skills and training. Workers in these jobs can be paid a premium wage reflecting their unique skills and training. Employees can be promoted from general jobs into specialty jobs. Those already in specialty jobs can be promoted to other, more difficult specialty jobs. Finally, those with enough specialty experience can be promoted into troubleshooting jobs that allow them to rotate among specialties as needed to help make the work process operate smoothly and more productively.

Organizations should take an active role in managing new computer technologies. Knowing more about the positive and negative potential effects or influences of computerization on management factors is an important first step in improving the management of computer technologies.

4.4. An International Perspective

In order to increase the market for their products and services and thus gain increasing profitability and, where appropriate, shareholder value, corporations are penetrating the international market. This requires a number of adjustments and considerations by corporations, including consideration of standard of living, prevailing economies, government incentives and public policies, and practices in the country where products and services will be marketed. In addition, it is important to consider the characteristics of the individuals in the country where products and services will be utilized, such as differences in anthropometric (body size), social, and psychological considerations. Table 13 illustrates with regard to computer products designed in the United States and the changes that need to be made for Chinese in Mainland China (Choong and Salvendy 1998, 1999; Dong and Salvendy 1999a,b). If both versions of the product were produced, both U.S. and Chinese users would be expected to achieve the fastest possible performance time with the lowest error rates. Identifying a local expert and following international standards (Cakir and Dzida 1997) can assist in identifying the modifications required to ensure a product is well suited to each international target market.

5. ITERATIVE DESIGN

Interactive systems are meant to make work more effective and efficient so that employees can be productive, satisfied, and healthy. Good design improves the motivation of employees to work toward the betterment of the employer. The consideration of ergonomic, cognitive, social, organizational, and management factors of interactive system design must be recognized as an iterative design process. By considering these factors in an iterative manner, system designs can evolve until the desired level of performance and safety are achieved. Additional modifications and resources expenditure will then be unnecessary. This allows valuable resources to be saved or applied to other endeavors. Table 14 provides a list of general guidelines as to how these factors can be designed to create an effective, productive, healthy, and satisfying work environment.

The concept of balance is very important in managing the design, introduction, and use of computer technologies (Smith and Sainfort 1989). Negative effects or influences can be counteracted by positive aspects. For instance, if the physical design of the technology cannot be changed and is known to be flawed, decision makers and users could counteract the negative influences of such design by, for instance, providing more control over the work–rest schedule. By having control over their work–rest schedule, workers could relieve some of the physical stresses imposed by the technology by moving around. If management expects layoffs due to the introduction of computers, actions should be taken to ensure that workers are aware of these changes. Sharing information and getting valuable training or skills could be positive ways to counteract the negative effects linked to layoff. Carayon (1994) has shown that office and computer jobs can be characterized by positive and negative aspects and that different combinations of positive and negative aspects are related to different strain outcomes. A job with high job demands, but positive aspects such as skill utilization, task clarity, job control and social support, led to low boredom and a medium level of daily life stress. A job with many negative aspects of work led to high boredom, workload dissatisfaction, and daily life stress.

Another important aspect of the iterative design process is time. Changes in the workplace occur at an increasing pace, in particular with regard to computer technology. The term *continuous change* has been used to characterize the fast and frequent changes in computer technology and its impact on people and organizations (Korunka et al. 1997; Korunka and Carayon 1999). The idea behind this is that technological change is rarely, if ever, a one-time shot. Typically, technology changes are closer to continuous rather than discrete events. This is because rapid upgrades and reconfigurations to make the systems work more effectively are usually ongoing. In addition to technological changes, time has other important effects on the entire work system (Carayon 1997). In particular, the aspects of the computerized work system that affect people may change over time. In a longitudinal study

TABLE 13 Differences in Design Requirements between Chinese and American Users

Attribute	Chinese	American
Knowledge regeneration	Abstract	Concrete
Base menu layout	Thematic	Functional
Menu layout	Vertical	Horizontal
Cognitive style	Relational–conceptual	Inferential–categorical
Thinking	Relational	Analytical
Field	Dependent	Independent
Translation from English to Chinese	Dynamics	N/A

TABLE 14 Guidelines for Designing Effective Interactive Systems

Helpful and supportive supervision
Allow job task characteristics to define ergonomic interventions
Appropriate consideration of ergonomic factors, including the technology, sensory environment, thermal environment, workstation design, and work practices
Consider light-related environmental factors, device use and posture factors, environmental factors, job design factors, individual user factors
Flexibility of hardware and software design
Focusing on target user groups in design
Employee involvement and participation in decision making (design, purchasing, and implementation of hardware and software)
Implementation of sound design practices, including the use of requirements definition, user profile development, tasks analysis, and task allocation
Setting usability objectives that focus around effectiveness, intuitiveness, and subjective perception
Iterative usability evaluation
Identification of users' mental models
Identification of appropriate metaphors
Effective integration of tasks into design via use scenarios, use sequences, use flow diagrams, use workflows, and use hierarchies
Design of reward systems (administrative, social, and monetary rewards)
Setup of workload standards (scientific basis, fairness, employee involvement, and acceptance)
Job enlargement and enrichment
Electronic monitoring of worker performance (fairness, feedback to employees, supervisory style, privacy)
Continuous communication between employees and management
Implementation of change
Development of career paths
Systemic approach (organizational culture, past experiences, long-term vs. short-term approach)
Balanced approach
Monitoring of changes (continuous data collection and monitoring of employee attitudes and performance)

of computer users, Carayon and her colleagues have shown that the job characteristics related to worker strain change over time (Carayon et al. 1995). Therefore, any iterative design strategy for improving the design of interactive systems should take into account temporal factors.

The idea of balancing the negative aspects of the work system by the positive aspects implies an active role from the part of all actors involved in the process. An active role characterized by information gathering, planning, and looking for alternatives can be much more effective than a passive role in achieving efficient use of computer technologies (Haims and Carayon 1998; Wilson and Haines 1997).

This chapter has presented information and data on how to design human–computer interfaces effectively from the physical, cognitive, and social points of view. Each of these has been presented separately, but there is a definite interaction among these three aspects. For example, Eberts et al. (1990) concluded that in group computer work, when the job design was enriched, the individuals in the group better understood the other group members' cognitive style than when the job was simplified. The better understanding resulted in more effective group performance than when the cognitive styles of other group members were less understood. This illustrates an interaction effect between social and cognitive factors in human–computer interaction.

The effects of physical and cognitive interaction in human–computer interaction have been documented by Karwowski et al. (1994). They demonstrated, as a result of a highly controlled study in computer-based mail sorting, that the mode of information presentation on a computer screen and the cognitive response requirement of the user affected and changed their physical posture. Thus, if designers consider both factors in interactive system design they can optimize their interaction.

Cook and Salvendy (1999) have documented, in computer-based work, the interrelationship between social and cognitive factors. They found that increased job enrichment and increased mental workload are some of the most important variables affecting job satisfaction. This raises an interesting dilemma for designers since the cognitive scientist would argue to minimize or optimize mental workload in order to minimize training time and maximize performance. The industrial engineer would argue for minimizing mental workload because that simplifies the work and thus decreases the rate of pay that a company needs to pay for the job. And the social scientist would argue that increasing the mental workload on the job would result in increased job satisfaction. The increased job satisfaction would, in turn, be expected to yield decreased labor turn over and decreased absenteeism, frequently resulting in increased productivity.

These interactions illustrate the type of dilemmas system developers can encounter during interactive system design. Involving a multidisciplinary team in the development process allows such opposing requirements to be addressed better. The team must be supported by ergonomists who understand physical requirements, human factors engineers who understand cognitive requirements, and management that believes in the competitive edge that can be gained by developing user-centered interactive systems. Close collaboration among these team members can lead to the development of remarkably effective and highly usable systems that are readily adopted by users.

Acknowledgment

This material is based, in part, upon work supported by the Naval Air Warfare Center Training Systems Division (NAWC TSD) under contract No. N61339-99-C-0098. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or the endorsement of NAWC TSD.

REFERENCES

Allen, R. B. (1997), "Mental Models and User Models," in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 49–63.

Armstrong, T. J., Foulke, J. A., Martin, B. J., and Rempel, D. (1991), "An Investigation of Finger Forces in Alphanumeric Keyboard Work," in *Design for Everyone Volume,* Y. Queinnec and F. Daniellou, Eds., Taylor & Francis, New York, pp. 75–76.

Armstrong, T. J., Foulke, J. A., Martin, B. J., Gerson, J., and Rempel, D. M. (1994), "Investigation of Applied Forces in Alphanumeric Keyboard Work," *American Industrial Hygiene Association Journal,* Vol. 55, pp. 30–35.

Armstrong, T. J., Martin, B. J., Franzblau, A., Rempel, D. M., and Johnson, P. W. (1995), "Mouse Input Devices and Work-Related Upper Limb Disorders," in *Work with Display Units 94,* A. Grieco, G. Molteni, E. Occhipinti, and B. Piccoli, Eds., Elsevier, Amsterdam, pp. 375–380.

Aronsson, G. (1989), "Changed Qualification Demands in Computer-Mediated Work," *Applied Psychology,* Vol. 38, pp. 57–71.

Aronsson, G., and Johanson, G. (1987), "Work Content, Stress and Health in Computer-mediated Work," in *Work with Display Units 86,* B. Knave and P.-G. Wideback, Eds., Elsevier, Amsterdam, pp. 732–738.

Asakura, T., and Fujigaki, Y. (1993), "The Impact of Computer Technology on Job Characteristics and Workers' Health," in *Human–Computer Interaction: Application and Case Studies,* M. J. Smith and G. Salvendy, Eds., Elsevier, Amsterdam, pp. 982–987.

Aydin, C. E. (1989), "Occupational Adaptation to Computerized Medical Information Systems," *Journal of Health and Social Behavior,* Vol. 30, pp. 163–179.

American National Standards Institute (ANSI) (1973), "American National Standard Practice for Office Lighting (A. 132.1-1973), ANSI, New York.

American National Standards Institute (ANSI) (1988), "American National Standard for Human Factors Engineering of Visual Display Terminal Workstations (ANSI/HFS Standard No. 100-1988)," Human Factors and Ergonomics Society, Santa Monica, CA.

Attewell, P., and Rule, J. (1984), "Computing and Organizations: What We Know and What We Don't Know," *Communications of the ACM,* Vol. 27, pp. 1184–1192.

Bailyn, L. (1989), "Toward the Perfect Workplace?" *Communications of the ACM,* Vol. 32, pp. 460–471.

Bammer, G., and Blignault, I. (1988), "More Than a Pain in the Arms: A Review of the Consequences of Developing Occupational Overuse Syndromes (OOSs)," *Journal of Occupational Health and Safety—Australia and New Zealand,* Vol. 4, No. 5, pp. 389–397.

Blackler, F., and Brown, C. (1987), "Management, Organizations and the New Technologies, in *Information Technology and People: Designing for the Future,* F. Blackler and D. Oborne, Eds., British Psychological Society, London, pp. 23–43.

Blackler, F. (1988), "Information Technologies and Organizations: Lessons from the 1980s and Issues for the 1990s," *Journal of Occupational Psychology,* Vol. 61, pp. 113–127.

Brown, C. M. L. (1988), *Human–Computer Interface Design Guidelines,* Ablex, Norwood, NJ.

Buchanan, D. A., and Boddy, D. (1982), "Advanced Technology and the Quality of Working Life: The Effects of Word Processing on Video Typists," *Journal of Occupational Psychology,* Vol. 55, pp. 1–11.

Buie, E. (1999), "HCI Standards: A Mixed Blessing," *Interactions,* Vol. VI2, pp. 36–42.

Bullinger, H. J., Kern, P., and Braun, M. (1977), "Controls," in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Slavendy, Ed., John Wiley & Sons, New York.

Cakir, A., and Dzida, W. (1997), "International Ergonomic HCI Standards," in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 407–420.

Campion, M. A., and Thayer, P. W. (1985), "Development and Field Evaluation of an Interdisciplinary Measure of Job Design," *Journal of Applied Psychology,* Vol. 70, pp. 29–43.

Carayon, P. (1993), "Effect of Electronic Performance Monitoring on Job Design and Worker Stress: Review of the Literature and Conceptual Model," *Human Factors,* Vol. 35, No. 3, pp. 385–395.

Carayon, P. (1994a), "Effects of Electronic Performance Monitoring on Job Design and Worker Stress: Results of Two Studies," *International Journal of Human–Computer Interaction,* Vol. 6, No. 2, pp. 177–190.

Carayon, P. (1994b), "Stressful Jobs and Non-stressful Jobs: A Cluster Analysis of Office Jobs," *Ergonomics,* Vol. 37, pp. 311–323.

Carayon, P. (1997), "Temporal Issues of Quality of Working Life and Stress in Human–Computer Interaction," *International Journal of Human–Computer Interaction,* Vol. 9, No. 4, pp. 325–342.

Carayon, P., and Karsh, B. (2000), "Sociotechnical Issues in the Implementation of Imaging Technology," *Behaviour and Information Technology,* Vol. 19, No. 4, pp. 247–262.

Carayon, P., Yang, C. L., and Lim, S. Y. (1995), "Examining the Relationship Between Job Design and Worker Strain over Time in a Sample of Office Workers," *Ergonomics,* Vol. 38, No. 6, pp. 1199–1211.

Carayon, P., Smith, M. J., and Haims, M. C. (1999), "Work Organization, Job Stress, and Work-related Musculoskeletal Disorders," *Human Factors,* Vol. 41, No. 4, pp. 644–663.

Carayon-Sainfort, P. (1992), "The Use of Computers in Offices: Impact on Task Characteristics and Worker Stress," *International Journal of Human–Computer Interaction,* Vol. 4, No. 3, pp. 245–261.

Card, S. K., Moran, T. P., and Newell, A. L. (1983), *The Psychology of Human–Computer Interaction,* Erlbaum, Hillsdale, NJ.

Carroll, J. M., and Mack, R. L. (1985), "Metaphor, Computing Systems, and Active Learning," *International Journal of Man–Machine Studies,* Vol. 22, pp. 39–57.

Carroll, J. M., and Olson, J. R. (1988), "Mental Models in Human–Computer Interaction, in *Handbook of Human–Computer Interaction,* M. Helander, Ed., North-Holland, Amsterdam, pp. 45–65.

Carroll, J. M., and Thomas, J. C. (1982), "Metaphor and the Cognitive Representation of Computing Systems," *IEEE Transactions on System, Man, and Cybernetics,* Vol. 12, pp. 107–116.

Centers for Disease Control (CDC) (1980), "Working with Video Display Terminals: A Preliminary Health-Risk Evaluation," *Morbidity and Mortality Weekly Report,* Vol. 29, pp. 307–308.

Choong, Y.-Y., and Salvendy, G. (1998), "Design of Icons for Use by Chinese in Mainland China," *Interacting with Computers,* Vol. 9, No. 4, pp. 417–430.

Choong, Y.-Y., and Salvendy, G. (1999), "Implications for Design of Computer Interfaces for Chinese Users in Mainland China," *International Journal of Human–Computer Interaction,* Vol. 11, No. 1, pp. 29–46.

Clement, A., and Gotlieb, C. C. (1988), "Evaluation of an Organizational Interface: The New Business Department at a Large Insurance Firm, in *Computer-Supported Cooperative Work: A Book of Readings,* I. Greif, Ed., Morgan Kaufmann, San Mateo, CA, pp. 609–621.

Cook, J., and Salvendy, G. (1999), "Job Enrichment and Mental Workload in Computer-Based Work: Implication for Adaptive Job Design," *International Journal of Industrial Ergonomics,* Vol. 24, pp. 13–23.

Crozier, M. (1964), *The Bureaucratic Phenomenon,* University of Chicago Press, Chicago.

Czaja, S. J., and Sharit, J. (1993), "Stress Reactions to Computer-Interactive Tasks as a Function of Task Structure and Individual Differences," *International Journal of Human–Computer Interaction,* Vol. 5, No. 1, pp. 1–22.

Danziger, J. N., Kraemer, K. L., Dunkle, D. E., and King, J. L. (1993), "Enhancing the Quality of Computing Service: Technology, Structure and People," *Public Administration Review,* Vol. 53, pp. 161–169.

Diaper, D. (1989), *Task Analysis for Human–Computer Interaction,* Ellis Horwood, Chichester.

Dong, J., and Salvendy, G. (1999a), "Improving Software Interface Usabilities for the Chinese Population through Dynamic Translation," *Interacting with Computers* (forthcoming).

Dong, J., and Salvendy, G. (1999b), "Design Menus for the Chinese Population: Horizontal or Vertical?" *Behaviour and Information Technology,* Vol. 18, No. 6, pp. 467–471.

Duffy, V., and Salvendy, G. (1999), "Problem Solving in an AMT Environment: Differences in the Knowledge Requirements for an Inter-discipline Team," *International Journal of Cognitive Ergonomics,* Vol. 3. No. 1, pp. 23–35.

Eberts, R. E. (1994), *User Interface Design,* Prentice Hall, Englewood Cliffs, NJ.

Eberts, R., Majchrzak, A., Payne, P., and Salvendy, G. (1990), "Integrating Social and Cognitive Factors in the Design of Human–Computer Interactive Communication," *International Journal of Human–Computer Interaction,* Vol. 2, No.1. pp. 1–27.

Egan, D. E. (1988), "Individual Differences in Human–Computer Interaction," in *Handbook of Human–Computer Interaction,* M. Helander, Ed., North-Holland, Amsterdam, pp. 231–254.

Erickson, T. D. (1990), "Working with Interface Metaphors," in *The Art of Human–Computer Interaction,* B. Laurel, Ed., Addison-Wesley, Reading, MA, pp. 65–73.

Ericsson, K. A., and Simon, H. A. (1980), "Verbal Reports as Data," *Psychological Review,* Vol. 87, pp. 215–251.

Eveland, J. D., and Bikson, T. K. (1988), "Work Group Structure and Computer Support: A Field Experiment," *AMC Transactions on Office Information Systems,* Vol. 6, pp. 354–379.

Faucett, J., and Rempel, D. (1994), "VDT-Related Musculoskeletal Symptoms: Interactions between Work Posture and Psychosocial Work Factors," *American Journal of Industrial Medicine,* Vol. 26, pp. 597–612.

Fischer, G. (1991), "The Importance of Models in Making Complex Systems Comprehensible," in *Mental Models and Human Computer Interaction,* Vol. 2, M. J. Tauber and D. Ackermann, Eds., North-Holland, Amsterdam, pp. 3–36.

Flynn, F. M. (1989), "Introducing New Technology into the Workplace: The Dynamics of Technological and Organizational Change," in *Investing in People—A Strategy to Address America's Workforce Crisis,* U.S. Department of Labor, Commission on Workforce Quality and Labor Market Efficiency, Washington, DC, pp. 411–456.

Fogelman, M., and Brogmus, G. (1995), "Computer Mouse Use and Cumulative Trauma Disorders of the Upper Extremities," *Ergonomics,* Vol. 38, No. 12, pp. 2465–2475.

Gentner, D. (1983), "Structure-Mapping: A Theoretical Framework for Analogy," *Cognitive Science,* Vol. 7, pp. 155–170.

Gentner, D., and Clement, C. (1983),"Evidence for Relational Selectivity in the Interpretation of Analogy and Metaphor," *Psychology of Learning and Motivation,* Vol. 22, pp. 307–358.

Gerard, M. J., Armstrong, T. J., Foulke, J. A., and Martin, B. J. (1996), "Effects of Key Stiffness on Force and the Development of Fatigue while Typing," *American Industrial Hygiene Association Journal,* Vol. 57, pp. 849–854.

Gould, J. D., Boies, S. J., and Ukelson, J. (1997), "How to Design Usable Systems," in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 231–254.

Grandjean, E. (1979), *Ergonomical and Medical Aspects of Cathode Ray Tube Displays,* Swiss Federal Institute of Technology, Zurich.

Grandjean, E. (1984), "Postural Problems at Office Machine Workstations," in *Ergonomics and Health in Modern Offices,* E. Grandjean, Ed., Taylor & Francis, London.

Grandjean, E. (1988), *Fitting the Task to the Man,* Taylor & Francis, London.

Gray, W. D., John, B. E., and Atwood, M. E. (1993), "Project Ernestine: Validating a GOMS Analysis for Predicting Real-World Task Performance," *Human–Computer Interaction,* Vol. 6, pp. 287–309.

Greif, I., Ed. (1988), *Computer-Supported Cooperative Work: A Book of Readings,* Morgan Kaufmann, San Mateo, CA.

Greif, I., and Sarin, S. (1988), "Data Sharing in Group Work," in *Computer-Supported Cooperative Work,* I. Greif, Ed., Morgan Kaufmann, San Mateo, CA, pp. 477–508.

Guggenbuhl, U., and Krueger, H. (1990), "Musculoskeletal Strain Resulting from Keyboard Use," in *Work with Display Units 89,* L. Berlinguet and D. Berthelette, Eds., Elsevier, Amsterdam.

Guggenbuhl, U., and Krueger, H. (1991), "Ergonomic Characteristics of Flat Keyboards," in *Design for Everyone Volume,* Y. Queinnec and F. Daniellou, Eds., Taylor & Francis, London, pp. 730–732.

Hackos, J. T., and Redish, J. C. (1998), *User and Task Analysis for Interface Design,* John Wiley & Sons, New York.

Hagberg, M. (1995), "The 'Mouse–Arm Syndrome' Concurrence of Musculoskeletal Symptoms and Possible Pathogenesis among VDT Operators," in *Work with Display Units 94,* A. Grieco, G. Molteni, E. Occhipinti, and B. Piccoli, Eds., Elsevier, Amsterdam, pp. 381–385.

Hagberg, M., Silverstein, B., Wells, R., Smith, M. J., Hendrick, H. W., Carayon, P., and Perusse, M. (1995), *Work Related Musculoskeletal Disorders (WMDSs): A Reference Book for Prevention,* Taylor & Francis, London.

Haims, M. C., and Carayon, P. (1998), "Theory and Practice for the Implementation of 'In-house', Continuous Improvement Participatory Ergonomic Programs," *Applied Ergonomics,* Vol. 29, No. 6, pp. 461–472.

Helander, M. G. (1982), *Ergonomic Design of Office Environments for Visual Display Terminals,* National Institute for Occupational Safety and Health (DTMD), Cincinnati, OH.

Ilg, R. (1987), "Ergonomic Keyboard Design," *Behaviour and Information Technology,* Vol. 6, No. 3, pp. 303–309.

Jeffries, R. (1997), "The Role of Task Analysis in Design of Software," in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 347–359.

Johansson, G., and Aronsson, G. (1984), "Stress Reactions in Computerized Administrative Work," *Journal of Occupational Behaviour,* Vol. 5, pp. 159–181.

Kahn, H., and Cooper, C. L. (1986), "Computing Stress," *Current Psychological Research and Reviews,* Summer, pp. 148–162.

Karat, J. (1988), "Software Evaluation Methodologies," in *Handbook of Human–Computer Interaction,* M. Helander, Ed., North-Holland, Amsterdam, pp. 891–903.

Karat, J., and Dayton, T. (1995), "Practical Education for Improving Software Usability, in *CHI '95 Proceedings, pp. 162–169.*

Karlqvist, L., Hagberg, M., and Selin, K. (1994), "Variation in Upper Limb Posture and Movement During Word Processing with and without Mouse Use," *Ergonomics,* Vol. 37, No. 7, pp. 1261–1267.

Karwowski, W., Eberts, R., Salvendy, G., and Norlan, S. (1994), "The Effects of Computer Interface Design on Human Postural Dynamics," *Ergonomics,* Vol. 37, No. 4, pp. 703–724.

Kay, A. (1990), "User Interface: A Personal View," in *The Art of Human–Computer Interaction,* B. Laurel, Ed., Addison-Wesley, Reading, MA, pp. 191–207.

Kearsley, G. (1989), "Introducing New Technology into the Workplace: Retraining Issues and Strategies," in *Investing in People—A Strategy to Address America's Workforce Crisis,* U.S. Department of Labor, Commission on Workforce Quality and Labor Market Efficiency, Washington, DC, pp. 457–491.

Kiesler, S., Siegel, J., and McGwire, T. W. (1984), "Social Psychological Aspects of Computer-Mediated Communication," *American Psychologist,* Vol. 39, pp. 1123–1134.

Kim, H., and Hirtle, S. C. (1995), "Spatial Metaphors and Disorientation in Hypertext Browsing," *Behaviour and Information Technology,* Vol. 14, No. 4, pp. 239–250.

Kirwan, B., and Ainsworth, L. K., Eds. (1992), *A Guide to Task Analysis,* Taylor & Francis, London.

Korunka, C., Weiss, A., and Zauchner, S. (1997), "An Interview Study of 'Continuous' Implementations of Information Technology," *Behaviour and Information Technology,* Vol. 16, No. 1, pp. 3–16.

Korunka, C., and Carayon, P. (1999), "Continuous Implementation of Information Technology: The Development of an Interview Guide and a Cross-Sectional Comparison of Austrian and American Organizations," *International Journal of Human Factors in Manufacturing,* Vol. 9, No. 2, pp. 165–183.

Korunka, C., Huemer, K. H., Litschauer, B., Karetta, B., and Kafka-Lutzow, A. (1996), Working with New Technologies—Hormone Excretion as Indicator for Sustained Arousal," *Biological Psychology,* Vol. 42, pp. 439–452.

Korunka, C., Weiss, A., Huemer, K. H., and Karetta, B. (1995), "The Effects of New Technologies on Job Satisfaction and Psychosomatic Complaints," *Applied Psychology: An International Review,* Vol. 44, No. 2, pp. 123–142.

Korunka, C., Weiss, A., and Karetta, B. (1993), "Effects of New Technologies with Special Regard for the Implementation Process per Se," *Journal of Organizational Behaviour,* Vol. 14, pp. 331–348.

Kroemer, K. H. E. (1972), "Human Engineering the Keyboard," *Human Factors,* Vol. 14, pp. 51–63.

Kroemer, K. H. E., and Grandjean, E. (1997), *Fitting the Task to the Human,* Taylor & Francis, London.

Levi, L. (1972), "Conditions of Work and Sympathoadrenomedullary Activity: Experimental Manipulations in a Real Life Setting, in *Stress and Distress in Response to Psychosocial Stimuli,* L. Levi, Ed., *Acta Medica Scandinavica,* Vol. 191, Suppl. 528, pp. 106–118.

Lim, S. Y., and Carayon, P. (1995), "Psychosocial and Work Stress Perspectives on Musculoskeletal Discomfort," in *Proceedings of PREMUS 95,* Institute for Research on Safety and Security (IRSST), Montreal.

Lim, S. Y., Rogers, K. J. S., Smith, M. J., and Sainfort, P. C. (1989), "A Study of the Direct and Indirect Effects of Office Ergonomics on Psychological Stress Outcomes," in *Work with Computers: Organizational, Management, Stress and Health Aspects,* M. J. Smith and G. Salvendy, Eds., Elsevier, Amsterdam, pp. 248–255.

Lindstrom, K., and Leino, T. (1989), "Assessment of Mental Load and Stress Related to Information Technology Change in Baking and Insurance," in *Man–Computer Interaction Research, MACINTER-II,* F. Klix, N. A. Streitz, Y. Waern, and H. Wandke, Eds., Elsevier, Amsterdam, pp. 523–533.

Linton, S. J., and Kamwendo, K. (1989), "Risk Factors in the Psychosocial Work Environment for Neck and Shoulder Pain in Secretaries," *Journal of Occupational Medicine,* Vol. 31, No. 7, pp. 609–613.

Marcus, A. (1997), "Graphical User Interfaces," in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 423–440.

Martel, A. (1998), "Application of Ergonomics and Consumer Feedback to Product Design at Whirlpool," in *Human Factors in Consumer Products,* N. Stanton, Ed., Taylor & Francis, London, pp. 107–126.

Mayhew, D. J. (1992), *Principles and Guidelines in Software User Interface Design,* Prentice Hall, Englewood Cliffs, NJ.

Mayhew, D. J. (1999), *The Usability Engineering Lifecycle: A Practitioner's Handbook for User Interface Design,* Morgan Kaufmann, San Francisco.

McLeod, R. W., and Sherwood-Jones, B. M. (1993), "Simulation to Predict Operator Workload in a Command System," in *A Guide to Task Analysis,* B. Kirwan, and L. K. Ainsworth, Eds., Taylor & Francis, London, pp. 301–310.

Mountford, S. J. (1990), "Tools and Techniques for Creative Design," in *The Art of Human–Computer Interaction,* B. Laurel, Ed., Addison-Wesley, Reading, MA, pp. 17–30.

Nardi, B. A. (1997), "The Use of Ethnographic Methods in Design and Evaluation," in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 361–366.

National Academy of Sciences (NAS) (1983), *Visual Displays, Work and Vision,* National Academy Press, Washington, DC.

Neale, D. C., and Carroll, J. M. (1997), "The Role of Metaphors in User Interface Design," in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 441–462.

Newell, A. F., Arnott, J. L., Carter, K., and Cruikshank, G. (1990), "Listening Typewriter Simulation Studies," *International Journal of Man–Machine Studies,* Vol. 33, pp. 1–19.

Nielsen, J. (1993), *Usability Engineering,* Academic Press Professional, Boston.

National Institute for Occupational Safety and Health (NIOSH) (1981), *Potential Health Hazards of Video Display Terminals,* Cincinnati.

National Institute for Occupational Safety and Health (NIOSH) (1990), *Health Hazard Evaluation Report—HETA 89-250-2046—Newsday, Inc.,* U.S. Department of Health and Human Services, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1992), *Health Hazard Evaluation Report—HETA 89-299-2230—US West Communications,* U.S. Department of Health and Human Services, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1993), *Health Hazard Evaluation Report—HETA 90-013-2277—Los Angeles Times,* U.S. Department of Health and Human Services, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1997), *Alternative Keyboards,* DHHS/NIOSH Publication No. 97-148, National Institute for Occupational Safety and Health, Cincinnati, OH.

Norman, D. A. (1988), *The Design of Everyday Things,* Doubleday, New York.

Norman, D. A. (1990), "Why Interfaces Don't Work," in *The Art of Human–Computer Interaction,* B. Laurel, Ed., Addison-Wesley, Reading, MA, pp. 209–219.

Office of Technology Assessment (OTA) (1985), *Automation of America's Offices,* OTA-CIT-287, U.S. Government Printing Office, Washington, DC.

Ortony, A. (1979), "Beyond Literal Similarity," *Psychological Review,* Vol. 86, No. 3, pp. 161–180.

Parunak, H. V. (1989), "Hypermedia Topologies and User Navigation," in *Hypertext '89 Proceedings,* ACM Press, New York, pp. 43–50.

Phizacklea, A., and Wolkowitz, C. (1995), *Homeworking Women: Gender, Racism and Class at Work,* Sage, London.

Punnett, L., and Bergqvist, U. (1997), *Visual Display Unit Work and Upper Extremity Musculoskeletal Disorders,* National Institute for Working Life, Stockholm.

Rada, R., and Ketchel, J. (2000), "Standardizing the European Information Society," *Communications of the ACM,* Vol. 43, No. 3, pp. 21–25.

Rempel, D., and Gerson, J. (1991), "Fingertip Forces While Using Three Different Keyboards," in *Proceedings of the 35th Annual Human Factors Society Meeting,* Human Factors and Ergonomics Society, San Diego.

Rempel, D., Dennerlein, J. T., Mote, C. D., and Armstrong, T. (1992), "Fingertip Impact Loading During Keyboard Use," in *Proceedings of NACOB II: Second North American Congress on Biomechanics* (Chicago).

Rempel, D., Dennerlein, J., Mote, C. D., and Armstrong, T. (1994), "A Method of Measuring Fingertip Loading During Keyboard Use," *Journal of Biomechanics,* Vol. 27, No. 8, pp. 1101–1104.

Rice, R. E., and Case, D. (1983), "Electronic Message Systems in the University: A Description of Use and Utility," *Journal of Communication,* Vol. 33, No. 1, pp. 131–152.

Robey, D. (1987), "Implementation and the Organizational Impacts of Information Systems," *Interfaces,* Vol. 17, pp. 72–84.

Rouse, W. B. (1991), *Design for Success: A Human-Centered Approach to Designing Successful Products and Systems,* John Wiley & Sons, New York.

Sauter, S. L., Gottlieb, H. S., Rohrer, K. N., and Dodson, V. N. (1983), *The Well-Being of Video Display Terminal Users,* Department of Preventive Medicine, University of Wisconsin-Madison, Madison, WI.

Schleifer, L. M., and Amick, B. C., III (1989), "System Response Time and Method of Pay: Stress Effects in Computer-Based Tasks," *International Journal of Human–Computer Interaction,* Vol. 1, pp. 23–39.

Sheridan, T. B. (1997a), "Task Analysis, Task Allocation and Supervisory Control," in *Handbook of Human–Computer Interaction* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 87–105.

Sheridan, T. B. (1997b), "Supervisory Control," in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1295–1327.

Shneiderman, B. (1992), *Designing the User Interface,* 2nd Ed., Addison-Wesley, Reading, MA.

Smith, M. J. (1984a), "Health Issues in VDT Work," in *Visual Display Terminals: Usability Issues and Health Concerns,* J. Bennett, D. Case, and M. J. Smith, Eds., Prentice Hall, Englewood Cliffs, NJ, pp. 193–228.

Smith, M. J. (1984b), "The Physical, Mental and Emotional Stress Effects of VDT Work," *Computer Graphics and Applications,* Vol. 4, pp. 23–27.

Smith, M. J. (1995), "Behavioral Cybernetics, Quality of Working Life and Work Organization in Computer Automated Offices," in *Work With Display Units 94,* A. Grieco, G. Molteni, E. Occhipinti, and B. Piccoli, Eds., Elsevier, Amsterdam, pp. 197–202.

Smith, M. J., and Carayon, P. (1995), "Work Organization, Stress and Cumulative Trauma Disorders," in *Beyond Biomechanics: Psychosocial Aspects of Cumulative Trauma Disorders,* S. Moon and S. Sauter, Eds., Taylor & Francis, London.

Smith, M. J., and Sainfort, P. C. (1989), "A Balance Theory of Job Design for Stress Reduction," *International Journal of Industrial Ergonomics,* Vol. 4, pp. 67–79.

Smith, M. J., Cohen, B. G. F., Stammerjohn, L. W., and Happ, A. (1981), "An Investigation of Health Complaints and Job Stress in Video Display Operations," *Human Factors,* Vol. 23, pp. 387–400.

Smith, M. J., Carayon, P., Sanders, K. J., Lim, S. Y., and LeGrande, D. (1992a), ‘‘Employee Stress and Health Complaints in Jobs with and without Electronic Performance Monitoring,’’ *Applied Ergonomics,* Vol. 23, No. 1, pp. 17–27.

Smith, M. J., Salvendy, G., Carayon-Sainfort, P., and Eberts, R. (1992b), ‘‘Human–Computer Interaction,’’ in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York.

Smith, M. J., Karsh, B., Conway, F., Cohen, W., James, C., Morgan, J., Sanders, K., and Zehel, D. (1998), ‘‘Effects of a Split Keyboard Design and Wrist Rests on Performance, Posture and Comfort,’’ *Human Factors,* Vol. 40, No. 2, pp. 324–336.

Sproull, L., and Kiesler, S. (1988), ‘‘Reducing Social Context Cues: Electronic Mail in Organizational Communication,’’ in *Computer-Supported Cooperative Work: A Book of Readings,* I. Greif, Ed., Morgan Kaufmann, San Mateo, CA, pp. 683–712.

Sproull, L. and Kiesler, S. (1991), *Connections,* MIT Press, Cambridge, MA.

Stanney, K. M., Maxey, J., and Salvendy, G. (1997), ‘‘Socially-Centered design,’’ in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 637–656.

Stanton, N. (1988), ‘‘Product Design with People in Mind,’’ in *Human Factors in Consumer Products,* N. Stanton, Ed., Taylor & Francis, London, pp. 1–17.

Stephanidis, C., and Salvendy, G. (1998), ‘‘Toward an Information Society for All: An International Research and Development Agenda,’’ *International Journal of Human–Computer Interaction,* Vol. 10, No. 2, pp. 107–134.

Swanson, N. G., Galinsky, T. L., Cole, L. L., Pan, C. S., and Sauter, S. L. (1997), ‘‘The Impact of Keyboard Design on Comfort and Productivity in a Text-Entry Task,’’ *Applied Ergonomics,* Vol. 28, No. 1, pp. 9–16.

Takahashi, D. (1998), ‘‘Technology Companies Turn to Ethnographers, Psychiatrists,’’ *The Wall Street Journal,* October 27.

Theorell, T., Ringdahl-Harms, K., Ahlberg-Hulten, G., and Westin, B. (1991), ‘‘Psychosocial Job Factors and Symptoms from the Locomotor System—A Multicausal Analysis,’’ *Scandinavian Journal of Rehabilitation Medicine,* Vol. 23, pp. 165–173.

Trankle, U., and Deutschmann, D. (1991), ‘‘Factors Influencing Speed and Precision of Cursor Positioning Using a Mouse,’’ *Ergonomics,* Vol. 34, No. 2, pp. 161–174.

Turk, M., and Robertson, G. (2000), ‘‘Perceptual User Interfaces,’’ *Communications of the ACM,* Vol. 43, No. 3, pp. 33–34.

Vanderheiden, G. C. (1997), ‘‘Design for People with Functional Limitations Resulting from Disability, Aging, or Circumstance,’’ in *Handbook of Human Factors and Ergonomics,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 2010–2052.

Verhoef, C. W. M. (1988), ‘‘Decision Making of Vending Machine Users,’’ *Applied Ergonomics,* Vol. 19, No. 2, pp. 103–109.

Vermeeren, A. P. O. S. (1999), ‘‘Designing Scenarios and Tasks for User Trials of Home Electronics Devices,’’ in *Human Factors in Product Design: Current Practice and Future Trends,* W. S. Green and P. W. Jordan, Eds., Taylor & Francis, London, pp. 47–55.

Vertelney, L., and Booker, S. (1990), ‘‘Designing the Whole Product User Interface,’’ in *The Art of Human–Computer Interaction,* B. Laurel, Ed., Addison-Wesley, Reading, MA, pp. 57–63.

Vicente, K. J. (1999), *Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-Based Work,* Erlbaum, Mahwah, NJ.

Wells, R., Lee, I. H. and Bao, S. (1997), ‘‘Investigations of Upper Limb Support Conditions for Mouse Use,’’ in *Proceedings of the 29th Annual Human Factors Association of Canada.*

Wilson, J. R., and Haines, H. M. (1997), ‘‘Participatory Ergonomics,’’ in *Handbook of Human Factors and Ergonomics* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 490–513.

Wixon, D., and Wilson, C. (1997), ‘‘The Usability Engineering Framework for Product Design and Evaluation, in *Handbook of Human–Computer Interaction,* 2nd Ed., M. Helander, T. K. Landauer, and P. V. Prabhu, Eds., North-Holland, Amsterdam, pp. 653–688.

Yang, C.-L. (1994), ‘‘Test of a Model of Cognitive Demands and Worker Stress in Computerized Offices,’’ Ph.D. Dissertation, Department of Industrial Engineering, University of Wisconsin-Madison.

Yang, C. L., and Carayon, P. (1995), ‘‘Effect of Job Demands and Social Support on Worker Stress: A Study of VDT Users,’’ *Behaviour and Information Technology,* Vol. 14, No. 1, pp. 32–40.

SECTION IV

MANAGEMENT, PLANNING, DESIGN, AND CONTROL

A. Project Management
B. Product Planning
C. Manpower Resource Planning
D. Systems and Facilities Design
E. Planning and Control
F. Quality
G. Supply Chain Management and Logistics

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

IV.A

Project Management

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 45

Project Management Cycle: Process Used to Manage Project (Steps to Go Through)

AVRAHAM SHTUB
Technion—Israel Institute of Technology

1. INTRODUCTION **1242**
  1.1. Projects and Processes 1242
  1.2. The Project Life Cycle 1242
  1.3. An Example of a Project Life Cycle 1243
2. PROJECT-MANAGEMENT PROCESSES **1243**
  2.1. Definition of a Process 1243
  2.2. Process Design 1243
  2.3. The PMBOK and Processes in the Project Life Cycle 1243
3. PROJECT INTEGRATION MANAGEMENT **1244**
  3.1. Processes 1244
  3.2. Description 1244
4. PROJECT SCOPE MANAGEMENT **1244**
  4.1. Processes 1244
  4.2. Description 1245
5. PROJECT TIME MANAGEMENT **1245**
  5.1. Processes 1245
  5.2. Description 1245
6. PROJECT COST MANAGEMENT **1245**
  6.1. Processes 1245
  6.2. Description 1246
7. PROJECT QUALITY MANAGEMENT **1246**
  7.1. Processes 1246
  7.2. Description 1247
8. PROJECT HUMAN RESOURCE MANAGEMENT **1247**
  8.1. Processes 1247
  8.2. Description 1247
9. PROJECT COMMUNICATIONS MANAGEMENT **1248**
  9.1. Processes 1248
  9.2. Description 1248
10. PROJECT RISK MANAGEMENT **1248**
  10.1. Processes 1248
  10.2. Description 1248
11. PROJECT PROCUREMENT MANAGEMENT **1249**
  11.1. Processes 1249
  11.2. Description 1249
12. THE LEARNING ORGANIZATION: CONTINUOUS IMPROVEMENT IN PROJECT MANAGEMENT **1250**
  12.1. Individual Learning and Organizational Learning 1250
  12.2. Workflow and Process Design as the Basis of Learning 1251
REFERENCES **1251**

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

1.1. Projects and Processes

A project is an organized endeavor aimed at accomplishing a specific nonroutine or low-volume task (Shtub et al. 1994). Due to sheer size (number of man-hours required to perform the project) and specialization, teams perform most projects. In some projects the team members belong to the same organization, while in many other projects the work content of the project is divided among individuals from different organizations.

Coordination among individuals and organizations involved in a project is a complex task. To ensure success, integration of deliverables produced at different geographical locations by different individuals from different organizations at different times is required. Projects are typically performed under a time pressure, limited budgets, tight cash flow constraints, and uncertainty. Thus, a methodology is required to support the management of projects. Early efforts in developing such a methodology focused on tools. Tools for project scheduling such as the Gantt chart and the critical path method (CPM) were developed along with tools for resource allocation, project budgeting and project control (Shtub et al. 1994). The integration of different tools into a complete framework that supports project management efforts throughout the entire project life cycle (see Section 1.2 below) was achieved by the introduction of project-management processes.

A project-management process is a collection of tools and techniques that are used on a predefined set of inputs to produce a predefined set of outputs. Processes are connected to each other as the input to some of the project-management processes is created (is an output) by other processes. The collection of interrelated project-management processes forms a methodology that supports the management of projects throughout their life cycle, from the initiation of a new project to its (successful) end.

This chapter presents a collection of such interrelated processes. The proposed framework is based on the Project Management Body of Knowledge (PMBOK), developed by the Project Management Institute (PMI) (PMI 1996). The purpose of this chapter is to present the processes and the relationship between them. A detailed description of these processes is available in the PMBOK. PMI conducts a certification program based on the PMBOK. A Project Management Professional (PMP) certificate is earned by passing an exam and accumulating relevant experience in the project-management discipline.

1.2. The Project Life Cycle

Since this is a temporary effort designed to achieve a specific set of goals, it is convenient to define phases that the project goes through. The collection of these phases is defined as the project life cycle. Analogous to a living creature, a project is born (initiated), performed (lives), and terminated (dies), always following the same sequence. This simple life-cycle model of three phases is conceptually helpful in understanding the project-management processes because each process can be defined with respect to each phase. However, this simple life-cycle model is not detailed enough for implementation (in some projects each phase may span several months or years). Thus, more specific life-cycle models for families of similar projects were developed. A specific life-cycle model is a set of stages or phases that a family of projects goes through. The project's phases are performed in sequence or concurrently. The project life cycle defines the steps required to achieve the project goals as well as the content of each step. Thus, the literature on software projects is based on specific life-cycle models, such as the spiral model developed by Muench (1994), and the literature on construction projects is based on construction project life-cycle models, such as the one suggested by Morris (1981).

In the Morris (1981) model, a project is divided into four stages performed in sequence:

Stage I—feasibility. This stage terminates with a go/no-go decision for the project. It includes a formulation of the project, feasibility studies, strategy design, and strategy approval for the project.

Stage II—planning and design. This stage terminates with Major contracts Let. It includes base design, cost and schedule planning, contract definitions, and detailed planning of the project.

Stage III—production. This stage terminates with the installation substantially complete. It includes manufacturing, delivery, civil works, installation, and testing.

Stage IV—turnover and startup. This stage terminates with full operation of the facility. It includes final testing and maintenance.

Clearly this model does not fit R&D projects or software projects, while it may be very helpful for many construction projects.

With the integration of the ideas of project processes and the project life cycle, a methodology for project management emerges. The methodology is a collection of processes and a definition of the part of each process that should be performed within each phase of the project life cycle. The responsibility to perform each process (or part of a process) can be allocated to specific individuals trained in the required tools and techniques. Furthermore, the information (input) required for each process can be delivered to the individuals responsible for the project, ensuring a well-coordinated flow of information and thus good communication between the project participants.

Templates or models of life cycles are useful for project management. When each phase is terminated with one or more completed documentable deliverables, the project-life cycle model is a simple yet very effective tool for monitoring and control of the project throughout its entire life cycle.

1.3. An Example of a Project Life Cycle

The Department of Defense uses a simple yet very popular life-cycle model for Defense acquisition (U.S. Department of Defense 1993):

Phase Number	Phase Description	Corresponding Milestone
	Determination of mission needs	Concept studies approval
0	Concept exploration and definition	Concept demonstration approval
1	Demonstration and validation	Development approval
2	Engineering and manufacturing development	Production approval
3	Production and deployment	Major modification approval as required
4	Operations and support	

2. PROJECT-MANAGEMENT PROCESSES

2.1. Definition of a Process

A process is a group of activities designed to transform a set of inputs into required outputs. There are three elements in the transformation of inputs into outputs:

1. Data and information
2. Decision making
3. Implementation and actions

A well-defined set of processes, supported by an appropriate information system (composed of a database and a model base) and implemented by a team trained in performing the processes is a cornerstone in the competitive edge of organizations.

2.2. Process Design

The design of a process is aimed at defining the following:

- Data required to support decisions, including:
  - The data sources
  - How the data should be collected
  - How the data should be stored
  - How the data should be retrieved
  - How the data should be presented as information to decision makers
- Models required to support decisions by transforming data into useful information, including:
  - Models that support routine decisions
  - Models that support ad hoc decisions
  - Models used for process control
- Data and models integration:
  - How data from the database is analyzed by the models
  - How information generated by the models is transferred and presented to decision makers

2.3. The PMBOK and Processes in the Project Life Cycle

A well-defined set of processes that applies to a large number of projects is discussed in the PMBOK (PMI 1996). Although some of the PMBOK processes may not apply to some projects, while other

PMBOK processes may need modifications in certain projects, the PMBOK is a well-accepted source of information and therefore the following definition of processes is based on the PMBOK.

The PMBOK classifies processes in two ways:

1. By sequence
   - Initiating processes
   - Planning processes
   - Executing processes
   - Controlling processes
   - Closing processes
2. By management function:
   - Processes in integration management
   - Processes in scope management
   - Processes in time management
   - Processes in cost management
   - Processes in quality management
   - Processes in human resource management
   - Processes in communication management
   - Processes in risk management
   - Processes in procurement management

The application of these processes in a specific organization requires customization, development of supporting tools, and training.

3. PROJECT INTEGRATION MANAGEMENT

3.1. Processes

Project integration management involves three processes:

1. Project plan development
2. Project plan execution
3. Overall change control

The purpose of these three processes is to ensure coordination among the various elements of the project. Coordination is achieved by getting inputs from other processes, integrating the information contained in these inputs, and providing integrated outputs to the decision makers and to other processes.

The project life-cycle model plays an important role in project integration management because project plan development is performed during the early phases of the project while project plan execution and project change control are performed during the other phases of the project. With a well-defined life-cycle model, it is possible to define the data, decision making, and activities required at each phase of the project life cycle and consequently train those responsible for performing the processes adequately.

3.2. Description

The project plan and its execution are the major outputs of this process. The plan is based on inputs from other planning processes (discussed later) such as scope planning, schedule development, resource planning, and cost estimating along with historical information and organizational policies. The project plan is continuously updated based on corrective actions triggered by approved change requests and analysis of performance measures.

Execution of the project plan produces work results—the deliverables of the project.

4. PROJECT SCOPE MANAGEMENT

4.1. Processes

Project scope management involves five processes:

1. Initiation
2. Scope planning

3. Scope definition
4. Scope verification
5. Scope change control

The purpose of these processes is to ensure that the project includes all work (and only that work) required for its successful completion.

In the following discussion, scope relates to the product scope (defined as the features and functions to be included in the product or service) and the project scope (defined as the work that must be done in order to deliver the product scope).

4.2. Description

The scope is defined based on a description of the needed product or service. Alternative products or services may exist. Based on appropriate selection criteria and a selection methodology, the best alternative is selected and a project charter is issued along with a nomination of a project manager. The project manager should evaluate different alternatives to produce the selected product or service and implement a methodology such as cost–benefit analysis to select the best alternative. Once an alternative is selected, a work breakdown structure (WBS) is developed. The WBS is a hierarchical presentation of the project scope in which the upper level is the whole project and at which the lowest-level work packages are defined. Each work package is assigned to a manager (organizational unit) responsible for its scope.

5. PROJECT TIME MANAGEMENT

5.1. Processes

Project Time Management involves five processes:

1. Activity definition
2. Activity sequencing
3. Activity duration estimating
4. Schedule development
5. Schedule control

The purpose of time management is to ensure timely completion of the project. The main output of time management is a schedule that defines what is to be done, when, and by whom. This schedule is used throughout the project execution to synchronize between people and organizations involved in the project and as a basis for control.

5.2. Description

Each work package in the WBS is decomposed into the activities required to complete its predefined scope. A list of activities is constructed and the time to complete each activity is estimated. Estimates can be deterministic (point estimates) or stochastic (distributions). Precedence relations among activities are defined, and a model such as a Gantt chart, activity on arc (AOA), or activity on nodes (AON) network is constructed (Shtub et al. 1994). An initial schedule is developed based on the model. This unconstrained schedule is a basis for estimating required resources and cash. Based on the constraint imposed by due dates, cash and resource availability, and resource requirements of other projects, a constrained schedule is developed. Further tuning of the schedule may be possible by changing the resource combination assigned to activities (these resource combinations are known as modes).

The schedule is implemented by the execution of activities. Due to uncertainty, a schedule control is required to detect deviations and decide how to react to such deviations and change requests. The schedule control system is based on performance measures such as actual completion of deliverables (milestones), actual starting time of activities, and actual finishing time of activities. Changes to the baseline schedule are required whenever a change in the project scope is implemented.

6. PROJECT COST MANAGEMENT

6.1. Processes

Project cost management involves four processes:

1. Resource planning
2. Cost estimating

3. Cost budgeting
4. Cost control

These processes are designed to provide an estimate of the cost required to complete the project scope, develop a budget based on management policies and strategy, and ensure that the project is completed within the approved budget.

6.2. Description

To complete the project activities, different resources are required. Labor, equipment, and information are examples of resources required to perform in-house activities, while money is required for outsourcing. The estimated amount of required resources as well as the timing of resource requirements are based on the activity list and the schedule. Resource allocation is performed at the lowest level of the WBS—the work package level—and requirements are aggregated to the project level and the whole-organization level. A comparison of resource requirements and resource availability is the basis of finite resource-allocation procedures (models) that assign available resources to projects and activities based on management's strategy and priorities. Resource planning results in a detailed plan specifying what resources are required for each work package. Applying the resource rates to the resource plan and adding overhead and outsourcing expenses allows a cost estimate of the project to be developed. The cost estimate is the basis for budgeting. Based on the schedule, cost estimates are time phased to allow for cash flow analysis. Furthermore, additional allocations are made, such as the management reserve required to buffer against uncertainty. The resulting budget is the baseline for project cost control.

Due to uncertainty, cost control is required to detect deviations and decide how to react to such deviations and change requests. The cost-control system is based on performance measures, such as actual cost of activities or deliverables (milestones) and actual cash flows. Changes to the baseline budget are required whenever a change in the project scope is implemented.

7. PROJECT QUALITY MANAGEMENT

7.1. Processes

Project quality management involves three processes:

1. Quality planning
2. Quality assurance
3. Quality control

The purpose of these processes is to ensure that the project will satisfy the needs for which it was undertaken. These needs are multidimensional—Garvin (1987) suggests that quality has eight dimensions or performance measures:

1. *Performance:* This dimension refers to the product or service's primary characteristics, such as the acceleration, cruising speed, and comfort of an automobile or the sound and picture clarity of a TV set. The understanding of performance required by the customer and the design of the service or product to achieve the required performance level are key factors in quality-based competition.
2. *Features:* This is a secondary aspect of performance—the characteristics that supplement the basic functioning. Garvin (1987) defines features as "the bells and whistles" of the product or service. The flexibility a customer has to select desired options from a large list of such options contributes to the quality of the product or service.
3. *Reliability:* This performance measure reflects the probability of a product malfunctioning or failing within a specified time period. It reflects on both the cost of maintenance and on downtime of the product.
4. *Conformance:* This is the degree to which the product or service design and operating characteristics meet established standards.
5. *Durability:* This is a measure of the economic and technical service duration of a product. It relates to the amount of use one can get from a product before it has to be replaced due to technical or economical considerations.
6. *Serviceability:* This measure reflects the speed, courtesy, competence, and ease of repair should the product fail. The reliability of a product and its serviceability complement each other. A

reliable product that rarely fails, and on those occasions fast and inexpensive service is available, has a lower downtime and better serves its owner.

7. *Aesthetics:* This is a subjective performance measure related to how the product feels, tastes, looks, or smells. It reflects individual preferences.
8. *Perceived quality:* This is another subjective measure related to the reputation of product or a service. This reputation may be based on past experience and partial information, but in many cases the customer's decisions are based on perceived quality because exact information about the other performance measures listed above is not readily available.

7.2. Description

Quality planning starts with the definition of standards or performance levels for each of the dimensions of quality. Based on the scope of the project, quality policy, standards, and regulations, a quality management plan is developed. The plan describes "the organizational structure, responsibilities, procedures, processes, and resources needed to implement quality management" (ISO 9000), that is, how the project management team will implement its quality policy to achieve the required quality levels. Checklists and metrics or operational definitions are also developed for each performance measure so that actual results and performance can be evaluated against specified requirements.

To provide confidence that the project will achieve the required quality level, a quality assurance process is implemented. By continuously reviewing (or auditing) the actual implementation of the plan developed during quality planning, quality assurance systematically seeks to increase the effectiveness and efficiency of the project and its results.

Actual results are monitored and controlled. This quality-control process provides input to quality assurance as well as a firm basis for acceptance (or rejection) decisions.

8. PROJECT HUMAN RESOURCE MANAGEMENT

8.1. Processes

Project human resource management involves three processes:

1. Organizational planning
2. Staff acquisition
3. Team development

These processes deal with the management of human resources during the project life cycle. The processes are aimed at making the most effective use of people involved with the project. The temporary nature of project organizations, the multidisciplinary teams required in many projects, and the participation of people from different organizations in the same project require special attention to team building, motivation, leadership, and communication in order to succeed.

8.2. Description

The work content of the project is allocated to the performing organizations by integrating the work breakdown structure (WBS) with the organizational breakdown structure (OBS) of the project. At the lowest level, these two hierarchical structures define work packages—specific work content assigned to specific organizational units. The managers of work packages are responsible for managing the building blocks of the projects. Each work package is an elementary project with a specific scope, schedule, budget, and quality requirements. Organizational planning activities are required to ensure that the total work content of the project is assigned and performed by the work packages and integration of the deliverables produced by the work packages into the final product results is possible according to the project plans. The organizational plan defines roles and responsibilities as well as staffing requirements and the OBS of the project.

Based on the organizational plan, staff assignment is performed. Availability of staff is compared to the requirements and gaps identified. These gaps are filled by staff-acquisition activities. The assignment of available staff to the project and the acquisition of new staff result in the creation of a project team that may be a combination of employees assigned full-time to the project, full-time employees assigned part-time to the project, and part-timers.

The assignment of staff to the project is the first step in the team-development process. To succeed in achieving the project goals, a team spirit is needed. The transformation of a group of people assigned to a project into a high-performance team requires leadership, communication skills, and negotiation skills as well as the ability to motivate people, coach and mentor them, and deal with conflicts in a professional yet effective way.

9. PROJECT COMMUNICATIONS MANAGEMENT

9.1. Processes

Project communications management involves four processes:

1. Communications Planning
2. Information Distribution
3. Performance reporting
4. Administrative closure

These processes are required to ensure ‘‘timely and appropriate generation, collection, dissemination, storage, and ultimate disposition of project information’’ (PMI 1996). The processes are tightly linked with organizational planning. The communication within the project team, with stakeholders, and with the external environment can take many forms, including formal and informal communication, written or verbal communication, and planned or ad hoc communication. The decision regarding communication channels, the information that should be distributed, and the best form of communication for each type of information is crucial in supporting teamwork and coordination.

9.2. Description

Communications planning is the process of selecting the communication channels, the modes of communication and the contents of the communication among project participants, stakeholders, and the environment. Taking into account the information needs, the available technology, and constraints on the availability and distribution of information, the communications-management plan specifies the frequency and the methods by which information is collected, stored, retrieved, transmitted, and presented to the parties involved in the project. Based on the plan, data-collection as well as data-storage and retrieval systems can be implemented and used throughout the project life cycle. The project communication system, which supports the transmission and presentation of information, should be designed and established early on to facilitate the transfer of information.

Information distribution is based on the communication-management plan. It is the process of implementing the plan throughout the project life cycle to ensure proper and timely information to the parties involved. In addition to the timely distribution of information, historical records are kept to enable analysis of the project records to support organizational and individual learning. Information related to performances of the project is important. Performance reporting provides stakeholders with information on the actual status of the project, current accomplishments, and forecasts of future project status and progress. Performance reporting is essential for project control because deviations between plans and actual progress trigger control actions.

To facilitate an orderly closure of each phase and of the entire project, information on actual performance levels of the project phases and product is collected and compared to the project plan and product specifications. This verification process ensures an ordered formal acceptance of the project products by the stakeholders and provides a means for record keeping that supports organizational learning.

10. PROJECT RISK MANAGEMENT

10.1. Processes

Project risk management involves four processes:

1. Risk identification
2. Risk quantification
3. Risk response development
4. Risk response control

These processes are designed to evaluate the possible risk events that might influence the project and integrate proper measures to handle uncertainty in the project-planning monitoring and control activities.

10.2. Description

A risk event is a discrete random occurrence that (if occurring) affects the project. Risk events are identified based on the difficulty to achieve the required project outcome (the characteristics of the product or service), constraints on schedules and budgets, and the availability of resources. The environment in which the project is performed is also a potential source of risk. Historical information is an important input in the risk-identification process—knowledge gaps are a common source of risk in projects. Risks are generated by different sources, such as technology—an effort to develop, use, or integrate new technologies in a project creates a knowledge gap and consequently risks. External risks such as new laws or a strike in government agencies may generate project risks. Internal sources within the project or its stakeholders may also do so. The probability of risk events and the magnitude of their affect on the project success are estimated during the risk-quantification process. This process is aimed at an effort to rank risks in order of the probability of occurrence and the level of impact on the project. Thus, a high risk is an event that is highly probable and may cause substantial damage to the project.

Based on the magnitude of risk associated with each risk event, a risk response is developed. Several responses are used in project management, including:

- *Risk elimination:* In some projects it is possible to eliminate some risks altogether by using a different technology, a different supplier, etc.
- *Risk reduction:* If risk elimination is too expensive or impossible, risk reduction may be used by reducing the probability of a risk event or its impact or both. A typical example is redundancy in R&D projects when two mutually exclusive technologies are developed in parallel to reduce the risk that a failure in development will harm the project. Although only one of the alternative technologies will be used, the redundancy reduces the probability of a failure.
- *Risk sharing:* It is possible in some projects to share risks (and benefits) with some stakeholders such as suppliers, subcontractors, partners, or even the client. Another form of risk sharing is with an insurance company.
- *Risk absorption:* If a decision is made to absorb the risk, buffers in the form of management reserve or extra time in the schedule can be used. In addition, contingency plans may be appropriate tools to help in coping with the consequences of risk events.

Since information is collected throughout the life cycle of a project, additional information is used to continuously update the risk-management plan. Risk-response control is a continuous effort to identify new sources of risk, update the estimates regarding probabilities and impacts of risk events, and activate the risk-management plan when needed. Constantly monitoring the project progress in an effort to update the risk-management plan and activate it when necessary can reduce the impact of uncertainty and increase the probability of successful project completion.

11. PROJECT PROCUREMENT MANAGEMENT

11.1. Processes

Project procurement management involves six processes:

1. Procurement planning
2. Solicitation planning
3. Solicitation
4. Source selection
5. Contract administration
6. Contract closeout

These processes are required to acquire goods and services from outside the performing organization (from consultants, subcontractors, suppliers, etc.). The decision to acquire such goods and services (the make or buy decision) has a short-term or tactical level (project-related) impact as well as a long-term or strategic level (organization-related) impact. At the strategic level, core competencies should rarely be outsourced even when outsourcing can reduce the project cost, shorten its duration, reduce its risk, or provide higher quality. At the tactical level, outsourcing can elevate resource shortages, help in closing knowledge gaps, and increase the probability of project success.

Management of the outsourcing process from supplier selection to contract closeout is an important part of the management of many projects.

11.2. Description

The decision on what parts in the project scope and product scope to purchase from outside the performing organization, how to do it, and when to do it is critical to the success of most projects. This is not only because significant parts of many project budgets are candidates for outsourcing, but because the level of uncertainty and consequently the risk involved in outsourcing are quite different from the levels of uncertainty and risk of activities performed in-house. In order to gain a competitive advantage from outsourcing, well-defined planning, execution, and control of outsourcing processes supported by data and models (tools) are needed.

The first step in the process is to consider what parts of the project scope and product scope to outsource. These are decisions regarding sources of capacity and know-how that can help the project in achieving its goal. A conflict may exist between the goals of the project and other goals of the stakeholders. For example, subcontracting may help a potential future competitor develop know-how and capabilities. This was the case with IBM when it outsourced the development of the Disk Operating System (DOS) for the IBM PC to Microsoft and the development of the CPU to Intel. The analysis should take into account the cost, quality, speed, risk, and flexibility of in-house vs. outsourcing. Outsourcing decisions should also take into account the long-term or strategic factors discussed earlier.

Once a decision is made to outsource, a solicitation process is required. Solicitation planning deals with the exact definition of the goods or services required, estimates of the cost and time required, and preparation of a list of potential sources. During solicitation planning, a decision model can be developed, such as a list of required attributes with a relative weight for each attribute and a scale for measuring the attributes of the alternatives. A simple scoring model, as well as more sophisticated decision support models prepared at the solicitation-planning phase, can help in reaching consensus among stakeholders and making the process more objective.

Solicitation can take many forms: a request for proposal (RFP) advertised and open to all potential sources is one extreme, while a direct approach to a single preferred (or only) source is another extreme—with many variations in between, such as the use of electronic commerce. The main output of the solicitation process is one or more proposals for the goods or services required. A well-planned solicitation-planning process followed by a well-managed solicitation process is required to make the next step, source selection, efficient and effective.

Source selection is required whenever more than one adequate source is available. If a proper selection model is developed during the solicitation-planning process and all the data required for the model are collected from the potential suppliers during the solicitation process, source selection is easy, efficient, and effective. Based on the evaluation criteria and organizational policies, proposals are evaluated and ranked. Negotiations with the highest-ranked suppliers can take place to get the best and final offer, and the process is terminated when a contract is signed. If, however, solicitation planning and the solicitation process do not yield a clear set of criteria and a selection model, source selection may become a difficult and time-consuming process; it may not end with the best supplier selected or the best possible contract signed. It is difficult to compare proposals that are not structured according to clear RFP requirements; in many cases important information is missing in the proposals.

Throughout the life cycle of a project, contracts are managed as part of the execution and change control efforts. Work results are submitted and evaluated, payments are made, and, when necessary, change requests are issued. When these are approved, changes are made to the contracts. Contract management is equivalent to the management of a work package performed in-house, and therefore similar tools are required during the contract-administration process.

Contract closeout is the closing process that signifies formal acceptance and closure. Based on the original contract and all the approved changes, the goods or services provided are evaluated and, if accepted, payment is made and the contract closed. Information collected during this process is important for future projects and supplier selection because effective management is based on such information.

12. THE LEARNING ORGANIZATION: CONTINUOUS IMPROVEMENT IN PROJECT MANAGEMENT

12.1. Individual Learning and Organizational Learning

Excellence in project management is based on the ability of individuals to initiate, plan, execute, control, and terminate the project scope and product scope successfully. The ability of individuals to master product- and project-related processes is the foundation on which organizational learning is built. Individual learning can take many forms, including the learning of verbal knowledge, intellectual skills, cognitive strategies, and attitudes. The learning mechanism can also take many forms, including learning by imitation of other people or learning by repetition of a process.

The ability of groups to improve performances by learning is also very important. Katzenbach and Smith (1993) explain how important it is to combine individual learning with team building.

A project team must combine these two learning processes in order to succeed. As it is important for each individual to learn and master his part in the product scope and in the project scope, it is equally important for the group to learn how to work as a team. Team building and organizational learning are important in the project environment. Establishing clear processes in which the input to each process is well defined and the individuals responsible for the process master the tools and techniques required to do the process right and to produce the desired output enables excellence in project management to be achieved.

12.2. Workflow and Process Design as the Basis of Learning

Successful project management requires successful planning, execution, and control of project scope and the product scope. The one-time, nonrepetitive nature of projects makes uncertainty a major factor affecting a project's success. In addition, the ability to learn by repetition is limited because most projects are unique. A key to project-management success is the exploitation of the repetitive parts of project scope. Identifying repetitive processes (repetitiveness within the project as well as repetitiveness between projects) and building an environment that supports learning and data collection enhances competitiveness in project management.

A key tool in building a learning–supporting environment is the design and implementation of a workflow-management system–a system that defines, manages, supports, and executes information-processing and decision-making processes. Each of the processes discussed in this chapter should be studied, defined, and implemented within the workflow management system. The definition includes the trigger (which initiates the process) of the process, inputs to the process, the participants in the process, the activities performed and required data processing, models used, the order or sequence of processing, and finally, process termination conditions and the process results or deliverables. The workflow-management system employs a workflow-enactment system or workflow-process engines that can create, manage, and execute multiple process instances.

By identifying the repetitive processes shared by many projects performed by an organization, it is possible to implement a workflow system that supports and even automates the repetitive processes. Automation means that the routing of each process is defined along with the input information, processing, and output information. Thus, although the product scope may vary substantially from project to project, the execution of the project scope is supported by an automatic workflow system that reduces the level of uncertainty (processes are clearly defined and the flow of information required to support these processes is automatic) and enables learning. The well-structured process can be taught to new employees or learned by repetition. In projects performed by the organization, the same processes are repeated, the same formats are used to present information, and the same models support decision making.

Definition of processes and the support of these processes by a workflow-management system are key to the success of organizations dealing repeatedly with projects.

REFERENCES

Garvin, A. D. (1987), "Competing on the Eight Dimensions of Quality," *Harvard Business Review,* Vol. 65, No. 6, November–December, pp. 107–109.

Katzenbach, R. J., and Smith K. D. (1993), *The Wisdom of Teams,* Harvard Business School Press, Boston.

Project Management Institute (PMI) (1996), *A Guide to the Project Management Body of Knowledge,*, PMI, Upper Darby, PA.

Morris, P. W. G. (1981), "Managing Project Interfaces: Key Points for Project Success," in *Project Management Handbook,* 2nd Ed., Cleland, D. I., and King, W. R., Eds., Prentice Hall, Englewood Cliffs, NJ.

Muench, D. (1994), *The Sybase Development Framework,* Sybase, Oakland, CA.

Shtub, A., Bard, J., and Globerson, S. (1994), *Project Management Engineering, Technology, and Implementation,* Prentice Hall, Englewood Cliffs, NJ.

U.S. Department of Defense (1993), DoD Directive 5000.2.

CHAPTER 46

Computer-Aided Project Management

Carl N. Belack
Oak Associates, Inc.

1. INTRODUCTION 1252

2. THE HISTORY OF COMPUTER-AIDED PROJECT MANAGEMENT 1253

3. THE PROJECT CONCENTRIC CIRCLE MODEL 1253
  - 3.1. Project Management Core Processes 1254
  - 3.2. Project Management Support Processes 1254
  - 3.3. Senior Management Leadership 1255

4. THE CAPM PLATFORM 1255
  - 4.1. Automating the Project Management Core Processes 1256
    - 4.1.1. Scope, Time, Cost, and Resource Management 1256
    - 4.1.2. Risk Management 1257
    - 4.1.3. Change Management 1259
    - 4.1.4. Communications Management 1259
  - 4.2. Automating the Organizational Support Processes 1260

5. IMPLEMENTING CAPM 1260

6. CAPM IN THE 21st CENTURY 1261

REFERENCES 1262

ADDITIONAL READING 1262

APPENDIX: TRADEMARK NOTICES 1262

1. INTRODUCTION

Although the art of managing projects has existed for thousands of years, it is only relatively recently that modern project management techniques were developed, codified, and implemented in any consistent, methodological manner. Perhaps it is not totally coincidental that these techniques came about during the same period as the beginning of the commercial availability of computers as we know them today. As of the writing of this book, we have arrived at a point where hundreds of commercial software applications that address various aspects of the project management processes are available for purchase. For one who is unfamiliar with the wide variety of these tools on the market, the process of selecting and implementing these tools can be a daunting task. The purpose of this chapter is to help make that process somewhat more manageable.

The reader should note that the title of this chapter, "Computer-Aided Project Management (CAPM)," was specifically intended to allow for the discussion of the different types of tools available to facilitate the entire process of project management. To many people, the consideration of project management tools is unfortunately limited to what has become known as *project management software*. Granted, this type of tool was among the first applications developed with project management in mind, and it does address an important aspect of project management, namely scheduling and tracking. This author believes, however, that such a tool is but one among many in what we will refer to as a *computer-aided project management platform*. A CAPM platform is one that includes applications that automate many of the project management processes, not just scheduling and tracking, and it is this platform that we will discuss below.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition. Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

The following pages will address the history of the development of CAPM, the use of computers in project management, and the implementation of CAPM platforms in business organizations, and will look at what we might expect of these tools in the years to come. It is hoped that this examination will help the reader better understand the use of CAPM tools while facilitating their selection and successful implementation.

2. THE HISTORY OF COMPUTER-AIDED PROJECT MANAGEMENT

While it may be difficult for some to believe (particularly for those to whom the slide rule is merely a curious artifact from an ancient civilization), modern project management techniques were at one time employed by those who did not have the advantage of using computers. In fact, some companies still don't use computers for project management in any organized fashion. And, as laborious as the work is, it is perfectly possible to put together a good project plan without the use of computers. However, having spent some time doing just that, the author can attest to the enormous amount of time and resources such an undertaking consumes. And, once changes are introduced to the initial project plan, the incremental use of time and resources expands exponentially.

In the late 1950s and early 1960s, a few companies began to develop their own internal software tools for managing projects, some of which are still in use today. With the apparent increasing need for such tools, commercial applications began to appear in the marketplace. Most of these tools were initially used by U.S. government contractors (who were required by the Department of the Energy or the Department of Defense to adhere to rules for managing government contracts) and were implemented on mainframe computers—the only computers commercially available at that time. These applications were the predecessors to tools that are still available today (although in a much different form), such as Artemis and Primavera. At the time, these tools were both difficult to learn and cumbersome to use. Since graphical user interfaces did not yet exist, command language was used to interface with the application. Nonetheless, since they enabled some automation of scheduling, tracking, and reporting activities, they were a welcome change for most project managers, who were used to performing these same tasks by hand.

The advent of commercially available PCs brought about the development of project management tools specifically aimed at the PC market. These tools (which we will refer to as low-end tools, as distinguished from the high-end tools that run on mainframes and minicomputers) were much less expensive than their high-end counterparts and were much easier to learn and to use, but also had far fewer capabilities. These tools also allowed the user to interface with the application through a rudimentary graphical user interface (GUI). These tools included software applications such as Harvard Project Manager, SuperProject, Project Manager's Workbench, and Microsoft Project. These tools were primarily aimed at the IBM-compatible marketplace. There were fewer tools available for Apple Macintosh computers, such as MacProject.

Over the past few years, the manufacturers of the high-end tools have incorporated GUIs and other devices to make their tools user friendly. At the same time, the makers of the low-end tools began building more capabilities into their applications that had previously been available only in the high-end tools. Some formerly low-end tools, such as Project Workbench, have migrated into the realm of high-end tools. And a number of high-end tool manufacturers have produced low-end tools for managing individual projects whose files can then be integrated into the high-end tools (such as Sure Trak for Primavera). As confusing as this all sounds, all of these software manufacturers have been trying to achieve the same end: to develop a tool that balances ease of learning and use with ever-increasing capabilities.

As the profession of project management began to gain acceptance in the workplace, additional applications became commercially available. These range from tools that automate other project management processes (such as risk-management tools like @Risk) to tools that help manage areas that are ancillary to, but have a direct impact upon, project management (such as multiproject resource management tools like ResSolution). With the availability of all of these different types of tools, it is often a difficult proposition deciding which tools, if any, are appropriate for a specific organization. In the next sections, we will discuss the processes that are involved in, or have an impact upon, project management and see how the use of computer tools can facilitate these processes.

3. THE PROJECT CONCENTRIC CIRCLE MODEL

In any given organization, generally two types of work activities take place. The first type, the one with which most people are familiar, is *operations* work. The activities in operations work have the following characteristics:

- They are repetitive (they occur over and over from fiscal quarter to fiscal quarter and from fiscal year to fiscal year).
- Their end result is essentially the same (production of financial reports, operations reports, etc.).

The second type, which we are addressing in this chapter, is *project* work. As one might expect, project work is characterized by work that is (1) not repetitive, but rather time limited (it has a specific beginning and end), and (2) produces a unique product or service. Project management is the set of activities involved in managing project work.

When looking at the set of processes involved in or surrounding project management, it is useful to use a framework that the author has called the Project Concentric Circle Model.* This model is depicted in Figure 1.

The model consists of three concentric circles. Each circle represents a level at which project management processes, or processes affecting the project management processes, take place. Each will be briefly discussed below.

3.1. Project Management Core Processes

The center circle represents the project management core processes, or processes that function within individual projects. The reader can find a detailed description of these processes in (PMI 1996) (PMBOK Guide). In brief, these are areas that address the following project management activities at the individual project level:

- Time management
- Scope management
- Cost management
- Risk management
- Quality management
- Human resources management
- Procurement management
- Communications management
- Integration management

The *PMBOK Guide* also describes the five project management processes throughout which activities in each above-noted area of management need to be performed. These processes are portrayed in Figure 2. It is these management areas and processes that most organizations associate with project management. And some assume that attending to these alone will ensure successful organizational project work. That, however, is a fatal error for many organizations. In order to achieve a high level of competence in project management, two other levels must also be addressed.

3.2. Project Management Support Processes

The second circle represents the project management support processes level and includes processes that occur outside of the day-to-day activities of the individual project teams. The activities within these processes generally comprise operational activities, not project activities, that support project

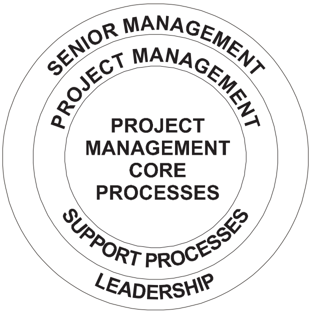


Figure 1 Project Concentric Circle Model.

---

* The reader is advised that the Project Concentric Circle Model is a copyright of Oak Associates, Inc. Any reproduction or use of the model without the express consent of Oak Associates, Inc. and the publisher is strictly prohibited.

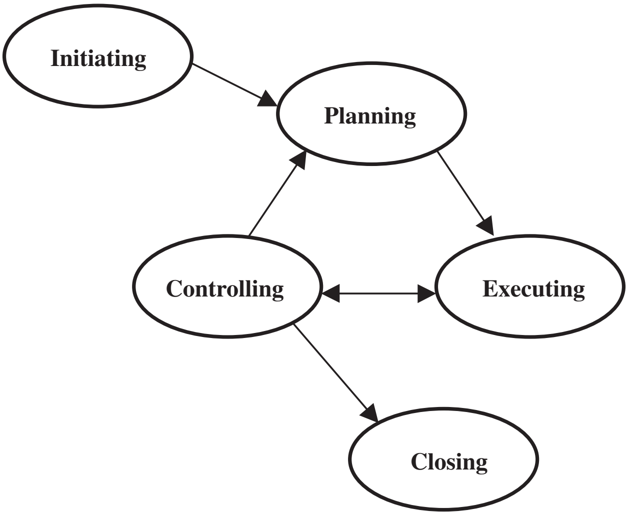


Figure 2 Project Management Processes.

work and project teams. In fact, these can best be described as processes that establish and maintain an environment in which project work can be successfully performed.

These processes can be grouped into two categories:

1. *Organizational development:* Activities that occur in these processes include assessing organizational and individual competency; development and updating of job descriptions for people at all levels of the project team; design of career paths for project managers and project team members; project manager selection and competency development; project team member selection and support; and training and mentoring.
2. *Communications and knowledge systems:* Activities that occur in the development and maintenance of these systems focus on interproject and intraproject communications; project reporting systems (for project team members, customers, and senior management); in-progress and postproject reviews and audits; development and maintenance of historical activity and estimating databases; capacity planning; project change management; and project and technical document/software configuration management.

3.3. Senior Management Leadership

The outermost circle represents processes that senior management must undertake in order to promote project-friendly corporate environments. This involves:

- Championing project management within the organization: This is done by understanding project management and project work within the organizational context; leading the change effort to enhance the role of projects and project management; prioritizing project work to enable effective resource management; and managing the portfolio of projects to ensure alignment with corporate goals.
- Creating and enabling the culture of project success: This includes fostering open and honest communication; promoting rational risk taking; supporting the need for project planning; valuing the differences of project and functional management; and encouraging “quiet” projects (and discouraging “heroic” projects).

Now that we have constructed a framework for the processes that need to work effectively in order for projects to be successful, let us look at the types of software applications that could automate many of these processes.

4. THE CAPM PLATFORM

An organization needs to have available for project managers, project teams, line managers, and senior managers a tool set that facilitates the activities of the management processes noted above.

An example of one type of tool frequently used in project management is a list of items that, when completed, would signify the completion of a project deliverable—a punch list is one such list that is regularly used to this day in construction projects. More and more, these tools are being incorporated into computer applications. In this section, we will take a look at tools that are available, or are being constructed, to automate the concentric circle processes.

4.1. Automating the Project Management Core Processes

Before proceeding, a brief word of caution is in order. It is the mistaken belief of many that in order to manage projects effectively, one merely needs to purchase a project management tool and become trained in use of the tool (the ‘‘buy ’em a tool and send ’em to school’’ approach). This is possibly the worst approach that could be taken to improve the effectiveness of project management in an organization. We have already noted that project management predated the commercially available tools to aid that endeavor, so we know that it is possible to manage projects effectively without the use of automation. The single most important thing to remember about these tools is that it is not the tool, but rather the people using the tool, who manage the projects. *In order for people to use the tools properly, they must first master the techniques upon which these tools are based.*

As an example, to develop useful data for scope, time, and cost management, the successful tool user must have a working knowledge of scope statement development; work definition (through work breakdown structures or other such techniques): activity estimating (three-point estimating of both effort and duration); precedence diagramming method (also known as project network diagramming); and progress-evaluation techniques (such as earned value). Expecting success through the use of a tool without a thorough prior grounding in these techniques is like expecting someone who has no grounding in the basics of writing (grammar, syntax, writing technique) to use a word-processing application to produce a novel. Some novels on the market notwithstanding, it just does not happen that way. With this firmly in mind, let’s look at the types of tools one might use in modern project management.

4.1.1. Scope, Time, Cost, and Resource Management

The preponderance of tools on the market today are those that aid project managers in time and cost management (commonly called schedule and budget management). In addition, many of these tools include resource management. These tools can be helpful in:

- Developing activity lists (project scope) and displaying work breakdown structures
- Noting activity estimates (in some cases, calculating ‘‘most likely’’ estimates for three-point estimating techniques)
- Assigning dependencies (precedence structure) among activities
- Calculating and displaying precedence diagrams (PERT charts)
- Calculating and displaying project schedules (Gantt charts)
- Assigning individual or group resources
- Setting and displaying calendars (both for the project and for individual resources)
- Calculating project costs (for various types of resources)
- Entering time card and resource usage data

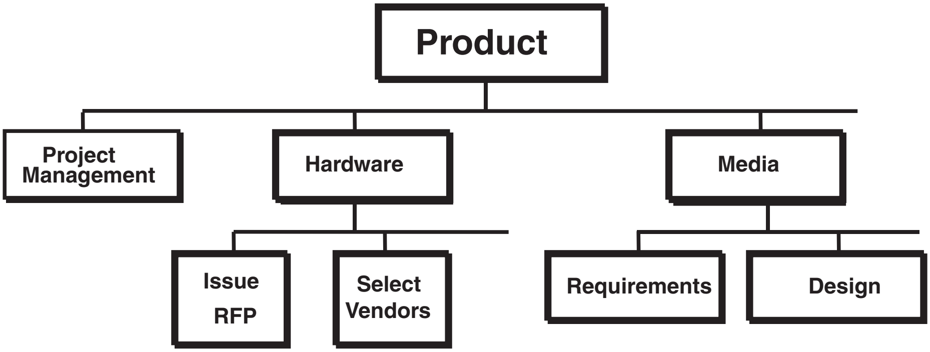


Figure 3 Work Breakdown Structure.

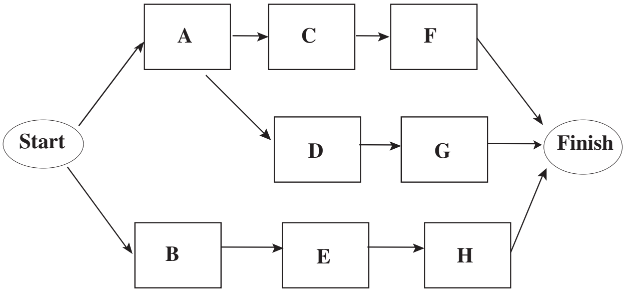


Figure 4 Precedence Diagram.

- Tracking project cost, schedule, and resource data
- Rescheduling and recalculating schedule and cost data after input of activity actual data
- Calculating and displaying project progress
- Leveling resources
- Displaying resource histograms
- Sorting and filtering for various scenarios
- Generating reports for use by various project stakeholders

These are just some of the capabilities that can be found in these tools (see the section on selecting tools below for a citation of an extended list of these capabilities). The tools that can be used for such an effort are too numerous to list. Examples of these are Microsoft Project 98, PS7, and Artemis. For some of the low-end tools (particularly for MS Project 98), there is an after-market of tools that can be used in conjunction with the primary tool to help it do its job more effectively. These range from tools like GRANEDA Dynamic (which provides an excellent graphical interface to print professional-looking precedence diagrams, Gantt charts, and work breakdown structures) to tools such as Project 98 Plus (which provides a very user-friendly interface for sorting and filtering for MS Project 98).

4.1.2. Risk Management

Since two characteristics that we have attributed to project work are its unique nature and its time limitations, projects are inherently risky. Many projects run into problems or fail altogether because an inadequate job was done around risk management. Project risk management is a three-step process that involves:

1. Identifying, assessing, and documenting all potential project risks
2. Developing risk avoidance and mitigation plans
3. Implementing these plans when the risks occur

Clearly, this is not a process that ends once the project-planning activities have been completed. Rather, project managers need to monitor and assess potential project risk throughout the entire conduct of the project.

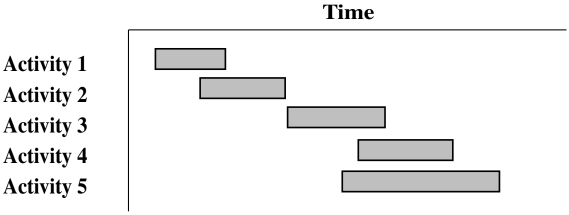
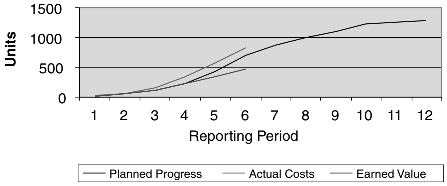


Figure 5 Gantt Chart.



	1	2	3	4	5	6	7	8	9	10	11	12
Planned Progress	20	50	110	225	425	695	875	1005	1105	1230	1260	1280
Actual Costs	30	60	151	337	567	827						
Earned Value	20	50	110	225	345	475						

Figure 6 Earned Value S Curves.

One type of risk that all project managers face is that associated with project schedules. A typical method for handling this risk is to run Monte Carlo simulations on the project precedence diagram (PERT chart). This is done by (1) assigning random durations (within predefined three-point activity estimates) to individual activities, (2) calculating project duration over and over for hundreds (sometimes thousands) of repetitions, and (3) analyzing the distribution of probable outcomes of project duration. There are a number of tools on the market that perform these tasks. Two of the most popular are @Risk and Risk+. Other fine tools are also available that perform similarly. These tools can perform simulations on practically any project calculations that lend themselves to numerical analysis. The output of these tools is the analysis of probable project durations in both numerical and graphical formats (see Figure 7).

Why is it important to use tools like these to help us manage risk? Quite simply, single-point project estimates are rarely, if ever, met. Project managers need to understand the probable range of outcomes of both project cost and duration so they can make informed decisions around a host of project issues (e.g., setting project team goals, deciding when to hire project personnel). They also need this information to set proper expectations and conduct intelligent discussions with the project team members, senior managers, and customers. The correct use of such tools can help project managers do just that.

In addition to these tools, other tools are available to help track the status of potential risk events over the course of a project. One such tool, Risk Radar, was designed to help managers of software-intensive development programs. Regardless of the intended target audience, this tool can be quite helpful for any type of project risk-tracking effort. With the proper input of risk data, it displays a

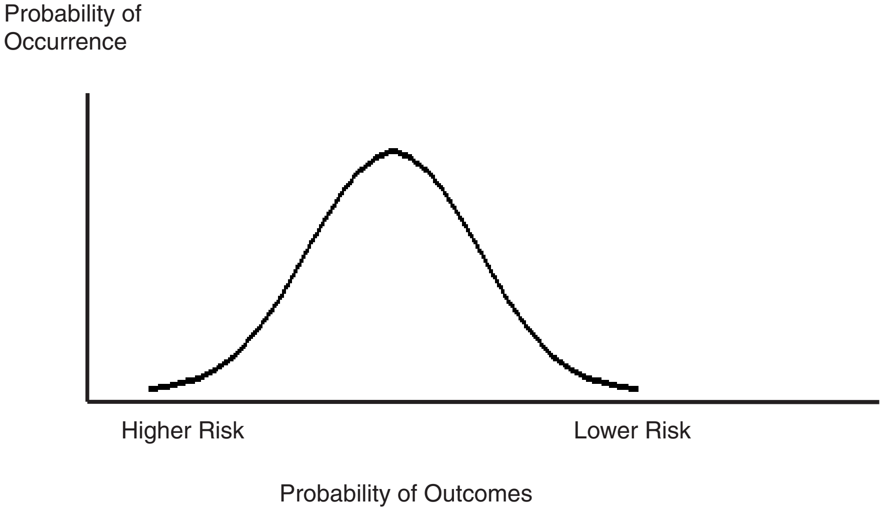


Figure 7 Project Outcome Probability Curve.

graphic depicting the number of risk events with similar risk exposure and lays them out on an easily understood grid. This is a common way to track risk. An example of a similar grid is shown below (see Figure 8).

4.1.3. Change Management

As noted earlier, there are two types of changes with which project managers need be concerned. The first is a change in the scope of work of the project. Most projects encounter scope changes during the evolution of work on the project. Since scope changes almost always result in budget and schedule changes, it is very important to track them accurately. This can usually be done by using the scope, time, cost, and resource-management software discussed above.

The second type of change is one that addresses changes in technical project documentation. Technical drawings, quality documents, and electrical wiring diagrams are examples of such documents. There are a number of tools available for these efforts, and they are as diverse as the technical functions that might employ them. They are generically known as configuration management tools. While these will not be addressed in this chapter, project functional teams should make every effort to select tools like these that will help them manage these documents so that current versions are available to all project members who need them.

4.1.4. Communications Management

Communications skills are arguably the most important skill of project management. Similarly, communications tools can be considered among the most important project tools. As noted in the *PMBOK Guide,*

> Project Communications Management includes the processes required to ensure timely and appropriate generation, collection, dissemination, storage, and ultimate disposition of project information. It provides the critical links among people, ideas, and information that are necessary for success. Everyone involved in the project must be prepared to send and receive communications in the project ‘‘language’’ and must understand how the communications they are involved in as individuals affect the project as a whole.

Tools that aid managers in project communications are not terribly different from those that are used in operations communications. They include:

- Word processors (e.g., WordPerfect, MS Word)
- Presentation tools (e.g., MS PowerPoint, Corel PRESENTS)
- Spreadsheets (e.g., Lotus 1-2-3, MS Excel)
- Individual and workgroup communications tools (e.g., e-mail, Lotus Notes)

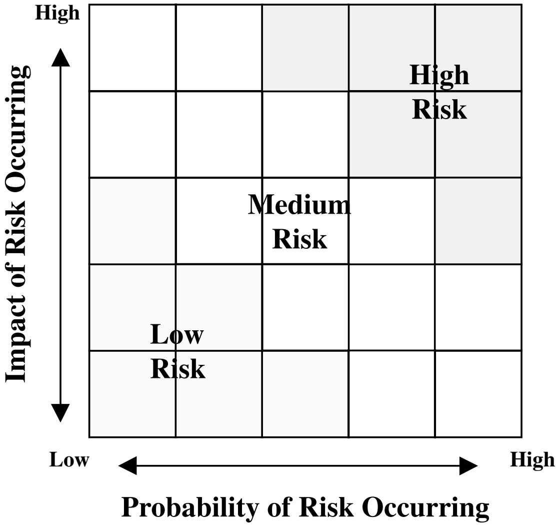


Figure 8 Project Risk Grid.

With the availability of Internet communications and the advent of tools similar to those noted above, tool sets such as these should be readily available for use by project teams.

4.2. Automating the Organizational Support Processes

Like the communications tools discussed above, tools that are useful in the organizational support processes are those that have been used for some time in operations management. Operations management processes are, after all, operations work (line or functional management) as opposed to project work (project management). Since operations management has been taught in business schools for decades, there are tools on the market that can aid in various aspects of these endeavors. While these tools are too numerous to discuss here, some have been designed with the express purpose of supporting project management activity. Among them are:

- *Multiproject resource-management tools:* These tools help line managers manage scarce resources across the many projects in which their organizations are involved. They include tools such as ResSolution* and Business Engine.
- *Project portfolio management tools:* These are tools that help senior managers balance the accomplishment of their organizational goals across the range of projects, both ongoing and potential, in their organizations. They address issues such as budget, benefits, market, product line, probability of success, technical objectives, and ROI to help them prioritize and undertake projects. One such tool that does this is the project portfolio module of Portfolio Plus.
- *Activity and project historical databases:* These are tools that help a project team and all managers more accurately estimate the outcomes of their projects. Among the many problems that arise in projects, an unrealistic expectation about project outcomes is one of the most flagrant. One reason for these unrealistic expectations is the poor quality of activity-level estimates. One way to increase the accuracy of these estimates is to employ three-point estimating techniques, which have been referred to above. An even better way of increasing the accuracy of estimates is to base them on historical data. Were one to examine the activities that are performed over time in an organization's projects, it would become apparent that many of the same types of activities are performed over and over from one project to another. In some cases, nearly 80% of these activities are repeated from one project to another. Unfortunately, in many organizations, such historical data is rarely available for project team members to use for estimating. Consequently, three-point estimating techniques need to be universally employed, project after project. Once organizations develop, maintain, and properly employ accurate activity historical databases, the need for the relatively less accurate three-point estimates (remember that single-point estimates are much less accurate than three-point estimates) will be reduced, thereby resulting in more accurate estimates at both the activity and project levels.

Finally, we should mention that while integration of all the types of tools discussed is probably technologically possible, it is not always either necessary or desirable. In fact, it is the author's belief that in some instances, particularly in the case of multiple-project resource management, it is better to do detailed management of project resources within the context of the center circle and less detailed management at a higher level within the context of the outer circles without daily integration of the two activities.

5. IMPLEMENTING CAPM

The selection, implementation, and use of these tools are not tasks to be taken lightly. And while the process may at first seem daunting, there are ways to make it easier. A number of sources can aid in the selection process. At least two publications presently do annual software surveys in which they compare the capabilities of various project management tools. Many of these tools perform the functions discussed above. These publications are *IIE Solutions,* a monthly publication of the Institute of Industrial Engineers, and *PMnet,* a monthly publication of the Project Management Institute. The National Software Testing Laboratories (NTSL) also tests and compares software programs. It makes these comparisons in over 50 categories of tool capabilities for project management software. The major areas of comparison include:

- Performance
- Versatility

---

* In the interest of fairness and full disclosure, the author must acknowledge that the organization in which he is a principal is a reseller of both ResSolution and Project 98 Plus software, both of which are cited in this chapter.

- Quality
- Ease of learning
- Ease of use

Individuals responsible for such selection need to ask the following types of questions:

- What is my organization trying to accomplish with this software? Will the software tools being considered meet those needs?
- How many people will be using the software—one person, a group of people, or an entire organization?
- Have the users of the software previously used any other type of project management software before? If so, what were the tools, and were they similar to any of the tools presently being considered?
- Have the users of the software been trained in project management methods, tools, and techniques?
- Are the tools being considered both easy to learn and easy to use?
- Can the tool be used as is or are modifications required?
- What type of postinstallation support is required? Will the vendor do the support or does it require an in-house support group?
- Does the tool need to be integrated with other tools being used in the organization? If so, how difficult will that integration be?
- What are the implications of introducing software of this sort into my organization? Do I need to develop a formal change management plan to get organizational buy-in for its introduction and use?

The answers to all of these questions can have a profound impact on the success of the tool in an organization. One needs to be especially careful in considering the last question. The human implications of introducing software tools in an organization are frequently underestimated. This underestimation has caused organizations to be unsuccessful in the introduction and implementation of these tools, resulting in wasted effort and dollars and in the frustration of those project stakeholders who were affected by the failed effort.

For many reasons, tool-selection processes can at times resemble religious wars. Participation in such a process is not for the fainthearted. Anyone contemplating the introduction of these tools into an organization would be well advised to develop a detailed project plan. Included in this project plan should be a plan to ease the introduction of the tool into the organization, thereby allowing for the greatest probability of a successful introduction and implementation. As with any project, a competent project team needs to be assembled with a specific individual assigned responsibility to manage the project. There should be senior management support and involvement appropriate to the effort. Expectations of all organizational stakeholders need to be set and met throughout the conduct of the project. These expectations should include (1) a detailed description of what will be done during the project, (2) who needs to be involved, and (3) how the implementation of the tool will affect members of the organization. Once project execution has begun, and throughout the course of the project, frequent progress reviews need to take place to ensure that the implementation is on schedule, on budget, and meets the needs of the project stakeholders. These efforts will go far in aiding in the integration of the tool into regular use in project and operations work

6. CAPM IN THE 21st CENTURY

With the rapid development and introduction of software onto the marketplace, some of what has been described above may soon be out of date. One thing that will surely not vanish, however, is the ever-increasing need of project managers and organizations for tools to help them accomplish their complex and difficult jobs. While once just nice to have, these tools are now a necessity. So what does the future hold for computer-aided project management?

More and more tools are expanding from those aimed at individual use to those available for workgroups. Projects are, after all, team endeavors. MS Project 2000 includes an Internet browser module called Microsoft Project Central that is aimed at allowing the collaborative involvement of project stakeholders in planning and tracking projects, as well as access to important project information. With the ever-increasing demand for accurate project information, coupled with the cross-geographical nature of many project efforts, Web-based project communications tools will surely also become a requirement and not just a convenience. The author has worked with a few companies that have already developed these tools for their internal use. It is also inevitable that at some point in the not too distant future, complete tool sets that incorporate and integrate many of the varied capabilities described in the paragraphs above will also become available for commercial use. It is only a matter of time before such software applications will be developed and appear on the shelves of your electronic shopping sites.

However, despite the advances in technology that will inevitably lead to this availability, the age-old problems of successful selection, introduction, and implementation of these tools will remain. If organizations take the time to accomplish these tasks in a cogent and supportive way, the tools will continue to be a significant benefit in the successful implementation of the project management processes.

REFERENCES

Project Management Institute (PMI) (1996), *A Guide to the Project Management Body of Knowledge,* PMI, Upper Darby, PA.

ADDITIONAL READING

Graham, R. J., and Englund, R. L., *Creating an Environment for Successful Projects,* Jossey-Bass, San Francisco, 1977.

Kerzner, H., *Project Management: A Systems Approach to Planning, Scheduling, and Controlling,* 5th Ed., Van Nostrand Reinhold, New York, 1995.

Martin, J., *An Information Systems Manifesto*, Prentice Hall, Englewood Cliffs, NJ, 1984.

National Software Testing Laboratory (NTSL), ''Project Management Programs,'' *Software Digest,* Vol. 7, No. 16, 1990.

Oak Associates, Inc., *Winning Project Management*, Course Notebook, Maynard, MA, 1999.

APPENDIX
Trademark Notices

Artemis is a registered trademark of Artemis Management Systems.
Business Engine is a registered trademark of Business Engine Software Corp.
GRANEDA Dynamic is a registered trademark of Netronic Software GMBH.
Harvard Project Manager is a registered trademark of Harvard Software, Inc.
Macintosh and *MacProject* are registered trademarks of Apple Computer Corp.
Microsoft Project, *Microsoft Project 98*, and *Microsoft Project 2000* are registered trademarks of Microsoft Corp.
Portfolio Plus is a registered trademark of Strategic Dynamics, Ltd.
ResSolution is a registered trademark of, and *Project 98 Plus* is a product of, Scheuring Projektmanagement.
Primavera and *SureTrak* are registered trademarks of Primavera Systems, Inc.
PS7 is a registered trademark of Scitor Corp.
Project Workbench is a registered trademark of Applied Business Technology Corp.
@Risk is a registered trademark of Palisade Corp.
Risk+ is a registered trademark of ProjectGear, Inc.
Risk Radar is a product of Software Program Managers Network.
SuperProject is a registered trademark of Computer Associates, Inc.

CHAPTER 47

Work Breakdown Structure

BOAZ GOLANY
AVRAHAM SHTUB
Technion—Israel Institute of Technology

1. **INTRODUCTION: DIVISION OF LABOR AND ORGANIZATIONAL STRUCTURES** **1264**
   1.1. Introduction 1264
   1.2. Organizational Structures 1264
   1.3. The Functional Structure 1265
   1.4. The Project Structure 1265
   1.5. The Matrix Structure 1265
2. **HIERARCHIES IN THE PROJECT ENVIRONMENT: THE NEED FOR A WORK BREAKDOWN STRUCTURE** **1266**
   2.1. Introduction 1266
   2.2. The Scope 1266
   2.3. Implementing Division of Labor in Projects 1266
   2.4. Coordination and Integration 1266
3. **THE RELATIONSHIP BETWEEN THE PROJECT ORGANIZATION AND THE WORK BREAKDOWN STRUCTURE** **1267**
   3.1. Introduction 1267
   3.2. Responsibility 1267
   3.3. Authority 1268
   3.4. Accountability 1268
4. **THE DESIGN OF A WORK BREAKDOWN STRUCTURE AND TYPES OF WORK BREAKDOWN STRUCTURES** **1268**
   4.1. Introduction 1268
   4.2. A WBS Based on Technology 1269
   4.3. A WBS Based on Project Life Cycle 1269
   4.4. A WBS Based on Geography 1269
   4.5. Other WBS Designs 1271
   4.6. Discussion 1271
5. **WORK PACKAGES: DEFINITION AND USE** **1272**
   5.1. Introduction 1272
   5.2. Definition of Work Packages 1272
   5.3. Definition of Cost Accounts 1272
6. **USING THE WORK BREAKDOWN STRUCTURE: EXAMPLES** **1273**
   6.1. R&D Projects 1273
   6.2. Equipment-Replacement Projects 1273
   6.3. Military Projects 1273
7. **CHANGE CONTROL OF THE WORK BREAKDOWN STRUCTURE** **1274**
   7.1. Introduction 1274
   7.2. Change Initiation 1276
   7.3. Change Approval 1276
   7.4. Change Implementation 1276
8. **THE WORK BREAKDOWN STRUCTURE AND THE LEARNING ORGANIZATION** **1276**
   8.1. Introduction 1276
   8.2. WBS as a Dictionary for the Project 1277
   8.3. Information Storage and Retrieval 1277

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

8.4. The Learning Organization 1277
REFERENCES 1277
ADDITIONAL READING 1278
APPENDIX: TABLE OF CONTENTS FOR A SOW DOCUMENT 1278

1. INTRODUCTION: DIVISION OF LABOR AND ORGANIZATIONAL STRUCTURES

1.1. Introduction

Division of labor is a management approach based on the breaking up of a process into a series of small tasks so that each task can be assigned to a different worker. Division of labor narrows the scope of work each worker has to learn enabling workers to learn new jobs quickly and providing an environment where each worker can be equipped with special tools and techniques required to do his job.

Some advantages of division of labor and specialization are:

- The fast development of a high degree of skill (specialization)
- The saving of set-up time required to change from one type of work to another
- The use of special-purpose, usually very efficient, machines, tools, and techniques developed for specific tasks.

These benefits do not come for free. Division of labor requires integration of the outputs produced by the different workers into the final product. Thus, some of the efficiency gained by specialization is lost to the additional effort of integration management required.

A common way to achieve integration is by a proper organizational structure, a structure that defines roles and responsibilities of each person as well as the inputs required and the tools and techniques used to produce that person's outputs.

This chapter discusses division of labor in projects. Section 1 deals with different organizational structures. Section 2 focuses on the work breakdown structure (WBS) as a tool that supports division of labor in projects. Section 3 discusses the relationship between the project organizational structure and the WBS, and Section 4 presents types of work breakdown structures along with a discussion on the design of a WBS. Section 5 discusses the building blocks of a WBS, known as work packages, and Section 6 discusses how the WBS should be used and managed in projects. Finally, Sections 7 and 8 present the issues of individual learning and organizational learning in the context of the support a WBS can provide to the learning process.

1.2. Organizational Structures

Organizations are as old as mankind. Survival forced people to organize into families, tribes, and communities to provide for basic needs (security, food, shelter, etc.) that a single person had difficulty providing. Kingdoms and empires of the ancient world emerged as more formal organizations. While these organizations had long-term goals, other organizations were created to achieve specific unique goals within a limited time frame. Some ambitious undertakings that required the coordinated work of many thousands of people, like the construction of the Pyramids, Great Wall of China, or the Jewish Temple, motivated the development of ad hoc organizations.

As organizations grew larger and spread geographically, communication lines and clear definitions of roles became crucial. Formal organizations based on a hierarchical structure were established. The hierarchical structure emerged due to the limited ability of a supervisor to manage too many subordinates. This phenomenon, known as the limited span of control, limits the number of subordinates one can supervise effectively. The role, responsibility, and authority of each person in the organization were defined. A typical example is documented in the Bible where Moses divided the people of Israel into groups of 10 and clustered every 10 of these basic groups into larger groups of 100, and so on. The underlying assumption in this case is that the span of control is 10. Clustering was based on family relationships. Formal authority was defined and lines of communication were established to form a hierarchical organizational structure. The idea of a formal organization, where roles are defined and communication lines are established, is a cornerstone in the modern business world. Division of labor and specialization are basic building blocks in modern organizations. There is a large variety of organizational designs; some are designed to support repetitive (ongoing) operations, while others are designed to support unique one-time efforts. A model known as the organizational structure is frequently used to represent lines of communication, authority, and responsibility in business organizations.

1.3. The Functional Structure

The functional organization is designed to support repetitive activities over a long (indefinite) period of time. It is a hierarchical structure in which roles are based on the function or specialization of the workers involved. Functions like marketing, finance, human resources, engineering, production, and logistics are common. In large organizations, each function is subdivided further, to the point that a proper span of control is achieved. For example, the marketing department can be divided geographically: marketing in Europe, the United States, Asia, and Africa. Engineering departments can be subdivided into electrical engineering, mechanical engineering, and industrial engineering. In a functional organization, the role of each organizational unit is to deal with the work content related to its function. Although fine tuning is required to define the exact border lines and interfaces between the different functions, division of labor is (naturally) along the functional lines.

An advantage of functional organization stems from the pooling together of similar resources: when all the electrical engineers are pooled together into one organizational unit, efficiency and effectiveness are achieved. Furthermore, workers in the same organizational unit (same function) share similar knowledge, education, and experience. They can learn from each other, and the flow of information within organizational units is natural. The stability of this organizational structure promotes career development to the point that people spend their entire career with the same organization moving up the hierarchical ladder while gaining expertise in their profession.

A disadvantage of this structure is its rigidity in dealing with complex tasks where different functions (or disciplines) must collaborate and the difficulty in introducing change. The flow of information between (different functions') organizational units may be difficult, causing difficulty in integration. Furthermore, customers frequently have to interact with several functions—they do not have a single point of contact.

1.4. The Project Structure

The project structure is designed to handle one-time, unique, and nonrecurrent endeavors. It is based on a task force assembled for a limited time to achieve a predefined goal. The members of the project team may come from different organizational units and have different educations and backgrounds. They have a common goal—the project success; and a common leader—the project manager. Organizations dealing with projects may adopt a flexible structure in which only a core group has a permanent structure while most of the organization is assigned to project groups.

An advantage of the project structure is its flexibility; the project team can be assembled exactly according to the task at hand. Another advantage is the creation of a single point of contact for the customer—the project manager has complete responsibility for the project and for customer satisfaction. Teamwork and coordination between people coming from different disciplines is easier to achieve when they belong to the same project, share a common goal, and have the same project manager.

The disadvantages of the project structure are related to its temporary nature—resources are not pooled and thus efficiency and effectiveness are hard to achieve. The limited life of the project's organizational structure creates anxiety and uncertainty about the future role of the team members, mainly at the final stages of the project, and information between project teams is not flowing easily.

A major problem in the project structure is division of labor. Unlike the functional organization, in which division of labor is natural because it is based on the specialization of each function, in a project there is no natural division of labor. It is important to allocate the work among the project participants in a very precise way so that the schedule and budget constraints will not be violated and resources will be efficiently and effectively utilized but not overloaded. Most importantly, it should be possible to integrate the parts of the work performed by different individuals and organizations participating in the project and to produce the deliverables required by the customers.

In the functional organizations where division of labor is based on specialization, each function performs the same set of tasks repeatedly. Due to repetition, learning is built into the process. In a project, division of labor can take many different forms and has to be designed carefully because a project is a one-time effort and improvement by repartition is not built into it.

The work breakdown structure is the tool used to divide the project work content among individuals and organizations so that efficiency and effectiveness will be achieved while ensuring the integration of work efforts to produce the project-required deliverables.

1.5. The Matrix Structure

Organizations involved in ongoing operations and multiple projects simultaneously develop hybrid structures that mix the functional organizational structure with the project structure. Although a large variety of such structures exist, most of these structures are based on a permanent functional skeleton and temporary project structures. Each project has a project manager (or coordinator) that serves as a point of contact for the customers and is responsible for the project success. A team that (typically) combines some members who are employed full time by the project and other members that belong to a functional unit and employed part time on one or more projects is assigned to the projects.

While the tasks assigned to each functional unit are repetitive and can be learned by repetition, the work content of each project must be defined and properly allocated to individuals and organizations participating in the project. The work breakdown structure (WBS) is the tool commonly used to ensure proper division of labor and integration of the project deliverables.

2. HIERARCHIES IN THE PROJECT ENVIRONMENT: THE NEED FOR A WORK BREAKDOWN STRUCTURE

2.1. Introduction

As discussed in Section 1, the natural division of labor that exists in a functional organization is missing in projects. It is important to divide the total scope of the project (all the work that has to be done in the project) among the individuals and organizations that participate in it in a proper way, a way that ensures that all the work that has to be done in the project (the project scope) is allocated to participants in the project while no other work (i.e., work that is not in the project scope) is being done. A framework composed of two hierarchical structures known as the work breakdown structure (WBS) and the organizational breakdown structure (OBS) is used for dividing the project scope amongst the participating individuals and organizations in an efficient and effective way, as discussed next.

2.2. The Scope

In a project context the term scope refers to:

- The product or service scope, defined as the features and functions to be included in the product of service
- The project scope, defined as the work that must be done in order to deliver a product or service with the specified features and functions

The project total scope is the sum of products and services it should provide. The work required to complete this total scope is defined in a document known as the statement of work, or scope of work (SOW). All the work that is required to complete the project should be listed in the SOW along with explanations detailing why the work is needed and how it relates to the total project effort.

An example of a table of contents of a SOW document is given in the Appendix. This example may be too detailed for some (small) projects, while for other (large) projects it may not cover all the necessary details. In any case, a clearly written SOW establishes the foundation for division of labor and integration.

2.3. Implementing Division of Labor in Projects

The SOW is translated into a hierarchical structure called the work breakdown structure (WBS). There are many definitions of a WBS:

1. The Project Management Institute (PMI) defines the WBS as follows: “A deliverable–oriented grouping of project elements which organizes and defines the total scope of the project. Each descending level represents an increasingly detailed definition of a project component. Project components may be products or services” (PMI 1996).
2. MIL-STD-881A defines WBS as “a product-oriented family tree composed of hardware, services and data which result from project engineering efforts during the development and production of a defense material item, and, which completely defines the project, program. A WBS displays and defines the product(s) to be developed or produced and relates the elements of work to be accomplished to each other and to the end product” (U.S. Department of Defense 1975).

Whatever definition is used, the WBS is a hierarchical structure in which the top level represents the total work content of the project while at the lowest level there are work elements or components. By allocating the lower-level elements to the participating individuals and organization, a clear definition of responsibility is created. The WBS is the tool with which division of labor is defined. It should be comprehensive—that is, cover all the work content of the project and logical—to allow clear allocation of work to the participating individual and organizations as well as integration of the deliverables produced by the participants into the project-required deliverables.

2.4. Coordination and Integration

Division of labor is required whenever the work content of the project exceeds what a single person can complete within the required time frame or when there is no single person who can master all the knowledge and abilities required for the project. However, the following two reasons that promote division of labor may lead to the failure of the project:

1. Coordination of the work performed by different individuals and organizations is required because outputs (deliverables) of some participants provide inputs to the work of other participants in the project. For example, in a construction project, civil engineers and architects produce the design while construction workers perform construction work. However, without the plans and drawings produced by the design team, construction workers cannot do their work.
2. The ability to integrate the deliverables produced by different participants is crucial. Thus, for example, the fact that in a new car development process one team developed an outstanding new body for the car and another team developed a state-of-the-art engine does not guarantee a project's success. Only a successful integration of the engine with the car that results in a vehicle that satisfies all the requirements and specifications of the project constitutes a success. For example, if the car becomes unstable after engine assembly due to a high center of gravity caused by the location of the assembled engine, the fact that the car body is excellent and the engine performs very well does not make the project a success.

In addition to defining the division of labor in the project, the WBS should support integration and coordination. Properly designed WBS is the tool for division of labor, integration, and coordination.

3. THE RELATIONSHIP BETWEEN THE PROJECT ORGANIZATION AND THE WORK BREAKDOWN STRUCTURE

3.1. Introduction

As explained earlier in this chapter, the WBS is designed to support the division of the project scope (work content) amongst the individuals and organizations participating in the project, which is accomplished by combining the WBS with the project organizational breakdown structure (OBS). The combined framework of OBS and WBS allocates each component of the project scope defined at the lowest WBS level to an organizational unit or a specific individual responsible for it in the OBS. The emerging framework of two hierarchical structures integrated at the lowest level provides an important tool for project-planning execution and control.

3.2. Responsibility

To support division of labor, the WBS should integrate with the organizational breakdown structure (OBS) of the project. The OBS is a hierarchical structure that depicts the relationship among the organizations and individuals participating in the project. At the top level of the OBS is the project manager, and at the bottom are individuals responsible for the accomplishment of specific work content in the WBS. These subprojects allocated to responsible managers are known as work packages. The manager of a work package is an expert in its product scope and project scope. Thus, all the project-management processes at the work package level are the responsibility of the work package manager. These include tasks such as scheduling the activities of the work packages, assigning resources, estimating cost, and monitoring and control. The work package tasks related to the product scope are also the responsibility of the work package manager. These tasks are specific to the work package and may include such activities as design, manufacturing, training, testing, and support.

The assignment of responsibility to the work package managers should be based on a clear definition of the work content of the work package, including:

- Deliverables and the delivery time of each
- Required inputs to the work package (data, output from other work packages, etc.)
- Required resources to perform the work package
- Cost of performing the work package
- Tests and design reviews

When a work package is subcontracted, the definition is part of the contract. However, when the work package is performed internally, it is important to define the content of the work package as well as all other points listed above to avoid misunderstanding and a gap in expectations between the performing organization and the project manager. A special tool called the responsibility assignment matrix (RAM) relates the project organization structure to the WBS to help ensure that each element in the project scope of work is assigned to an individual. As an example, consider a project in which six work packages, A, B, C, D, E, and F, are performed by an organization with three departments, I, II, and III. Assuming that in addition to the project manager the three department heads, the CEO, and the controller are involved in the project, an example RAM follows:

Work Package Person	**A**	**B**	**C**	**D**	**E**	**F**
CEO	S					
Controller	R	R	R	R	R	R
Project manager	A	S	S	S	S	S
Head Department I	P		I	P	I	A
Head Department II	I	P	A	P	A	
Head Department III	P	A	I	A	P	I

Legend:

P:	Participant
A:	Accountable
R:	Review required
I:	Input required
S:	Sign-off required

3.3. Authority

Along with the responsibility for on-time completion of the content of the work package and performance according to specifications, the work package managers must have proper authority. Authority may be defined as a legal or a rightful power to command or act. A clear definition of authority is important for both project scope and product scope. For example, the work package managers may have the authority to delay noncritical activities within their slack but have no authority to delay critical activities—this is the authority of the project manager only. In a similar way, the work package manager may have the authority to approve changes that do not affect the product form fit or function, while approval of all other changes is by the project manager. The authority of work package managers may differ according to their seniority in the organization, the size of the work package they are responsible for, geographical location and whether they are from the same organization as the project manager. Clear definition of authority must accompany the allocation of work packages in the WBS to individuals and organizations.

3.4. Accountability

Accountability means assuming liability for something either through a contract or by one's position of responsibility. The project manager is accountable for his own performances as well as the performances of other individuals to whom he delegates responsibility and authority over specific work content—the managers of work packages. The integration of the WBS with the OBS through the responsibility assignment matrix (RAM) is a major tool that supports the mapping of responsibility, authority, and accountability in the project.

To make the WBS an effective tool for project management, it should be properly designed and maintained throughout the project life cycle. The project's work content may be presented by different WBS models, and the decision which one to select is an important factor affecting the probability of project success.

4. THE DESIGN OF A WORK BREAKDOWN STRUCTURE AND TYPES OF WORK BREAKDOWN STRUCTURES

4.1. Introduction

The WBS serves as the taxonomy of the project. It enables all the project stakeholders—customers, suppliers, the project team itself, and others—to communicate effectively throughout the life cycle of the project. For each project, one can design the WBS in several different ways, each emphasizing a particular point of view. However, different WBS patterns call for different organizational structures or management practices during the implementation of the project. Thus, the design of the WBS at the early stage of the project life cycle may have a significant impact on the project success. Often the individuals who prepare the WBS are not aware of the crucial role they play in determining future coordination and understanding among the operational units who eventually execute the work packages. A mismatch among the WBS, the OBS, and the management style of the project manager may lead to a poor project-completion record. Such difficulties are compounded if different parties that are involved in the project have produced different WBSs. In this section, we present alternative WBS patterns and explain their possible impact on OBS and management practices. We use an example project to illustrate different patterns and indicate their strengths and weaknesses. The example project assumes that a large multinational corporation operating in the semiconductor business has just finished evaluating existing and future markets and obtained forecasts on the demand for its products in the next five years. Based on these forecasts, the firm has decided it will need five new plants (also known as FABs) in addition to the nearly dozen it currently operates. Labor availability, wage levels, and tax regulations were chief considerations affecting the decision to construct the plants in three countries.

The various WBS formats shown below can all be useful in describing the expansion project. We denote them as WBS based on technology, life cycle, geography, and so on, according to the focus of the second level in the WBS hierarchy. By choosing the focus of that crucial level, the WBS designer determines the fundamental structure of the project. Still, the designer has to make similar decisions at the third level, fourth level, and so on, but these are secondary choices compared with the second level.

4.2. A WBS Based on Technology

Projects that are characterized by a relatively high degree of specialization, especially those associated with the high-tech sector of the economy, typically require the assignment of a leading professional to lead all the project activities that are related to a particular technology. This professional is expected to maintain the same standards of quality and performance among the different facilities. Thus, this WBS format would fit especially well organizations that are structured in a functional hierarchy (see Section 1.2). This type of WBS will be a favorite for managers preferring strong central control of the project because every activity in the different locations is reported to the headquarters (where the professionals heading the various technologies are based). Figure 1 illustrates a WBS by technology in our case.

4.3. A WBS Based on Project Life Cycle

Organizing the WBS by the various stages of the project life cycle (or, more generally, by time) is not a particularly common practice. Still, it may fit certain organizations that elect to orchestrate their activities by timing. For example, the FABs construction project may be outsourced to a number of subcontractors, starting with a subcontractor in charge of preparing detailed floor plans and construction programs, followed by another contractor charged with all the infrastructure activities, and so on. This will lead to the WBS presented in Figure 2. The work content is first broken by the major stages of the project (from design to delivery). Then each stage is further broken down to its relevant categories. This process is repeated, sometimes to 7–10 levels or even more, until we reach the final level of the work packages.

4.4. A WBS Based on Geography

Breaking the work by geography lends itself quite easily to the assignment of five plant managers, each responsible for the entire work required for establishing his plant. In a way, this amounts to breaking the project into five identical subprojects, each duplicating the activities undertaken by the others. Obviously, this will be the preferred mode when the circumstances (culture, language, type of government, law system, etc.) are dramatically different in the different countries. This type of WBS will fit decentralized management practices in which local plant managers are empowered with full authority (and responsibility) for the activities relating to their respective plants.

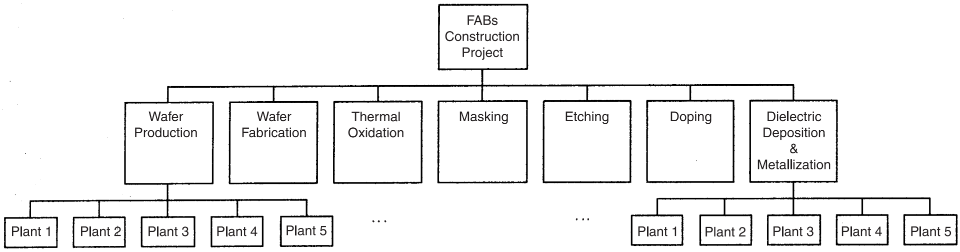


Figure 1 WBS by Technology.

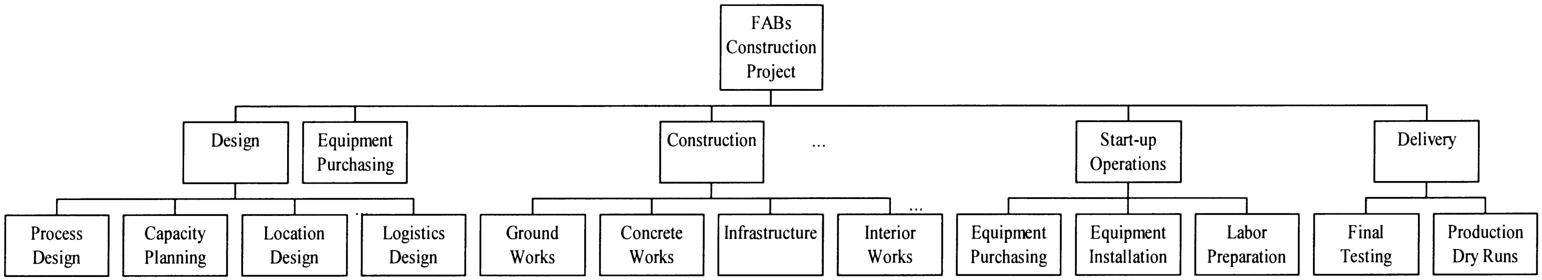

FABs Construction Project
Design
Equipment Purchasing
Construction
...
Start-up Operations
Delivery
Process Design
Capacity Planning
Location Design
Logistics Design
Ground Works
Concrete Works
Infrastructure
Interior Works
...
Equipment Purchasing
Equipment Installation
Labor Preparation
Final Testing
Production Dry Runs

Figure 2 WBS by Project Life Cycle.

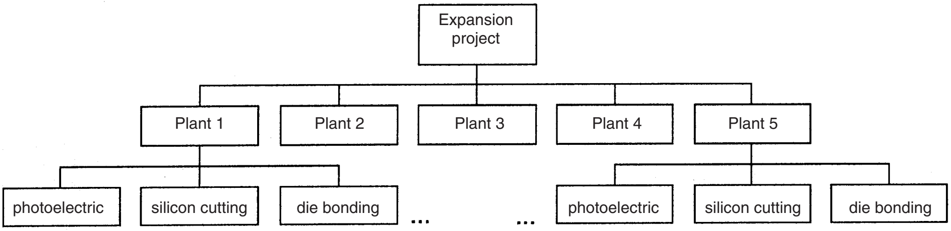

Expansion project
Plant 1
Plant 2
Plant 3
Plant 4
Plant 5
photoelectric
silicon cutting
die bonding
...
...
photoelectric
silicon cutting
die bonding

Figure 3 WBS by Geography.

4.5. Other WBS Designs

There are many other possible orientations in which a WBS can be designed. The choice among them depends on the organization charged with the project execution. For example, the growing recognition of the importance of supply chain management has caused some organizations to adopt structures that are logistics oriented. In such cases, we may find at the second level of the WBS a breakdown by logistics functions as illustrated in Figure 4. Other organizations favor structures oriented towards subsystems. That is, the entire system is divided into its major subsystems. In our case, a FAB can be divided into the warehouse subsystem (receiving and checking raw materials, packing and shipping finished goods), shop-floor subsystem (scheduling and dispatching jobs), quality control subsystem (testing components and finished units), and so on. These subsystems serve as the entities in the second level of the WBS.

4.6. Discussion

We conclude this section with a summary of the pros and cons in using a WBS to plan a project.

- *Advantages:*
  - The WBS reflects the project objectives. By listing all the activities required to accomplish these objectives, it prevents confusion and doubts as to the aim of the project.
  - The WBS creates a common database and a dictionary of common notation that serves as a reference point for all involved parties.
  - The WBS, in conjunction with the OBS, defines the way the project is to be managed. It relates each work activity to the corresponding organizational unit that is responsible for delivering the work.
  - The WBS enables smooth communications among the project team members and between them and customers, suppliers, regulators, etc.
  - The WBS serves as an archive that can later facilitate knowledge transfer to other projects or learning by new members of the workforce.
  - The WBS is an effective tool for resource management.
- *Disadvantages:*
  - The WBS requires a significant amount of effort to build and maintain.
  - The WBS encourages rigid structure for the project. Thus, it reduces managerial flexibility to initiate and lead changes during the project life cycle.

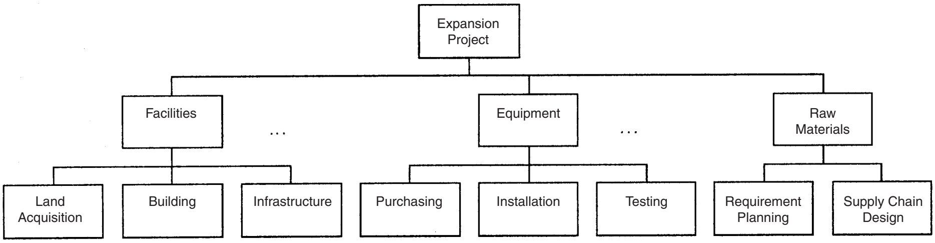

Expansion Project
Facilities
...
Equipment
...
Raw Materials
Land Acquisition
Building
Infrastructure
Purchasing
Installation
Testing
Requirement Planning
Supply Chain Design

Figure 4 WBS by Logistics.

- There are many legitimate ways to view a project, and occasionally, depending on circumstances, one approach may be preferred to others. Yet the WBS forces the project designer to choose one approach and remain with it throughout the project life cycle.

5. WORK PACKAGES: DEFINITION AND USE

5.1. Introduction

At the lowest levels of the WBS and OBS, integration between the two hierarchical structures takes place. The assignment of specific work content (project scope) to a specific individual or organization creates the building bocks of the project-management framework: the work packages. In the following section, a detailed discussion of the definition and meaning of work packages is presented along with a discussion of the cost accounts that accompany each work package, translating its work content into monetary values for the purpose of budgeting and cost control.

5.2. Definition of Work Packages

The PERT Coordinating Group (1962) defined a work package (WP) as "The work required to complete a specific job or process, such as a report, a design, a documentation requirement or portion thereof, a piece of hardware, or a service." PMI (1996) states: "[A] work package is a deliverable at the lowest level of the WBS." Unfortunately, there is no accepted definition of the WPs nor accepted approach to link them with other related structures (foremost among them the OBS). In most cases, the WPs are defined in an informal, intuitive manner and without proper feedback loops to verify their definition.

One of the difficulties in defining the WPs is the trade-off between the level of detail used to describe the WPs and the managerial workload that is involved. On the one hand, one wishes to provide the project team with an unambiguous description of each work element to avoid unnecessary confusion, overlap, and so on. On the other hand, each WP requires a certain amount of planning, reporting, and control. Hence, as we increase the level of detail, we also increase the overhead in managing the project. To overcome this problem, some organizations set guidelines in terms of person-hours, dollar-value, or elapsed time to assist WBS designers in sizing the WPs. These guidelines are typically set to cover a broad range of activities, and therefore they ignore the specific content of each WP. Hence, they should be applied with care and with appropriate adjustments in places where the work content requires them.

Planning the work by the WPs and structuring it through the WBS is closely related to another important planning activity—costing the project. By dividing the project into small, clearly defined activities—the WPs—we provide a better information basis to estimate the costs involved. For example, consider the activity of design of the FAB processes. It is much easier to estimate its components when they are considered separately (designing the silicon melting and cooling process, silicon cutting, grounding and smoothing, etc.). Furthermore, the separate components may require different cost-estimation procedures or expertise.

Another consideration is related to the statistical nature of the cost-estimation errors. The estimation of the cost for each WP involves a random error that, assuming no particular bias, can be either positive or negative. As the cost estimates are aggregated up the WBS hierarchy, some of these errors cancel each other and the relative size of the aggregated error decreases. This observation holds as long as there is no systematic bias in the estimation procedure. If such a bias exists (e.g., if all the time and cost estimates were inflated to protect against uncertainties), then further decomposition of the WPs may eventually have a negative effect on the overall cost estimate.

In practice, in many scenarios there are limits to the precision that can be achieved in time and cost estimations. Beyond these limits, the errors remain constant (or may even grow). Thus, from the precision perspective, division into smaller WPs should be carried out as long as it improves the estimation accuracy, and not beyond that point.

5.3. Definition of Cost Accounts

Cost accounts are a core instrument used in planning and managing the financial aspects of a project. Three fundamental processes depend on the cost accounts: costing individual activities and aggregating them to the project level for the purpose of preparing project cost estimates; budgeting the project; and controlling the expenses during the project execution.

The first issue, costing the project and its activities, requires the project planner to choose certain costing procedures as well as cost classification techniques. Costing procedures range from the traditional methods to state-of-the-art techniques. For example, absorption cost accounting, a traditional method that is still quite popular, relates all costs to a specific measure (e.g., absorbing all material, equipment, energy, and management cost into the cost of a person-hour) and cost new products or services by that measure. An example of a more advanced cost accounting technique is activity-based costing (ABC), which separately analyzes each activity and measures its contribution to particular products or services.

Cost classification can be done in many ways. Each organization builds its own hierarchy of cost accounts, which is also known as the cost breakdown structure (CBS). In many cases, the CBS is closely linked to the OBS. This means that each organizational unit at the bottom level of the OBS is associated with a cost account. All the expenses planned for the various activities are accounted for through these accounts. Often we find that these cost accounts are further broken down along general accounting principles (e.g., variable vs. fixed costs or manpower, material, equipment, and subcontracting costs). Some organizations prefer to construct the CBS according to the WBS. That is, each WP is associated with a unique cost account. The latter method enables easier control over the individual activities, therefore lending itself more easily to project structure. The former approach might fit better functional structures because it is geared to maintain control over functions rather than activities. It is possible to combine these two approaches by defining the cost accounts at the lowest level of the OBS–WBS level. Then one can aggregate these accounts either by the OBS or by the WBS structures and still obtain consistent estimates at the project-wide level.

Other organizations create the CBS according to the project life cycle. Each of the major life-cycle stages (conceptual design, detailed design, production, operation, divestment) is a major cost account that is further broken down into finer accounts according to secondary criteria (e.g., detailed schedules, functional association). This form of CBS allows the most straightforward control of cost accumulation over time.

The second process, budgeting, uses the cost accounts as a vehicle to generate the project budget. A popular way to generate a budget is through a bottom-up aggregation. The cost accounts associated with individual WPs are aggregated towards a complete budget. Along the way, management may intervene in many ways that may alter the original cost estimates. For example, a ‘‘crashing’’ policy may be adopted in order to expedite certain activities as a result of exogenous considerations. This will make the respective budget line larger than the original estimate stated in the cost account. Similarly, a decision to hold certain amounts as ‘‘management reserve’’ (a common practice) will also inflate the budget above the original cost accounts. Thus, gaps may exist between the budget and the cost estimate of WPs and the WBS as a whole. However, even with these gaps, the cost accounts are the basis for building and maintaining the budget for every project.

Based on cost estimates, allocated budget, and other considerations (primarily competitive pressure), the pricing of the project is established. The project price may be above or below its cost or its budget, depending on management policies and extraneous constraints.

The third process, financial control of the project, is again based on the cost accounts. The basic control method is an ongoing comparison between actual and planned cost accumulation. Methods such as the earned value technique develop ratio measures that help the controller to analyze the schedule and cost deviations over time and employ control limits as triggers for corrective action. The control is usually performed at the WP cost account level.

6. USING THE WORK BREAKDOWN STRUCTURE: EXAMPLES

6.1. R&D Projects

Managing R&D projects is among the toughest areas in project management. These projects are characterized by a high degree of uncertainty, and consequently a large proportion of them is never completed. The importance of careful planning in this environment cannot be overstated.

The diagram in Figure 5 illustrates a WBS planned for an R&D project aimed at developing a new product. The second level of this WBS is organized by the project life cycle, and the third level corresponds to functional departments that are involved in the project.

6.2. Equipment-Replacement Projects

Every technology-intensive firm is challenged from time to time with equipment-replacement projects. This type of project is especially common in the high-tech sector, where the frequency of such projects is now measured in months rather than years. The WBS presented in Figure 6 focuses at its second level on the division among activities related to the facility and its infrastructure (requiring civil engineering expertise), activities related to the equipment itself (requiring mechanical engineering expertise), and activities related to manpower (requiring human resource expertise).

Unlike the previous example, the third level is not identical across the three entities of the second level. A greater level of detail is needed to describe the equipment-related activities, and so the corresponding WBS branch is more developed.

6.3. Military Projects

To demonstrate the wide-range applicability of the WBS concept, we close this section with an example of a military operation. An army unit (say, a brigade) is faced with a mission to capture a riverbank, construct a bridge, and secure an area (bridgehead) across the river, thus enabling the movement of a larger force in that direction. Figure 7 illustrates how this mission can be planned through WBS principles. The second level of the WBS is arranged by the major military functions

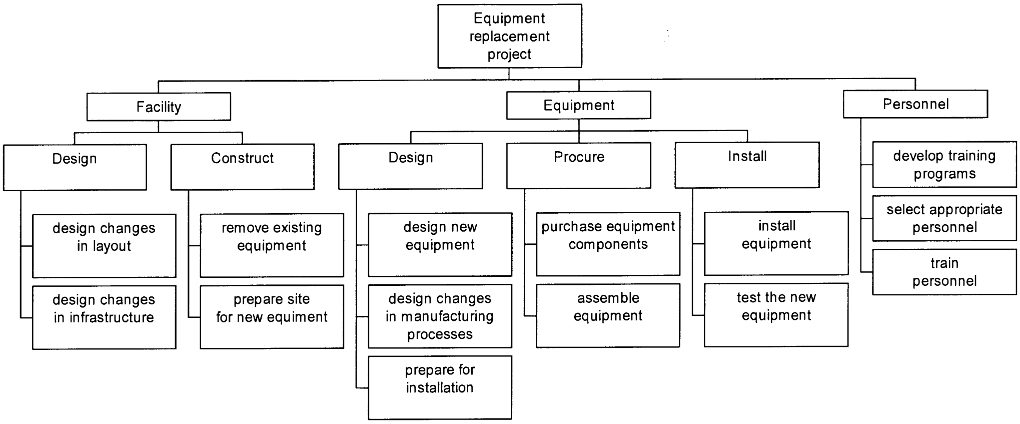

Equipment replacement project
Facility
Equipment
Personnel
Design
Construct
Design
Procure
Install
design changes in layout
design changes in infrastructure
remove existing equipment
prepare site for new equiment
design new equipment
design changes in manufacturing processes
prepare for installation
purchase equipment components
assemble equipment
install equipment
test the new equipment
develop training programs
select appropriate personnel
train personnel

Figure 5 WBS for an R&D Project.

that are involved, and the third level is arranged by the life-cycle (illustrated here with a basic distinction between all precombat activities and during-combat activities).

7. CHANGE CONTROL OF THE WORK BREAKDOWN STRUCTURE

7.1. Introduction

Projects are often done in a dynamic environment in which technology is constantly updated and advanced. In particular, projects in high-tech organizations go through several major changes and many minor changes during their life cycle. For example, the development of a new fighter plane may take over a decade. During this time, the aircraft goes through many changes as the technology that supports it changes rapidly. It is quite common to see tens of thousands of change proposals submitted during such projects with thousands of them being implemented. Without effective control over this process, all such projects are doomed to chaos. Changing elements in the WBS (deleting or adding work packages or changing the contents of work packages) belong to an area known as configuration management (CM). CM defines a set of procedures that help organizations in maintaining information on the functional and physical design characteristics of a system or project and support the control of its related activities. CM procedures are designed to enable keeping track of what has been done in the project until a certain time, what is being done at that time, and what is planned for the future. CM is designed to support management in evaluating proposed technological

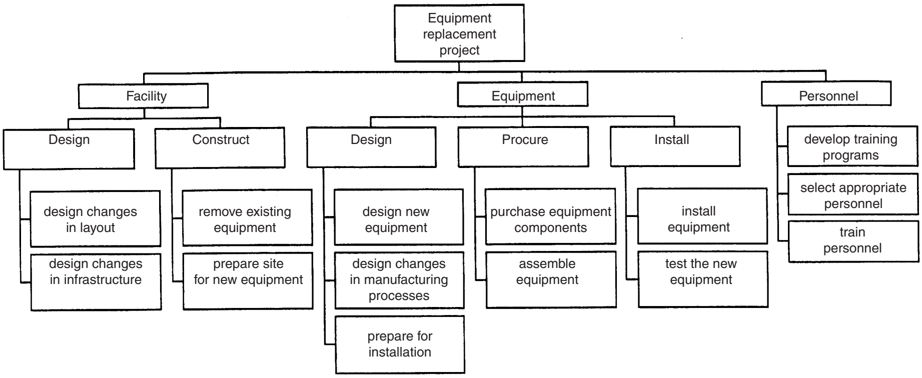

Equipment replacement project
Facility
Equipment
Personnel
Design
Construct
Design
Procure
Install
design changes in layout
design changes in infrastructure
remove existing equipment
prepare site for new equipment
design new equipment
design changes in manufacturing processes
prepare for installation
purchase equipment components
assemble equipment
install equipment
test the new equipment
develop training programs
select appropriate personnel
train personnel

Figure 6 WBS for an Equipment-Replacement Project.

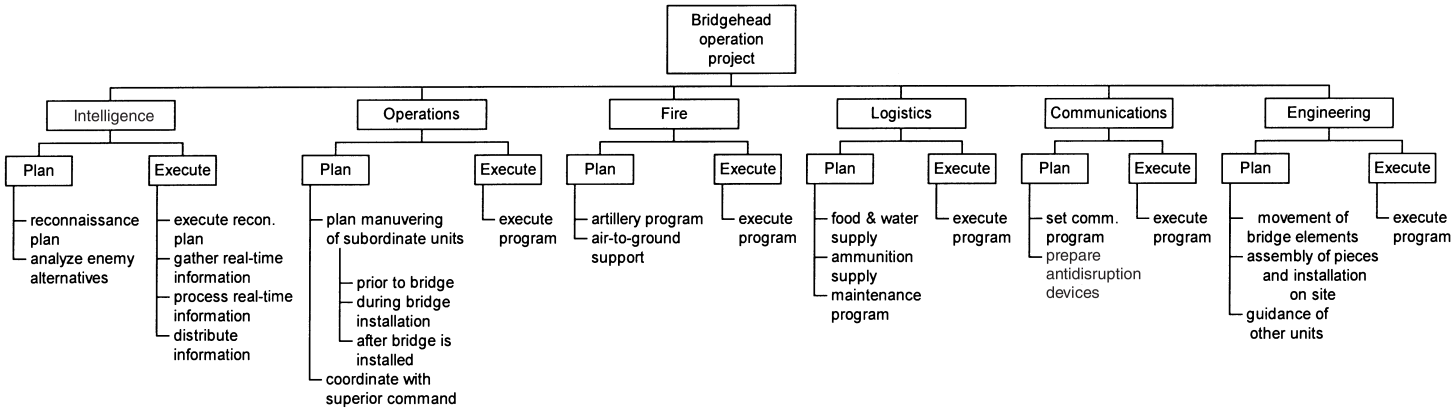

Bridgehead operation project
Intelligence
Plan
reconnaissance plan
analyze enemy alternatives
Execute
execute recon. plan
gather real-time information
process real-time information
distribute information
Operations
Plan
plan manuvering of subordinate units
prior to bridge
during bridge installation
after bridge is installed
coordinate with superior command
Execute
execute program
Fire
Plan
artillery program
air-to-ground support
Execute
execute program
Logistics
Plan
food & water supply
ammunition supply
maintenance program
Execute
execute program
Communications
Plan
set comm. program
prepare antidisruption devices
Execute
execute program
Engineering
Plan
movement of bridge elements
assembly of pieces and installation on site
guidance of other units
Execute
execute program

Figure 7 WBS of Military Operation.

changes. It relies on quality assurance techniques to ensure the integrity of the project (or product) and lends itself easily to concurrent engineering, where certain activities are done in parallel to shorten time-to-market and secure smooth transitions among the design, production, and distribution stages in the project life cycle. Change control involves three major procedures, which are outlined below.

7.2. Change Initiation

Changes can be initiated either within the project team or from outside sources. Either way, to keep track of such initiatives and enable organizational learning, a formal procedure of preparing and submitting a change request is necessary. A complete change request should include the following information:

- *Pointers and identifiers:* These will enable the information system to store and retrieve the data in the request. Typical data are request i.d.; date, name or originator; project i.d.; configuration item affected (e.g., work packages i.d., part number i.d.).
- *Description of change:* A technical description (textual, diagrammatic, etc.) that provides full explanation on the content of the proposed change, the motivation for the proposal, the type of change (temporary or permanent), and suggested priority.
- *Effects:* A detailed list of all possible areas that might be affected (cost, schedule, quality, etc.) along with the estimated extent of the effect.

7.3. Change Approval

Change requests are forwarded to a team of experts who are capable of evaluating and prioritizing them. This team, often known as the change control board (CCB), evaluates each proposed change, taking into account its estimated effects on the various dimensions that are involved. Foremost among these criteria are the cost, schedule, and quality (or performance) issues. However, other criteria, such as contractual agreements and environmental (or collateral) effects, are also considered. The review is done both in absolute and relative terms. Thus, a proposed change may be rejected even if it leads to overall improvement in all relevant areas if there are other changes that promise even better effects. If the board approves a change, it needs to specify whether the change is temporary or permanent. Example of temporary changes are construction of a partial pilot product for the purpose of running some tests that were not planned when the project was launched, releasing an early version of a software to a "beta" site to gain more insights on its performance, and so on. It can be expected that approval of temporary changes will be obtained more easily and in shorter time spans than the approval of permanent changes.

The CCB is responsible for accumulating and storing all the information on the change requests and the outcomes of the approval process. Maintaining a database that contains all the relevant information on the changes usually facilitates this process. The database should enable easy access to future queries, thus facilitating continuous organizational learning.

7.4. Change Implementation

Changes that were approved, on either a temporary or a permanent basis, are to be integrated into the project. The information on approved changes is usually disseminated to all involved parties through an engineering change order. This form contains all the information that might be required by the various functions (engineering, manufacturing, quality assurance, logistics). Proper change implementation requires the creation of feedback loops that will provide information on the impact of the implemented change. There is a need to verify that this impact is consistent with the estimated effects that were analyzed during the approval stage. These feedback mechanisms alert the system to any departure from the planned effects and help management to identify potential troubles before they actually occur. As before, the information that flows in these loops is recorded in the CM database to support further learning and improvement.

8. THE WORK BREAKDOWN STRUCTURE AND THE LEARNING ORGANIZATION

8.1. Introduction

In addition to supporting division of labor and integration, the WBS–OBS framework is an effective tool for the accumulation, storage, and retrieval of information at the individual and organizational levels. By using templates of work breakdown structures as the project dictionary, it is possible to accumulate information about the actual cost duration and risks of project activities and work packages. This information is the basis for a continuous learning process by which, from one project to the next, a database is developed to support better project planning and management as well as the training of individuals and groups. The next section discusses individual and organizational learning in the project environment.

8.2. WBS as a Dictionary for the Project

The division of labor among the parties participating in a project supports specialization. It is also an answer to the need to finish the project work content within a predetermined schedule, which is not determined by the amount of work to be performed. Due to the division of labor, it is possible to perform each part of the work content of the project by the best experts within the required time frame.

These benefits of the division of labor do not come for free—they carry the risks associated with integration. Integration of information, knowledge, and deliverables produced by the different work packages must be based on a common language to ensure a smooth and fault-free process. This common language is based on the WBS.

A well-planned WBS serves as a dictionary of a project. Because each work package is defined in terms of its work content, its deliverables, its required inputs (data, information, resources, etc.), and its relationship to other work packages within the WBS, all the stakeholders have a common reference or a common baseline. Furthermore, careful management of changes to the WBS throughout the project life cycle provides a continuous update to the project dictionary.

Learning a common language is easier if the same language is used over a long period of time and becomes a standard. Thus, organizations involved in similar projects should strive to develop a WBS template that can serve most of the projects with minor modifications. This is easier if the projects are repetitive and similar to each other. However, if there are major differences among projects, the WBS representing the project scope (as opposed to the product scope) can be standardized if the processes used for project management are standardized in the organization. Thus, by developing a standard set of project-management processes and supporting these standards by appropriate information technology tools, it is possible to standardize the project part of the WBS and make it easier to learn. A standard WBS ensures that project-management knowledge and practices are transferred between projects and become common knowledge in the organization.

8.3. Information Storage and Retrieval

The flow of information within and between projects is a key to the organizational learning process. Information generated in one project can serve other projects either by transferring people between the projects, assuming that these people carry information with them, or by a carefully planned method of information collection, storage, and retrieval. A library-like system is required to support the transfer of information, which is not based on human memory. A coding system that supports an efficient search and retrieval of information or data for the estimation of cost, duration, risks, and so on. is required. In the extreme, such a system can help the planner of a new project to identify parts of historical projects similar to the new project he or she is involved with. Such subprojects that were performed in past projects can serve as building blocks for a new project. A carefully planned WBS is a natural coding system for information collection, storage, and retrieval. Work packages performed on past projects can serve as templates or models for work packages in new projects if the same WBS is used.

Developing WBS templates for the types of projects performed by the organization enables a simple yet effective information storage and retrieval system to be developed. Even if some of the projects are unique, a good coding system based on WBS templates can help in identifying similar parts in projects, such as parts related to the project scope. The ability to retrieve complete work packages and use them as building blocks or parts of work packages and as input data for estimation models enhances the ability of organizations to compete in cost, time, and performance.

8.4. The Learning Organization

The transfer of information within or between projects or the retrieval of information from past projects provides an environment that supports the learning organization. However, in addition to these mechanisms, a system that seeks continuous improvement from project to project is required. This can be done if the life cycle of each project is examined at its final stage and conclusions are drawn regarding the pros and cons of the management tools and practices used. Based on a thorough investigation of each project at its completion, current practices can be modified and improved, new practices can be added, and, most importantly, a new body of knowledge can be created. This body of knowledge can, in turn, serve as a basis for teaching and training new project managers in how to manage a project right.

REFERENCES

PERT Coordination Group (1962), *DoD and NASA Guide: PERT Cost Systems Design.*

Project Management Institute (PMI) (1996), *A Guide to the Project Management Body of Knowledge,* PMI, Upper Darby, PA.

U.S. Department of Defense (1975), "A Work Breakdown Structure for Defense Military Items," MIL-STD 881, U.S. Department of Defense, Washington, DC.

ADDITIONAL READING

Globerson, S., "Impact of Various Work Breakdown Structures on Project Conceptualization," *International Journal of Project Management,* Vol. 12, No. 3, 1994, pp. 165–179.

Raz, T., and Globerson, S., "Effective Sizing and Content Definition of Work Packages," *Project Management Journal,* Vol. 29, No. 4, 1998, pp. 17–23.

Shtub, A., Bard, J., and Globerson, S., *Project Management Engineering, Technology and Implementation,* Prentice Hall, Englewood Cliffs, NJ, 1994.

APPENDIX
Table of Contents for a SOW document

- Introduction: project scope and general description
- Type of project (development, production, construction, maintenance, etc.)
- A description of final deliverables, intermediate deliverables and delivery schedule
- Main phases in project life cycle: a short description of each phase, its deliverables and its work content
- Applicable documents
- Operational requirements
- Technical specifications
- Applicable standards
  - Applicable laws and regulations
  - Applicable procedures
  - Other applicable documents
- Development
  - Conceptual design
  - Functional design
  - Detailed design
  - Prototypes required
  - Deliverables of development phase
- Production and construction
  - Quantities to be delivered and delivery schedule
  - Accompanying documents (production documents, product documents)
- Markings and signs
  - Mechanical markings and signs
  - Electrical markings and signs
  - Other markings and signs
- Purchasing
  - Purchasing in state
  - Purchasing abroad
  - Subcontractors management
  - Suppliers management
  - Purchasing of critical and long-lead items
- Testing
  - Master plan for testing
  - Detailed plans for each test
  - Performance approval testing
  - Functional tests
  - Environmental conditions testing
  - Commercial production testing
  - Acceptance testing at the supplier's site
  - Acceptance testing at the customer's site

- Prototypes
- Reliability and maintainability
  - Anticipated reliability (calculations of MTBF)
  - Fault tree analysis (FTA)
  - Anticipated maintainability (calculations of MTTR)
  - Maintainability analysis and level of repair
- Adaptability
  - Electromagnetic (RFI/EMI) requirements and tests
  - Adaptability requirements to existing hardware and software systems.
- Integrated logistics support (ILS)
  - Engineering support (maintenance plans, maintenance during warranty, engineering support after warranty period)
  - Spare parts (during warranty period, recommended spare parts list, spare parts supply after warranty period)
  - Training (initial training in maintenance and operation, training literature and other aids, training during system life cycle)
  - Documentation: description of documentation to be supplied for training, maintenance, operation, and logistic support.
  - Project documentation: design reviews test documents, etc.
  - Data and its storage: means for data storage and retrieval during and after the project
- System's acceptance
  - Acceptance tests
  - Milestones for acceptance and payments
  - Support during acceptance and commissioning
  - Spare parts supply during acceptance
  - Support equipment for acceptance and testing
  - Packaging and shipment
- Installation at customer's site
  - Packaging and shipping requirements
  - Permissions, export and import licensing
  - Constraints in the installation site
  - Master plan for installation
  - Logistical, technical, and operational responsibility for installation
  - Acceptance testing after installation
  - Training at customer's site
  - Maintenance during installation and commissioning
- Project management: planning and control
  - Project phases
  - Deliverables of each phase
  - Work breakdown structure
  - Organizational breakdown structure
  - Work packages
  - Schedule and milestones
  - Progress reports and updates
  - Human resources, equipment, and materials
  - Data required and data supplied
  - Project budget and budget control
  - Risk identification, quantification, and management
  - Configuration management and change control
  - Milestones and related payments
  - Project monitoring system, regular reports, exception reports, meetings, and design reviews
  - Approval of subcontractors and suppliers
  - Software for project management

- Quality assurance and reliability
  - Quality assurance plan
  - Quality procedures
  - Quality and reliability analysis
  - Quality reports
  - Quality control and reviews
- Documentation
- Operational requirements
  - Technical specifications
  - Engineering reports
  - Testing procedures
  - Test results
  - Product documentation
  - Production documentation
  - Drawings
  - Definition of interfaces
  - Operation, maintenance, and installation instructions
  - Software documentation (in the software itself and in its accompanying documents)
- Warranty
- Customer-furnished equipment

IV.B

Product Planning

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 48

Planning and Integration of Product Development

HANS-JÖRG BULLINGER, JOACHIM WARSCHAT, JENS LEYH, AND THOMAS CEBULLA
Fraunhofer Institute of Industrial Engineering

1. INTRODUCTION 1283
  1.1. Overview 1283
  1.2. New RP Technologies 1283
  1.3. Communication Technologies 1284
2. CHARACTERISTICS OF RAPID PRODUCT DEVELOPMENT 1284
  2.1. The Life Cycle 1284
  2.2. The Organization 1285
  2.3. The Process 1286
  2.4. The Human and Technical Resources 1286
  2.5. The Product 1286
3. ELEMENTS OF RAPID PRODUCT DEVELOPMENT 1287
  3.1. Process Planning 1287
  3.2. Physical Prototyping 1288
  3.3. Digital Prototyping 1288
  3.4. The Engineering Solution Center 1290
4. KNOWLEDGE ENGINEERING 1291
  4.1. Communication and Cooperation 1291
  4.2. Knowledge Integration 1293
5. SUMMARY AND PERSPECTIVE 1293
REFERENCES 1294

1. INTRODUCTION

1.1. Overview

Today's market is characterized by keen international competition, increasingly complex products, and an extremely high innovation dynamic. Parallel to the shortening of innovation cycles, the life cycles of products and the time until investments pay off are decreasing.

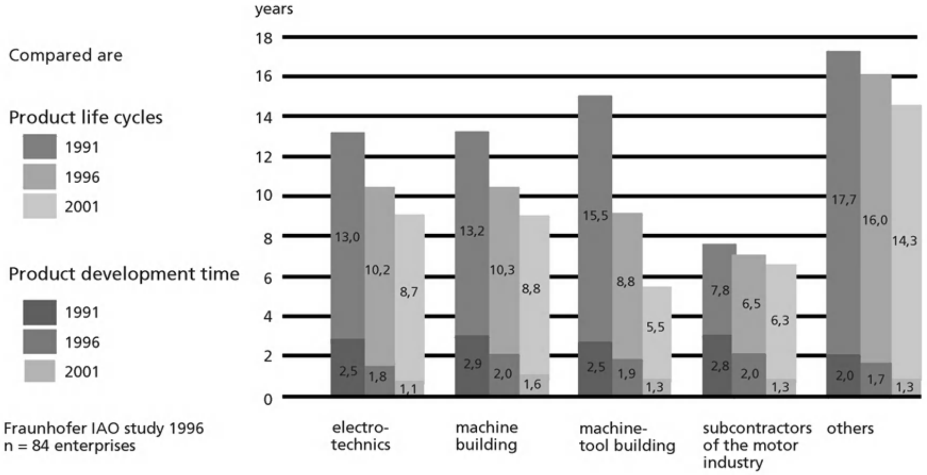
Thus, time is presently the most challenging parameter. Fast, successful positioning of new products on the market has become vital for a company, and the development of innovative products needs to be accelerated. The production of prototypes is significant for a rapid product development (RPD) process.

One feature of this process is the coordination of work tasks within the distributed teams. It will become increasingly important to bring together the different experts. Effective and efficient project management is the basis for the way a team functions. The early integration of different experts serves as a means to develop innovative products. This is also an important factor concerning costs because the main part of the end costs are determined in the early phases of product development. To facilitate the integration of different experts and enhance the efficiency of the iterative phases, prototypes are used as cost-efficient visual models.

1.2. New RP Technologies

Generative prototyping technologies, such as stereolithography (STL), reduce prototyping lead times from a few hours to up to three months, depending on the quality required. These prototypes can serve as visual models or as models for follow-up technologies such as casting.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.



Fraunhofer IAO study 1996
n = 84 enterprises

Figure 1 Product life cycles and Development Times.

New, powerful CAD technologies make it possible to check design varieties in real time employing virtual reality tools. The use of virtual prototypes, especially in the early phases of product development, enables time- and cost-efficient decision making.

1.3. Communication Technologies

ATM (asynchronous transfer mode) networks (Ginsburg 1996) and gigabit ethernet networking (Quinn and Russell 1999) enable a quick and safe exchange of relevant data and thus support the development process tremendously. The Internet provides access to relevant information from all over the world in no time, such as via the World Wide Web or e-mail messages. Communication and cooperation are further supported by CSCW tools (Bullinger et al. 1996) like videoconferencing and e-mail. The distributed teams need technical support for the development of a product to enable synchronous and asynchronous interactions. Furthermore, the Internet provides a platform for the interaction of distributed experts within the framework of virtual enterprises. The technologies used support the consciousness of working temporarily in a network, which includes, for example, the possibility of accessing the same files. All these new technologies have been the focus of scientific and industrial interest for quite a while now. However, the understanding how these new technologies can be integrated into one continuous process chain has been neglected. Combining these technologies effectively enables the product-development process to be reduced decisively. Rapid product development is a holistic concept that describes a rapid development process achieved mainly by combining and integrating innovative prototyping technologies as well as modern CSCW (computer-supported cooperative work) tools.

Objectives of the new concept of RPD are to:

- Shorten the time-to-market (from the first sketch to market launch)
- Develop innovative products by optimizing the factors of time, cost, and quality
- Increase quality from the point of view of the principle of completeness

2. CHARACTERISTICS OF RAPID PRODUCT DEVELOPMENT

2.1. The Life Cycle

Simultaneous engineering (SE) considers the complete development process and thus carries out the planning on the whole. RPD, on the other hand, considers single tasks and the respective expert team responsible for each task. SE sets up the framework within which RPD organizes the rapid, result-oriented performance of functional activities. The mere application of SE organization on the functional level leads to a disproportionate coordination expenditure.

The overall RPD approach is based on the idea of an evolutionary design cycle (Bullinger et al. 1996). In contrast to traditional approaches with defined design phases and respective documents, such as specification lists or concept matrices, the different design phases are result oriented.

The whole cycle is subject to constraints from the project environment, such as market developments, legislation, and new technologies. Late changes in customer requirements (Thomke and Reinertsen 1998) make necessary research and development (R&D) management that is able to handle these uncertainties efficiently. Furthermore, the execution of the cycle is not necessarily sequential. For example, results from the generation of prototypes can be directly incorporated into a new design phase.

The idea of evolutionary design means that previously unrecognized product requirements or technological progress be considered and incorporated. This issue leads to an important feature of RPD, namely, the abandonment of a homogeneus definition of a product throughout the project. Each product module has an individual definition. The initial concept is conceptualized for the complete product as well as the final integration of the modules. In between, changes are made through new design methods and tools.

The RPD approach will become more transparent by comparing it to the concept of SE (Bullinger and Warschat 1996). The influencable and controllable parameters of a company will serve as a frame for the comparison (see Table 1):

- Organization
- Processes
- Human and technical resources
- Product

2.2. The Organization

Organizational changes, rearrangement of processes, investment in new machines, and training of staff, as well as new solutions for product structures, are necessary to increase the effectiveness and efficiency of the product-development process.

The organization of a company defines its structures, such as the formation of organizational units and the coordination between the units. Project management, as a method that uses certain tools, influences organizational change to a large extent. Whereas SE exhibits a more or less formalized frame with milestones, RPD requires a reactive project-management methodology. The apparent plan precision within phase-oriented approaches such as SE becomes increasingly inaccurate with proceeding project progress. Hence, it will be replaced by a result-oriented approach, where the plan inaccuracy decreases with the progress of the project. For both development approaches, integration of tasks is needed, with labor being planned, controlled, and steered by one responsible person or team.

TABLE 1 Simultaneous Engineering vs. RPD

Parameter	Element	SE	RPD
Organization	Project management	Formalized	Reactive, individual
	Planning	Product-neutral plan with decreasing accuracy	Result-oriented plan with increasing accuracy
	Labor	Integrated approach	
Process	Structure	Full-process orientation	
	Innovation source	Initial product concept	Continuous improvement and redefinition of concepts
	Development cycles	Avoidance strategy	Active process element
Resources	Data integration	Static	Dynamic
	Communication and coordination media	Short paths and SE teams	Short paths and SE teams and CSCW and ASN
Product	Documents	Unique approval by responsible source	Continuous testing and redefinition of concepts
	Definition	Homogeneous according to modularization	Individual according to project progress
	Data management	Standardized product and process data (STEP)	
	Learning/experiences	For next/from previous project	Within the project

In the early phases of product development, those decisions are made that are relevant for the creation of value. The organization needs to be flexible enough to provide appropriate competencies for decisions and responsibilities.

2.3. The Process

RPD concentrates on the early phases of product development. SE already achieved the reduction of production times. RPD now aims to increase the variety of prototypes through an evolutionary iterative process in order to enable comprehensive statements to be made about the product. The interdisciplinary teams work together from the start. The key factors here are good communication and coordination of everyone involved in the process. Thus, the time for finding a solution to the following problems is reduced:

- There are no customer demands. Therefore, without any concrete forms many solutions are possible, but not the solution that is sought.
- The potential of new or alternative technologies results from the integration of experts, whose knowledge can influence the whole process.
- The changing basic conditions of the development in the course of the process make changes necessary in already finished task areas (e.g. risk estimation of market and technology).

These possible basic conditions have to be integrated into the RPD process to reduce time and costs of the whole process.

The application of processes determines the product development and its effectiveness and efficiency. Product data generation and management process can be distinguished. Hence, it is important for the SE as well as the RPD approach to achieve a process orientation in which both product data generation and management process are aligned along the value chain. In a traditional SE approach, innovation results from an initial product concept and product specification, whereas the RPD concept will be checked and redefined according to the project progress. RPD therefore makes it possible to integrate new technologies, market trends, and other factors for a much longer period. Thus, it leads to highly innovative products. Design iterations are a desirable and therefore promoted element of RPD. The change of design concepts and specifications is supported by a fitting framework, including the testing and the most-important evaluation of the design for further improvement.

2.4. The Human and Technical Resources

Common SE approaches are based on standardized and static product data integration, whereas RPD requires dynamic data management in semantic networks in order to enable short control cycles. Short paths and multidisciplinary teams for quick decisions are essential for both approaches. Moreover, RPD requires team-oriented communication systems, which open up new ways of cooperation. They need to offer support not only for management decisions, but also for decision making during the generation of product data.

In RPD, the people and machines involved are of great importance. The people involved need free space for the development within the framework of the evolutionary concept, and well as the will to use the possibilities for cooperation with other colleagues. This means a break with the Taylorized development process. The employees need to be aware that they are taking part in a continually changing process. The technical resources, especially machines with hardware and software for the production of digital and physical prototypes, have to meet requirements on the usability of data with unclear features regarding parameters. They have to be able to build first prototypes without detailed construction data. The quality of the statements that can be made by means of the prototypes depends on how concrete or detailed they are. For optimal cooperation of the single technologies, it is important to use data that can be read by all of them.

2.5. The Product

The results of the product-development process are the documents of the generated product, such as product models, calculations, certificates, plans, and bills of materials as well as the respective documents of the process, such as drawings of machine tools, process plans, and work plans. The aim of all documentation is to support information management. A documentation focusing on product and process data guarantees project transparency for all the persons involved. The standardization of the whole product data is a basic prerequisite for evolutionary and phase-oriented approaches. STEP (standard for the exchange of product model data), as probably the most promising attempt to standardize product data and application interfaces, offers applicable solutions for quite a few application fields, such as automotive and electronic design, rapid prototyping, and ship building. Documents reflecting parts of the complete product data generated, such as specifications, bills of materials, and process data, represent an important difference between SE and RPD. Whereas in an SE documents are synchronized at a certain time (e.g., milestones), the RPD process documents are subject to persistent alteration until a certain deadline. Thus, figures can be changed or agreed upon and boundaries narrowed. The RPD approach only sets rough boundaries within which the modules mature individually. This yields in specific project-management questions, such as (re)allocation of resources or synchronization of the overall process, which are presently still subject to research (Malone and Crowston 1994). Therefore, the RPD process focusses specifically on the management of variants and versions.

3. ELEMENTS OF RAPID PRODUCT DEVELOPMENT

3.1. Process Planning

The goals of RPD are to speed up these iteration cycles, on the one hand, and promote learning within the cycle, on the other. The whole development process involves cooperating development teams that are increasingly distributed globally. The functioning of the cooperation between these teams is essential for the success of the development process. This can only be realized by effective coordination of the partial plans of each of the distributed development teams that are part of the overall product-development chain.

The decentralization of decisions and responsibilities enhances the flexibility and responsiveness of development teams significantly. Hence, planning tools used to coordinate the tasks of development teams have to fit the characteristics of a development process. Consequently, a tool that is designed for central planning approaches will not fit the requirements of a decentralized structure. Specifically, issues needed for RPD, such as coordination of decentralized teams or learning within each cycle, are not supported.

Based on the understanding of planning as the initial step for scheduling diverse processes, the planning of processes involved in complex R&D projects must be possible. The planning system has to be suitable for use in surroundings that are characterized by decentralized and multisited operations. A high grade of expression of the generated plans is based on the ability to process of incomplete and inconsistent data. Therefore, support of documentation and planning has to be integrated, too. Because separate development teams have different tasks, adaptability to the type of the development task and consideration of specific situations have to be ensured. For this reason, open and standardized data storage is fundamental for the planning system. Therefore, the team-oriented project planning system (TOPP) has been developed.

In order to ensure a high grade of expression of the plans, time relations as proposed by Allen (1991) have been used for the phase of plan definition. Logical and time-connected relations between tasks to be planned have to be described within the plan definition phase. Based on 13 Allen time relations, the results of each task are described dependent on the relations to other tasks. Therewith all necessary constraints between related tasks can be represented. This is why TOPP differs from critical path methods. This method only uses the start and the finish to describe time relations.

Each distributed team can define the relations between the tasks to be planned within their responsibility by Allen time relations (internal relations). External relations (interfaces) to other distributed development teams can also be defined. These external relations are additional constraints for the backtracking-planning algorithm that forms the basis for calculating the optimal plan.

Further, the planner uses disjunctive relations to define the constraints between tasks in order to take the requirements of uncertainty into account. For example, the planner can determine whether a task A has to start at the same time as task B, or whether task B can start at the same time as task A is finished.

If all other constraints, such as available resources, can be met, each disjunctive relation will lead to one possible plan alternative. The required resources and the expected duration of the task are added to the task to be planned in order to consider the limits of resources adequately.

The first reason for this approach is the high uncertainty and complexity of R&D and RPD projects. The definition of rules forms the basis for the use of automatic resource assignments. Therefore, abstractions and simplifications are needed, which cannot easily be obtained from complex systems such as R&D or RPD projects. Second, planners are subject to cognitive and mental limitations. Hence, the planning system has to support the planner by giving him the ability to compare plan alternatives under various circumstances.

A typical problem in multiattributive decision making is the proposed selection of one plan out of a limited number of plans characterized by specific figures. Since the figures show ordinal quality, the process of selecting the optimum can be supported by using the precedence sum method.

Planning as a complex task can normally not be solved optimally, due to the limited mental capacities of human planners. The use of models to plan projects offers many advantages:

- From a statistic point of view, the probability of finding the optimal plan increases with the number of plans.
- The comparison of plans based on specific figures offers possibilities for finding advantages and disadvantages of each plan. Additionally, the planner is not limited to one plan.

- Failures within obscure and complex structures typical of RPD are detected rather than anticipated.
- Since sensitivity for the different figures increases, there is a support mechanism with regard to the knowledge about the situation.
- The evaluation of plans based on quantifiable figures contributes to the achievement of plans.

Five different scenarios have been implemented in TOPP. A particular planning aspect is emphasized by each scenario. The scenario is defined via characteristic figures such as process coordination, process risk, and process logics. Hence, the planner is given the ability to judge all possible plans from a specific point of view.

According to the scenario, the calculation of the order of precedence always takes place in the same manner. First plans are evaluated in view of first-order criteria (FOC). If plans still show the same ranking, second-order criteria (SOC) are taken to refine the order. If a final order of precedence is still not possible, the ideal solution, defined by the best characteristic numbers of all plans, determines the order. The plan with the least difference from the optimal plan will be preferred.

Decentralized planning within rapid product development involves more than simple distribution of partial goals. Since development teams are distributed and are responsible for achieving their partial goals, different coordination mechanisms are necessary. The coordination of TOPP is based on planning with consistency corridors, an integration of phase-oriented and result-oriented planning and task-oriented planning.

By the use of characteristic numbers and planning scenarios, a new approach has been presented to support the selection of the optimal plan within complex R&D projects and rapid product development (Wörner 1998).

In general, TOPP offers a way to support planners coordinating global engineering projects of rapid product development and R&D.

3.2. Physical Prototyping

3.2.1. Rapid Prototyping

In addition to the conventional manufacturing of physical prototypes (e.g., CNC milling) the rapid prototype technologies (RPT) are gaining more and more importance. RPT makes it possible to produce a physical artifact directly from its CAD model without any tools. Thus, it is possible to build the prototype of a complex part within a few days rather than the several weeks it would take with conventional prototyping.

In the past, great effort has been put into developing RPTs, improving their processes, and increasing the accuracy of the produced parts. The most common techniques today, like stereolithography (STL), selective laser sintering (SLS), solid ground curing (SGC), and fused deposition modelling (FDM), are mainly used to produce design or geometrical prototypes. They are used primarily for aesthetic, ergonomic, and assembly studies or as pattern masters for casting or molding processes. However, up to now current materials and process limitations have hindered their use as technical or functional prototypes. To accelerate the development process, technical and functional prototypes are of great importance. Therefore, it is necessary to develop powerful technologies for rapid production of prototypes with nearly serial characteristics, for example, material or surface quality. In addition to new or improved RPTs, there are promising developments in the field of coating technologies and sheet metal and solid modeling, which will be a valuable contribution.

3.2.2. Rapid Tooling

In addition to rapid prototyping, rapid tooling has become increasingly important in recent years. It offers the possibility of building functional prototypes. Here, the material and the process of the series product is used. With rapid tooling it is possible to build tools rapidly and inexpensively for prototypes parallel to the product development process. Rapid tooling technologies help to make the process from the first sketch to the final product more efficient. A range of technologies is available, from cutting to generative methods and from milling to the direct or indirect metal laser-sintering process.

3.3. Digital Protoyping

Physical prototypes are often time and cost intensive and thus need to be reduced to a minimum. By the combining of CAD technologies, rapid prototyping, virtual reality, and reverse engineering, prototypes can be produced faster and more cheaply then before. The employment of virtual prototypes in the early phases of product development, in particular, optimizes the whole development process (Thomke and Fujimoto 1998). The strategic advantage of digital prototyping is the advancement of decisions from the test phase with physical prototypes to the early phases of product development with digital prototypes. Thus, the process of product development and testing can be considerably ameliorated. The digital demonstration allows early modification and optimization of the prototype. Furthermore, it leads to a cost-saving increase in the variety of prototypes. By means of virtual prototypes product features can be easily verified and thus development times can be reduced enormously. Also, faults concerning fabrication or the product itself can be detected in the early development phases and thus be eliminated without great expenditures. This makes it possible to start product planning at an early stage. Due to the early overlapping of development and fabrication, additional synergy effects can be expected. Prerequisites for digital prototyping are the following three areas: CAD, simulation, and virtual reality. Simulation (Rantzau and Thomas 1996) and CAD data produce quantifiable results, whereas the connection with VR technologies enables a qualitative evaluation of the results (Figure 2).

An important component of digital prototyping is the digital mock-up (DMU), a purely digital test model of a technical product. The objective of the DMU is the current and consistent availability of multiple views of product shape, function, and technological coherences. This forms the basis on which the modeling and simulation (testing) can be performed and communicated for an improved configuration of the design. This primary digital design model is also called the virtual product. The virtual product is the reference for the development of a new product, specifically in the design and testing phase. The idea is to test the prototype regarding design, function, and efficiency before producing the physical prototype. Thus, effects of the product design can be detected in a very early phase of product development. This way, possible weaknesses of the physical prototype can be detected and corrected in the design phase, before the physical prototype is built. An enormous advantage of the DMU is the shortening of iteration cycles. The decisive changes in the digital prototype are carried out while the physical prototype is being built. During this period, the DMU process can achieve almost 100% of the required quality by means of corrections resulting from the simulation processes. The development process without DMU, on the contrary, requires further tests

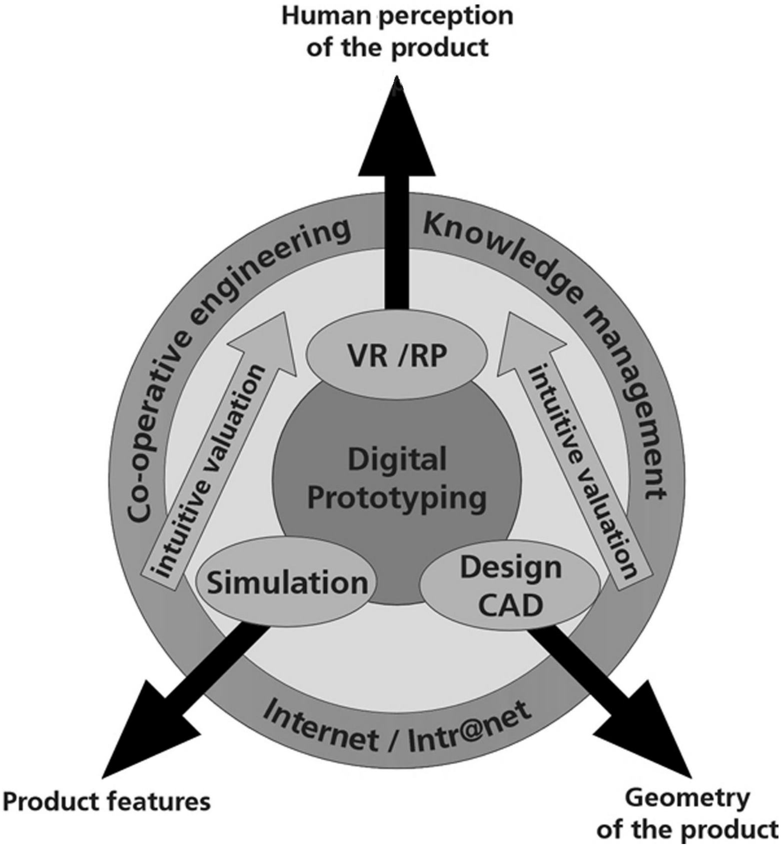


Figure 2 Application Triangle.

with several physical prototypes before the end product can be produced. This means that employing the DMU considerably reduces the time-to-market. The DMU platform also offers the possibility for a technical integration of product conception, design, construction, and packaging.

Digital prototyping offers enormous advantages to many different applications, such as aircraft construction, shipbuilding, and the motor industry. Fields of application for digital prototyping in car manufacturing are, for example:

- Evaluation of components by visualization
- Evaluation of design variations
- Estimation of the surface quality of the car body
- Evaluation of the car's interior
- Ergonomic valuation with the aid of virtual reality

To sum up, creating physical or virtual prototypes of the entire system is of utmost importance, especially in the early phases of the product-development process. The extensive use of prototypes provides a structure, a discipline and an approach that increases the rate of learning and integration within the development process.

3.4. The Engineering Solution Center

The use of recent information and communication technology, interdisciplinary teamwork, and an effective network is essential for the shortening of development times, as we have demonstrated. The prerequisites for effective cooperative work are continuous, computer-supported process chains and new visualization techniques. In the engineering solution center (ESC), recent methods and technologies are integrated into a continuous process chain, which includes all phases of product development, from the first CAD draft to the selection and fabrication of suitable prototypes to the test phase.

The ESC is equipped with all the necessary technology for fast and cost-efficient development of innovative products. Tools, like the Internet, CAD, and FEM simulations, are integrated into the continuous flow of data. Into already existing engineering systems (CAD, knowledge management, databases, etc.) computer-based information and communication technologies are integrated that support the cooperative engineering effectively. Thus, the engineering solution center offers, for example, the complete set of tools necessary for producing a DMU. A particular advantage here is that these tools are already combined into a continuous process chain. All respective systems are installed, and the required interfaces already exist. An important part of the ESC is the power wall, a recent, very effective, and cost-efficient visualization technology. It offers the possibility to project 3D CAD models and virtual prototypes onto a huge canvas. An unlimited number of persons can view the 3D simultaneously. The power wall is a cost-efficient entrance into large 3D presentations because it consists of only one canvas.

Another essential component of the ESC is the engineering/product-data management (EDM/PDM) system. The EDM encompasses holistic, structured, and consistent management of all processes and the whole data involved in the development of innovative products, or the modification

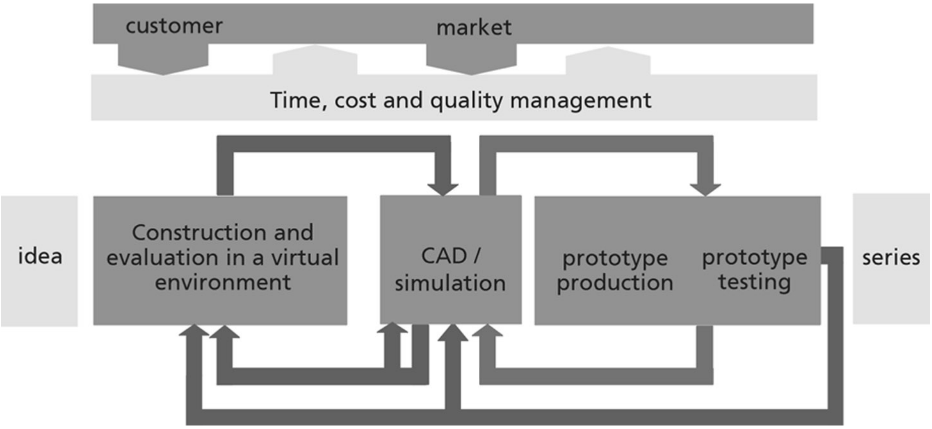

customer
market
Time, cost and quality management
idea
Construction and evaluation in a virtual environment
CAD / simulation
prototype production
prototype testing
series

Figure 3 Digital Prototyping in the Product-Development Process.

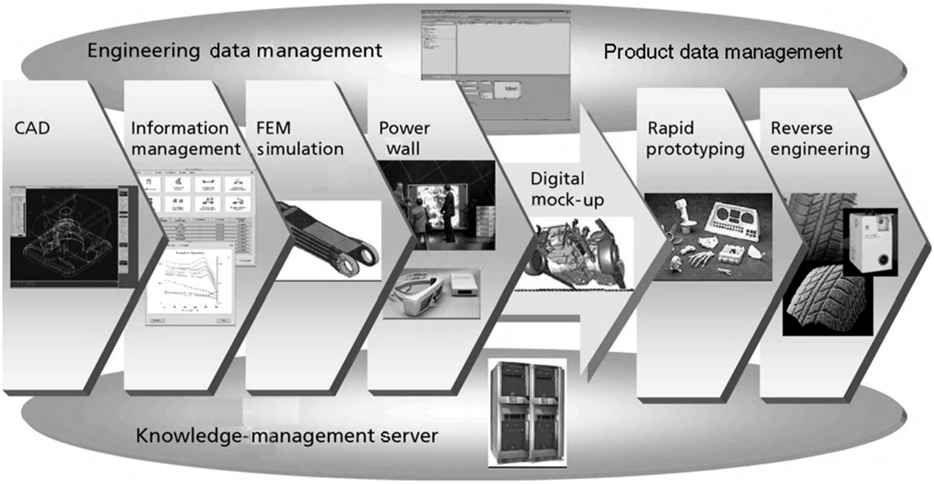


Figure 4 The Engineering Solution Center (ESC).

of already existing products, for the whole product life cycle. The EDM systems manage the processing and forwarding of the produced data. Thus, these systems are the backbone of the technical and administrative information processing. They provide interfaces to CAD systems and other computer-aided applications (CAX), such as computer-aided manufacturing (CAM), computer-aided planning (CAP), and computer-aided quality assurance (CAQ). This way, these systems enable a continuous, company-wide data flow. Inconsistent or obsolete information stocks are reduced to a minimum through the use of EDM.

The innovative approach realized here is what makes the engineering solution center so special. The ESC integrates recent technologies into a continuous process chain. By the use of virtual prototypes the time- and cost-intensive production of physical prototypes can be considerably reduced. The interplay of all methods and technologies makes it possible to achieve high development quality from the first. The virtual product, together with all the applications of virtual technologies and methods in product development and testing, is a necessary reaction to the rapidly changing requirements of the market.

4. KNOWLEDGE ENGINEERING

4.1. Communication and Cooperation

Communication is a further basis for cooperation. It guarantees the continuous exchange of data, information, and knowledge. Particularly dynamic processes, like the development of innovative products, demand great willingness to communicate from the development partners, especially when the partners have not worked together before. Project partners often complain about the great expenditure of time for meetings, telephone calls, and the creation and exchange of documents. If partners are not located nearby, resource problems mean that small and medium-sized companies can arrange personal meetings at short notice only with great difficulty. Nevertheless, face-to-face communication is important because it helps the partners to build up trust and confidence. An information exchange via phone or fax is an insufficient substitute. Especially for small companies, trust is an important factor because it gives them the ability to reduce burdensome expenditures such as frequent coordination and comprehensive documentations. The dynamic in a network of cooperating partners requires a higher degree of communication than most companies are used to because spontaneous agreements concerning the further development process are often necessary. Above all, the manner of communication between the partners has to be altered. Continuous advancement in knowledge and the time pressure put on the development of new products make quick feedback necessary if any deviations from the original planning emerge.

How can communication be improved for the movement of knowledge? There is a difference between explicit knowledge, which is documented, such as on paper or electronically, as language, sketch, or model, and implicit knowledge, which exists only in the heads of the employees.

Another distinction to be made regarding communication is that between the synchronous exchange of knowledge, where partners can communicate at the same time, and the asynchronous exchange, where the transmission and the reception of the message do not happen simultaneously, such as the sending of documents via e-mail.

In most cases of cooperation, exactly this relation is made: implicit knowledge is exchanged synchronously via phone or face to face, and explicit knowledge is exchanged asynchronously via documents. As a consequence, implicit knowledge remains undocumented and explicit knowledge is not annotated. This is a great obstacle to rapid reception of knowledge. Support here is offered by telecooperation systems, which allow communication from computer to computer. Besides documents, pictures, sketches, CAD models, and videos, language can also be exchanged. This way, documents, pictures, and so on can be explained and annotated. By using special input devices, it is possible to incorporate handwritten notes or sketches. This facilitates the documentation of implicit knowledge.

These systems have a further advantage for the integration of knowledge. Learning theory tells us that the reception of new knowledge is easier if several input channels of the learner are occupied simultaneously.

Knowledge-intensive cooperation processes need coordination that is fundamentally different from conventional regulations and control mechanisms. A particular feature of knowledge-intensive processes is that they can hardly be planned. It is impossible to know anything about future knowledge—it is not available today. The more knowledge advances in a project, the higher the probability that previous knowledge will be replaced by new knowledge. Gaining new knowledge may make a former agreement obsolete for one partner. As a consequence, it will sometimes be necessary to give up the previous procedure, with its fixed milestones, work packages, and report cycles, after a short time. The five modules of RPD can be of great help here:

1. Plan and conceive
2. Design
3. Prototyping
4. Check and
5. Evaluate

However, they do not proceed sequentially, as in traditional models. A complete product is not in a certain, exactly-defined development phase. These development projects have an interlocked, networked structure of activities. Single states are determined by occurrences, like test results, which are often caused by external sources, sometimes even in different companies. Instead of a sequential procedure, an iterative, evolutionary process is initiated. But this can function only if the framework for communication is established as described above.

Traditional product-development processes aspire to a decrease in knowledge growth with preceding development time. According to the motto “Do it right from the start,” one aim is usually to minimize supplementary modifications. New knowledge is not appreciated; it might provoke modifications of the original concept. RPD, on the other hand, is very much knowledge oriented. Here, the process is kept open for new ideas and changes, such as customer demands and technical improvements. This necessitates a different way of thinking and alternative processes.

Knowledge-integrated processes are designed according to standards different from those usually applied to business process reengineering. The aim is not a slimming at all costs, but rather the robustness of the process. The motto here is “If the knowledge currently available to me is not enough to reach my target, I have enough time for the necessary qualification and I have the appropriate information technology at my disposal to fill the gap in my knowledge.”

The knowledge-management process has to be considered a direct component of the value-added process. According to Probst et al. (1997), the knowledge-management process consists of the following steps: setting of knowledge targets, identification of knowledge, knowledge acquisition, knowledge development, distribution of knowledge, use of knowledge, and preservation and evaluation of knowledge.

For each of these knowledge process modules, the three fields of organization, human resource management, and information technology are considered. After recording and evaluating the actual state, the modules are optimized with regard to these mentioned fields. The number of iterations is influenced by the mutual dependencies of the modules (Prieto et al. 1999).

From the experiences already gained from R&D cooperation projects the following conclusion can be drawn: if the knowledge-management modules are integrated into the development process and supported by a suitable IT infrastructure, the exchange of knowledge between team members becomes a customer–supplier relationship, as is the case in the value-added process itself. This enables effective and efficient coordination of the project.

4.2. Knowledge Integration

For distributed, interdisciplinary teams it is of great significance that the different persons involved in a project base their work on a common knowledge basis. The cooperating experts have to be able to acquire a basic knowledge of their partners' work contents and processes and their way of thinking in only a short time. Function models and design decisions cannot be communicated without a common basis. Knowledge integration is therefore the basis of communication and coordination in a cooperation. Without a common understanding of the most important terms and their context, it is not possible to transport knowledge to the partners. As a consequence a coordination of common activities becomes impossible. Growing knowledge integration takes time and generates costs. On the other hand, it meliorates the cooperation because few mistakes are made and results are increasingly optimal in a holistic sense. A particular feature of knowledge integration in R&D projects is its dynamic and variable character due to turbulent markets and highly dynamic technological developments. Experiences gained in the field of teamwork made clear that the first phase of a freshly formed (project) team has to be devoted to knowledge integration, for the reasons mentioned above. The task of knowledge integration is to systematize knowledge about artifacts, development processes, and cooperation partners, as well as the respective communication and coordination tools, and to make them available to the cooperation partners. The significance of knowledge integration will probably increase if the artifact is new, as in the development of a product with a new functionality or, more commonly, a highly innovative product. If the project partners have only a small intersection of project-specific knowledge, the integration of knowledge is also very important because it necessitates a dynamic process of building knowledge.

To find creative solutions, it is not enough to know the technical vocabulary of the other experts involved. Misunderstandings are often considered to be communication problems, but they can mostly be explained by the difficult process of knowledge integration.

Which knowledge has to be integrated? First, the knowledge of the different subject areas. Between cooperating subject areas, it is often not enough simply to mediate facts. In this context, four levels of knowledge have to be taken into consideration. In ascending order, they describe an increasing comprehension of coherence within a subject area.

1. Knowledge of facts (“know-what”) forms the basis for the ability to master a subject area. This knowledge primarily reflects the level of acquiring “book knowledge.”
2. Process knowledge (“know-how”) is gained by the expert through the daily application of his knowledge. Here, the transfer of his theoretical knowledge takes place. To enable an exchange on this level, it is necessary to establish space for experiences, which allows the experts to learn together or from one another.
3. The level of system understanding (“know-why”) deals with the recognition of causal and systemic cohesion between activities and their cause and effect chains. This knowledge enables the expert to solve more complex problems that extend into other subject areas.
4. Ultimately, the level of creative action on one's own initiative, the “care-why,” has to be regarded (e.g., motivation). The linkage of the “care-why” demands a high intersection of personal interests and targets.

Many approaches to knowledge integration concentrate mainly on the second level. Transferred to the knowledge integration in a R&D cooperation, this means that it is not enough to match the know-what knowledge. The additional partial integration of the know-how and the know-why is in most cases enough for a successful execution of single operative activities. The success of the whole project, however, demands the integration of the topmost level of the care-why knowledge. A precondition here is common interests between the project partners. This is also a major difference between internal cooperation and cooperation with external partners. In transferring internal procedures to projects with external partners, the importance of the care-why is often neglected because within a company the interests of the cooperation partners do not differ very much; it is one company, after all (Warschat and Ganz 2000).

The integration process of the four levels of knowledge described above is based on the exchange of knowledge between project partners. Some knowledge is documented as CAD models or sketches, reports, and calculations. This is explicit knowledge, but a great share of knowledge is not documented and based on experience; it is implicit.

This model, developed by Nonaka and Takeuchi (1997), describes the common exchange of knowledge as it occurs in knowledge management.

5. SUMMARY AND PERSPECTIVE

The RPD concept is based fundamentally on the early and intensive cooperation of experts from different disciplines. This concept therefore makes it possible to bring together the various sources of expert knowledge in the early phases of product development. Thus, all available sources of information can be used right from the beginning. The initial incomplete knowledge is incrementally completed by diverse experts. Cooperation within and between the autonomous multifunctional teams is of great importance here. The selection and use of suitable information and communication technology are indispensable.

Information exchange is considerably determined by the local and temporal situation of cooperation partners. If the cooperating team members are situated at one place, ordinary, natural communication is possible and sensible. Nevertheless, technical support and electronic documentation might still be helpful. In case cooperation partners are located at different places technical support is indispensable. For this, CSCW and CMC (computer-mediated communication) tools are applied, such as shared whiteboard applications, chat boxes, electronic meeting rooms, and audio/videoconferencing. The currently existing systems make it possible to bridge local barriers. However, they neglect the requirements of face-to-face communication and cooperation. For instance, it is necessary to establish appropriate local and temporal relations among team members. The communication architecture, therefore, should enable the modeling of direct and indirect interactions between individuals. Because of the dynamic of the development process, these relations change. The system should therefore possess sufficient flexibility to enable keeping track of the modifications. Furthermore, the communication basis should be able to represent information not as isolated, but as in the relevant context.

During product development, especially within creative sectors, frequent and rather short ad hoc sessions are preferred. This form of spontaneous information exchange between decentralized development teams requires computer-mediated communication and cooperation techniques, which permit a faster approach and lead to closer cooperation between experts. This results in a harmonized product development, which maintains the autonomy of decentralized teams.

Along with the short iteration cycles, the interdisciplinary teams are an essential feature of the RPD concept. They operate autonomously and are directly responsible for their respective tasks. Additionally, the increasing complexity of products and processes requires an early collaboration and coordination. Thus, it is necessary to make knowledge of technology, design, process, quality, and costs available to anyone involved in the development process.

Conventional databases are not sufficient for an adequate representation of the relevant product and process knowledge. On the one hand, current systems do not consider the dynamic of the development process sufficiently. On the other hand, it is not possible to assess the consequences of one's definition. However, this is a fundamental prerequisite for effective cooperation.

To cope with the given requirements, it is necessary to represent the knowledge in the form of an active semantic network (ASN). This is characterized by active independent objects within a connected structure, which enables the modeling of cause-and-effect relations. The objects in this network are not passive, but react automatically to modifications. This fact provides the possibility of an active and automatic distribution of modifications throughout the whole network. In contrast to conventional systems, ASN contains, in addition to causal connections, representations of methods, communication, and cooperation structures as well as the knowledge required to select the suitable manufacturing technique. Furthermore, negative and positive knowledge (rejected and followed-up alternatives) are stored therein. These acquired perceptions will support the current and future development process. The ASN should exhibit following functions and characteristics:

- Online dialog capability
- Dynamicness
- Robustness
- Version management
- Transparency

All in all, the ASN makes it possible to represent and to manage the design, quality, and cost knowledge together with the know-how of technologies and process planning in the form of the dynamic chains of cause and effect explained here. Thus, the ASN forms the basis for the concept of RPD.

REFERENCES

Allen, J. F. (1991), "Temporal Reasoning and Planning," in *Temporal Reasoning and Planning,* M. B. Morgan and Y. Overton, Eds., Morgan Kaufmann, San Francisco, pp. 1–68.

Bullinger, H.-J., and Warschat, J. (1996), *Concurrent Simultaneous Engineering Systems: The Way to Successful Product Development,* Springer, Berlin.

Bullinger, H.-J., Warschat, and J., Wörner, K. (1996), "Management of Complex Projects as Cooperative Task," in *Proceedings of the 5th International Conference on Human Aspects of Advanced*

Manufacturing Agility an Hybrid Automation (Maui, HI, August) R. J. Koubek and W. Karwowski, Eds., IEA Press, Louisville, KY. pp. 88–96.

Ginsburg, D. (1996), *ATM: Solutions for Enterprise Internetworking,* Addison-Wesley, Reading, MA.

Malone, T., and Crowston, K. (1994), "The Interdisciplinary Study of Coordination," *ACM Computing Survey,* Vol. 26, No. 1, pp. 87–119.

Nonaka, I., and Takeuchi, H. (1997), *Die Organisation des Wissens: wie japanische Unternehmen eine brachliegende Ressource nutzbar machen,* Frankfurt am Main, Campus.

Prieto, J., Hauss, I., Prenninger, J., Polterauer, A., and Röhrbor, D. (1999), "MaKe-IT SME: Managing of Knowledge Using Integrated Tools," in *Proceedings of International Conference on Practical Aspects of Knowledge Management* (London).

Probst, G., Raub, S., and Romhardt, K. (1997), *Wissen managen: wie Unternehmen ihre wertvollste Ressource optimal nutzen,* Gabler, Wiesbaden.

Quinn, L. B., and Russell, R. G. (1999), *Gigabit Ethernet Networking,* Macmillan, New York.

Rantzau, D., and Thomas, P. (1996), "Parallel CFD-Simulations in a Collaborative Software Environment Using European ATM Networks," in *Proceedings of the Parallel CFD '96* (Capri, May 20–23).

Thomke, S., and Fujimoto, T. (1998), "The Effect of 'Front-Loading' Problem-Solving on Product Development Performance," Working Paper, Harvard Business School, Cambridge, MA.

Thomke, S., and Reinertsen, D. (1998), "Agile Product Development: Managing Development Flexibility in Uncertain Environments," *California Management Review,* Vol. 41, No. 1, pp. 8–30.

Warschat, J., and Ganz, W. (2000), "Wissensaustausch über F&E Kooperationen," *io Management* (forthcoming).

Wörner, K., (1998), "System zur dezentralen Planung von Entwicklungsprojekten im Rapid Product Development," Doctoral dissertation, Stuttgart University, Springer, Berlin.

CHAPTER 49

Human-Centered Product Planning and Design

William B. Rouse
Enterprise Support Systems

1. INTRODUCTION 1297
2. DESIGN OBJECTIVES 1297
3. DESIGN ISSUES 1297
4. DESIGN METHODOLOGY 1298
  4.1. Design and Measurement 1298
  4.2. Measurement Issues 1298
    4.2.1. Viability 1299
    4.2.2. Acceptability 1299
    4.2.3. Validity 1299
    4.2.4. Evaluation 1299
    4.2.5. Demonstration 1299
    4.2.6. Verification 1299
    4.2.7. Testing 1299
  4.3. A Framework for Measurement 1299
    4.3.1. Naturalist Phase 1299
    4.3.2. Marketing Phase 1300
    4.3.3. Engineering Phase 1300
    4.3.4. Sales and Service Phase 1300
    4.3.5. The Role of Technology 1300
    4.3.6. Organization for Measurement 1301
5. NATURALIST PHASE 1301
  5.1. Identifying Stakeholders 1301
    5.1.1. Stakeholder Populations 1301
    5.1.2. Designers as Stakeholder Surrogates 1301
    5.1.3. Elusive Stakeholders 1301
  5.2. Methods and Tools for Measurement 1302
    5.2.1. Magazines and Newspapers 1302
    5.2.2. Databases 1302
    5.2.3. Questionnaires 1302
    5.2.4. Interviews 1302
    5.2.5. Experts 1304
  5.3. Summary 1304
6. MARKETING PHASE 1304
  6.1. Methods and Tools for Measurement 1304
    6.1.1. Questionnaires 1304
    6.1.2. Interviews 1305
    6.1.3. Scenarios 1305
    6.1.4. Mock-ups 1305
    6.1.5. Prototypes 1305
  6.2. Summary 1306
7. ENGINEERING PHASE 1306
  7.1. Four-Step Process 1306
  7.2. Objectives Document 1306
  7.3. Requirements Document 1307
  7.4. Conceptual Design Document 1307
  7.5. Detailed Design Document 1307
  7.6. Summary 1307
8. SALES AND SERVICE PHASE 1308
  8.1. Sales and Service Issues 1308
  8.2. Methods and Tools for Measurement 1308
    8.2.1. Sales Reports 1308
    8.2.2. Service Reports 1308
    8.2.3. Questionnaires 1308
    8.2.4. Interviews 1310
  8.3. Summary 1310
9. CONCLUSIONS 1310
REFERENCES 1310

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

This chapter is concerned with designing products and systems using a methodology called human-centered design (Rouse 1991, 1994). Human-centered design is a process of ensuring that the concerns, values, and perceptions of all stakeholders in a design effort are considered and balanced. Stakeholders include users, customers, evaluators, regulators, service personnel, and so on.

Human-centered design can be contrasted with user-centered design (Billings, 1996; Booher 1990; Card et al. 1983; Norman and Draper 1986). The user is a very important stakeholder in design, often the primary stakeholder. However, the success of a product or system is usually strongly influenced by other players in the process of design, development, fielding, and ongoing use of products and systems. Human-centered design is concerned with the full range of stakeholders.

Considering and balancing the concerns, values, and perceptions of such a broad range of people presents difficult challenges. Ad hoc approaches do not consistently work—too much drops through the cracks. A systematic framework, which is comprehensive but also relatively easy to employ, is necessary for human-centered design to be practical. This chapter presents such a framework.

2. DESIGN OBJECTIVES

There are three primary objectives within human-centered design. These objectives should drive much of the designers' thinking, particularly in the earlier stages of design. Discussions in later sections illustrate the substantial impact of focusing on these three objectives.

The first objective of human-centered design is that it should enhance human abilities. This dictates that humans' abilities in the roles of interest be identified, understood, and cultivated. For example, people tend to have excellent pattern-recognition abilities. Design should take advantage of these abilities—for instance, by using displays of information that enable users to respond on a pattern-recognition basis rather than requiring more analytical evaluation of the information.

The second objective is that human-centered design should help overcome human limitations. This requires that limitations be identified and appropriate compensatory mechanisms be devised. A good illustration of a human limitation is the proclivity to make errors. Humans are fairly flexible information processors, but this flexibility can lead to ''innovations'' that are erroneous in the sense that undesirable consequences are likely to occur.

One way of dealing with this problem is to eliminate innovations, perhaps via interlocks and rigid procedures. However, this is akin to throwing out the baby with the bathwater. Instead, mechanisms are needed to compensate for undesirable consequences without precluding innovations. Such mechanisms represent a human-centered approach to overcoming the human limitation of occasional erroneous performance.

The third objective of human-centered design is that it should foster human acceptance. This dictates that stakeholders' preferences and concerns be explicitly considered in the design process. While users are certainly key stakeholders, there are other people who are also central to the process of designing, developing, and operating a system. For example, purchasers or customers are important stakeholders who often are not users. The interests of these stakeholders also have to be considered to foster acceptance by all the people involved.

3. DESIGN ISSUES

This chapter presents an overall framework and systematic methodology for pursuing the above three objectives of human-centered design. There are four design issues of particular concern within this framework.

The first concern is formulating the right problem—making sure that system objectives and requirements are right. All too often, these issues are dealt with much too quickly. There is a natural tendency to ''get on with it,'' which can have enormous negative consequences when requirements are later found to be inadequate or inappropriate.

The second issue is designing an appropriate solution. All well-engineered solutions are *not* necessarily appropriate. Considering the three objectives of human-centered design, as well as the broader context within which systems typically operate, it is apparent that the excellence of the technical attributes of a design are necessary but not sufficient to ensure that the system design is appropriate and successful.

Given the right problem and an appropriate solution, the next concern is developing it to perform well. Performance attributes should include operability, maintainability, and supportability—that is, using it, fixing it, and supplying it. Supportability includes spare parts, fuel, and, most importantly, trained personnel.

The fourth design concern is ensuring human satisfaction. Success depends on people using the system and achieving the benefits for which it was designed. However, before a system can be used, it must be purchased, usually by other stakeholders, which in turn depends on it being technically approved by yet other stakeholders. Thus, a variety of types of people have to be satisfied.

4. DESIGN METHODOLOGY

Concepts such as user-centered design, user-friendly systems, and ergonomically designed systems have been around for quite some time. Virtually everybody endorses these ideas, but very few people know what to do in order to realize the potential of these concepts. What is needed, and what this chapter presents, is a methodological framework within which human-centered design objectives can be systematically and naturally pursued.

4.1. Design and Measurement

What do successful products and systems have in common? The fact that people buy and use them is certainly a common attribute. However, sales are not a very useful measure for designers. In particular, using *lack* of sales as a way to uncover poor design choices is akin to using airplane crashes as a method of identifying design flaws—this method works, but the feedback provided is a bit late.

The question, therefore, is one of determining what can be measured early that is indicative of subsequent poor sales. In other words, what can be measured early to find out if the product or system is unlikely to fly? If this can be done early, it should be possible to change the characteristics of the product or system so as to avoid the predicted failure.

This section focuses on the issues that must be addressed and resolved for the design of a new product or system to be successful. Seven fundamental measurement issues are discussed and a framework for systematically addressing these issues is presented. This framework provides the structure within which the remainder of this chapter is organized and presented.

4.2. Measurement Issues

Figure 1 presents seven measurement issues that underlie successful design (Rouse 1987). The "natural" ordering of these issues depends on one's perspective. From a nuts-and-bolts engineering point of view, one might first worry about testing (i.e., getting the system to work), and save issues such as viability until much later. In contrast, most stakeholders are usually first concerned with viability and only worry about issues such as testing if problems emerge.

A central element of human-centered design is that designers should address the issues in Figure 1 in the same order that stakeholders address these issues. Thus, the *last* concern is, "Does it run?" The *first* concern is, "What matters?," or, "What constitutes benefits and costs?"

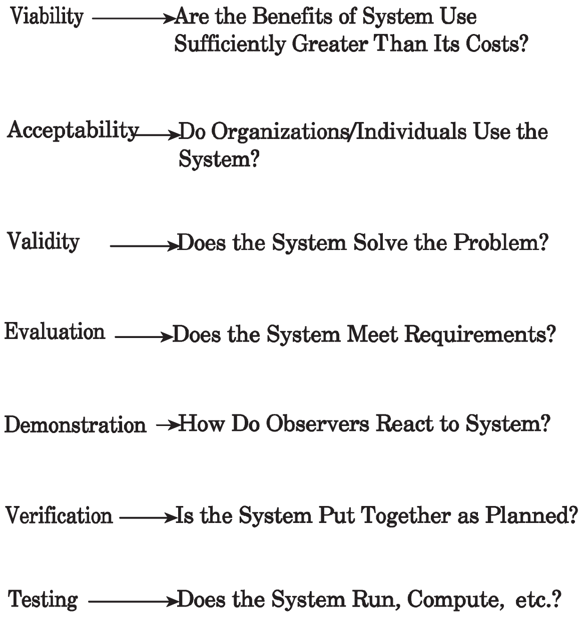

Viability → Are the Benefits of System Use Sufficiently Greater Than Its Costs?
Acceptability → Do Organizations/Individuals Use the System?
Validity → Does the System Solve the Problem?
Evaluation → Does the System Meet Requirements?
Demonstration → How Do Observers React to System?
Verification → Is the System Put Together as Planned?
Testing → Does the System Run, Compute, etc.?

Figure 1 Measurement Issues.

4.2.1. Viability

Are the benefits of system use sufficiently greater than the costs? While this question cannot be answered empirically prior to having a design, one can determine how the question is likely to be answered. How do stakeholders characterize benefits? Are they looking for speed, throughput, an easier job, or appealing surroundings? What influences their perceptions of these characteristics? How do stakeholders characterize costs? Is it simply purchase price? Or do costs include the costs of maintenance and, perhaps, training? Are all the costs monetary?

4.2.2. Acceptability

Do organizations/individuals use the system? This is another question that cannot be answered definitively prior to having the results of design. However, one can determine in advance the factors that are likely to influence the answer. Most of these factors relate to the extent to which a product or system fits into an organization's philosophy, technology, and so on.

4.2.3. Validity

Does the product or system solve the problem? This, of course, leads to the question, what is the problem? How would you know if the problem was solved, or not solved? The nature of this question was discussed earlier in this chapter.

4.2.4. Evaluation

Does the system meet requirements? Formulation of the design problem should result in specification of requirements that must be satisfied for a design solution to be successful. Examples include speed, accuracy, throughput, and manufacturing costs.

4.2.5. Demonstration

How do observers react to the system? It is very useful to get the reactions of potential stakeholders long before the product or system is ready for evaluation. It is important, however, to pursue demonstration in a way that does not create a negative first impression.

4.2.6. Verification

Is the system put together as planned? This question can be contrasted with a paraphrase of the validation question—is the plan any good? Thus, verification is the process of determining that the system was built as intended, but it does not include the process of assessing whether or not it is a good design.

4.2.7. Testing

Does the system run, compute, and so on? This is a standard engineering question. It involves issues of physical measurement and instrumentation for hardware, and runtime inspection and debugging tools for software.

4.3. A Framework for Measurement

The discussion thus far has emphasized the diversity of measurement issues from the perspectives of both designers and stakeholders. If each of these issues were pursued independently, as if they were ends in themselves, the costs of measurement would be untenable. Yet each issue is important and should not be neglected.

What is needed, therefore, is an overall approach to measurement that balances the allocation of resources among the issues of concern at each stage of design. Such an approach should also integrate intermediate measurement results in a way that provides maximal benefit to the evolution of the design product. These goals can be accomplished by viewing measurement as a process involving the four phases shown in Figure 2.

4.3.1. Naturalist Phase

This phase involves understanding the domains and tasks of stakeholders from the perspective of individuals, the organization, and the environment. This understanding includes not only people's activities, but also prevalent values and attitudes relative to productivity, technology, and change in general. Evaluative assessments of interest include identification of difficult and easy aspects of tasks, barriers to and potential avenues of improvement, and the relative leverage of the various stakeholders in the organization.

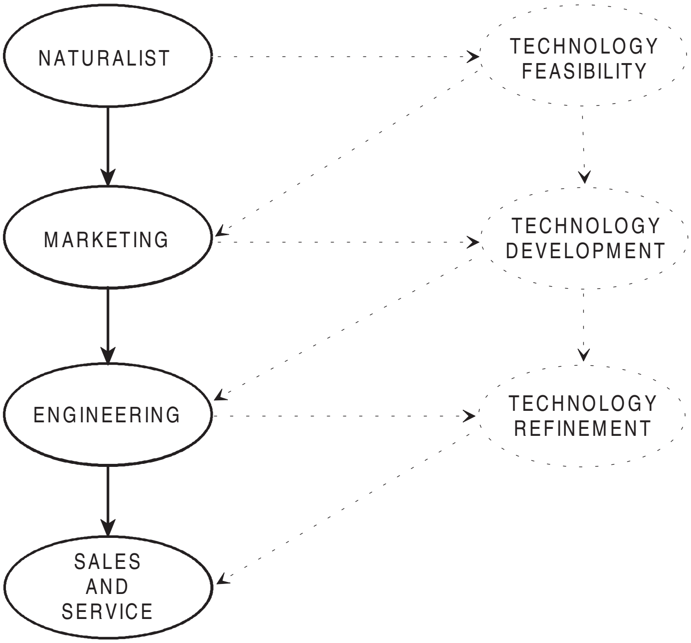


Figure 2 A Framework for Measurement.

4.3.2. *Marketing Phase*

Once one understands the domain and tasks of current and potential stakeholders, one is in a position to conceptualize alternative products or systems to support these people. Product concepts can be used for initial marketing in the sense of determining how users react to the concepts. Stakeholders' reactions are needed relative to validity, acceptability, and viability. In other words, one wants to determine whether or not people perceive a product concept as solving an important problem, solving it in an acceptable way, and solving it at a reasonable cost.

4.3.3. *Engineering Phase*

One now is in a position to begin trade-offs between desired conceptual functionality and technological reality. As indicated in Figure 2, technology development will usually have been pursued prior to and in parallel with the naturalist and marketing phases. This will have at least partially ensured that the product concepts shown to stakeholders were not technologically or economically ridiculous. However, one now must be very specific about how desired functionality is to be provided, what performance is possible, and the time and dollars necessary to provide it.

4.3.4. *Sales and Service Phase*

As this phase begins, the product should have successfully been tested, verified, demonstrated, and evaluated. From a measurement point of view, the focus is now on validity, acceptability, and viability. It is also at this point that one ensures that implementation conditions are consistent with the assumptions underlying the design basis of the product or system.

4.3.5. *The Role of Technology*

It is important to note the role of technology in the human-centered design process. As depicted in Figure 2, technology is pursued in parallel with the four phases of the design process. In fact, technology feasibility, development, and refinement usually consume the lion's share of the resources in a product or system design effort. However, technology should not drive the design process. Human-centered design objectives should drive the process and technology should support these objectives.

4.3.6. Organization for Measurement

Table 1 illustrates how the seven measurement issues should be organized, or sequenced, in the four phases. *Framing* an issue denotes the process of determining what an issue means within a particular context and defining the variables to be measured. *Planning* is concerned with devising a sequence of steps and schedule for making measurements. *Refining* involves using initial results to modify the plan, or perhaps even rethink issues and variables. Finally, *completing* is the process of making outcome measurements and interpreting results.

Table 1 provides a useful context in which to discuss typical measurement problems. There are two classes of problems of interest. The first class is *planning too late,* where, for example, failure to plan for assessing acceptability can preclude measurement prior to putting a product into use. The second class of problems is *executing too early,* where, for instance, demonstrations are executed prior to resolving test and verification issues, and potentially lead to negative initial impressions of a product or system.

5. NATURALIST PHASE

The purpose of the naturalist phase is gaining an understanding of stakeholders' domains and tasks. This includes assessing the roles of individuals, their organizations, and the environment. Also of interest is identifying barriers to change and avenues for change.

The result of the naturalist phase is a formal description of stakeholders, their tasks, and their needs. This description can take many forms, ranging from text to graphics and from straightforward descriptions to theories and hypotheses regarding stakeholders' behaviors.

This section elaborates and illustrates the process of developing descriptions of stakeholders, tasks, and needs. The descriptions resulting from the naturalist phase are the starting point for the marketing phase.

5.1. Identifying Stakeholders

Who are the stakeholders? This is the central question with which a human-centered design effort should be initiated. The answer to this question is not sufficient for success, but it is certainly necessary.

5.1.1. Stakeholder Populations

The key issue is identifying a set of people whose tasks, abilities, limitations, attitudes, and values are representative of the total population of interest. It is often necessary to sample multiple organizations to gain this understanding of the overall population. An exception to this guideline occurs when the total population of stakeholders resides in a single organization.

5.1.2. Designers as Stakeholder Surrogates

Rather than explicitly identifying stakeholders, it is common for designers to think, perhaps only tacitly, that they understand stakeholders and, therefore, can act as their surrogates. To the extent that designers are former stakeholders, this approach has some merit. However, it is inherently limited from capturing the abilities, attitudes, and aspirations of current or potential stakeholders, as well as the current or potential impact of their organizations.

5.1.3. Elusive Stakeholders

It is often argued, particularly for advanced technology efforts, that the eventual stakeholders for the product or system of interest do not yet exist—there are no incumbent stakeholders. This is very

TABLE 1 Organization of Measurement Process

Phase Issue	Naturalist	Marketing	Engineering	Sales and Service
Viability	Frame	Plan	Refine	Complete
Acceptability	Frame	Plan	Refine	Complete
Validity	Frame	Plan	Refine	Complete
Evaluation	–	Frame/plan	Refine/complete	–
Demonstration	–	Frame/plan	Refine/complete	–
Verification	–	Frame/plan	Refine/complete	–
Testing	–	Frame/plan	Refine/complete	–

seldom true, because there are actually extremely few products and systems that are designed from scratch. Even when designing the initial spacecraft, much was drawn from previous experiences in aircraft and submarines.

5.2. Methods and Tools for Measurement

How does one identify stakeholders, and in particular, how does one determine their needs, preferences, values, and so on? Observation is, of course, the necessary means. Initially, unstructured direct observations may be appropriate. Eventually, however, more formal means should be employed to assure unbiased, convincing results. Table 2 lists the methods and tools appropriate for answering these types of questions.

5.2.1. Magazines and Newspapers

To gain an initial perspective on what is important to a particular class of stakeholders or a particular industry, one should read what they read. Trade magazines and industry newspapers publish what interests their readers. One can capitalize on publishers' insights and knowledge by studying articles for issues and concerns. For example, is cost or performance more important? Is risk assessment, or equivalent, mentioned frequently?

One should pay particular attention to advertisements because advertisers invest heavily in trying to understand customers' needs, worries, and preferences. One can capitalize on advertisers' investments by studying the underlying messages and appeals in advertisements.

It is useful to create a file of articles, advertisements, brochures, and catalogs that appear to characterize the stakeholders of interest. Many of these types of materials can also be found on Internet websites. The contents of this file can be slowly accumulated over a period of many months before it is needed. This accumulation might be initiated in light of long-term plans to move in new directions. When these long-term plans become short-term plans, this file can be accessed, the various items juxtaposed, and an initial impression formed relatively quickly. For information available via Internet websites, this file can be readily updated when it is needed.

5.2.2. Databases

Many relatively inexpensive sources of information about stakeholders are available via online databases. This is especially true for Internet-based information sources. With these sources, a wide variety of questions can be answered. How large is the population of stakeholders? How are they distributed, organizationally and geographically? What is the size of their incomes? How do they spend it?

Such databases are also likely to have information on the companies whose advertisements were identified in magazines and newspapers. What are their sales and profits? What are the patterns of growth? Many companies make such information readily available on their Internet websites.

By pursuing these questions, one may be able to find characteristics of the advertisements of interest that discriminate between good and poor sales growth and profits. Such characteristics might include leading-edge technology, low cost, and/or good service.

5.2.3. Questionnaires

Once magazines, newspapers, and databases are exhausted as sources of information, attention should shift to seeking more specific and targeted information. An inexpensive approach is to mail, e-mail, or otherwise distribute questionnaires to potential stakeholders to assess how they spend their time, what they perceive as their needs and preferences, and what factors influence their decisions.

Questions should be brief, have easily understandable responses, and be straightforward to answer. Multiple-choice questions or answers in terms of rating scales are much easier to answer than open-ended essay questions, even though the latter may provide more information.

Low return rate can be a problem with questionnaires. Incentives can help. For example, those who respond can be promised a complete set of the results. In one effort, an excellent response rate was obtained when a few randomly selected respondents were given tickets to Disney World.

Results from questionnaires can sometimes be frustrating. Not infrequently, analysis of the results leads to new questions that one wishes had been on the original questionnaire. These new questions can, however, provide the basis for a follow-up agenda.

5.2.4. Interviews

Talking with stakeholders directly is a rich source of information. This can be accomplished via telephone or even e-mail, but face to face is much better. The use of two interviewers can be invaluable in enabling one person to maintain eye contact and the other to take notes. The use of two interviewers also later provides two interpretations of what was said.

TABLE 2 Methods and Tools for the Naturalist Phase

Methods and Tools	Purpose	Advantages	Disadvantages
Magazines and newspapers	Determine customers' and users' interests via articles and advertisements.	Use is very easy and inexpensive.	Basis and representativeness of information may not be clear.
Databases	Access demographic, product, and sales information.	Coverage is both broad and quantitative.	Available data may only roughly match information needs.
Questionnaires	Query large number of people regarding habits, needs, and preferences.	Large population can be inexpensively queried.	Low return rates and shallow nature of responses.
Interviews	In-depth querying of small number of people regarding activities, organization, and environment.	Face-to-face contact allows in-depth and candid interchange.	Difficulty of gaining access, as well as time required to schedule and conduct.
Experts	Access domain, technology, and/or methodological expertise.	Quickly able to come up to speed on topics.	Cost of use and possible inhibition on creating in-house expertise.

TABLE 3 Methods and Tools for the Marketing Phase

Methods and Tools	Purpose	Advantages	Disadvantages
Questionnaires	Query large number of people regarding preferences for product's functions.	Large population can be inexpensively queried.	Low return rates and shallow nature of responses.
Interviews	In-depth querying of small number of people regarding reactions to and likely use of product's functions.	Face-to-face contact allows in-depth exploration of nature and perceptions of product's functions and benefits.	Difficulty of gaining access, as well as time required to schedule and conduct.
Scenarios	Provide feeling for using product in terms of how functions would likely be used.	Inexpensive approach to providing rich impression of product's functions and benefits.	Written scenarios are not as compelling as visual presentation and require users' willingness to read.
Mock-ups	Provide visual look and feel of product.	Strong visual image can be created and reinforced with photographs.	Necessarily emphasize surface features which are not always product's strength.
Prototypes	Provide ability to use product, typically in a fairly limited sense.	Very powerful and compelling approach to involving potential users.	Relatively expensive and not fully portable; sometimes lead to inflated expectations.

Usually, interviewees thoroughly enjoy talking about their jobs and what types of products and systems would be useful. Often one is surprised by the degree of candor people exhibit. Consequently, interviewees usually do not like their comments tape-recorded.

It is helpful if interviewees have first filled out questionnaires, which can provide structure for the interview as they explain and discuss their answers. Questionnaires also ensure that they will have thought about the issues of concern prior to the interview. In the absence of a prior questionnaire, the interview should be carefully structured to avoid unproductive tangents. This structure should be explained to interviewees prior to beginning the interview.

5.2.5. Experts

People with specialized expertise in the domain of interest, the technology, and/or the market niche can be quite valuable. People who were formerly stakeholders within the population of interest tend to be particularly useful. These people can be accessed via e-mail or informal telephone calls (which are surprisingly successful), gathered together in invited workshops, and/or hired as consultants.

While experts' knowledge can be essential, it is very important that the limitations of experts be understood. Despite the demeanor of many experts, very few experts know everything! Listen and filter carefully.

Further, it is very unlikely that one expert can cover a wide range of needs. Consider multiple experts. This is due not only to a need to get a good average opinion. It is due to the necessity to cover multiple domains of knowledge.

5.3. Summary

The success of all of the above methods and tools depends on one particular ability of designers—the ability to listen. During the naturalist phase, the goal is understanding stakeholders rather than convincing them of the merits of particular ideas or the cleverness of the designers. Designers will get plenty of time to talk and expound in later phases of the design process. At this point, however, success depends on listening.

6. MARKETING PHASE

The purpose of the marketing phase is introducing product concepts to potential customers, users, and other stakeholders. In addition, the purpose of this phase includes planning for measurements of viability, acceptability, and validity. Further, initial measurements should be made to test plans, as opposed to the product, to uncover any problems before proceeding.

It is important to keep in mind that the product and system concepts developed in this phase are primarily for the purpose of addressing viability, acceptability, and validity. Beyond that which is sufficient to serve this purpose, minimal engineering effort should be invested in these concepts. Beyond preserving resources, this minimalist approach avoids, or at least lessens, ‘‘ego investments’’ in concepts prior to knowing whether or not the concepts will be perceived to be viable, acceptable, and valid.

These types of problem can also be avoided by pursuing more than one product concept. Potential stakeholders can be asked to react to these multiple concepts in terms of whether or not each product concept is perceived as solving an important problem, solving it in an acceptable way, and solving it at a reasonable cost. Each person queried can react to all concepts, or the population of potential stakeholders can be partitioned into multiple groups, with each group only reacting to one concept.

The marketing phase results in an assessment of the relative merits of the multiple product concepts that have emerged up to this point. Also derived is a preview of any particular difficulties that are likely to emerge later. Concepts can be modified, both technically and in terms of presentation and packaging, to decrease the likelihood of these problems.

6.1. Methods and Tools for Measurement

How does one measure the perceptions of stakeholders relative to the viability, acceptability, and validity of alternative product and system concepts? Table 3 (see page 1303) lists the appropriate methods and tools for answering this question, as well as their advantages and disadvantages.

6.1.1. Questionnaires

This method can be used to obtain the reactions of a large number of stakeholders to alternative functions and features of a product or system concept. Typically, people are asked to rate the desirability and perceived feasibility of functions and features using, for example, scales of 1 to 10. Alternatively, people can be asked to rank order functions and features.

As noted when questionnaires were discussed earlier, low return rate can be a problem. Further, one typically cannot have respondents clarify their answers, unless telephone or in person follow-ups are pursued. This tends to be quite difficult when the sample population is large.

Questionnaires can present problems if they are the only methods employed in the marketing phase. The difficulty is that responses may not discriminate among functions and features. For example, respondees may rate as 10 the desirability of all functions and features.

This sounds great—one has discovered exactly what people want! However, another interpretation is that the alternatives were not understood sufficiently for people to perceive different levels of desirability among the alternatives. Asking people to rank order items can eliminate this problem, at least on the surface. However, questionnaires are usually not sufficiently rich to provide people with real feelings for the functionality of the product or system.

6.1.2. Interviews

Interviews are a good way to follow up questionnaires, perhaps for a subset of the population sampled, if the sample was large. As noted earlier, questionnaires are a good precursor to interviews in that they cause interviewees to organize their thoughts prior to the interviews. In-person interviews are more useful than telephone or e-mail interviews because it is much easier to uncover perceptions and preferences iteratively during face-to-face interaction.

Interviews are a good means for determining people's a priori perceptions of the functionality envisioned for the product or system. It is useful to assess these a priori perceptions independently of the perceptions that one may subsequently attempt to create. This assessment is important because it can provide an early warning of any natural tendencies of potential stakeholders to perceive things in ways other than intended in the new product or system. If problems are apparent, one may decide to change the presentation or packaging of the product to avoid misperceptions.

6.1.3. Scenarios

At some point, one has to move beyond the list of words and phrases that describe the functions and features envisioned for the product or system. An interesting way to move in this direction is by using stories or scenarios that embody the functionality of interest and depict how these functions might be utilized.

These stories and scenarios can be accompanied by a questionnaire within which respondents are asked to rate the realism of the depiction. Further, they can be asked to consider explicitly, and perhaps rate, the validity, acceptability, and viability of the product functionality illustrated. It is not necessary, however, to explicitly use the words "validity," "acceptability," and "viability" in the questionnaire. Words should be chosen that are appropriate for the domain being studied—for example, viability may be an issue of cost in some domains and not in others.

It is very useful to follow up these questionnaires with interviews, or at least e-mail queries, to clarify respondents' comments and ratings. Often the explanations and clarifications are more interesting and valuable than the ratings.

6.1.4. Mock-ups

Mock-ups are particularly useful when the form and appearance of a product or system are central to stakeholders' perceptions. For products such as automobiles and furniture, form and appearance are obviously central. However, mock-ups can also be useful for products and systems where appearance does not seem to be crucial.

For example, computer-based systems obviously tend to look quite similar. The only degree of freedom is what is on the display. One can exploit this degree of freedom by producing mock-ups of displays using photographs or even viewgraphs for use with an overhead projector.

One word of caution, however. Even such low-budget presentations can produce lasting impressions. One should make sure that the impression created is such that one wants it to last. Otherwise, as noted earlier, one may not get an opportunity to make a second impression.

6.1.5. Prototypes

Prototypes are a very popular approach and, depending on the level of functionality provided, can give stakeholders hands-on experience with the product or system. For computer-based products, rapid prototyping methods and tools have become quite popular because these methods and tools enable the creation of a functioning prototype in a matter of hours.

Thus, prototyping has two important advantages. Prototypes can be created rapidly and enable hands-on interaction. With these advantages, however, come two important disadvantages.

One disadvantage is the tendency to produce ad hoc prototypes, typically with the motivation of having something to show stakeholders. It is very important that the purpose of the prototype be kept in mind. It is a device with which to obtain initial measurements of validity, acceptability, and viability. Thus, one should make sure that the functions and features depicted are those for which these measurements are needed. One should not, therefore, put something on a display simply because it is intuitively appealing. This can be a difficult impulse to avoid.

The second disadvantage is the tendency to become attached to one's prototypes. At first, a prototype is merely a device for measurement, to be discarded after the appropriate measurements are made. However, once the prototype is operational, there is a tendency for people, including the creators of the prototype, to begin to think that the prototype is actually very close to what the final product or system should be like. In such situations, it is common to hear someone say, ''Maybe with just a few small changes here and there . . .''

Prototypes can be very important. However, one must keep their purpose in mind and avoid rabid prototyping! Also, care must be taken to avoid premature ego investments in prototypes. The framework for design presented in this chapter can provide the means for avoiding these pitfalls.

6.2. Summary

During the naturalist phase, the goal was to listen. In the marketing phase, one can move beyond just listening. Various methods and tools can be used to test hypotheses that emerged from the naturalist phase, and obtain potential stakeholders' reactions to initial product and system concepts.

Beyond presenting hypotheses and concepts, one also obtains initial measurements of validity, acceptability, and viability. These measurements are in terms of quantitative ratings and rankings of functions and features, as well as more free-flow comments and dialogue.

7. ENGINEERING PHASE

The purpose of the engineering phase is developing a final design of the product or system. Much of the effort in this phase involves using various design methods and tools in the process of evolving a conceptual design to a final design. In addition to synthesis of a final design, planning and execution of measurements associated with evaluation, demonstration, verification, and testing are pursued.

7.1. Four-Step Process

In this section, a four-step process for directing the activities of the engineering phase and documenting the results of these activities is discussed. The essence of this process is a structured approach to producing a series of design documents. These documents need not be formal documents. They might, for example, only be a set of presentation materials.

Beyond the value of this approach for creating a human-centered design, documentation produced in this manner can be particularly valuable for tracing design decisions back to the requirements and objectives that motivated the decisions. For example, suggested design changes are much easier to evaluate and integrate into an existing design when one can efficiently determine why the existing design is as it is.

It is important to note that the results of the naturalist and marketing phases should provide a strong head start on this documentation process. In particular, much of the objectives document can be based on the results of these phases. Further, and equally important, the naturalist and marketing phases will have identified the stakeholders in the design effort, and are likely to have initiated relationships with many of them.

7.2. Objectives Document

The first step in the process is developing the objectives document. This document contains three attributes of the product or system to be designed: goals, functions, and objectives.

Goals are characteristics of the product or system that designers, users, and customers would like the product or system to have. Goals are often philosophical choices, frequently very qualitative in nature. There are usually multiple ways of achieving goals. Goals are particularly useful for providing guidance for later choices.

Functions define what the product or system should do, but not how it should be done. Consequently, there are usually multiple ways to provide each function. The definition of functions subsequently leads to analysis of objectives.

Objectives define the activities that must be accomplished by the product or system in order to provide functions. Each function has at least 1, and often 5 to 10, objectives associated with it. Objectives are typically phrased as imperative sentences beginning with a verb.

There are two purposes for writing a formal document listing goals, functions, and objectives. First, as noted earlier, written documents provide an audit trail from initial analyses to the ''as-built'' product or system. The objectives document provides the foundation for all subsequent documents in the audit trail for the engineering phase. The second purpose of the objectives document is that it provides the framework—in fact, the outline—for the requirements document.

All stakeholders should be involved in the development of the objectives document. This includes at least one representative from each type of stakeholder group. This is important because this document defines what the eventual product or system will and will not do. All subsequent development assumes that the functions and objectives in the objectives document form a necessary and complete set.

The contents of the objectives document can be based on interviews with subject-matter experts, including operators, maintainers, managers, and trainers. Baseline and analogous systems can also be valuable, particularly for determining objectives that have proven to be necessary for providing specific functions.

Much of the needed information will have emerged from the marketing phase. At the very least, one should have learned from the marketing phase what questions to ask and who to ask. All the stakeholders in the process should have been identified and their views and preferences assessed.

The level of detail in the objectives document should be such that generality is emphasized and specifics are avoided. The activities and resulting document should concentrate on what is desirable. Discussion of constraints should be delayed; budgets, schedules, people, and technology can be considered later.

7.3. Requirements Document

Once all the stakeholders agree that the objectives document accurately describes the desired functions and objectives for the product or system, the next step is to develop the requirements document. The purpose of this document is to identify all information and control requirements associated with each objective in the objectives document.

For evolutionary designs, baseline and analogous systems can be studied to determine requirements. However, if the product or system being designed has no antecedent, subject-matter expertise can be very difficult to find. In this case, answers to the above questions have to come from engineering analysis and, if necessary, validated empirically.

The requirements document should be reviewed and approved by all stakeholders in the design effort. This approval should occur prior to beginning development of the conceptual design. This document can also be very useful for determining the functional significance of future design changes. In fact, the requirements document is often used to answer downstream questions that arise concerning why particular system features exist at all.

7.4. Conceptual Design Document

The conceptual design of a product or system should accommodate *all* information and control requirements as parsimoniously as feasible within the state of the art. The conceptual design, as embodied in the conceptual design document, is the first step in defining how the final system will meet the requirements of the requirements document. The conceptual design document should describe a complete, workable system that meets all design objectives.

Realistically, one should expect considerable disagreement as the conceptual design evolves. However, the conceptual design document should not reflect these disagreements. Instead, this document should be iteratively revised until a consensus is reached. At that point, all stakeholders should agree that the resulting conceptual design is a desirable and appropriate product or system.

7.5. Detailed Design Document

The fourth and final step in the design process involves synthesizing a detailed design. Associated with the detailed design is the detailed design document. This document describes the production version of the product or system, including block diagrams, engineering drawings, parts lists, and manufacturing processes.

The detailed design document links elements of the detailed design to the functionality within the conceptual design document, which are in turn linked to the information and control requirements in the requirements document, which are in turn linked to the objectives within the objectives document. These linkages provide powerful means for efficiently revising the design when, as is quite often the case, one or more stakeholders in the design process do not like the implications of their earlier choices. With the audit trail provided by the four-step design process, evaluating and integrating changes is much more straightforward. As a result, good changes are readily and appropriately incorporated and bad changes are expeditiously rejected.

7.6. Summary

In this section, the engineering phase has been described in terms of a documentation process, including the relationships among documents. As noted earlier, this documentation need not be overly formal. For instance, documents may be simply sets of presentation materials. The key is to make intentions, assumptions, and so on explicit and to ensure that key stakeholders, or their representatives, understand, agree with, and will support the emerging design solution.

Obviously, much of the engineering phase concerns creating the contents of the documents described here. Many of the other chapters in this Handbook provide detailed guidance on these engineering activities. Human-centered design is primarily concerned with ensuring that the plethora of engineering methods and tools discussed in this Handbook are focused on creating viable, acceptable, and valid design solutions.

8. SALES AND SERVICE PHASE

Initiation of the sales and service phase signals the accomplishment of several important objectives. The product or system will have been successfully tested, verified, demonstrated, and evaluated. In addition, the issues of viability, acceptability, and validity will have been framed, measurements planned, and initial measurements executed. These initial measurements, beyond the framing and planning, will have exerted a strong influence on the nature of the product or system.

8.1. Sales and Service Issues

In this phase, one is in a position to gain closure on viability, acceptability, and validity. One can make the measurements necessary to determining whether the product or system really solves the problem that motivated the design effort, solves it in an acceptable way, and provides benefits that are greater than the costs of acquisition and use. This is accomplished using the measurement plan that was framed in the naturalist phase, developed in the marketing phase, and refined in the engineering phase.

These measurements should be performed even if the product or system is presold—for example, when a design effort is the consequence of a winning proposal. In this case, even though the purchase is assured, one should pursue closure on viability, acceptability, and validity in order to gain future projects.

There are several other activities in this phase beyond measurement. One should ensure that the implementation conditions for the product or system are consistent with the assumed conditions upon which the design is based. This is also the point at which the later steps of stakeholder acceptance plans are executed, typically with a broader set of people than those who participated in the early steps of the plan. This phase also often involves technology-transition considerations in general.

The sales and service phase is also where problems are identified and remedied. To the greatest extent possible, designers should work with stakeholders to understand the nature of problems and alternative solutions. Some problems may provide new opportunities rather than indicating shortcomings of the current product or system. It is important to recognize when problems go beyond the scope of the original design effort. The emphasis then becomes one of identifying mechanisms for defining and initiating new design efforts to address these problems.

The sales and service phase also provides an excellent means for maintaining relationships. One can identify changes among stakeholders that occur because of promotions, retirements, resignations, and reorganizations. Further, one can lay the groundwork for, and make initial progress on, the naturalist phase, and perhaps the marketing phase, for the next project, product, or system.

8.2. Methods and Tools for Measurement

How does one make the final assessments of viability, acceptability, and validity? Further, how does one recognize new opportunities? Unstructured direct observation can provide important information. However, more formal methods are likely to yield more definitive results and insights. Table 4 lists the methods and tools appropriate for answering these types of questions.

8.2.1. Sales Reports

Sales are an excellent measure of success and a good indicator of high viability, acceptability, and validity. However, sales reports are a poor way of discovering a major design inadequacy. Further, when a major problem is detected in this manner, it is quite likely that one may not know what the problem is or why it occurred.

8.2.2. Service Reports

Service reports can be designed, and service personnel trained, to provide much more than simply a record of service activities. Additional information of interest concerns the specific nature of problems, their likely causes, and how stakeholders perceive and react to the problems. Stakeholders' suggestions for how to avoid or solve the problems can also be invaluable. Individuals' names, addresses, and telephone numbers can also be recorded so that they can be contacted subsequently.

8.2.3. Questionnaires

Questionnaires can be quite useful for discovering problems that are not sufficient to prompt service calls. They also can be useful for uncovering problems with the service itself. If a record is maintained of all stakeholders, this population can regularly be sampled and queried regarding problems, as well as ideas for solutions, product upgrades, and so on. As noted before, however, a primary disadvantage of questionnaires is the typical low return rate.

TABLE 4 Methods and Tools for the Sales and Service Phase

Methods and Tools	Purpose	Advantages	Disadvantages
Sales reports	Assess perceived viability of product or system.	The ultimate, bottom-line measure of success.	Information on lack of sales due to problems is likely to be too late to help.
Service reports	Assess typical problems and impact of solutions attempted.	Associate problems with customers and users and enable follow-up.	May be too late for major problems and may not explain cause.
Questionnaires	Query large number of customers and users regarding experiences with product.	Large population can be inexpensively queried.	Low return rates and shallow nature of responses.
Interviews	In-depth querying of small number of customers and users regarding experiences with product.	Face-to-face contact allows in-depth and candid interchange.	Difficulty of gaining access, as well as time required to schedule and conduct.

8.2.4. Interviews

Interviews can be a rich source of information. Stakeholders can be queried in depth regarding their experiences with the product or system, what they would like to see changed, and new products and systems they would like. This can also be an opportunity to learn how their organizations make purchasing decisions, in terms of both decision criteria and budget cycles.

While sales representatives and service personnel can potentially perform interviews, there is great value in having designers venture out to the sites where their products and systems are used. Such sorties should have clear measurement goals, questions to be answered, an interview protocol, and so on, much in the way that is described in earlier sections. If necessary, interviews can be conducted via telephone or e-mail, with only selected face-to-face interviews to probe more deeply.

8.3. Summary

The sales and service phase brings the measurement process full circle. An important aspect of this phase is using the above tools and methods to initiate the next iteration of naturalist and marketing phases. To this end, as was emphasized earlier, a primary prerequisite at this point is the ability to *listen*.

9. CONCLUSIONS

This chapter has presented a framework for human-centered design. Use of this framework will ensure a successful product or system in terms of viability, acceptability, validity, and so on. In this way, human-centered design provides the basis for translating technology opportunities into market innovations.

REFERENCES

Billings, C. E. (1996). *Aviation Automation: The Search for a Human-Centered Approach*, Erlbaum, Hillsdale, NJ.

Booher, H. R., Ed. (1990), *MANPRINT: An Approach to Systems Integration,* Van Nostrand Reinhold, New York.

Card, S. K., Moran, T. P., and Newell, A. (1983), *The Psychology of Human–Computer Interaction,* Erlbaum, Mahwah, NJ.

Norman, D. A., and Draper, S. W., Eds. (1986), *User Centered System Design: New Perspectives on Human–Computer Interaction,* Erlbaum, Hillsdale, NJ.

Rouse, W. B. (1987), ‘‘On Meaningful Menus for Measurement: Disentangling Evaluative Issues in System Design,’’ *Information Processing and Management,* Vol. 23, pp. 593–604.

Rouse, W. B. (1991), *Design for Success: A Human-Centered Approach to Designing Successful Products and Systems,* John Wiley & Sons, New York.

Rouse, W. B. (1994), *Best Laid Plans*, Prentice Hall, Englewood Cliffs, NJ.

CHAPTER 50

Design for Manufacturing*

C. Richard Liu*
Purdue University

Xiaoping Yang
Purdue University

1. INTRODUCTION **1311**
2. DESIGN AND DESIGN ALTERNATIVES **1313**
3. DRAWINGS **1314**
4. GENERAL PRINCIPLES FOR DESIGN FOR MANUFACTURABILITY **1315**
5. PROCESSES AND MATERIALS FOR PRODUCING THE DESIGN **1316**
6. DESIGN FOR BASIC PROCESSES—METAL **1316**
  6.1. Liquid State 1316
  6.2. Solid State 1317
  6.3. Other Basic Processes 1319
7. DESIGN FOR SECONDARY OPERATION **1320**
8. DESIGN FOR BASIC PROCESSES—PLASTICS **1324**
  8.1. Compression Molding 1324
  8.2. Transfer Molding 1324
  8.3. Injection Molding 1324
  8.4. Extrusion 1324
  8.5. Casting 1324
  8.6. Cold Molding 1325
  8.7. Thermoforming 1325
  8.8. Calendering 1325
  8.9. Blow Molding 1325
  8.10. Parameters Affecting the Selection of the Optimum Basic Process 1325
9. DESIGN FOR ASSEMBLY **1328**
10. COMPUTER SOFTWARE TOOLS: OBJECT-ORIENTED PROGRAMMING AND KNOWLEDGE-BASED SYSTEMS **1328**
11. ORGANIZATIONAL ISSUES **1329**
REFERENCES **1330**
ADDITIONAL READING **1330**

1. INTRODUCTION

The objective of design for manufacturability is to incorporate producibility issues early on in the product design stage so that the customers can be attracted and the needs of the customers can be satisfied in a short lead time and at a competitive cost. The customers' needs include satisfaction in the product with respect to its performance capabilities, quality, reliability, serviceability, aesthetics, and time of delivery.

---

* Parts of this chapter were originally published in Chapter 13 of the Second Edition of this Handbook authored by C. Richard Liu and the late Benjamin W. Niebel.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

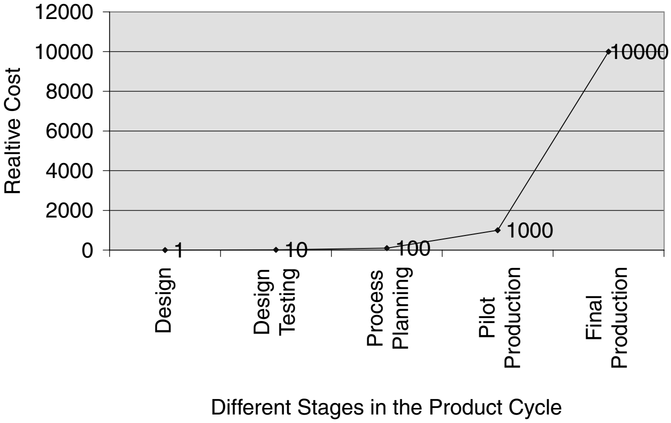


Figure 1 Comparative Cost of an Engineering Change in Different Stages in the Product Cycle. (*Source:* Shina 1991)

Conventional engineering practice in the past has resulted in separate, and sometimes isolated, activities between design and manufacturing that have proven to be time consuming and costly. A study compared the cost of any change in design in three different stages, namely, in production, manufacturing engineering, and design. The cost of a change in the production stage may be ap-

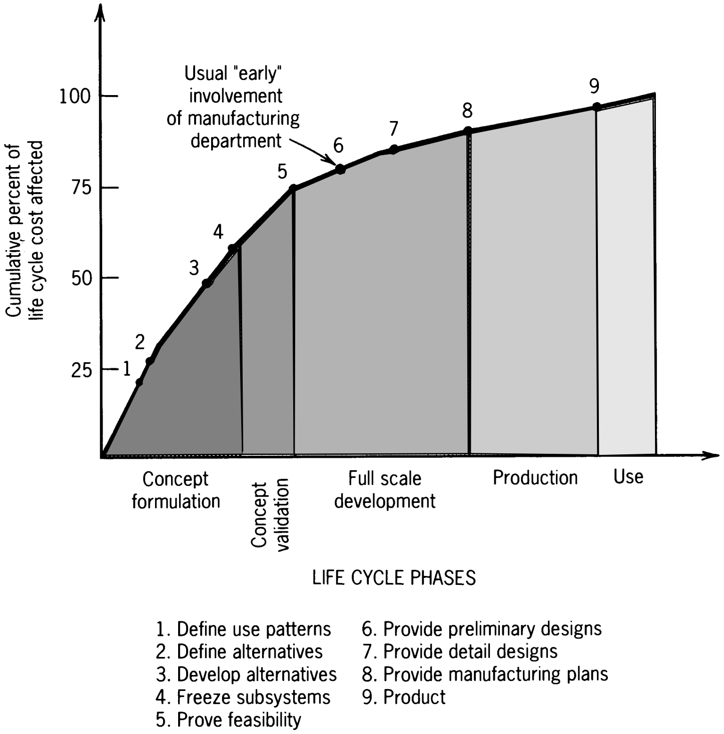


Figure 2 Life-Cycle Phases.

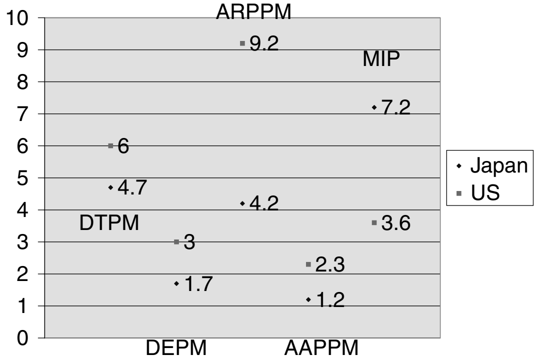


Figure 3 Comparison between Japanese and U.S. auto industry. DTPM: design time/model (multiply the number by 10 to convert the time unit into month); DEPM: design effort/model; ARPPM: average replacement period/model (years); AAPPM: average annual production/model (multiple the number by 100,000); and MIP: models in production (multiple the number by 10). (*Source:* Shina 1991).

proximately an order of magnitude more than the cost of the change made early in the manufacturing engineering stage. Figure 1 shows comparative cost of an engineering change in different stages in the product cycle. To avoid the costly changes due to manufacturability problems, factors related to manufacturing must be considered in all phases of the design process, starting with the design-conception phase. Another study (Nevins and Whitney 1989) further confirmed the importance of making the right decision early. This study indicated that in the concept formulation stage 60% of the life-cycle cost of a product has already been determined. Before full-scale development 75% of the life-cycle cost has been determined. This is illustrated in Figure 2. It is clear from the figure that the DFM needs to be considered in the early conceptual design stage to yield the maximized benefits.

The major issues in competitiveness have moved from cost to productivity to quality. The current and future major issue is time. The other issue are not less important, but the new frontier is speed: studies have shown that over 80% of market share in a new product category goes to the first two companies that get their products to market. Further studies have shown that a 20% cost overrun in the design stage of the product cycle will result in about 8% reduced profits over the lifetime of the product. A six-month overrun in time during the design stage today will result in about 34% loss over the life of the product (Brazier and Leonard 1990).

Figure 3 compares Japanese and U.S. auto design and product cycles. The competitive advantage of the Japanese auto industry results from concurrent engineering and design for manufacture (Shina 1991).

2. DESIGN AND DESIGN ALTERNATIVES

The essence of design is that it is a plan to achieve a purpose or to satisfy a need. In mechanical design, the plan is a representation, such as a set of drawings defining the configuration (geometry and material) of physical elements.

The immediate purpose of a specific set of physical elements or a specific design is the functional requirement. The design process, at this level, is to start with the known functional requirement to plan or search for the design configurations.

The design solution is almost always not unique. Conceptually, the design process can be considered a mapping process between the "purpose space" and the "functional space" and between the "functional space" and "configuration space." The ability to develop alternative physical designs is of fundamental importance to design for manufacturability.

Alternative physical designs may be developed by knowing the functional requirement. A design, in general, can be decomposed into subfunctional requirements for each of its subsystems. Each subfunctional requirement, again, can be used to characterize and develop the design alternatives of each subsystem. By repeating this process, a functional design hierarchy can be developed with the possible design alternatives at various levels of functional requirements, for the product (the assembly), the subassembly, and parts. This design hierarchy is shown in Figure 4 (Liu and Trappey 1989).

The properties of the design hierarchy are as follows:

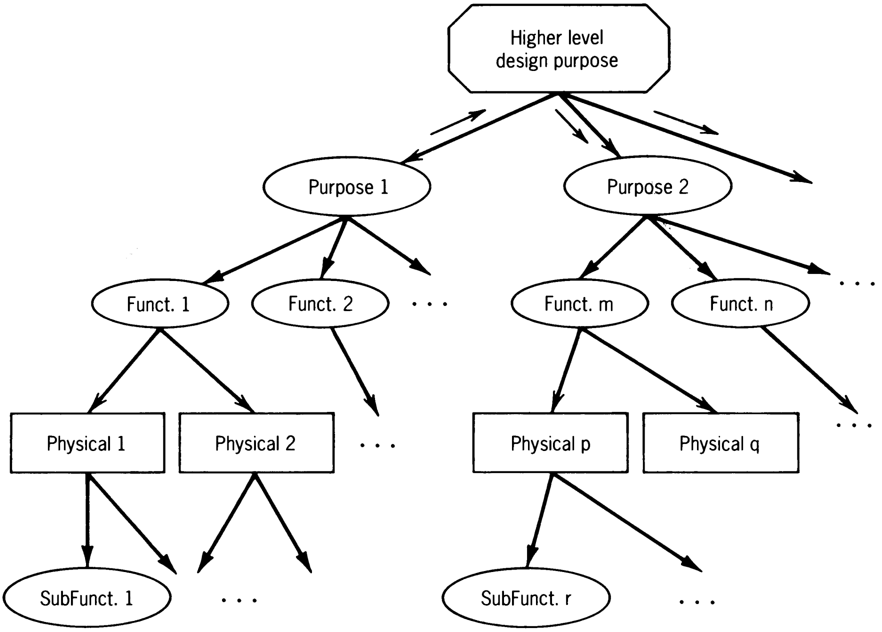


Figure 4 The Design Hierarchy with Upstream and Downstream Reasonings.

1. All the subfunctions are required to form a design, and only one among all the physical design alternatives is needed to satisfy a specific functional requirement.
2. Each and every possible physical design, for a system or a subsystem, has place in the hierarchy. Therefore, the hierarchy serves as a guide for design knowledge acquisition, as a structure for design knowledge storage, and as an indexing system for design knowledge retrieval. This has important application in serving as a software tool for concurrent engineering.
3. Upstream reasoning from the physical design can be conducted by answering the question "What is the design for?" Then the higher level functional requirement may be reached.
4. Downstream reasoning from functional requirement can be done by answering the question "How can the functional requirement be satisfied?" Then the physical design alternatives may be generated.
5. Upstream–downstream reasoning forces the designer to analyze the functional requirements and higher purposes. Thus, it can be used for managing the design process and yet, in the meantime, allow for the individual designer's creativity (see Figure 2).
6. The hierarchical system can serve as structured blackboard for design communication, consultation, and retrieval.

The application of the design hierarchy by one of the authors (Liu) has led to very significant product innovation. When the same method was applied by the students in his classes, general improvement in design creativity was observed. However, the results varied tremendously among the individuals.

More discussions and elaboration of the proposed functional–physical design hierarchy were done in Liu and Trappey (1989) and Trappey and Liu (1990).

3. DRAWINGS

Drawings represent the heart of design for manufacturing because they are the principal means of communication between the functional designer and the producer of the design. They alone control and completely delineate shape, form, fit, finish, and interchangeability requirements that lead to the most competitive procurement. An engineering drawing, when supplemented by reference specifications and standards, should permit a competent manufacturer to produce the part shown within the dimensional and surface tolerance specifications provided. It is the engineering drawing that should demonstrate the most creative design for manufacturing thinking.

Certain product specifications may not be included on the drawings in view of space constraints. Product specifications such as quality assurance checkpoints, inspection procedures, and general design criteria may be separately summarized but should always be cross-referenced on the engineering drawing. At all times the design engineer must remember that the end product drawing is the communication medium between the design engineer and the producer. It is the basis for interchangeability for repair parts; it provides the form, fit, and function to the manufacturing function.

Too often the language of drawings is incomplete. For example, chamfers may be indicated but not be dimensioned; worse yet, they may be desired but not even be shown. Frequently the finish desired is omitted. Complex coring may be incorrectly shown. The principal errors common to many designs are as follows:

1. Design is not conducive to the application of economic processing.
2. Designer has not taken advantage of group technology and creates a new design for an already existing item.
3. Design exceeds the manufacturing state of the art.
4. Design and performance specifications are not compatible.
5. Critical location surfaces have not been established.
6. Design specifies the use of inappropriate items.
7. Design specifications are not definitive.
8. Inadequate consideration has been given to measurement problems.
9. Tolerances are more restrictive than necessary.
10. Item has been overdesigned.

4. GENERAL PRINCIPLES FOR DESIGN FOR MANUFACTURABILITY

In this section we only stress some important concepts. In later sections we will review the design for basic processes. For more detailed information, see Bralla (1986) and Stillwell (1989).

1. Consider the entire product, including all its subsystems and components, and the entire spectrum of manufacturing–inspection–assembly activities. We should avoid producing improvement in one at the expense of another. For example, product design to specify the assembly operations may create difficulties in disassembling the product, thus hurting maintainability and serviceability of the product. Simplifying the component processing may create complexity in assembly.
2. Search for simplicity first in system designs, then in subsystem designs, and then in component designs. Considering simplicity in component level only will lead to missing the opportunities for significant improvement.
3. Ask whether the functional needs are absolutely necessary. Chances are the functional needs can be reduced, thus leading to significant simplifications of the configuration design and processing requirements. Example: A careful examination of the functional needs of a gear train system has led to a relaxation of the functional specifications that enables the use of a four-bar linkage as a replacement of the gear train. The impact on manufacturability is obviously very significant.

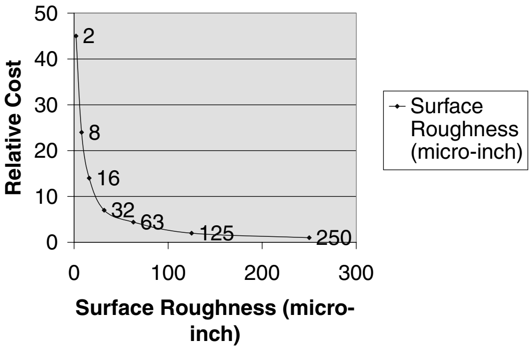


Figure 5 Relative Cost Corresponding to Different Surface Roughness. (*Source:* Bralla 1986)

4. Design for fewer parts, simpler shapes, least precision requirements, fewer manufacturing steps, and minimum information requirements. Figure 5 shows the relative cost corresponding to different surface roughness.
5. Apply the concept of design modularization and group technology. Reduce the varieties of sizes and shapes. Experience has shown that the number of hole sizes may be reduced significantly without affecting the function, thus reducing the number of sizes of drills needed.
6. Always consider standard materials, components, and subassemblies.
7. Design appropriate to the expected level of production and to fit the existing production facilities.
8. Select the shape of the raw material close to the finished designs.
9. Design for easy inspection.
10. Design for part orientation to maximize the value added in each setup.
11. Design for easy assembly and maintainability.

5. PROCESSES AND MATERIALS FOR PRODUCING THE DESIGN

The selection of the ideal processes and materials with which to produce a given design cannot be an independent activity. It must be a continuing activity that takes place throughout the design life cycle, from initial conception to production. Material selection and process selection need to be considered together; they should not be considered independently.

In considering the selection of materials for an application, it is usually possible to rule out entire classes of materials because of cost or their obvious inability to satisfy specific operational requirements. But even so, with the acceleration of material development there are so many options for the functional design engineer that optimum selection is at best difficult. The suggested procedure for organizing data related to material selection is to divide it into three categories: properties, specifications, and data for ordering.

The property category will usually provide the information that suggests the most desirable material. A property profile is recommended, where all information, such as yield point, modulus of elasticity, resistance to corrosion, and so on, is tabulated. Those materials that qualify because of their properties will stand out.

Each material will have its own specifications on the individual grades available and on their properties, applications, and comparative costs. The unique specifications of a material will distinguish it from all competing materials and will serve as the basis for quality control, planning, and inspection.

Finally, the data needed when physically placing an order need to be maintained. This includes minimum order size, quantity breakpoints, and sources of supply.

In the final selection of a material, cost of the proposed material needs to be considered—hence the need for close association between material selection and process selection in connection with design.

Design evaluation invariably is in terms of a proposed material cost, which may be derived by analyzing the involved processing steps, including setup and lead-time costs along with the preprocessed material cost.

6. DESIGN FOR BASIC PROCESSES—METAL

6.1. Liquid State

Early in the planning of the functional design, one must decide whether to start with a basic process that uses material in the liquid state, such as a casting, or in the solid state, such as a forging. If the engineer decides a part should be cast, he or she will have to decide simultaneously which casting alloy and process can most nearly meet the required dimensional tolerance, mechanical properties, and production rate at the least cost.

Casting has several distinct assets: the ability to fill a complex shape, economy when a number of similar pieces are required, and a wide choice of alloys suitable for use in highly stressed parts, where light weight is important or where corrosion may be a problem. There are inherent problems, too, including internal porosity, dimensional variations caused by shrinkage, and solid or gaseous inclusions stemming from the molding operation. However, most of these problems can be minimized by sound design for manufacturing.

Casting processes are basically similar in that the metal being formed is in a liquid or highly viscous state and is poured or injected into a cavity of a desired shape.

The following design guidelines will prove helpful in reducing casting defects, improving their reliability and assisting in their producibility:

1. When changes in sections are required, use smoothly tapered sections to reduce stress concentration. Where sections join, use generous fillets and blending radii.
2. Machining allowances should be detailed on the part drawing so as to ensure adequate stock and avoid excessive differences in casting thickness.
3. Remember that when design castings to be produced in a metal mold or die, convex forms are easy to mill but concave notches are both difficult and expensive.
4. Raised lettering is simple to cut into a metal mold or die; depressed lettering will cost considerably more.
5. Avoid the design of thin sections since they will be difficult to fill.
6. To facilitate the secondary operations of drilling and tapping, cored-through holes should have countersinking on both ends of the holes.
7. Avoid large, plain surfaces. Break up these areas with ribs or serration to avoid warpage and distortion.
8. For maximum strength, keep material away from the neutral axis. Endeavor to keep plates in tension and ribs in compression.

Table 1 identifies the important design parameters associated with the various casting processes and provides those limitations that should be incorporated by the functional designer to ensure producibility.

6.2. Solid State

A forging, as opposed to a casting, is usually used because of improved mechanical properties, which are a result of working metals into a desired configuration under impact or pressure loading. A refinement of the grain structure is another characteristic of the forging process. Hot forging breaks up the large dendritic grain structure characteristic of castings and gives the metal a refined structure, with all inclusions stretched out in the direction in which plastic flow occurs. A metal has greater load-carrying ability in the direction of its flow lines than it does across the flow lines. Consequently, a hot-formed part should be designed so that the flow lines run in the direction of the greatest load during service.

An extension of conventional forging known as precision forging can be used to acquire geometric configurations very close to the final desired shape, thus minimizing secondary machining operations.

Guidelines that should be observed in the design of forging in order to simplify its manufacturing and help ensure its reliability are as follows:

1. The maximum length of bar that can be upset in a single stroke is limited by possible buckling of the unsupported portion. The unsupported length should not be longer than three times the diameter of the bar or distance across the flats.
2. Recesses in depth up to their diameter can be easily incorporated in either or both sides of a section. Secondary piercing operations to remove the residual web should be utilized on through-hole designs.
3. Draft angle should be added to all surfaces perpendicular to the forging plane so as to permit easy removal of the forged part. Remember that outside draft angles can be smaller than inside angles since the outside surfaces will shrink away from the die walls and the inside surfaces will shrink toward bosses in the die.
4. Deeper die cavities require more draft than shallow cavities. Draft angles for hard-to-forge materials, such as titanium and nickel-base alloys, should be larger than when forging easy-to-forge materials.
5. Uniform draft results in lower-cost dies, so endeavor to specify one uniform draft on all outside surfaces and one larger draft on all inside surfaces.
6. Corner and fillet radii should be as large as possible to facilitate metal flow and minimize die wear. Usually 6 mm (0.24 in.) is the minimum radius for parts forged from high-temperature alloys, stainless steels, and titanium alloys.
7. Endeavor to keep the parting line in one plane since this will result in simpler and lower-cost dies.
8. Locate the parting lines along a central element of the part. This practice avoids deep impressions, reduces die wear, and helps ensure easy removal of the forged part from the dies.

Table 2 provides important design for manufacturing information for the major forging processes.

TABLE 1 Important Design Parameters Associated with Various Casting Processes

Design Parameter	Casting Process							
	Sand Casting							
	Green	Dry/Cold/Set	Shell	Plaster (Preheated Mold)	Investment (Preheated Mold)	Premanent Mold (Preheated Mold)	Die (Preheated Mold)	Centrifugal
Weight	100 g to 400 MT	100 g to 400 MT	100 g to 100 kg	100 g to 100 kg	Less than 1 g to 50 kg	100 g to 25 kg	Less than 1 g to 30 kg	Grams to 200 kg
Minimum section thickness	3 mm	3 mm	1.5 mm	1 mm	0.5 mm	3 mm	0.75 mm	6 mm
Allowance for machining	Ferrous—2.5–9.5 mm; nonferrous—1.5–6.5	Ferrous—2.5–9.5 mm; nonferrous—1.5–6.5 mm	Often not required; when required, 2.5–6.5 mm	0.75 mm	0.25–0.75 mm	0.80–3 mm	0.80–1.60 mm	2.50–6.5 mm
General tolerance	±0.4 ~ 6.4 mm	±0.4 ~ 6.4 mm	±0.08 ~ ±1.60 mm	±0.13 ~ ±0.26 mm	±0.05–±1.5 mm	±0.25–±1.5 mm	±0.025–±0.125 mm	±0.80–±3.5 mm
Surface finish (μrms)	6.0–24.0	6.0–24.0	1.25–6.35	0.8–1.3	0.5–2.2	2.5–6.35	0.8–2.25	2.5–13.0
Process reliability	90	90	90	90	90	90	90	90
Cored holes	Holes $<$ 6 mm	Holes $<$ 6 mm	Holes $<$ 6 mm	Holes $<$ 12 mm	Holes as small as 0.5 mm diameter	Holes as small as 5 mm diameter	Holes as small as 0.80 mm diameter	Holes as small 25 mm diameter; no undercuts
Minimum lot size	1	1	100	1	20	1000	3000	100
Draft allowances	1°–3°	1°–3°	$\frac{1}{4}$°–1°	$\frac{1}{2}$°–2°	0°–$\frac{1}{2}$°	2°–3°	2°–5°	0°–3°

TABLE 2 Important Design Parameters Associated with Various Forging Processes

Design Parameter	Forging Process: Open Die	Conventional Utilizing Preblocked	Closed Die	Upset	Precision Die
Size or weight	500 g to 5000 kg	Grams to 20 kg	Grams to 20 kg	20–250 mm bar	Grams to 20 kg
Allowance for finish machining	2–10 mm	2–10 mm	1–5 mm	5–10 mm	0–3 mm
Thickness tolerance	±0.6 mm −0.2 mm to +3.00 mm −1.00 mm	+0.4 mm −0.2 mm to +2.00 mm −0.75 mm	+0.3 mm −0.15 mm to +1.5 mm − 0.5 mm	—	+0.2 mm −0.1 mm to +1 mm 0.2 mm
Filet and corners	5–7 mm	3–5 mm	2–4 mm	—	1–2 mm
Surface finish (μrms)	3.9–4.5	3.8–4.5	3.2–3.8	4.5–5.0	1.25–2.25
Process reliability	95	95	95	95	95
Minimum lot size	25	1000	1500	25	2000
Draft allowance	5°–10°	3°–5°	2°–5°	—	0°–3°
Die wear tolerance	±0.075 mm/kg weight of forging	±0.075 mm/kg weight of forging	±0.075 mm/kg weight of forging		±0.075 mm/kg weight of forging
Mismathcing tolerance	±.25 mm ±0.01 mm/3 kg weight of forging	±0.25 mm ±0.01 mm/3 kg weight of forging	±0.25 mm ±0.01 mm/3 kg weight of forging	—	±0.25 mm ±0.01 mm/3 kg weight of forging
Shrinkage tolerance	±0.08 mm	±0.08 mm	±0.08 mm	—	±0.08 mm

6.3. Other Basic Processes

In addition to casting and forging, several other processes that may be considered basic since they impart the approximate finished geometry to material that is in the powdered, sheet, or rod-shape form. Notable among these are powder metallurgy, cold heading, extrusion, roll forming, press forming, spinning, electroforming, and automatic screw machine work.

In powdered metallurgy, powdered metal is placed in a die and compressed under high pressure. The resulting cold-formed part is then sintered in a furnace to a point below the melting point of its major constituent.

Cold heading involves striking a segment of cold material up to 25 mm (1 in.) in diameter in a die so that it plastically deformed to the die configuration.

Extrusion is performed by forcing heated metal through a die having an aperture of the desired shape. The extruded lengths are then cut into the desired length. From the standpoint of producibility, the following design features should be observed:

1. Very thin sections with large circumscribing area should be avoided.
2. Any thick wedge section that tapers to a thin edge should be avoided.
3. Thin sections that have close space tolerance should be avoided.
4. Sharp corners should be avoided.
5. Semiclosed shapes that necessitate dies with long, thin projections should be avoided.
6. When a thin member is attached to a heavy section, the length of the thin member should not exceed 10 times its thickness.

In roll forming, strip metal is permanently deformed by stretching it beyond its yield point. The series of rolls progressively changes the shape of the metal to the desired shape. In design, the extent of the bends in the rolls, allowance must be made for springback.

In press forming, as in roll forming, metal is stretched beyond its yield point. The original material remains about the same thickness or diameter, although it will be reduced slightly by drawing or ironing. Forming is based upon two principles:

1. Stretching and compressing material beyond the elastic limit on the outside and inside of a bend.
2. Stretching the material beyond the elastic limit without compressing the material beyond the elastic limit without stretching.

Spinning is a metal-forming process in which the work is formed over a pattern, usually made of hard wood or metal. As the mold and material are spun, a tool (resting on a steady rest) is forced against the material until the material contacts the mold. Only symmetrical shapes can be spun. The manufacturing engineer associated with this process is concerned primarily with blank development and proper feed pressure.

In electroforming, a mandrel having the desired inside geometry of the part is placed in an electroplating bath. After the desired thickness of the part is achieved, the mandrel pattern is removed, leaving the formed piece.

Automatic screw machine forming involves the use of bar stock, which is fed and cut to the desired shape.

Table 3 provides important design for manufacturing information for these basic processes.

7. DESIGN FOR SECONDARY OPERATION

Just as there should be careful analysis in the selection of the ideal basic or primary process, so must there be sound planning in the specification of the secondary processes. The parameters associated with all process planning include the size of the part, the geometric configuration or shape required, the material, the tolerance and surface finished needed, the quantity to be produced, and of course the cost. Just as there are several alternatives in the selection of a basic process, so there are several alternatives in determining how a final configuration can be achieved.

With reference to secondary removal operations, several guidelines should be observed in connection with the design of the product in order to help ensure its producibility.

1. Provide flat surfaces for entering of the drill on all holes that need to be drilled.
2. On long rods, design mating members so that male threads can be machined between centers, as opposed to female threads, where it would be difficult to support the work.
3. Always design so that gripping surfaces are provided for holding the work while machining is performed and ensure that the held piece is sufficiently rigid to withstand machining forces.
4. Avoid double fits in design for mating parts. It is much easier to maintain close tolerance when a single fit is specified.
5. Avoid specifying contours that require special form tools.
6. In metal stamping, avoid feather edges when shearing. Internal edges should be rounded, and corners along the edge of the strip stock should be sharp.
7. In metal stamping of parts that are to be subsequently press formed, straight edges should be specified, if possible, on the flat blanks.
8. In tapped blind holes, the last thread should be at least 1.5 times the thread pitch from the bottom of the hole.
9. Blind-drilled holes should end with a conical geometry to allow the use of standard drills.
10. Design the work so that diameters of external features increase from the exposed face and diameters of internal features decrease.
11. Internal corners should indicate a radius equal to the cutting tool radius.
12. Endeavor to simplify the design so that all secondary operations can be performed on one machine.
13. Design the work so that all secondary operations can be performed while holding the work in a single fixture or jig.

Table 4 provides a comparison of the basic machining operations used in performing the majority of secondary operations.

TABLE 3 Important Design Parameters Associated with Manufacturing Information for Basic Processes

Design Parmeter	Powder Metallurgy	Cold Headng	Extrusion	Roll Forming	Press Forming	Spinning	Electroforming	Automatic Screw Machine
Size	Diameter—1.5–300 mm; length —3–225 mm	Diameter—0.75–20 mm; length —1.50–250 mm	1.5–250 mm diameter	1–2000 mm	Up to 6 mm diameter	6–4000 mm	Limited to size of plating tanks	0.80 mm diameter by 1.50 mm; length to 200 mm diameter by 900 mm length
Minimun thickness	1 mm	—	1 mm	0.075 mm	0.075 mm	0.1 mm	0.0025 mm	—
Allowance for finish maching	To size	To size	To size	To size	To size	To size	To size	To size
Tolerance	Diameter—±0.025–0.125 mm; length—±0.25–0.50 mm	Diameter—−0.05–0.125 mm; length—±0.75–2.25 mm	Flatness—±0.01 mm/in. of width; wall thickness—±0.15–0.25 mm; cross section—±0.15–0.20	Cross section—±0.050–0.35 mm; length—±1.5 mm	±0.25 mm	Length—±0.12 mm; thickness—±0.05 mm	Wall thicknesses—±0.025 mm; dimension—±0.005 mm	Diameter—±0.01–.06 mm; length—±0.04–0.01 mm; 0.10 concentricity—±0.06 mm
Surface finish (μrms)	0.125–1.25	2.2–2.6	2.5–3	2.2–2.6	2.2–4.0	0.4–2.2	0.125–0.250	0.30–2.5
Process reliability	95	99	99	99	99	90–95	99	98
Minimum lot size	1000	5000	500 ft	10,000 ft	1500	5	25	1000
Draft allowance	0	—	—	—	$0°–\frac{1}{4}°$	—	—	—
Bosses permitted	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Undercuts permitted	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inserts permitted	Yes	No	No	No	No	No	No	No
Holes permitted	Yes	No	Yes	Yes	Yes	No	Yes	Yes

TABLE 4 Machining Operations Used in Performing Secondary Operation

Process	Shape Produced	Machine	Cutting Tool	Tolerance	Surface Finish (πrms)	Relative Motion	
						Tool	Work
Turning (external)	Surface of revolution (cylindrical)	Lath, boring machine	Single point	±0.005–±0.025 mm	0.8–6.4	↔ ↕	↻
Boring (external)	Cylindrical (enlarges holes)	Boring machine	Single point	±0.005–±0.025 mm	0.4–5.0	↻ ↔	↻ ↔
Shaping and planing	Flat surface or slots	Shaper, planer	Single point	±0.025–±0.050 mm	0.8–6.4	↕	→
Milling (end, form, slab)	Flat and contoured surfaces and slots	Milling machine—horizontal, vertical, bed-type	Multiple points	±0.025–±0.050 mm	0.8–6.4	↔ ↻	✣
Drilling	Cylindrical (originating holes (0.1–100 mm diameter)	Drill press	Twin-edge drill	±0.050–±0.100 mm	2.5–6.4	↻ ↕	Fixed
Grinding (cylindrical surface, plunge)	Cylindrical, flat and formed	Griding machine—cylindrical, surface, thread	Multiple points	±0.0025–±0.0075 mm	0.2–3.2	↻ ✛	
Reaming	Cylindrical (enlarging and improving finish of holes)	Drill press, turret lathe	Multiple points	±0.0125–±0.0500 mm	0.8–2.5	↻ ↕	Fixed
Broaching	Cylindrical, flat, slots	Broaching machine, press	Multiple points	±0.005–±0.0150 mm	0.8–2.5	↕	Fixed
Electric discharge machining	Variety of shapes depending on shape of electrode	Electric discharge machine	Single-point electrode	±0.050 mm	0.8–5.0	↕	Fixed

Electrochemical machining	Variety of shapes; usually odd-shaped cavities of hard material	Electrochemical machine	Dissolution process	±0.050 mm	0.3–1.5	Andoic dissolution; tool is cathode	Workpiece is anode
Chemical machining	Variety of shapes; usually blanking of intricate shapes, printed circuit etching, or shallow cavities	Chemical machining machine	Chemical attack of exposed surfaces	±0.050 mm	0.6–1.8	Chemical attack of exposed surfaces	Fixed
Laser machining	Cylindrical holes as small as 5 μm	Laser beam machine	Single-wavelength beam of light	Holes are reproducible within ±3%	0.6–2.5	Fixed	Fixed
Ultrasonic machining	Same shape as tool	Machine equipped with magnetostrictive transducer, generator power supply	Shaped tool and abrasive powder	±0.025 mm	0.3–0.9	↕	Fixed
Electron beam machining	Cylindrical slots	Electron beam machine equipped with vacuum of 10^{-4} mm of mercury	High-velocity electrons focus on workpiece	±0.025 mm	0.6–1.8	↕	Fixed
Gear generating	Eccentric cams, ratchets, gears	Gear shaper	Single-point reciprocating	±0.013–±0.025 mm	1.8–3.8	↕ ↻	↻
Hobbing	Any form that regularly repeats itself on periphery of circular part	Hobbing machine	Multiple points	±0.013–±0.025 mm	1.8–3.8	↻	↻
Treplanning	Large through holes, circular grooves	Lathe-like machine	One or more single-point cutters revolving around a center	±0.13 mm	2.5–6.4	↻	↻

8. DESIGN FOR BASIC PROCESSES—PLASTICS

There are more than 30 distinct families of plastic, from which evolve thousands of types and formulations that are available to the functional designer. However, in the fabrication of plastics, either thermoplastic or thermosetting, only a limited number of basic processes are available. These processes include compression molding, transfer molding, injection molding, extrusion, casting, cold molding, thermoforming, calendering, rotational molding, and blow molding. The functional designer usually gives little thought to how the part will be made. He or she is usually concerned primarily with the specific gravity, hardness, water absorption, outdoor weathering, coefficient of linear thermal expansion, elongation, flexural modulus, izod impact, deflect temperature under load, and flexural yield, tensile, shear, and compressive strengths.

8.1. Compression Molding

In compression molding, an appropriate amount of plastic compound (usually in powder form) is introduced into a heated mold, which is subsequently closed under pressure. The molding material, either thermoplastic or thermosetting, is softened by the heat and formed into a continuous mass having the geometric configuration of the mold cavity. If the material is thermoplastic, hardening is accomplished by cooling the mold. If the material is thermosetting, further heating will result in the hardening of the material.

Compression molding offers the following desirable features:

1. Thin-walled parts (less than 1.5 mm) are readily molded with this process with little warpage or dimensional deviation.
2. There will be no gate markings, which is of particular importance on small parts.
3. Less shrinkage, and more uniform, is characteristic of this molding process.
4. It is especially economical for larger parts (those weighing more than 1 kg).
5. Initial costs are less since it usually costs less to design and make a compression mold than a transfer or injection mold.
6. Reinforcing fibers are not broken up as they are in closed-mold methods such as transfer and injection. Therefore, the fabricated parts under compression molding may be both stronger and tougher.

8.2. Transfer Molding

Under transfer molding, the mold is first closed. The plastic material is then conveyed into the mold cavity under pressure from an auxiliary chamber. The molding compound is placed in the hot auxiliary chamber and subsequently forced in a plastic state through an orifice into the mold cavities by pressure. The molded part and the residue (cull) are ejected upon opening the mold after the part has hardened. Under transfer molding, there is no flash to trim; only the runner needs to be removed.

8.3. Injection Molding

In injection molding, the raw material (pellets, grains, etc.) is placed into a hopper, called the barrel, above a heated cylinder. The material is metered into the barrel every cycle so as to replenish the system for what has been forced into the mold. Pressure up to 1750 $\mathrm{kg/cm^2}$ forces the plastic molding compound through the heating cylinder and into the mold cavities. Although this process is used primarily for the molding of thermoplastic materials, it can also be used for thermosetting polymers. When molding thermosets, such as phenolic resins, low barrel temperatures should be used (65–120°C). Thermoplastic barrel temperatures are much higher, usually in the range of 175–315°C.

8.4. Extrusion

Like the extrusion of metals, the extrusion of plastics involves the continuous forming of a shape by forcing softened plastic material through a die orifice that has approximately the geometric profile of the cross-section of the work. The extruded form is subsequently hardened by cooling. With the continuous extrusion process, such products as rods, tubes, and shapes of uniform cross-section can be economically produced. Extrusion to obtain a sleeve of the correct proportion almost always precedes the basic process of blow molding.

8.5. Casting

Much like the casting of metals, the casting of plastics involves introducing plastic materials in the liquid form into a mold that has been shaped to contour of the piece to be formed. The material that is used for making the mold is often flexible, such as rubber latex. Molds may also be made of nonflexible materials such as plaster. Epoxies, phenolics, and polyesters are plastics that are frequently fabricated by the casting process.

8.6. Cold Molding

Cold molding takes place when thermosetting compounds are introduced into a room-temperature steel mold that is closed under pressure. The mold is subsequently opened, and the formed article is transferred to a heating oven, where it is baked until it becomes hard.

8.7. Thermoforming

Thermoforming is restricted to thermoplastic materials. Here sheets of the plastic material are heated and drawn over a mold contour so that the work takes the shape of the mold. Thermosetting may also be done by passing the stock between a sequence of rolls that produce the desired contour. Most thermoplastic materials become soft enough for thermoforming between 135 and 220°C. The plastic sheet that was obtained by calendering or extrusion can be brought to the correct thermoforming temperature by infrared radiant heat, electrical resistance heating, or ovens using gas or fuel oil.

8.8. Calendering

Calendering is the continuous production of a thin sheet by passing thermoplastic compounds between a series of heated rolls. The thickness of the sheet is determined by adjusting the distance between the rolls. After passing between the final set of rolls, the thin plastic sheet is cooled before being wound into large rolls for storage.

8.9. Blow Molding

In blow molding, a tube of molten plastic material, the parison, is extruded over an apparatus called the blow pipe and is then encased in a split mold. Air is injected into this hot section of extruded stock through the blow pipe. The stock is then blown outward, where it follows the contour of the mold. The part is then cooled, the mold opened, and the molded part ejected. In very heavy sections, carbon dioxide or liquid nitrogen may be used to hasten the cooling. This process is widely used in molding high- and low-density polyethylene, nylon, polyvinyl chloride (PVC), polypropylene, polystyrene, and polycarbonates.

8.10. Parameters Affecting the Selection of the Optimum Basic Process

Selecting the optimum basic process in the production of a given plastic design will have a significant bearing on the success of that design. The principal parameters that should be considered in the selection decision include the plastic material to be used, the geometry or configuration of the part, the quantity to be produced, and the cost.

If the functional designer cannot identify the exact plastic material that is to be used, he or she should be able to indicate whether a thermoplastic or thermosetting resin is being considered. This information alone will be most helpful. Certainly both thermoforming and blow molding are largely restricted to thermosetting resins, as is transfer molding. Injection molding is used primarily for producing large-volume thermoplastic moldings, and extrusion for large-volume thermoplastic continuous shapes.

Geometry or shape also has a major impact on process selection. Unless a part has a continuous cross-section, it would not be extruded; unless it were relatively thin walled and bottle shaped, it would not be blow molded. Again, calendering is restricted to flat sheet or strip designs, and the use of inserts is restricted to the molding processes.

The quantity to be produced also has a major role in the selection decision. Most designs can be made by simple compression molding, yet this method would not be economical if the quantity were large and material were suitable for injection molding.

The following design for manufacturing points apply to the processing of plastics:

1. Holes less than 1.5 mm diameter should not be molded but should be drilled after molding.
2. Depth of blind holes should be limited to twice their diameter.
3. Holes should be located perpendicular to the parting line to permit easy material removal from the mold.
4. Undercuts should be avoided in molded parts since they require either a split mold or a removable core section.
5. The section thickness between any two holes should be greater than 3 mm.
6. Boss heights should not be more than twice their diameter.
7. Bosses should be designed with at least a 5° taper on each side for easy withdrawal from the mold.
8. Bosses should be designed with radii at both the top and the base.
9. Ribs should be designed with at least a 2–5° taper on each side.

TABLE 5 Basic Processes Used to Fabricate Plastics and Their Principal Parameters

	Parameter									
Process	Shape Produced	Machine	Mold or Tool	Material	Typical Tolerance	Minimum Wall Thickness	Ribs	Draft	Inserts	Minimim Quantity Requirements
Calendering	Continuous sheet for film	Multiple-roll calender	None	Thermoplastic	0.05–0.200 mm depending on material		None	None	None	Low
Extrusion	Continuous form such as rods, tubes, filaments, and simple shapes	Extrusion press	Hardened steel die	Thermoplastic	0.01–0.30 mm depending on material		None	None	Possible to extrude over or around wire insert	Low (tooling is inexpensive)
Compression molding	Simple outlines and plain cross sections	Compression press	Hardened tool steel mold	Thermoplatic or thermosetting	0.04–0.25 mm depending on material	1.25			Yes	Low
Transfer molding	Complex geometrics possible	Transfer press	Hardened tool steel mold	Thermosetting	0.04–0.25 mm depending on material	1.5 mm	3°–5° taper; height < 3 times wall thickness	$\frac{1}{2}$°–5°	Yes	High
Injection molding	Complex geometrics possible	Injection press	Hardened tool steel mold	Thermoplastic or thermosetting	0.04–0.25 mm depending on material	1.25 mm	2°–5° taper; height = $1\frac{1}{2}$ times wall thickness; width $\frac{1}{2}$ of wall thickness	$\frac{1}{4}$°–4°	Yes	High
Casting	Simple outlines and plain cross sections	None	Metal mold or expoxy mold	Thermosetting	0.10–0.50 mm depending on material	2.0 mm			Yes	Low to medium depending on mold
Cold molding	Simple outlines and plain cross sections	None	Mold of wood, plaster, or steel	Thermosetting	0.01–0.05 mm depending on material	2.0 mm			Yes	Low

Blow molding	Thin walled and bottle shaped	Pneumatic blow molding machine	Tool steel mold	Thermoplastic				No	High
Rotational molding	Full enclosures or semienclosures (hollow objects)	Rotomolding system	Cast aluminum or fabricated metal	Thermoplastic, limited thermosetting	0.03–0.60 mm	3.0 mm		No	Medium
Filament winding	Tubes, piping, tanks	Filament winding machine	Must have axis about which the filament can be wound	Sing-end continuous strand glass fiber and thermoplastic	0.20–0.50 mm	3.0 mm			Medium

10. Ribs should be designed with radii at both the top and the base.
11. Ribs should be designed at a height of 1.5 times the wall thickness. The rib width of the base should be half the wall thickness.
12. Outside edges at the parting line should be designed without a radius. Fillets should be specified at the base ribs and bosses and on corners and should be not less than 0.8 mm.
13. Inserts should be at right angles to the parting line and of a design that allows both ends to be supported in the mold.
14. A draft or taper of 1–2° should be specified on the vertical surfaces or walls parallel with the direction of mold pressure.
15. Cavity numbers should be engraved in the mold. The letters should be 2.4 mm high and 0.18 mm deep.
16. Threading below 8 mm diameter should be cut after molding.

Table 5 identifies the major parameters associated with basic processes used to fabricate thermoplastic and thermosetting resins.

9. DESIGN FOR ASSEMBLY

The goal of DFA is to ease the assembly of the product. Boothroyd et al. (1994) propose a method for DFA that involves two principal steps:

- Designing with as few parts as possible. This is accomplished by analyzing parts pairwise to determine whether the two parts can be created as a single piece rather than as an assembly.
- Estimating the costs of handling and assembling each part using the appropriate assembly process to generate costs figures to analyze the cost savings through DFA.
- In addition to the assembly cost reductions through DFA, there are reductions in part costs that are more significant. Other benefits of DFA include improved reliability and reduction in inventory and production control costs. Consequently, DFA should be applied regardless of the assembly cost and product volume.

10. COMPUTER SOFTWARE TOOLS: OBJECT-ORIENTED PROGRAMMING AND KNOWLEDGE-BASED SYSTEMS

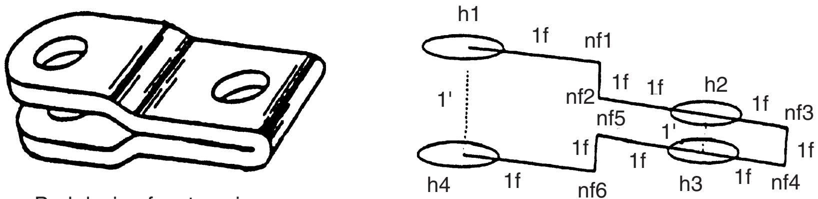
Modern CAD/CAM systems and computer-aided processing planning systems for machining are well known and are very important for integrating design and manufacturing. However, more work is needed to develop them into tools for helping design for manufacturability. We need a software system that can be easily modulized, expanded, alternated in its structures and contents, and integrated partially or fully. The key technology is a recently developed style and structure of programming called object-oriented programming (OOP).

Object-oriented programming supports four unique object functions or properties:

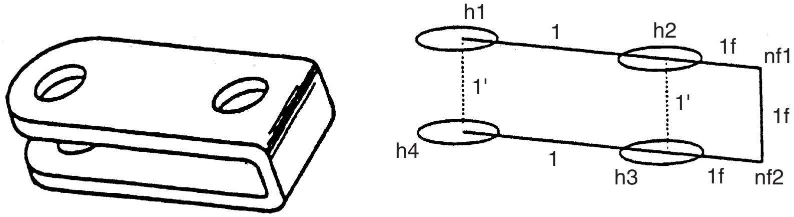
1. *Abstraction:* Abstraction is done by the creation of a ‘‘class protocol description’’ that defines the properties of any object that is an instance of that class.
2. *Encapsulation:* An object encapsulates all the properties (data and messages) of the specific instance of the class.
3. *Inheritance:* Some classes are subordinate to others and are called subclasses. Subclasses are considered to be special cases of the class under which they are grouped in the hierarchy. The variables and methods defined in the higher-level classes will be automatically inherited by the lower-level classes.
4. *Polymorphism:* Allows us to send the same message to different objects in different levels of class hierarchy. Each object responds in a way that is inherited or redefined with respect to the object’s characteristics.

With these properties, integrated and expandable software for supporting designs, including design for manufacturability, can be developed. An example is shown in Trappey and Liu (1990), who developed a system shell for design using the object-oriented programming language, SMALLTALK-80 (Goldberg 1984).

Another key software technology for design for manufacturability, such as automated rule checking, is knowledge-based systems, or expert systems. The general methodology for building these systems roughly consists of five steps: identification, conceptualization, formalization, implementation, and testing (Hayes-Roth et al. 1983). An example of this approach for fixture design for machining is shown in Ferreira et al. (1985).



Bad design for stamping



Good design for stamping

(*a*) (*b*)

Figure 6 Parts and Their Sketching Abstractions. (*a*) Examples of stamping parts. (*b*) Parts-sketching abstraction that facilitates manufacturability evaluation in conceptual design. (Reprinted from *Robotics and Computer-Integrated Manufacturing,* Vol. 13, No. 3, A. Mukherjee and C. R. Liu, "Conceptual Design, Manufacturability Evaluation and Preliminary Process Planning Using Function–Form Relationships in Stamped Metal Parts," p. 258, Copyright 1997, with permission from Elesevier Science)

Current CAD systems have been oriented to detail design, while the application of DFM guidelines to early design stages yields the largest benefits. Research is needed to lay the foundation for the CAD system for conceptual design so that DFM and CAD can be integrated successfully. Mukherjee and Liu (1995) propose a promising method. In the proposed representation, called sketching abstraction, the discretionary geometry of the part having functional relevance is captured using functional features, while the nondiscretionary geometry is represented using a linkage mechanism. The functional features are related to the part function using data structures called function–form matrices. They attempt to bridge the gap between function and form representations and provide the designer with a tool that can be used for generating design alternatives. Figure 6 is an example of this representation, which facilitates manufacturability evaluation in conceptual design (Mukherjee and Liu 1997).

11. ORGANIZATIONAL ISSUES

Design for manufacturability is to be implemented in an industrial environment. Therefore, we are concerned with (1) increasing the individual engineer's knowledge in both the areas of design and manufacturing and (2) facilitating better and early communication between design and manufacturing groups. To increase the individual engineer's knowledge, training courses for manufacturability guidelines specific and nonspecific to the company concerned should be established. Examples, good and bad, are always helpful. Rotation of job responsibilities between design and manufacturing engineers, when feasible, is also a good way to increase an engineer's knowledge in design for manufacturability.

To facilitate better and early communication, product and process design should be managed in an integrated manner. For small companies, short product life cycle, or simple products, integrated product and process design task force may prove to be effective. For large companies, long product life cycle, or complex products, product and process engineering should be integrated within one organizational unit, or at least have a close working relationship. In large projects, computer tools may prove to be necessary. The computer tools now available are expert system software shells, CAD/CAM systems, and object-oriented programming tools, as discussed in Section 9.

In managing integrated product and process designs, there are several points worth considering:

1. Select a competent, strong project manager.
2. Quickly develop constraints for the product design and process selection at various levels by the effort of the entire team, that is, list the impossible and infeasible first.
3. Develop the product profile and specification through the team effort, remembering the purpose of a design, and list three other alternatives for every design, be it a subsystem or a component.
4. Aim high, recognizing that quality and cost need not be compromised when development time is compressed.
5. Give enough authorization to the team manager so that quick decisions can be made.

REFERENCES

Boothroyd, G., Dewhurst, P., and Knight, W. (1994), *Product Design for Manufacture and Assembly,* Marcel Dekker, New York.

Bralla, J. G. (1986), *Handbook of Product Design for Manufacturing,* McGraw-Hill, New York.

Brazier, D., and Leonard, M. (1990), "Concurrent Engineering: Participating in Better Designs," *Mechanical Engineering,* January.

Ferreira, P. M., Kochar, B., Liu, C. R., and Chandru, V. (1985), "AIFIX: An Expert System Approach to Fixture Design," *Computer Aided/Intelligent Process Planning, PED,* Vol. 19, in C. R. Liu, T. C. Chang, and R. Komanudari, Eds., ASME, New York.

Goldberg, A. (1984), SAMLLTALK-80, *The Interactive Programming Environment,* Addison-Wesley, Reading, MA.

Hayes-Roth, F., Waterman, D., and Lenat, D. (1983), *Building Expert Systems,* Addison-Wesley, Reading, MA, pp. 3–29.

Liu, C. R., and Trappey, J. C. (1989), "A Structured Design Methodology and MetaDesigner: A System Shell for Computer Aided Creative Design," *in Proceedings of ASME Design Automation Conference* (Montreal, September).

Mukherjee, A., and Liu, C. R. (1995), "Representation of Function–Form Relationship for Conceptual Design of Stamped Metal Parts," *Research in Engineering Design,* Vol. 7, pp. 253–269.

Mukherjee, A., and Liu, C. R. (1997), "Conceptual Design, Manufacturability Evaluation and Preliminary Process Planning Using Function–Form Relationships in Stamped Metal Parts," *Robotics and Computer-Integrated Manufacturing,* Vol. 13, No. 3, pp. 253–270.

Nevins, J. L. and Whitney, D. E., Eds. (1989), *Concurrent Design of Product and Processes,* McGraw-Hill, New York.

Shina, S. G. (1991), *Concurrent Engineering and Design for Manufacture of Electronic Products,* Van Nostrand Reinhold, New York.

Stillwell, H. R. (1989), *Electronic Product Design for Automated Manufacturing,* Marcel Dekker, New York.

Trappey, J. C., and Liu, C. R. (1990), "An Integrated System Shell Concept for Computer Aided Design and Planning," Technical Report TR-ERC-90-2, Purdue University, NSF Engineering Research Center, June.

ADDITIONAL READING

Doyle, L. E., *Manufacturing Processes and Materials for Engineers,* Prentice Hall, Englewood Cliffs, NJ 1969.

Edosmwan, J. A., and Ballakur, A., *Productivity and Quality Improvement in Electronic Assembly,* McGraw-Hill, New York, 1989.

Greenwood, D. C., *Product Engineering Design Manual,* McGraw-Hill, New York, 1959.

Greenwood, D. C., *Engineering Data for Product Design*, McGraw-Hill, New York, 1961.

Greenwood, D. C., *Mechanical Details for Product Design,* McGraw-Hill, New York, 1964.

Legrand, R., *Manufacturing Engineers' Manual,* McGraw-Hill, New York, 1971.

Niebel, B. W., and Baldwin, E. N., *Designing for Production,* Irwin, Homewood, IL, 1963.

Niebel, B. W., and Draper, A., *Product Design and Process Engineers,* McGraw-Hill, New York, 1974.

Niebel, B. W., Draper, A. B., and Wysk, R. A., *Modern Manufacturing process Engineering,* McGraw-Hill, New York, 1989.

Priest, J. W., *Engineering Design for Producibility and Reliability,* Marcel Dekker, New York, 1988.

Trucks, H. E., *Designing for Economical Production,* Society of Manufacturing Engineers, Dearborn, MI, 1974.

U.S. Army Material Command (AMC), *Design Guidance for Producibility,* Engineering Design Handbook, AMC, Washington, DC, 1971.

CHAPTER 51

Managing Professional Services Projects

Barry M. Mundt
The Strategy Facilitation Group

Francis J. Smith
Francis J. Smith Management Consultants

1. **PROJECT MANAGEMENT IN THE PROFESSIONAL SERVICES CONTEXT** **1333**
    1.1. Professional Services Defined 1333
    1.2. Project Management Defined 1333
    1.3. Managing Professional Services Projects 1333
2. **THE PROJECT MANAGER** **1334**
    2.1. The Project Manager's Role 1334
    2.2. Project Manager Characteristics 1334
    2.3. Identifying Project Manager Candidates 1334
3. **OVERVIEW OF THE PROJECT MANAGEMENT PROCESS** **1334**
    3.1. The Phases of Project Management 1334
    3.2. Relationship of Phases to the Delivery of Professional Services 1335
4. **PHASE I: PROJECT DEFINITION** **1335**
    4.1. Project Definition Purpose 1335
    4.2. The Project Definition Team 1335
    4.3. Project Definition Components 1335
        4.3.1. Project Objectives (Outcomes) 1336
        4.3.2. Scope 1336
        4.3.3. Deliverables (Outputs) 1336
        4.3.4. Project Approach 1336
        4.3.5. Resource and Infrastructure Requirements 1336
        4.3.6. High-Level Time/Cost Estimates 1336
        4.3.7. Project Risks 1336
5. **PHASE II: PROJECT PLANNING** **1338**
    5.1. Project Planning Purpose 1338
    5.2. The Project Planning Team 1338
    5.3. Project Planning Components 1338
        5.3.1. Confirm Objectives and Scope 1338
        5.3.2. Develop Work Breakdown Structure 1338
        5.3.3. Develop a Task and Deliverables List 1339
        5.3.4. Assign Personnel to Tasks 1339
        5.3.5. Develop Time Estimates and Preliminary Schedule of Tasks 1341
        5.3.6. Determine the Critical Path 1341
        5.3.7. Balance the Workplan 1341
    5.4. Prepare the Project Budget 1343
        5.4.1. Determine Personnel Costs 1343
        5.4.2. Add Support, Overhead, and Contingency Factors 1344
        5.4.3. Compile and Reconcile the Project Budget 1346
6. **PHASE III: PROJECT MONITORING AND CONTROL** **1346**
    6.1. Organizing for Project Implementation 1346

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

- 6.1.1. The Project Steering Committee 1346
- 6.1.2. The Project Office 1346
- 6.2. Project Monitoring 1347
  - 6.2.1. Informal Monitoring 1347
  - 6.2.2. Project Workplan and Budget Maintenance 1347
  - 6.2.3. Project Status Reporting 1347
  - 6.2.4. Status Meetings 1347
- 6.3. Project Control 1347
  - 6.3.1. Identifying Out-of-Control Conditions 1347
  - 6.3.2. Developing Corrective Actions 1348
  - 6.3.3. Following up on Corrective Action Measures 1348

7. PHASE IV: PROJECT CLOSE 1348

- 7.1. Project Performance Assessment and Feedback 1348
  - 7.1.1. Time, Cost, and Quality Performance 1348
  - 7.1.2. Lessons Learned 1348
- 7.2. Final Status Reporting 1348
- 7.3. Performance Review of Project Team Members 1349
- 7.4. Archiving Project Documentation 1349
- 7.5. Disbanding the Project Organization 1349

8. AVOIDING PROJECT-MANAGEMENT PROBLEMS 1349

- 8.1. When and Why Project Management Problems Occur 1349
- 8.2. Tips for Avoiding Problems in Project Management 1349

ADDITIONAL READING 1350

1. PROJECT MANAGEMENT IN THE PROFESSIONAL SERVICES CONTEXT

1.1. Professional Services Defined

Professional services are knowledge- and experience-based activities, performed by appropriately qualified individuals or teams, that are intended to result in predefined, desired outputs and/or outcomes. Such services can be performed for a fee by one professional enterprise for the benefit of another enterprise (the external client), such as accounting, management consulting, engineering, legal, and marketing services; or they can be performed within an enterprise for the benefit of that enterprise (the internal client), such as new product development, strategic analysis, operations improvement, and systems development/implementation (for example, an industrial engineering department in a manufacturing company or financial institution would be considered a professional services organization in this context). The delivery of a professional service typically is supported by the application of information technology and the use of appropriate data/knowledge bases.

1.2. Project Management Defined

Project management is the planning, organizing, guiding, and monitoring of organizational resources that are necessary to successfully produce one or more desired outputs or outcomes (often called deliverables). It encompasses management of project risks, issues, and changes, as well as product/deliverable configuration and quality.

A project is:

- A unique venture with a defined beginning and end
- Carried out by people to meet a specific objective or set of objectives
- Defined within the parameters of scope, schedule, cost, and quality

Project management does not include the management of ongoing tasks and/or repetitive functions.

1.3. Managing Professional Services Projects

Most of the work done by professional service organizations is performed as, and can be defined in terms of, projects. Often the desired deliverables and outcomes of a professional services project can be somewhat fuzzy, conceptual, and less tangible when compared to, say, a more concrete construction project. This causes project time and cost to be more difficult to estimate during project planning; accordingly, progress during project execution tends to be more difficult to measure.

Regardless, a project team is formed to produce a definite set of deliverables within a certain time frame for a specified cost (budget). The project team is led by a project manager, who is responsible for ensuring that the objectives of the project are achieved on time and within budget.

2. THE PROJECT MANAGER

A project manager typically is someone who has a wide breadth and depth of knowledge and experience in a number of areas. He or she also is someone who is skilled in working with teams, leverages relationships judiciously, and is knowledgeable in the use of tools and techniques that aid in accomplishing his or her role.

2.1. The Project Manager's Role

The project manager facilitates the team-building process and collaborates with the project team to create and execute the project plan. The project manager also acts as the liaison between the team and the client. He or she continually monitors the progress of the project and reports project status periodically to both the client and other interested stakeholders. The project manager works to ensure that the client is satisfied and that the project is completed within the parameters of scope, schedule, cost, and quality.

2.2. Project Manager Characteristics

A project manager's characteristics differ from those of the typical functional manager. For example, functional managers usually:

- Are specialists
- Function as technical supervisors
- Are skilled at analysis and analytical approaches
- Maintain line control over team members

Project managers, on the other hand, typically:

- Are generalists with wide experience and knowledge
- Coordinate teams of specialists from a variety of technical areas
- Have technical expertise in one or two areas
- Are skilled at synthesis and the systems approach
- Do not have line control over the project team members

2.3. Identifying Project Manager Candidates

Project managers can be persons working for the professional enterprise who are reassigned to the project manager position for the life of a given project or for a specified period of time. Project manager candidates may also come from outside the organization. For example, they may be contracted by the organization to serve as the project manager for a specific project.

Project managers may come from almost any educational background, although the industrial engineering curriculum is probably the most relevant. In any case, a successful project manager candidate should be able to demonstrate significant experience in the position. Additional qualifications might include project manager certification, which is awarded to qualified persons by organizations such as the Project Management Institute.

3. OVERVIEW OF THE PROJECT MANAGEMENT PROCESS

3.1. The Phases of Project Management

The project management process includes four phases:

- Phase I: project definition
- Phase II: project planning
- Phase III: project monitoring and control
- Phase IV: project close

Each phase has its purpose and the phases are linked in order. In fact, Phases I through III tend to be iterative. For example, some level of project planning is required to develop reasonably accurate, high-level estimates of project time and cost during project definition. Likewise, during project execution, the monitoring and control process may identify situations that will require changes in the project plan and possibly even in the project definition.

3.3. Relationship of Phases to the Delivery of Professional Services

In the professional services context, the project definition serves as the proposal to the client or the statement of work (SOW) issued by the client. Subsequent to any negotiations, the client formally accepts the proposal (or the firm formally accepts the SOW); in many cases a formal contract is executed, often incorporating the proposal or SOW. Detailed project planning then begins.

The project plan sets out the work steps, schedule, resources, and budget necessary to successfully conduct the project. The plan then becomes the basis for routinely monitoring and measuring progress during project execution and applying appropriate controls, when necessary, to make sure the project stays on track.

The project close phase seeks to determine whether the client is satisfied with the results of the work and ensures that the client understands and agrees that the project has been completed.

4. PHASE I: PROJECT DEFINITION

4.1. Project Definition Purpose

Project definition is arguably the most important phase of a project. It entails defining the objectives, scope, and deliverables of the project; selecting the most appropriate approach; developing high-level estimates of time and cost; defining the project-management process; and identifying and addressing potential problems and risks. The project-definition phase ensures that the stakeholders and project team have a common understanding of the project's purpose, objectives, and benefits. Many failed projects have been linked to inadequate development of a project definition.

A sound project definition enables the organization to begin the project in a systematic manner; it sets the tone and establishes the project's direction, opens channels of communication, forms a basis for client and stakeholder trust, and tends to bring to the surface any client concerns or challenges. The risks of not having a sound project definition include false starts, inadequate communication, confusion among the participants, and failure to meet the client's and other stakeholders' expectations.

4.2. The Project Definition Team

A core team is established to prepare the project definition (and usually the subsequent project plan). The core team may include the project manager, functional personnel, technical personnel, and possibly other professionals drawn from outside the firm.

Expectations need to be communicated clearly to the core team members, including their accountability for the project deliverables and outcomes as well as their required time commitment throughout the life of the project. The time commitment, which might be full time during the development of the project definition and the project plan, should also be explained to the core team members' supervisors (performance managers).

The specific technical expertise required of the core team members is dependent upon the nature of the project. However, at a minimum, they should have a sufficient top-level understanding of the technical issues to be able to define and plan the project effectively as well as to manage the resources that may be recruited to work on the project.

4.3. Project Definition Components

A project definition should contain at least the following components:

- Project objectives (outcomes)
- Scope
- Deliverables (outputs)
- Approach
- Resource and infrastructure requirements
- High-level time and cost estimates
- Project risks

Each of these components is described below. Examples are provided based on a typical project to identify, evaluate, and select a manufacturing business planning and control software system. Software evaluation and selection assistance is a fairly common professional service provided by consulting firms to clients (in this example, ABC Manufacturing, Inc.). This example project will be carried throughout this chapter to illustrate the various aspects of professional services project management.

4.3.1. *Project Objectives (Outcomes)*

Objectives are the destination, target, or aim of the project. They are needed to clarify the client's and other stakeholder's expectations. They also help to:

- Establish a common vision
- Guide the team during project plan development and execution
- Keep the team focused as the project progresses
- Provide a basis for communications during the project
- Establish a means for assessing success at the completion of the project

Good objectives state what will be achieved and/or the results sought, not how the team will get there. They are specific, unambiguous, and measurable, containing a time frame for the intended achievements. Examples of outcome-oriented objectives include: select a business planning and control software system in six months; implement the selected business planning and control software system in 18 months; increase product throughput by 50% in two years; decrease inventory investment by 75% by year end.

4.3.2. *Scope*

The statement of scope sets the boundaries of the project, in that it defines the confines, the reach, and/or the extent of the areas to be covered. It clarifies what will be included in the project and, if necessary, states specifically what will *not* be included.

Scope may be defined in terms such as geographical coverage, range of deliverables, quality level, and time period. The statement of scope must be clear, concise, and complete, as it will serve as the basis for determining if and when out-of-scope work is being conducted during project execution. In the professional services field, performance of unauthorized, out-of-scope work on a project usually will result in a budget overrun, unrecovered fees and expenses from the client, and unsatisfactory project financial results. Potential out-of-scope work should be identified *before* it is performed and negotiated as additional work, along with its attendant cost and schedule requirements.

An example of a scope statement is: "Assist ABC Company in selecting an appropriate business planning and control software and hardware system and implementing the selected system. The assistance will include defining business system requirements, evaluating system alternatives, making a selection that will support manufacturing and accounting functions, and facilitating the implementation of the selected system."

4.3.3. *Deliverables (Outputs)*

A deliverable is anything produced on a project that supports achievement of the project objectives. It is any measurable, tangible, verifiable outcome, result, or item that is produced to complete a task, activity, or project. The term is sometimes used in a more narrow context when it refers to an external deliverable (handed over to a stakeholder or client and subject to approval).

Examples of deliverables are system requirements definition document; request for proposal (RFP) document; systems-evaluation criteria; software and hardware configuration design; system-implementation project plan; and facilitation assistance during the system-implementation process.

4.3.4. *Project Approach*

The project approach defines the general course of action that will be taken to accomplish the project objectives. For example, the project approach may be defined in such terms as the methodology to be used, the timing/phases of the project, and/or the types of technology and human resources to be applied. The approach section of the project definition explains, in general, how the project will be carried out.

An example of an approach statement for developing a system requirements definition document is: "We will conduct interviews with personnel in each functional area to develop and define the system requirements, based on a system requirements profile. We will provide advice in the definition of critical system requirements, such as system performance (timing, volumes, and the like). This phase will culminate with the development of a system requirements definition document."

4.3.5. *Resource and Infrastructure Requirements*

Resource and infrastructure requirements for professional service projects typically fall into any of three categories: human resources, facilities and equipment, and information technology (including knowledge bases). Human resource requirements, often the major cost of a project, should be defined in terms of the roles, responsibilities, and skills that are needed for the project to be successful. The roles and responsibilities then are translated into a depiction and/or description of the planned organization structure for the project.

Arguably the most important role in the project is that of the project sponsor, the person who authorizes or ‘‘legitimizes’’ the project (often referred to simply as ‘‘the client’’). If the sponsor is not committed to the project, the chances of successful completion are reduced. Sponsorship often can be extended by the formation of a project steering committee, which, if constructed properly, can be particularly useful in clearing barriers that are likely to come up during project execution.

An example of a resource and infrastructure requirements statement might be: ‘‘We will require designation of a project team leader to assist in scheduling interviews; arranging for the collection of information, reports, and documentation; and assisting in the coordination of administrative support. Knowledge of existing systems is also important to provide valuable background information. In addition, office space with a telephone and computer access to the internet is required.’’

4.3.6. High-Level Time/Cost Estimates

The purpose of high-level time/cost estimates is to gauge and validate project size. A high-level estimate sets out resource and staffing levels by project phase or activity and elapsed time by activity.

High-level time/cost estimates are top-down estimates made while developing the project definition. A high-level estimate can be developed by reviewing and drawing upon estimates from previous similar projects and any estimates developed in project definition work sessions. Assumptions, initial estimates, and associated calculations should be documented. During project planning (Phase II), detail calculations are summed and compared to the high-level figures as a way of validating the estimates.

An example of a high-level time/cost estimate is: ‘‘It is estimated that it will take 14 weeks to reach the point at which software has been selected. Once the hardware and software are installed, it will take approximately 10 months to implement the system. Based on the scope of work and the approach, it is estimated that the project will cost $350,000 to $375,000.’’

4.3.7. Project Risks

A project risk is any factor that challenges the project team’s ability to manage the budget, resources, time, and quality of the project deliverables and acceptance of the deliverables. Risks would include any factors that could disrupt the project. They are uncertainties or vulnerabilities that could cause the team to deviate from the project plan.

Risks may be managed through a risk management plan. Establishment of such a plan entails identifying the risks, determining the potential impacts, defining preventive actions, estimating the costs (both monetary and nonmonetary) required to reduce the risks to acceptable levels, developing contingency plans, and obtaining management’s commitment to the risk management plan.

Risk management is valuable in that it minimizes the unfavorable impact of unplanned incidents on a project. It enhances the probability of successful project implementation, creates a sense of urgency when unplanned incidents occur, and facilitates their timely and effective resolution.

Risk management typically involves assessing a number of dimensions of each identified risk. These dimensions include:

- The impact of the risk if it were to occur
- The likelihood that the risk will occur
- How difficult it would be to detect the risk

Each of these dimensions can be assessed separately for a specific area of risk and assigned a numerical low-to-high rating. The three dimension ratings then can be combined (e.g., averaged or added) to produce a relative risk value for the area, which can be compared with the ratings of other identified project risk areas to determine where emphasis needs to be placed in the risk management plan.

One way of judging overall project risk is through a constraint matrix, where three levels of flexibility (low, medium, high) are charted against the elements of schedule, scope, and cost. The purpose of a constraint matrix is to assess the degree of relative flexibility within a given project and the related risk. For example, a project profile that exhibits low flexibility in all three elements (schedule, scope, and cost) is a profile associated with high-risk projects and one that will demand extra careful management.

An example of a project risk management statement is: ‘‘A mechanism is provided for identifying and assessing any adverse consequences of a tentative system selection so their effects can be controlled during implementation. A structured approach will be used for developing relative priorities of individual requirements. Shortfalls in the functionality of a tentative selection will be evaluated. Alternative solutions to providing the functionality will be considered and the impact on the effectiveness of the overall system will be evaluated prior to making a final selection. Additionally, arrangements for holding the source code in escrow will ensure the availability of the source code in the event the software company cannot sustain ongoing viability.''

5. PHASE II: PROJECT PLANNING

5.1. Project Planning Purpose

The purpose of project planning is to confirm the project scope and objectives; develop the project organization, schedule, and budget; secure the necessary resources; and create clear expectations about the project organization, timing, budget, and resources.

The project workplan should clearly identify the deliverables that will be prepared and the tasks that need to be performed in order to prepare them. The project planning team uses the project definition as the beginning point for preparing the project workplan. The workplan is typically broken down into phases, activities, and tasks. Deliverables and resources are usually defined at the task level.

5.2. The Project Planning Team

A key member of the project planning team is the project manager because he or she will have primary responsibility for executing the project plan. The project planning team may also include one or more members of the project definition team to ensure that the thinking that went into defining the project is reflected in the project plan. If the project definition team is not represented on the project planning team, a draft of the project plan should be reviewed by one or more project definition team members.

Other members of the project planning team might include appropriate technical specialists and others who may be team members during project execution.

5.3. Project Planning Components

There are seven main steps in creating a project workplan:

1. Confirm objectives and scope.
2. Develop work breakdown structure.
3. Develop a detail task list.
4. Assign personnel to tasks.
5. Develop time estimates and a preliminary schedule of tasks.
6. Determine the critical path.
7. Balance the detailed workplan.

Each of these steps is described below.

5.3.1. Confirm Objectives and Scope

Often there can be a significant time period between the completion of project definition and the initiation of detailed project planning, which may result in divergent views as to the purpose of the project. It is important that there be full agreement regarding the objectives and scope of the project before the workplan is prepared. The project manager should seek confirmation of the objectives and scope based on input from the sponsor and/or the steering committee as well as the project-definition team members. If there are differences, the project manager should rely on the sponsor to settle them.

5.3.2. Develop Work Breakdown Structure

Developing a work breakdown structure entails expanding the project phases or deliverables into the major activities that need to occur to complete each phase and defining the tasks that need to occur to complete each activity.

Steps for developing a work breakdown structure and examples are presented in Table 1.

The work breakdown structure can encompass more than the three levels shown in Table 1, depending on the nature and complexity of the work to be done. For example, if the work effort has been done many times previously and/or is routine, it may not require more than three levels of detail (phase/activity/task). Conversely, work that is less familiar or more complex may require additional levels of detail to gain a full understanding of the work that must be done.

A work statement (often called a work package) then is prepared to describe the effort for each task or subtask at the lowest level of the work breakdown structure. Each work statement should be designed to ensure that the related task or subtask:

TABLE 1 Developing a Work Breakdown Structure

Steps in Developing a Work Breakdown Structure	Examples
1. Draw out the phases of the project or organize it by major project deliverables.	Phases: • Analysis and implementation preparation • Coordination of required disciplines and controls • Training of personnel • Implementation tailoring • Online implementation • New system break-in
2. Detail the major activities that need to occur to complete each phase.	Activities within the ‘‘Analysis’’ phase: • Define system requirements. • Develop a request for proposal document. • Develop evaluation criteria and evaluate alternatives. • Select a software and hardware configuration. • Develop an implementation project plan.
3. Detail all the tasks that will need to occur to complete each activity.	Tasks within the ‘‘Define systems requirements’’ activity: • Define and document business system objectives. • Document performance objectives. • Define and document anticipated benefits. • Document functional requirements. • Conduct interviews. • Document special considerations and constraints. • Assemble requirements documentation.

- Is measurable
- Has tangible results/outputs (deliverables)
- Has identifiable and readily available inputs
- Is a finite, manageable unit of work
- Requires a limited number of resources
- Fits into the natural order of work progression

A completed work breakdown structure will include the assembled detail tasks and their relationship to respective activities. A work breakdown structure may be displayed according to Figure 1 (in the figure, level 1 corresponds to ‘‘phase,’’ level 2 to ‘‘activity,’’ and level 3 to ‘‘task’’).

5.3.3. Develop a Task and Deliverables List

The different levels of the work breakdown structure should be documented in a task list that identifies each phase, activity, and task (and subtask, as appropriate). Next, the name of the person to be responsible for each task (the task ‘‘owner’’) and a description of the deliverable(s) associated with each task can be added. An example of a detailed task and deliverables list is shown in Figure 2.

When the task and deliverables list is complete, the logical order in which tasks should be performed is defined. This is done by first determining task dependencies at the lowest level of the work breakdown structure. These dependency relationships can be portrayed in the form of a project network diagram. An example of task dependencies is shown in Figure 3.

5.3.4. Assign Personnel to Tasks

Each task must be assigned personnel resources to perform the work. The steps for assigning personnel to tasks include:

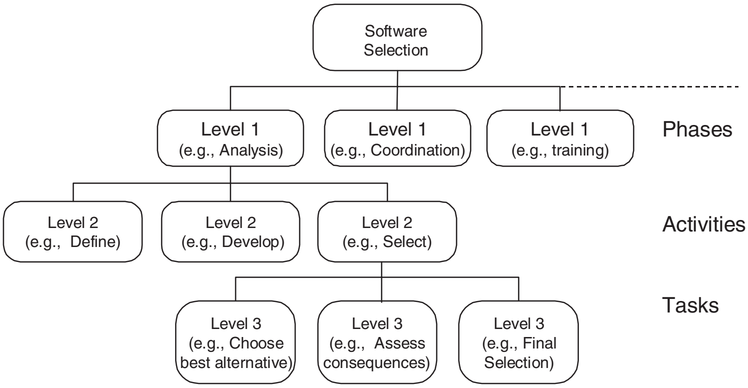

Software Selection
Level 1 (e.g., Analysis)
Level 1 (e.g., Coordination)
Level 1 (e.g., training)
Phases
Level 2 (e.g., Define)
Level 2 (e.g., Develop)
Level 2 (e.g., Select)
Activities
Tasks
Level 3 (e.g., Choose best alternative)
Level 3 (e.g., Assess consequences)
Level 3 (e.g., Final Selection)

Figure 1 Partial Work Breakdown Structure.

- List the skills required to complete each task.
- Identify candidate team members whose skills meet the task requirements.
- Develop a rough estimate of the time that will be required to complete each task, based on the experience of the candidates.
- Negotiate roles and responsibilities of candidate team members relative to each task.

Project Name: Software Selection and Implementation **Project Manager:** Joan Ryan

Date Prepared: 5/31/2000

Table of Detail Tasks and Deliverables Description List

ID	Detail Tasks List	Task Owner	Deliverable
2.0	**Define system requirements**		
2.1	Define and document business system objectives	Jim B.	Business system objectives documentation
2.2	Document performance objectives	Mary P.	Performance objectives documentation
2.3	Define and document anticipated benefits	Joan R.	Anticipated benefits documentation
3.0	**Develop Evaluation Criteria**		
3.1	Identify alternative systems or supplements	Bob S.	List of software solutions
3.2	Prioritize requirements according to importance	Guy R.	Priorities assigned
3.3	Evaluate each alternative against the absolute requirements	Marie S.	Go/no go decision
3.4	Calculate scores for each alternative	Bob S.	Score sheet
3.5	Assess each alternative's functions against requirements	Henry B.	Evaluation sheet
4.0	**Select an Alternative**		
4.1	Choose the best alternative as a tentative decision	Guy R.	Tentative selection
4.2	Assess adverse consequences and decide if an alternative should be selected	Marie S.	Adverse consequences list
4.3	Make final selection	Team	System selected
5.0	**Develop Implementation Project Plan**		
5.1	Meet with project team members	Team	Committee assignments
5.2	Develop detail tasks to be performed within each activity	Wendy L.	Detail tasks

Figure 2 Detailed Task and Deliverables List.

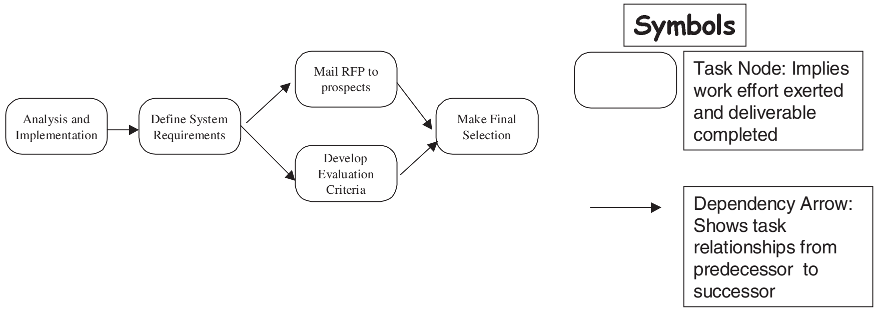

Symbols
Analysis and Implementation
Define System Requirements
Mail RFP to prospects
Develop Evaluation Criteria
Make Final Selection
Task Node: Implies work effort exerted and deliverable completed
Dependency Arrow: Shows task relationships from predecessor to successor

Figure 3 Task Dependencies.

- Gain commitment from the departments, performance managers, and candidates involved, particularly regarding the team members' availability to participate.
- Document the team members' project roles and responsibilities.

The project task assignments can be documented by extending the task and deliverables list (see Figure 2) to include an additional column entitled "Task Team Members."

5.3.5. Develop Time Estimates and Preliminary Schedule of Tasks

The next step is to develop time estimates for each task. The time estimates take two forms:

1. *Effort:* the actual work time required to complete each task, typically in terms of hours
2. *Duration:* the elapsed time between the start and finish of a task, typically in terms of days or weeks.

The effort estimates are used to determine the amount of time each team member is expected to spend on the project, as well as the related cost of their services. The task-duration estimates, when combined with the project network diagram (see Figure 4), provide the basis for setting a project schedule (i.e., task start and completion dates).

The effort and duration estimates should be developed from the bottom up, preferably by the persons responsible for their execution, in contrast to the top-down, high-level estimates prepared during the project definition phase. They should take into account past experiences with similar, standard processes, to the extent possible. As predictions, the estimates should be equally likely to be above or below the actual results rather than represent the minimum or maximum time. Figure 4 provides an example of task time estimates.

Size comparability of tasks can be achieved by applying the following rule of thumb: combine a task requiring less than seven hours of effort with another task; subdivide tasks that require more than 70 hours.

5.3.6. Determine the Critical Path

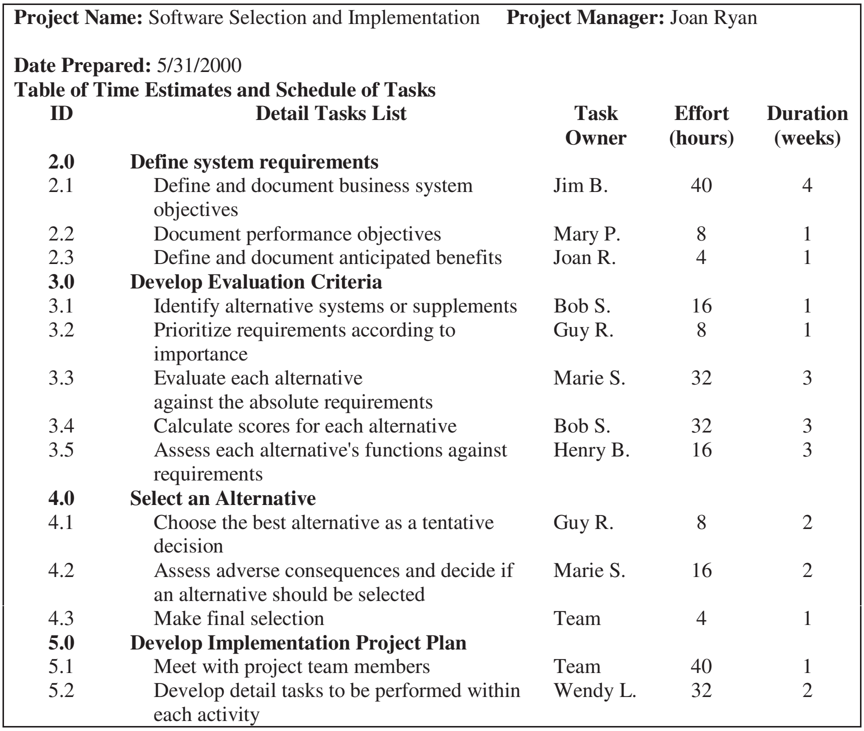
The critical path is the path in the project network diagram that consumes the longest elapsed time. It is the path in which there is no extra time available to accommodate delays. This is in contrast to other paths, where float exists and can be used to accommodate delays. A delay in the critical path will result in a delay of the entire project.

The critical path is identified for two reasons. First, if unexpected issues or changes occur to a given task after the project begins, the impact, if any, on the overall project schedule can be determined quickly. Second, knowing the critical path enables the project manager and team to consider quickly where the schedule can be compressed to accommodate project imperatives and unexpected changes.

Table 2 shows some schedule-compression options that can be considered, along with the associated caveats.

5.3.7. Balance the Workplan

Resource loading follows the critical path analysis. Resource loading specifies the resources that will be needed for each planning period (typically one week) on the project timeline. The benefit of resource loading is that it identifies conflicts in the schedule (i.e., resources assigned to different



Project Name: Software Selection and Implementation **Project Manager:** Joan Ryan

Date Prepared: 5/31/2000

Table of Time Estimates and Schedule of Tasks

ID	Detail Tasks List	Task Owner	Effort (hours)	Duration (weeks)
2.0	**Define system requirements**			
2.1	Define and document business system objectives	Jim B.	40	4
2.2	Document performance objectives	Mary P.	8	1
2.3	Define and document anticipated benefits	Joan R.	4	1
3.0	**Develop Evaluation Criteria**			
3.1	Identify alternative systems or supplements	Bob S.	16	1
3.2	Prioritize requirements according to importance	Guy R.	8	1
3.3	Evaluate each alternative against the absolute requirements	Marie S.	32	3
3.4	Calculate scores for each alternative	Bob S.	32	3
3.5	Assess each alternative's functions against requirements	Henry B.	16	3
4.0	**Select an Alternative**			
4.1	Choose the best alternative as a tentative decision	Guy R.	8	2
4.2	Assess adverse consequences and decide if an alternative should be selected	Marie S.	16	2
4.3	Make final selection	Team	4	1
5.0	**Develop Implementation Project Plan**			
5.1	Meet with project team members	Team	40	1
5.2	Develop detail tasks to be performed within each activity	Wendy L.	32	2

Figure 4 Task Time Estimates.

tasks simultaneously). It may reveal resources that are overcommitted or underutilized. It helps the project manager to determine whether tasks need to be rescheduled, work reprioritized, or additional time or resources renegotiated.

The workplan is balanced when all appropriate resources are confirmed and an acceptable completion date is determined. The preliminary schedule, resource availability, and required project-completion date all need to be brought into balance.

TABLE 2 Schedule-Compression Options

Compression Option	Caveat
Overlap tasks by using partial dependencies.	• If resource loading indicates there are enough resources available • If the interim deliverable is sufficiently complete
Break dependencies and resequence tasks.	• If the associated risk is acceptable
Break tasks into subtasks that can be done in parallel.	• If resource loading indicates there are enough resources available
Reallocate resources from paths with float to the critical path.	• If the task is resource driven • If the resources have the correct skills and available time • If the noncritical path(s) have not become critical.
Authorize overtime, add shifts, increase staffing, subcontract jobs.	• If there is approval for the additional budget expense
Remove obstacles.	• If priority is high enough
Reduce project scope.	• If the project sponsor approves

5.4. Prepare the Project Budget

The primary purpose of preparing a project budget is to estimate the total cost of accomplishing the project. If the budget amount is not acceptable to the project sponsor (the client), then the project workplan will need to be reworked or the project redefined until an acceptable figure is achieved. When cast in time intervals, such as biweekly or monthly, the budget serves as one basis for project monitoring during execution of the workplan.

There are three steps to preparing a project budget:

1. Determine personnel costs.
2. Add support, overhead, and contingency factors.
3. Compile and reconcile the project budget.

This process is not necessarily linear, and some tasks related to these steps may need to be performed as part of the development of high-level cost estimates during the project-definition phase (see Section 4).

5.4.1. Determine Personnel Costs

Project personnel costs are determined by totaling the hours required by each person to perform his or her work on all the tasks to which he or she has been assigned. The person's total project hours then are multiplied by their hourly billing or compensation rate, as determined by management, to

Project Name: Software Selection and Implementation **Project Manager:** Joan Ryan

Date Prepared: 5/31/2000

Table of Personnel Costs

ID	**Detail Tasks List**		**Hours**	
		Joan R. ($365/hr)	**Mary P. ($215/hr)**	**Bob S. ($150/hr)**
2.0	**Define system requirements**			
2.1	Define and document business system objectives	4	16	40
2.2	Document performance objectives	1	8	8
2.3	Define and document anticipated benefits	1	4	4
3.0	**Develop Evaluation Criteria**			
3.1	Identify alternative systems or supplements	1	8	16
3.2	Prioritize requirements according to importance	1	4	8
3.3	Evaluate each alternative against the absolute requirements	4	16	32
3.4	Calculate scores for each alternative	1	16	32
3.5	Assess each alternative's functions against requirements	4	8	16
4.0	**Select an Alternative**			
4.1	Choose the best alternative as a tentative decision	4	8	8
4.2	Assess adverse consequences and decide if an alternative should be selected	1	16	16
4.3	Make final selection	4	4	4
5.0	**Develop Implementation Project Plan**			
5.1	Meet with project team members	4	40	40
5.2	Develop detail tasks to be performed within each activity	1	32	32
	Totals	31	180	256
	Personnel Budget	**$11,315.00**	**$38,700.00**	**$38,400.00**
	Total Personnel Budget		**$88,415.00**	

Figure 5 Personnel Costs.

calculate his or her direct project cost. The costs for all personnel assigned to the project then are totaled to determine the project's personnel budget. Figure 5 provides an example of how personnel costs might be determined.

5.4.2. Add Support, Overhead, and Contingency Factors

Support tasks and overhead should be included in the detail workplan to account for their impact on project cost and duration. Support refers to all those tasks that facilitate production of the deliverables through better communication, performance, or management. It could be project-related training, meetings, administration, project and team management, report production, and quality assurance reviews.

Overhead is nonproductive time spent on tasks that do not support execution of the project workplan or production of the deliverables but can have considerable impact on the project schedule, the resource loading, and potentially the budget. Overhead could include travel time, holidays, vacation, professional development, or personal/sick time.

Nonpersonnel costs associated with the project are identified and estimated. Such costs may include travel expense, technology/knowledge acquisition, and contractor assistance.

Finally, contingency factors are considered to compensate for project risks and other potential project issues as well as to accommodate personnel learning curves. Contingency factors may be applied at the phase or activity level of the workplan/budget, although accuracy may be improved if applied at the detail task level.

Figure 6 extends the example from Figure 5 with nonpersonnel costs to arrive at a total project cost.

Project Name: Software Selection and Implementation **Project Manager:** Joan Ryan

Date Prepared: 5/31/2000

Table of Personnel and Other Costs

			Hours	
ID	**Detail Tasks List**	**Joan R. ($365/hr)**	**Mary P. ($215/hr)**	**Bob S. ($150/hr)**
2.0	**Define system requirements**			
2.1	Define and document business system objectives	4	16	40
2.2	Document performance objectives	1	8	8
2.3	Define and document anticipated benefits	1	4	4
3.0	**Develop Evaluation Criteria**			
3.1	Identify alternative systems or supplements	1	8	16
3.2	Prioritize requirements according to importance	1	4	8
3.3	Evaluate each alternative against the absolute requirements	4	16	32
3.4	Calculate scores for each alternative	1	16	32
3.5	Assess each alternative's functions against requirements	4	8	16
4.0	**Select an Alternative**			
4.1	Choose the best alternative as a tentative decision	4	8	8
4.2	Assess adverse consequences and decide if an alternative should be selected	1	16	16
4.3	Make final selection	4	4	4
5.0	**Develop Implementation Project Plan**			
5.1	Meet with project team members	4	40	40
5.2	Develop detail tasks to be performed within each activity	1	32	32
	Totals	31	180	256
	Personnel Budget	**$11,315.00**	**$38,700.00**	**$38,400.00**
	Total Personnel Budget		**$88,415.00**	
	Administration Support		**15,000.00**	
	Overhead		**30,000.00**	
	Contractor		**12,000.00**	
	subtotal		**$145,415.00**	
	Contingency @ 10%		**14,540.00**	
	Total Phase I Project		**$159,955.00**	

Figure 6 Total Project Cost.

Project Name: Software Selection and Implementation **Project Manager:** Joan Ryan

Time-phased Budget

ID	Detail Task List	Feb	Mar	Apr	May	Total
2.0	**Define system requirements**					
2.1	Define and document business system objectives	$10,900				$10,900
2.2	Document performance objectives		$3,285			$3,285
2.3	Define and document anticipated benefits		$1,825			$1,825
3.0	**Develop Evaluation Criteria**					$0
3.1	Identify alternative systems or supplements	$4,485				$4,485
3.2	Prioritize requirements according to importance		$2,425			$2,425
3.3	Evaluate each alternative against the absolute requirements		$9,700			$9,700
3.4	Calculate scores for each alternative			$8,605		$8,605
3.5	Assess each alternative's functions against requirements			$5,580		$5,580
4.0	**Select an Alternative**					$0
4.1	Choose the best alternative as a tentative decision				$4,380	$4,380
4.2	Assess adverse consequences and decide if an alternative should be selected				$6,205	$6,205
4.3	Make final selection				$2,920	$2,920
5.0	**Develop Implementation Project Plan**					$0
5.1	Meet with project team members				$16,060	$16,060
5.2	Develop detail tasks to be performed within each activity				$12,045	$12,045
	Totals	$15,385	$17,235	$14,185	$41,610	$88,415

Figure 7 Time-Phased Budget.

5.4.3. Compile and Reconcile the Project Budget

The budget is compiled by adding personnel costs and all other costs (including contingencies) to arrive at a total budget number. The budget is subdivided into time increments (weekly, biweekly, or monthly) for the expected life of the project, based on the expected allocation of resources and related costs to the time periods in the project schedule. An example of a time-phased budget is shown in Figure 7.

Because the budget is a projection of project costs, it is based on many assumptions. The compiled budget should be accompanied by a statement of the assumptions made regarding schedules, resource availability, overhead, contingency factors, nonpersonnel costs, and the like.

If the project budget is materially different from the high-level cost estimate in the project definition, a reconciliation process may need to occur. This may result in the need to rework/rebalance the project definition, the workplan, and the budget before the sponsor will approve execution of the project.

6. PHASE III: PROJECT MONITORING AND CONTROL

The project workplan and budget (see Section 5) are used as the basis for project monitoring and control. The three main purposes of project monitoring and control are to:

1. Manage the project within the constraints of budget, time, and resources
2. Manage changes that will occur
3. Manage communications and expectations

Project monitoring helps the project manager balance constraints, anticipate/identify changes, and understand expectations. A well-designed project monitoring process provides:

- Timely information regarding actual vs. planned results
- An early warning of potential project problems
- A basis for assessing the impact of changes
- An objective basis for project decision making

Project monitoring also is used to set up ongoing channels of communication among project stakeholders. The major deliverables are project progress reports and status updates; detailed workplans (updated as necessary); and cost and schedule performance reports.

6.1. Organizing for Project Implementation

An important element of project monitoring and control is the organization that is put in place to support it. Typically the project-implementation structure consists of at least two entities: the project steering committee and the project office.

6.1.1. The Project Steering Committee

The project steering committee is made up of the key stakeholders in the project. It usually is chaired by the project sponsor, and the membership is made up of individuals who are in a position to help move the project along when barriers are encountered or changes are necessary. The committee members typically are project supporters, although antagonists may also be included to ensure that their views are heard and, to the extent possible, accommodated. The project manager is often a member of the committee.

The steering committee has a number of roles. It provides direction to the project; reviews deliverables, as appropriate; receives periodic reports regarding project progress, status, difficulties, and near-future activities; helps clear roadblocks as they occur; and provides final approval that the project has been completed satisfactorily.

6.1.2. The Project Office

The project office is led by the project manager. Depending on the size and complexity of the project, the office may be staffed by appropriate technical and administrative personnel to provide assistance to the project manager.

The primary role of the project office is to ensure that the project is proceeding according to plan and that the deliverables are of acceptable quality. This is accomplished by periodic (e.g., weekly or biweekly) review of project progress with respect to plan, as well as review of deliverables as they are produced. The project office maintains and updates the master project workplan and budget and routinely reports progress and difficulties to interested stakeholders, including the steering committee.

The office also takes the lead in ensuring that any actions necessary to correct project problems are effected in a timely manner.

6.2. Project Monitoring

Project monitoring takes a number of forms, including:

- Informal monitoring
- Project workplan and budget maintenance
- Project status reporting
- Status meetings

6.2.1. Informal Monitoring

Informal project monitoring entails ‘‘walking the project’’ on a periodic basis, daily if possible. It may involve observing deliverables; holding ad hoc meetings with team members; and communicating informally and frequently with stakeholders. Much can be learned about how the project is doing simply by talking with project team members and other stakeholders.

6.2.2. Project Workplan and Budget Maintenance

Maintenance of workplans and budgets is a routine and ongoing activity. Project plans should be updated on a regular basis to reflect corrective actions and proactive strategies being implemented. Plan maintenance involves updating the detailed workplan’s latest estimate to reflect current status *and* the time/cost necessary to complete the project. Plan maintenance should occur at least biweekly and should not alter the baseline workplan and budget—unless variances have become large and persistent or the scope of the project has changed. If rebaselining is necessary, it should only be done with sponsor/steering committee approval and may require approval by the person who approved the original project workplan and budget if he or she is other than the sponsor/steering committee.

6.2.3. Project Status Reporting

Status reports provide project leaders and other interested parties with an objective picture of progress, variances, and trends, as well as an audit trail and record of project progress. These reports provide leaders with an opportunity to understand and rectify variances and formulate appropriate actions to identified strengths, vulnerabilities, and risks.

Status reports may be assembled in a variety of configurations, depending on the audience being served. Typical configurations of status report packages include project team leader reports, project manager reports, and steering committee reports.

6.2.4. Status Meetings

Status report packages typically are delivered at status meetings. Team leader report packages are delivered on a regular basis to the project manager to cover completed, current, and planned work. The team leader reports are consolidated by the project office and the overall project status report is presented by the project manager to the steering committee at its regular meeting. The report to the steering committee focuses on overall project status and specific issues or roadblocks that need to be cleared. A key role of the steering committee is to take the lead in resolving issues and clearing roadblocks so the project can proceed as planned.

6.3. Project Control

Project control involves three main activities:

1. Identifying out-of-control conditions
2. Developing corrective actions
3. Following up on corrective action measures

6.3.1. Identifying Out-of-Control Conditions

An activity or task is in danger of going out of control when its schedule or budget is about to be exceeded but the deliverable(s) are not complete. Adequate monitoring of project schedule and budget will provide an early warning that a potential problem exists. An activity or task that is behind schedule and is on the critical path requires immediate attention because it will impact the overall timetable for project completion, which ultimately will adversely impact the budget for the task, the activity, the phase, and the project. Oftentimes, exceeding the budget or missing the scheduled completion date for a particular task or activity may be an early warning sign that a significant problem is developing for the project. Either of these signs requires an immediate investigation on the part of the project manager to determine the underlying reasons.

6.3.2. Developing Corrective Actions

Once the project manager determines the cause of an overage in the budget or a slippage in the schedule, he or she must determine an appropriate response and develop a specific corrective action plan. Sometimes a problem is caused by an impediment that the project manager alone can not resolve. In these instances, the project manager should engage the help of the steering committee. One of the responsibilities of the steering committee is to provide ''air cover'' for a project manager when he or she encounters complex difficulties.

In other cases, the impediment may have been anticipated and a corrective action plan formulated as part of the project's risk-management plan. In these cases, all that may be required is to execute the corrective action specified in the plan.

6.3.3. Following up on Corrective Action Measures

To ensure that the desired outcome of the corrective action is being achieved, it is important to employ project monitoring techniques when executing a corrective action plan. The walking-the-project technique mentioned earlier is an example of an effective follow-up technique. More complex corrective actions may require a more formal status-reporting approach.

7. PHASE IV: PROJECT CLOSE

Successful completion of all the deliverables set out in the workplan does not, by itself, conclude the project. Several activities need to be accomplished before the project can be brought to a formal close, including:

- Project performance assessment and feedback
- Final status reporting
- Performance review of project team members
- Project documentation archiving
- Disbanding of the project organization

The primary purpose of the project close phase is to ensure that the expectations set throughout the project have been met.

7.1. Project Performance Assessment and Feedback

Project performance should be assessed in a structured manner, addressing the extent to which the objectives set out in the project definition and workplan have been achieved. Obtaining objectivity requires that the client's views be considered in the assessment.

7.1.1 Time, Cost, and Quality Performance

Time, cost, and quality performance are three key project parameters that should be subject to assessment. Time performance can be assessed by comparing the originally planned completion dates of deliverables, both interim and final, with the actual completion dates. Causes of any material schedule slippage should be determined and means for precluding them in future projects developed. A similar assessment of project cost can be conducted by comparing budgeted to actual expenditures and then examining any material favorable and unfavorable variances.

Quality-performance assessment, however, tends to be less quantitative than time and cost assessment. It usually relies on solicited or unsolicited input from the client and other stakeholders regarding how they view the project deliverables (e.g., did they receive what they expected?). Quality performance can also relate to how well the project team has communicated with the stakeholders and perceptions of how well the project has been managed and conducted.

7.1.2. Lessons Learned

The opportunity to identify and capture lessons learned from having done a particular project should not be missed. Lessons learned should be documented and any best practice examples relating to the project captured. Documentation of lessons learned and best practice examples should be made available to others in the organization who will be involved in future project design and execution efforts.

7.2. Final Status Reporting

A final project status report should be prepared and issued to at least the project steering committee, including the sponsor. The report does not need to be extensive, but should include:

- A statement of the project objectives and deliverables
- A recap of the approach and key elements of the workplan
- A brief discussion of any open matters that need to be addressed
- A statement that the project has (or has not) achieved its objectives
- A list of suggested next steps, if any
- Acknowledgment of any special contributions by personnel involved

Typically, the final status report is prepared by the project manager and presented at the final steering committee meeting.

7.3. Performance Review of Project Team Members

Individual performance reviews of team members should be conducted on a one-to-one basis by the project manager. Project staff should be evaluated against defined roles and expectations, with the focus on strengths and areas for improvement. Individual contributions to the project should be recognized. The information should be transmitted to each team member's performance manager in order to update developmental needs and provide helpful information for annual performance reviews. Likewise, subcontractors and vendors also should be provided with feedback on their performance.

7.4. Archiving Project Documentation

Archiving entails compilation and organization of appropriate project documentation and submitting it for filing in a designated repository. The project file should include at least the project definition, the project workplan/budget, copies of key deliverables, the final project status report, and the results of the project performance assessment.

7.5. Disbanding the Project Organization

Disbanding the project organization includes notifying the appropriate offices of the future availability of the participants who had been assigned to the project; returning the space and equipment to the issuing offices; and establishing a mechanism for following up and maintaining any deliverables, if necessary.

A project team close-out meeting should be held to reflect on the team members' interaction and identify areas for improvement in working with an extended project team. It also provides an opportunity to identify and discuss any areas of potential improvement in the project management process.

Consideration should also be given to celebrating the success of the project with the project team and extended team members. This would be a positive way to mark the end of the project, celebrate its success, and acknowledge the professional ties developed throughout the course of the work.

8. AVOIDING PROJECT MANAGEMENT PROBLEMS

8.1. When and Why Project Management Problems Occur

Project management problems typically don't manifest themselves until the project is well underway. But the basis for most problems is established early on, during the project definition and planning phases. In particular, unclear and poorly communicated statements of project objectives, scope, and deliverables will almost always result in a project workplan that, when implemented, does not fulfill the expectations of the client and other stakeholders. In the absence of clear project definition, the stakeholders will set their own expectations, which usually won't match those of the project manager and his or her team members.

During project implementation, regular and clear communication between the project participants and the project manager, as well as between the project manager and the sponsor/steering committee, will help raise issues to the surface before they become time, cost, or quality problems.

8.2. Tips for Avoiding Problems in Project Management

The following are some suggestions for avoiding problems on professional services projects:

- Invest time up front in carefully defining and communicating the project objectives, scope, and deliverables. This will save time and reduce frustration in later stages of the project.
- Know the subject matter of the project and stay within your professional skills. Seek help before you need it.
- Avoid overselling and overcommitting in the project definition. Include risk-reducing language when necessary and appropriate.
- Always be clear on project agreements and other matters. Do what you say you will do and stay within the parameters of the project definition.

- Be flexible. A project workplan documents the expected route and allows you to communicate the expected deliverables and path to the stakeholders. Make adjustments as necessary and use the plan and related expectations as the basis for explaining how the changes will affect the project.
- Effective project management is as critical as the tasks and milestones in the project. It keeps the team's efforts focused and aligned. Budget time for the project manager to perform the necessary project management tasks, including, but not limited to, frequent quality, schedule, and budget reviews.
- In large projects, consider the use of a project administrator to take responsibility for some of the crucial but time-consuming administrative tasks.
- Communicate informally and frequently with the client. Identify and communicate risks and problems as early as possible.
- Make sure the project is defined as closed at an appropriate time, typically when the objectives have been met, to avoid having the project drift on in an attempt to achieve perfection.

ADDITIONAL READING

Cleland, D. I., *Project Management: Strategic Design and Implementation,* 3rd Ed., McGraw-Hill, New York, 1999.

Duncan, W. R., *A Guide to the Project Management Body of Knowledge,* Project Management Institute, Newtown Square, PA, 1996.

KPMG Global Consulting, *Engagement Conduct Guide, Version 1.0,* KPMG International, Amstelveen, The Netherlands, 1999.

KPMG U.S., *Project Management Reference Manual,* KPMG U.S. Executive Office, Montvale, NJ, 1993.

Maister, D. H., *Managing the Professional Service Firm*, Free Press, New York, 1993.

IV.C

Manpower Resource Planning

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 52

Methods Engineering*

STEPHAN KONZ
Kansas State University

1. **OVERVIEW** **1353**
2. **GUIDELINES** **1353**
   - 2.1. Criteria of Job Design 1354
   - 2.2. Organization of Groups of Workstations 1354
   - 2.3. Individual Workstations 1357
   - 2.4. Musculoskeletal Disorders 1362
     - 2.4.1. Risk Factors 1362
     - 2.4.2. Solutions 1362
   - 2.5. Fatigue 1365
     - 2.5.1. Background 1365
     - 2.5.2. Fatigue Guidelines 1366
   - 2.6. Error-Reduction Guidelines 1368
3. **GATHERING/ORGANIZING INFORMATION** **1371**
   - 3.1. What to Study (Pareto) 1371
   - 3.2. Videotaping Jobs 1371
   - 3.3. Searching for Solutions 1373
   - 3.4. Between-Operations Analysis 1374
     - 3.4.1. Flow Diagrams 1374
     - 3.4.2. Multiactivity Charts 1376
     - 3.4.3. Arrangement (Layout) of Equipment 1379
     - 3.4.4. Balancing Flow Lines 1382
   - 3.5. Within-Operation Analysis 1385
     - 3.5.1. Fish Diagrams 1385
     - 3.5.2. Decision Structure Tables 1385
     - 3.5.3. Checklists 1385
4. **ENGINEERING DESIGN** **1387**

REFERENCES **1389**

ADDITIONAL READING **1390**

1. OVERVIEW

Section 2 starts with some criteria for job design. The emphasis is on health and safety; due to progress in technology, every worker will supervise a number of slaves (machines). Section 2 then gives guidelines for organization of groups of workstations, the workstation itself, musculoskeletal disorders, fatigue, and error reduction.

Section 3 discusses gathering and organizing information. After potential projects are prioritized, the existing situation can be recorded. The big picture (between-operations analysis) or the detailed picture (within-operation analysis) can then be analyzed.

Section 4 discusses how to use the information to design.

2. GUIDELINES

This section discusses (1) criteria of job design, (2) organization of groups of workstations, (3) individual workstations, (4) musculoskeletal disorders, (5) fatigue, and (6) error reduction.

---

* This is a concise version of material in Konz and Johnson (2000).

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2.1. Criteria of Job Design

Table 1 gives the six ergonomic criteria for job design:

1. Safety is first.
   A job design that endangers the worker's safety or health is not acceptable. However, life does not have infinite value. Management must take reasonable precautions. Naturally, the definition of "reasonable" is debatable. After designing for safety, design for performance, then worker comfort, and, finally, worker higher wants (social, ego, and self-actualization).
2. Make the machine user-friendly.
   Make the machine adjust to the worker, not the worker to the machine. If the system does not function well, redesign the machine or procedure rather than blame the operator.
3. Reduce the percentage excluded by the design.
   Maximize the number of people who can use the machine or procedure. As mandated by the Americans with Disabilities Act, gender, age, strength, and other personal attributes should not prevent people from using the design.
4. Design jobs to be cognitive and social.
   Physical and procedural jobs now can be done by machines—especially in the developed countries. Over 50% of all jobs in the developed countries are in offices.
5. Emphasize communication.
   We communicate with machines through controls; machines communicate to us by displays. Improved communication among people reduces errors and thus improves productivity.
6. Use machines to extend human performance.
   The choice is not whether to assign a task to a person or to a machine; it is which machine to use. Small machines (such as word processors and electric drills) tend to have costs (total of capital, maintenance, and power) of less than \$0.25/hr. Even large machines (such as lathes and automobiles) tend to cost only \$1–2/hr. Labor cost (including fringes) tends to be a minimum of \$10/hr, with \$15, \$20, or higher quite common. What is your wage rate (including fringes)/per hour?

Consider machines as "slaves." The real question then becomes how many slaves the human will supervise and how to design the system to use the output of the slaves.

2.2. Organization of Groups of Workstations

This section discusses organization of workstations (the big picture) while Section 3 discusses the workstation itself. Table 2 gives seven guidelines:

1. Use specialization even though it sacrifices versatility.
   Specialization is the key to progress. Seek the simplicity of specialization; distrust it thereafter, but first seek it. Use special-purpose equipment, material, labor, and organization.
   Special-purpose equipment does only one job but does it very well. For example, at McDonald's a special-purpose grill cooks the hamburger on both sides simultaneously.
   Special-purpose material typically trades off higher material cost vs. greater capability.
   Labor specialization improves both quality and quantity. More experience, combined with specialized tools, should yield both higher quality and higher quantity. The challenge is that this specialization gives monotonous work; with low pay, it is difficult to find workers. With high pay, finding employees is not a problem.

TABLE 1 Ergonomic Job-Design Criteria

Number	Criteria
1.	Safety is first.
2.	Make the machine user-friendly.
3.	Reduce the percentage excluded by the design.
4.	Design jobs to be cognitive and social.
5.	Emphasize communication.
6.	Use machines to extend human performance.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

TABLE 2 Guidelines for Organization of Workstations

Number	Guideline
1.	Use specialization even though it sacrifices versatility.
2.	Consider both progressive and nonprogressive assembly.
3.	Minimize material-handling cost.
4.	Decouple tasks.
5.	Make several identical items at the same time.
6.	Combine operations and functions.
7.	Vary environmental stimulation inversely with task stimulation.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

For organization specialization, we have been describing jobs that are rigidly structured, disciplined, machinelike. There is a need for a high volume of a standardized product. If you don't have the required volumes, use specialization as a goal. Use group technology to get the benefits of mass production from batch production.

2. Consider both nonprogressive and progressive assembly.

Consider an assembly with $N = 60$ elements with $m = 5$ people. Should each of the 5 workers do all 60 elements—a complete workstation, job enlargement, nonprogressive assembly? Or should each person do N/m elements—a flow line, job simplification, progressive assembly? Table 3 compares nonprogressive vs progressive assembly.

3. Minimize material-handling (MH) cost.

Material handling does not add value, just cost. Reduce this cost by analysis of its components.

- MH cost/yr = capital cost + operating cost
- Operating cost = (no. of trips/yr)(cost/trip)
- Cost/trip = fixed cost/trip + (variable cost/distance) (distance/trip)

Capital costs vary little with utilization. For low volumes, the lowest total cost may be for a system with high operating cost but low capital cost. For high volumes, the high capital-cost system may be best.

TABLE 3 Progressive (P) vs. Non-progressive (NP) Lines

Criterion	Comment
Advantages of nonprogressive	
1. Balance delay	No balance delay.
2. Scheduling flexibility	Can make multiple products at same time.
3. Shocks	Do not affect multiple stations.
4. Cumulative trauma	Less CT since greater variety in motions.
5. Satisfaction	Each worker does complete job.
Neutral	
6. Quality	NP has immediate feedback of errors vs. mixed feedback in P.
7. Material handling	Simpler in NP. But mechanization more difficult in NP.
8. Space	NP has more space/station; material handling may increase space.
9. Walking	NP has some walking; this is good.
Disadvantages	
10. Direct labor	Higher for NP (due to less specialization, cost more to learn).
11. Skill required	NP requires more skill, which may require higher pay.
12. Equipment capital cost	NP higher as equipment duplicated.
13. In-process inventory	Higher for NP as inventory at many stations.
14. Supervision	Supervision more difficult for NP (more variety, paperwork).

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

Reduce the number of trips by moving less often. Consider replacing transportation with communication (e.g., mail vs. e-mail).

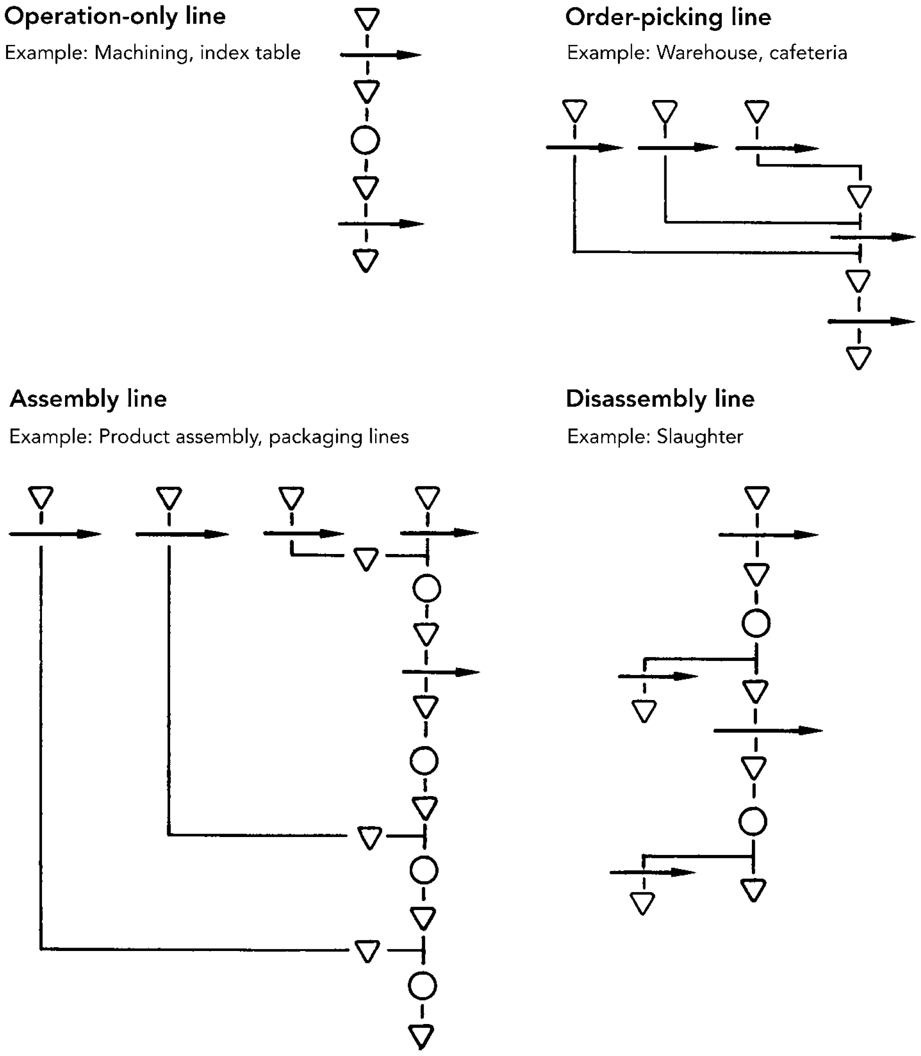
Fixed cost/trip includes information transfer; thus, consider computerization. Much transportation (and communication) is distance insensitive. Twice the distance does not cost twice as much. This affects plant layout because closer (or farther) locations make little difference.

Variable cost/distance should consider energy and labor costs. Distance/trip may be irrelevant, especially if a "bus system" can replace a "taxi system."

4. Decouple tasks.

Figure 1 shows different types of flow lines. Table 4 shows some alternatives for conveyor movement, work location, and operator posture.

Assuming progressive assembly, the elements are divided among the line's stations. In most cases, the mean amount of work time allocated to each workstation is not equal. But since



Operation-only line

Example: Machining, index table

Order-picking line

Example: Warehouse, cafeteria

Assembly line

Example: Product assembly, packaging lines

Disassembly line

Example: Slaughter

Figure 1 Flow Lines Can Be Operation-Only, Order-Picking, Assembly, or Disassembly Lines. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

TABLE 4 Alternatives for Conveyor Movement, Work Location, and Operator Posture

Conveyor Movement	Work Location	Operator Posture
Moves continuously	Removed from conveyor	Stands/sits in one spot
	Stays on conveyor	Stands/sits in one spot
	Stands/sits on moving system	
	Walks	
Starts/stops on timer	Stays on conveyor	Stands/sits in one spot
Starts/stops at operator discretion	Stays on conveyor	Stands/sits in one spot
	Removed from conveyor	Stands/sits in one spot

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

each workstation has the same cycle time, there is some idle time (cycle time − work time) at each station—this is called balance delay time.

Although it may be obvious, (1) the transport between stations need not be by conveyor (carts are a common alternative, either pushed manually or as automatically guided vehicles); (2) the transport between stations need not be at a fixed speed or time interval; and, very important, (3) there probably will be storages between the operations and transportations. This storage is known as a buffer; its purpose is to decouple the workstations (isolate stations of the line).

The two primary reasons for buffers are line balancing and shocks and disturbances.

Assume the mean times for stations A, B, C . . . are not equal. Assume A took 50 sec, B took 40 sec, and C took 60 sec. Then, without buffers, the line must be set at the speed of the slowest workstation, C. But, in addition, shocks and disturbances cause variability in the mean times. Thus, without buffers, the cycle time must be set at the slowest time of the slowest station, say 65 sec. Buffers give flexibility to the flow line.

There are two buffering techniques: change product flow and move operators. Changing product flow can be done by buffers at or between the stations, buffers due to carrier design, and buffers off-line. Moving operators can be done by (1) utility operator, (2) helping your neighbor, (3) *n* operators floating among *n* workstations, and (4) *n* operators floating among more than *n* stations. See Konz and Johnson (2000) for an extensive discussion of the various alternatives.

5. Make several identical items at the same time.

   Tasks can be divided into three stages: (1) get ready, (2) do, and (3) put away. Reduce cost/unit by prorating the get-ready and put-away stages over more units.

6. Combine operations and functions.

   Consider multifunction materials and equipment rather than single-function materials and equipment. For example, an invoice can be used with a window return envelope to reduce sorting the envelope when it is returned and speed processing of the bill. A farmer can combine fertilizing and plowing.

7. Vary environmental stimulation inversely with task stimulation.

   For low-stimulation tasks, (1) add physical movement to the task and (2) add stimulation to the environment. The easiest environmental stimulation solution is to allow operators to talk to each other (see Figure 2). Other alternatives are windows and making the break area attractive. For high-stimulation tasks, such as in offices, reduce stimulation by improving both visual and auditory privacy.

2.3. Individual Workstations

Table 5 gives 14 guidelines for workstation design:

1. Avoid static loads and fixed work postures.

   Static (isometric) load (from low variability in postures and movement) is bad for the blood supply/disposal of a specific muscle as well as the total body. We will discuss standing, sitting, head/neck, and hand/arm.

   - *Standing.* The veins store the body's blood. If the legs don't move, the blood from the heart tends to go down to the legs and stay there (venous pooling). This causes more work for the heart because, for a constant blood supply, when blood per beat is lower, there are more

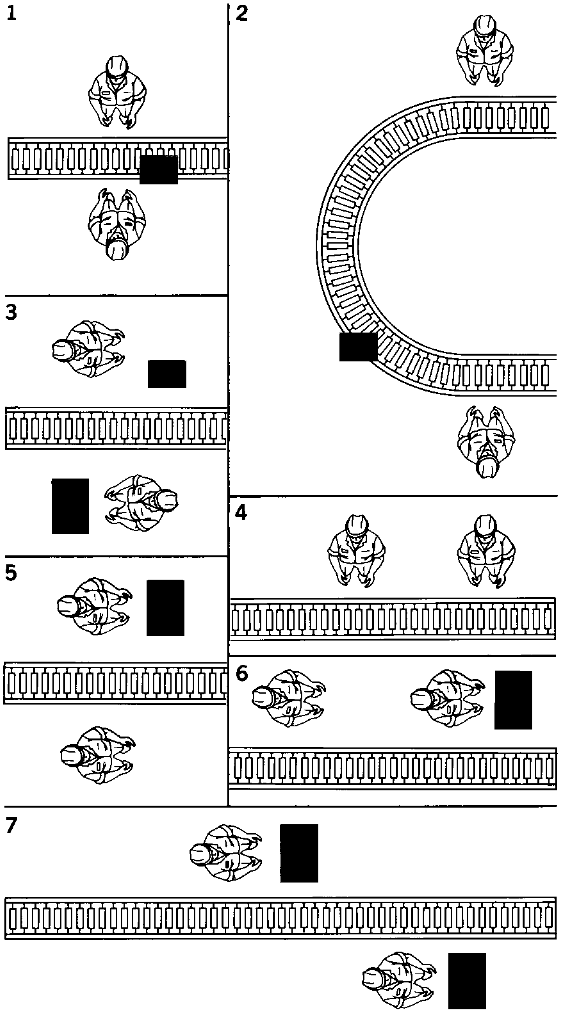


Figure 2 Vary Environmental Stimulation by Adjusting Orientation in Relation to Other People, Changing Distance, and Using Barriers Such as Equipment between Stations. The highest stimulation is 1, the lowest is 7. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

beats. Venous pooling causes swelling of the legs, edema, and varicose veins. After standing, walking about 10 steps changes ankle blood pressure to about 48 mm Hg (the same level as for sitting). Thus, standing jobs should be designed to have some leg movements. Consider a bar rail (to encourage leg movement while standing) and remote storage of supplies (to encourage walking).

- *Sitting.* Sitting is discussed in workstation guideline 4.
- *Head/neck.* The head weighs about 7.3% of body weight. For a 90 kg person, that is 6.6 kg (about the same as a bowling ball). Neck problems occur if the head leans forward or backward on the neck. Forward tilt occurs when the person reduces the distance to an object to improve visibility (inspection, fine assembly, VDT work); consider better lighting. Backward tilt may occur for people using bifocals at VDT workstations; consider single-vision glasses (work glasses). If the hands are elevated, head- or workstation-mounted magnifiers permit lowering the hands.

TABLE 5 Guidelines for Workstation Design

Number	Guideline
1.	Avoid static loads and fixed work postures.
2.	Reduce musculoskeletal disorders.
3.	Set the work height at 50 mm below the elbow.
4.	Furnish every employee with an adjustable chair.
5.	Use the feet as well as the hands.
6.	Use gravity; don't oppose it.
7.	Conserve momentum.
8.	Use two-handed motions rather than one-handed motions.
9.	Use parallel motions for eye control of two-handed motions.
10.	Use rowing motions for two-handed motions.
11.	Pivot motions about the elbow.
12.	Use the preferred hand.
13.	Keep arm motions in the normal work area.
14.	Let the small woman reach; let the large man fit.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

- *Hand/arm.* When an operator holds a workpiece, this not only causes static load but also reduces productivity (since the person is working with one hand, not two).

  Keep the upper arm vertical. Support the arm weight by chair arms or a workstation. Pad sharp edges of workstations. Supporting the arms reduces tremor and thus permits more accurate work with the hands.

2. Reduce musculoskeletal disorders.

See Section 2.4.

3. Set the work height at 50 mm below the elbow.

For manipulative work, the optimum is about 50 mm (2 in.) below the elbow. Output does not decrease more than a couple of percent within a range from 125 mm (5 in.) below to 25 mm (1 in.) above the elbow; beyond this, the penalty is greater. If a downward force is required (polishing, sanding), the optimum is probably with the lower arm at a 45° angle.

The three solution techniques are: (1) change work surface height, (2) adjust elbow height, and (3) adjust work height on the machine.

Work surface heights (such as conveyors) often are easily adjustable. Workstation table heights often can be adjusted—perhaps even with motors.

Adjusting elbow height while sitting involves use of an adjustable-height chair. For standing, consider movable wooden platforms for short operators (for tall operators, remove the platform).

Adjusting work height may be as simple as tilting (45 or 90°) a container box. Use containers with short sides. If the part/assembly is held in a fixture, design the fixture to permit vertical and horizontal adjustment. For example, cars are now assembled in rollover frames; the frames allow the operator to work on the car while it is tilted at various angles.

4. Furnish every employee with an adjustable chair.

A chair is a low-cost tool. Chair cost is entirely a capital cost because operating cost and maintenance costs are zero. Assuming a cost of $200/chair, a life of 11 years, and a one-shift operation of 1800 hr/yr, the cost/hour is $200/(11 × 1800) = 1 cent/hr. (Often the real comparison is between a chair without good ergonomics vs. one with ergonomic features, and the incremental cost of the good chair is less than $100—that is, less than 0.5 cent/hr.) The cost of labor (wages plus fringes) varies with the job but will be at least $10/hr, with $15 or $20 fairly common. Assuming a cost of $10/hr, a 0.1% change in productivity (30 sec/shift) equals 1 cent/hr.

Even with an ergonomic chair, many people will have back pain if they sit continuously. The reason is that the disks of the back are nourished by movement of the back—no movement, no nourishment. Thus, encourage seating posture variability by having the person walk occasionally (to answer the phone, discuss things with others, take breaks).

5. Use the feet as well as the hands.

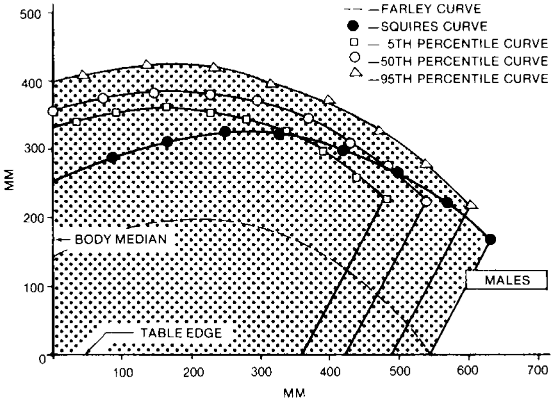


Figure 3 Normal Male Work Area for Right Hand. The left-hand area is mirrored. (From S. Konz and S. Goel, *The Shape of the Normal Work Area in the Horizontal Plane.* Reprinted with the permission of the Institute of Industrial Engineers, 25 Technology Park, Norcross, GA 30092, 770-449-0461. Copyright © 1969)

The feet can take some of the load off the hands in simple control tasks such as automobile pedals or on/off switches. For human-generated power, the legs have about three times the power of the arms.

6. Use gravity; don't oppose it.

Consider the weight of the body and the weight of the work. The weight of arm plus hand is typically about 4.9% of body weight; thus, keep arm motion horizontal or downward. Gravity also can be a "fixture"; consider painting the floor vs. the ceiling; welding below you vs. above you, and so on. Gravity also can be used for transport (chutes, wheel conveyors).

7. Conserve momentum.

Avoid the energy and time penalties of acceleration and deceleration motions. Avoid change of direction in stirring, polishing, grasping, transport, and disposal motions. For example, using the MTM predetermined time system, an 18 in. (450 mm) move to toss aside a part is an M18E and an RL1—a total of 0.63 sec. A precise placement would require an M18B, a P1SE, and an RL1—a total of 0.89 sec, an increase of 42%.

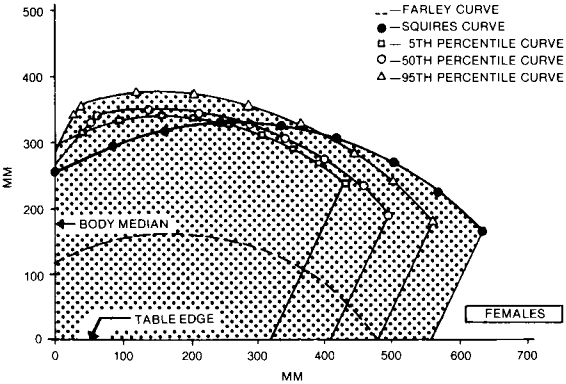


Figure 4 Normal Female Work Area for Right Hand. The left-hand area is mirrored. (From S. Konz and S. Goel, *The Shape of the Normal Work Area in the Horizontal Plane.* Reprinted with the permission of the Institute of Industrial Engineers, 25 Technology Park, Norcross, GA 30092, 770-449-0461. Copyright © 1969)

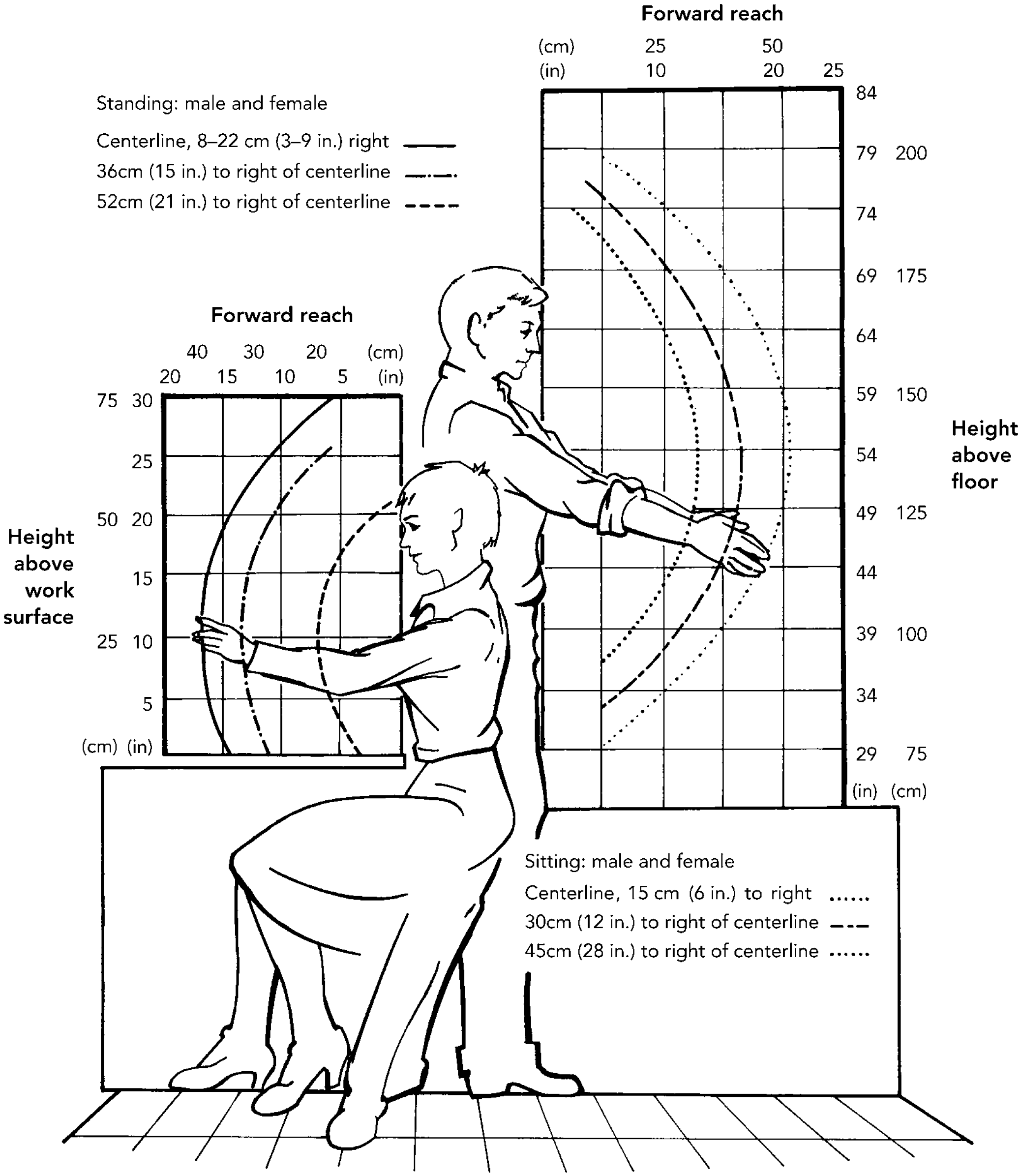

Forward reach
(cm) 25 50
(in) 10 20 25
Standing: male and female
Centerline, 8–22 cm (3–9 in.) right
36cm (15 in.) to right of centerline
52cm (21 in.) to right of centerline
84
79 200
74
69 175
64
59 150
54
49 125
44
39 100
34
29 75
(in) (cm)
Height above floor
Forward reach
40 30 20 (cm)
20 15 10 5 (in)
75 30
25
50 20
15
25 10
5
(cm) (in)
Height above work surface
Sitting: male and female
Centerline, 15 cm (6 in.) to right
30cm (12 in.) to right of centerline
45cm (28 in.) to right of centerline

Figure 5 Approximate Reach Distances for Average U.S. Male and Female Workers. (From V. Putz-Anderson, Ed., *Cumulative Trauma Disorders,* copyright © 1988 Taylor & Francis Books Ltd., by permission)

8. Use two-hand motions rather than one-hand motions.

   For movement of the hand/arm, two hands are better than one. Using two hands takes more time and effort, but more is produced so cost/unit is lower. When one hand (generally the left) is acting as a clamp, consider it as idle and use a mechanical clamp instead.

9. Use parallel motions for eye control of two-hand motions.

   When the two hands are both moving, minimize the distance between the hands rather than making the arm motions symmetrical about the body centerline.

10. Use rowing motions for two-hand motions.
    A rowing motion is superior to an alternating motion because there is less movement of the shoulder and torso.
11. Pivot motions about the elbow.
    For a specific distance, minimum time is taken by pivoting about the elbow. It takes about 15% more time to reach across the body (it also takes more energy since the upper arm is moved as well as the lower arm).
12. Use the preferred hand.
    The preferred hand is 5–10% faster and stronger than the nonpreferred hand. Unfortunately, since it is used more, it is more likely to have cumulative trauma. About 10% of the population uses the left hand as the preferred hand.
13. Keep arm motions in the normal work area.
    Figure 3 shows the area for males; Figure 4 shows the area for females. These are the distances to the end of the thumb with no assistance from movement of the back. Closer is better. Figure 5 shows reach distance above and below the horizontal plane.
14. Let the small woman reach; let the large man fit.
    The concept is to permit most of the user population to use the design. Alternative statements are "Exclude few," "Include many," and "Design for the tall; accommodate the small." What percentage should be excluded? The Ford–UAW design guide excludes 5% of women for reach, 5% of men for clearance, and 0% of men and women for safety.

Three alternatives are (1) one size fits all, (2) multiple sizes, and (3) adjustability. Examples of one size fits all are a tall door and a big bed. An example of multiple sizes is clothing. An example of adjustability is an automobile seat.

2.4. Musculoskeletal Disorders

2.4.1. Risk Factors

Safety concerns are for short-term (time frame of seconds) effects of physical agents on the body. An example is cutting of a finger. Toxicology generally deals with long-term (years, decades) effects of chemicals on body organs. An example is exposure to acetone for 10 years, causing damage to the central nervous system. Musculoskeletal disorders concern intermediate-term (months, years) effects of body activity upon the nerves, muscles, joints, and ligaments. An example is back pain due to lifting.

Nerves supply the communication within the body. Muscles control movements of various bones. Ligaments (strong, ropelike fibers) connect bones together.

The three main occupational risk factors are repetition/duration, joint deviation, and force. Vibration is also an important risk factor. Nonoccupational risk factors can be from trauma outside work or a nonperfect body. The lack of perfection can be anatomical (weak back muscles or weak arm due to an injury) or physiological (diabetes, insufficient hormones).

Problem jobs can be identified through (1) records/statistics of the medical/safety department, (2) operator discomfort (e.g., Figure 6), (3) interviews with operators, and (4) expert opinion (perhaps using checklists such as Table 6, Table 7, and Table 8).

2.4.2. Solutions

Decrease repetition/duration, joint deviation, force, and vibration.

In general, a job is considered repetitive if the basic (fundamental) cycle time is less than 30 sec (for hand/wrist motions) and several minutes (for back/shoulder motions). However, if the job is only done for a short time (say, 15 min/shift), there is relatively low risk of cumulative trauma due to the short duration. Consider short duration as <1 hr/shift, moderate as 1 to 2 hr, and long as >2 hr. Thus, repetition really concerns the repetitions/shift. Duration also assumes weeks, months, and years of repeated activity, not just a couple of days.

Ideally the joint should operate at the neutral position—that is, minimum joint deviation.

Force on a joint typically is multiplied by a lever arm (i.e., we are really talking about a torque). Reduce (1) the magnitude of the force, (2) the lever arm, and (3) the length of time the force is applied.

Vibration typically comes from a powered handtool. Vibration increases force because operators tend to grip the vibrating tool more tightly because it vibrates.

Solutions are divided into engineering and administrative.

2.4.2.1. Engineering Solutions The first possible approach is automation—that is, eliminate the person. No person means no possible injury.

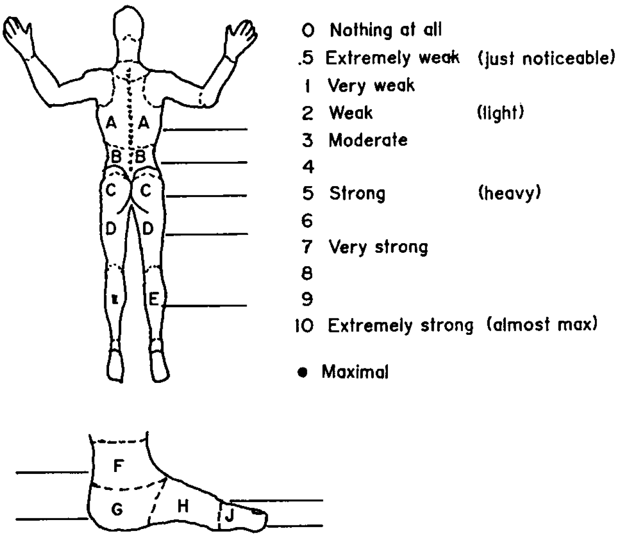


Figure 6 Body Discomfort Map. Maps can have more or less detail; this map includes the foot but other maps may emphasize the arm or hand. The worker gives the amount of discomfort at any location. One discomfort scale is the Borg category scale (CR-10) shown in the figure. The CR-10 scale gives 0.5 for "extremely weak" and 10 for "extremely strong"; people are permitted to go below 0.5 and above 10. Body discomfort maps are a popular technique of getting quantified input from operators. (From Corlett and Bishop 1976)

A second possible approach is to reduce the number of cycles or the difficulty of the cycle. Two alternatives are mechanization and job enlargement. In mechanization, the operator is still present but a machine does part of the work. Two examples are electric scissors (vs. manual scissors) and a bar code scanner (instead of information being keyed). In job enlargement, the same motions are done, but by a larger number of people. If the job takes four minutes instead of two, the repetitive motions per person per day are reduced.

The third approach is to minimize joint deviation. One guide is *Don't bend your wrist*. For example, place keyboards low to minimize backward bending of the hand. Another guide is *Don't lift your elbow*. An example was spraypainting the side of a truck by workers standing on the floor. The job was modified to have the workers stand on a platform, thus painting horizontally and downward. (This had the additional benefit of less paint settling back into the painters' faces.)

The fourth approach is to minimize force duration and amount. Consider an ergonomic pen (large grip diameter with a high-friction grip) to reduce gripping force. Can a clamp eliminate holding by a hand? Can a balancer support the tool weight? Can sliding replace lifting?

To reduce vibration, use tools that vibrate less; avoid amplified vibration (resonance). Maintain the equipment to reduce its vibration. Minimize gripping force (e.g., support the tool with a balancer or a steady rest). Keep the hands warm and dry; avoid smoking.

2.4.2.2. Administrative Solutions Administrative solutions (such as job rotation and part-time workers) seek to reduce exposure or increase operator's ability to endure stress (exercise, stress reduction, and supports).

In job rotation, people rotate jobs periodically during the shift. The concept is working rest—a specific part of the body rests while another part is working. Job rotation requires cross-trained workers (able to do more than one job), which allows more flexible scheduling; there is perceived fairness because everyone shares good and bad jobs.

Part-time workers reduce musculoskeletal disorders/person. Some other advantages are lower wage cost/hr, better fit to fluctuating demand, and (possibly) higher-quality people. (It is difficult to hire high-quality full-time people for repetitive low-paying jobs.) Some disadvantages are less time on the job (and thus less experience), possible moonlighting with other employers, and a high cost of hiring and training.

TABLE 6 General Ergonomic Checklist to Prioritize Potential Problems

After completing the effort (force), continuous effort time (duration), and efforts/min (repetition) columns, determine the priority for change.

Job Title ______________________ Analyst ______________________
Specific task ______________________ Phone ______________________
Job number ____________ Dept ____________ Date of analysis ____________
Location ______________________

Body part		Effort	Continuous Effort Time	Effort/Min	Priority
Back					
Neck/shoulders	R				
	L				
Arms/elbows	R				
	L				
Wrists/hands/fingers	R				
	L				
Legs/knees	R				
	L				
Ankles/feet/toes	R				
	L				

Effort
1 = light
2 = moderate
3 = heavy

Cont. Effort Time
1 = <6 sec
2 = 6 to 20 sec
3 = >20 sec

Efforts/min
1 = <1
2 = 1 to 5
3 = >5 to 15

Priority for change
332
331
323 Very high
322
321
313 High
223
312 Moderate
232
231
222
213
132
123

From S. A. Rodgers, ''Functional job analysis technique,'' *Occupational Medicine: State of the Art Reviews,* Vol. 7, Copyright © 1992. Hanley & Belfus. Reprinted by permission.

TABLE 7 Sample Checklist for Upper-Extremity Cumulative Trauma Disorders

No	Yes	Risk Factors
		Physical stress
______	______	1. Can the job be done without hand/wrist contact with sharp edges?
______	______	2. Is the tool operating without vibration?
______	______	3. Are the worker's hands exposed to temperatures >70°F (21°C)?
______	______	4. Can the job be done without using gloves?
		Force
______	______	1. Does the job require exerting less than 10 lb (4.5 kg) of force?
		2. Can the job be done without using a finger pinch grip?
		Posture
______	______	1. Can the job be done without wrist flexion or extension?
______	______	2. Can the tool be used without wrist flexion or extension?
______	______	3. Can the job be done without deviating the wrist from side to side?
______	______	4. Can the tool be used without deviating the wrist from side to side?
______	______	5. Can the worker be seated while performing the job?
______	______	6. Can the job be done without a clothes-wringing motion?
		Workstation hardware
______	______	1. Can the worksurface orientation be adjusted?
______	______	2. Can the worksurface height be adjusted?
______	______	3. Can the tool location be adjusted?
		Repetitiveness
______	______	1. Is the cycle time longer than 30 s?
		Tool design
______	______	1. Are the thumb and finger slightly over lapped in a closed grip?
______	______	2. Is the tool handle span between 2 and 2.75 in. (5 and 7 cm)?
______	______	3. Is the tool handle made from material other than metal?
______	______	4. Is the tool weight below 9 lb (4 kg)? Note exceptions to the rule.
______	______	5. Is the tool suspended?

Adapted from Y. Lifshitz and T. J. Armstrong, "A Design Checklist for Control and Prediction of Cumulative Trauma Disorder in Hand Intensive Manual Jobs," in *Proceedings of the Human Factors Society 30th Annual Meeting 1986.* Copyright © 1986. Used with permission from the Human Factors and Ergonomics Society. All rights reserved. A "no" indicates a risk factor. Checklist was tested only in an automobile final assembly plant.

Exercises should be tailored to the specific set of muscles, tendons, and ligaments that are stressed. If the work loads the muscles statically, the exercise should move them. If the work loads the muscles dynamically, the exercise should relax and stretch them.

Stress can be caused by social factors, both on and off the job. Supports for the body (armrests, back belts, wrist splints, etc.) are an appealing concept, but evidence is lacking concerning their benefits.

2.5. Fatigue

2.5.1. Background

The problem is to reduce fatigue so workers can maintain/increase productivity and have optimal stress. Fatigue can be divided into five factors:

1. Physical exertion (e.g., bicycle ergonometer work; descriptions such as "warm," "sweaty," "out of breath," "breathing heavily," "palpitations")
2. Physical discomfort (e.g., static load on small-muscle groups; descriptions such as "tense muscles," "aching," "numbness," "hurting," "stiff joints")
3. Lack of energy (mental plus physical; descriptions such as "exhausted," "spent," "overworked," "worn out," "drained")
4. Lack of motivation (mental; descriptions such as "lack of initiative," "listless," "passive," "indifferent," "uninterested")
5. Sleepiness (mental; descriptions such as "sleepy," "yawning," "drowsy," "falling asleep," "lazy").

TABLE 8 Posture Checklist for Neck, Trunk, and Legs

	Percent Time Posture Used in Job		
Job Studied	Never	<1/3	>1/3
Neck			
1. Mild forward bending (>20°)	0	0	X
2. Severe forward bending (>45°)	0	X	*
3. Backward bending (>20°)	0	X	*
4. Twisting or lateral bending (>20°)	0	X	*
Trunk			
5. Mild forward bending (>20°)	0	X	*
6. Severe forward bending (>45°)	0	*	*
7. Backward bending (>20°)	0	X	*
8. Twisting or lateral bending (>20°)	0	X	*
General body/legs			
9. Standing stationary (no walking or leaning)	0	0	X
10. Standing, using footpedal	0	X	*
11. Knees bent or squatting	0	X	*
12. Kneeling	0	X	*
13. Lying on back or side	0	X	*
Total X = ________	Total * = ________		
Comments:			

Reprinted from International Journal of Industrial Ergonomics, Vol. 9, M. Keyserling, M. Brouwer, and B. A. Silverstien, "A Checklist for Evaluating Ergonomic Risk Factors Resulting from Awkward Postures of the Legs, Trunk, and Neck," pp. 283–301, Copyright © 1992, with permission from Elsevier Science.
A zero indicates insignificant risk. An X indicates a potential risk. An asterisk indicates a significant risk.

Jobs will have different combinations of fatigue and the combinations often will vary during the shift. Fatigue generally is overcome by rest (recovery). Rest time can be classified as off-work (evenings, holidays, weekends, vacations) and at-work. At-work is further divided into formal breaks (lunch, coffee), informal breaks (work interruptions, training), microbreaks (short pauses of a minute or less), and working rest (a different task using a different part of the body, such as answering the phone instead of keying data). From the viewpoint of financial cost, some rest time is paid and some is unpaid; from a fatigue viewpoint, a rest is a rest.

In general, weekly, annual, and lifetime working hours are decreasing and thus present fewer health/productivity problems. But daily hours may be a problem, especially for people working more than 8 hr/day and without proper sleep.

2.5.2. Fatigue Guidelines

Table 9 gives seven guidelines concerning fatigue. Guidelines 1, 2, and 3 concern fatigue prevention; guidelines 4, 5, 6, and 7 concern fatigue reduction.

TABLE 9 Guidelines for Fatigue

Number	Guideline
Fatigue prevention	
1.	Have a work-scheduling policy.
2.	Optimize stimulation during work.
3.	Minimize the fatigue dose.
Fatigue reduction	
4.	Use work breaks.
5.	Use frequent short breaks.
6.	Maximize the recovery rate.
7.	Increase the recovery/work ratio.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

TABLE 10 Tips for Day Sleeping

- Develop a good sleeping environment (dark, quiet, cool, bed). Have it *dark* (e.g., use eyeshades or opaque curtains). Have it *quiet* since it is difficult to go back to sleep when daytime sleep is interrupted. Minimize noise volume. Consider earplugs, unplugging bedroom phones, turning down phone volume in other rooms, reducing TV volume in other rooms, Train your children. Have a *cool* sleeping area. The *bed* normally is OK but may be poor if the sleeper is not sleeping at home (e.g., is part of an ''augmented crew'' for trucks, aircraft). Then provide a good mattress and enough space.
- Plan your sleeping time. Tell others your schedule to minimize interruptions. Morning-to-noon bedtimes are the most unsuitable times to sleep. Consider sleeping in two periods (5–6 hr during the day and 1–2 hr in the late evening before returning to work). Less daytime sleep and some late evening sleep not only make it easier to sleep but also may give a better fit with family/social activities.
- Have a light (not zero or heavy) meal before sleep. Avoid foods that upset your stomach and thus wake you up. Avoid liquid consumption, as it increases the need to urinate (which wakes you up). Avoid caffeine. A warm drink before your bedtime (perhaps with family members starting their day) may help your social needs.
- If under emotional stress, relax before going to bed. One possibility is light exercise.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

1. Have a work scheduling policy.

The problem is insufficient rest. Two aspects are (1) too many work hours, and (2) work hours at the wrong time.

Too many hours can accrue through by not counting all the hours in duty time. For example, train and flight crews have waiting and preparation time before and after the primary job. Long hours may be from overtime due to absenteeism of another person and/or moonlighting. There probably should be some organizational restriction on prolonged overtime (say over 12 hr/day and over 55 hr/week).

Lack of sleep can result from conflicts with the circadian rhythm. Table 10 gives sleeping tips; Table 11 gives shiftwork recommendations.

2. Optimize stimulation during work.

Too much stimulation causes overload; too little causes boredom. Stimulation comes from the task and the environment.

For too much stimulation, reduce environmental stimulation (see Figure 2). For example, for office tasks, increase visual and auditory privacy.

For too little stimulation, increase either task or environmental stimulation or both. Increase task stimulation by adding physical work to sedentary jobs or by job rotation. Increase environmental stimulation by (1) conversation with others, (2) varying the auditory environment

TABLE 11 Design Recommendations for Shift Systems

- Avoid permanent night work. Full entrainment of physiological functions to night work is difficult. Permanent night workers have problems due to readapting to day cycles during weekends, holidays, and vacations.
- If shifts rotate, rapid rotation is preferable to slow (weekly) rotation. Shifts should rotate forward (day, evening, night).
- Avoid starting the morning shift before 7 a.m.
- Consider distribution of leisure time. Have sufficient time to sleep between shifts (e.g., between shift changeovers). Limit the number of consecutive working days to 5–7. For every shift system, have some nonworking weekends with at least 2 successive full days off.
- Shift durations of 12 hr have disadvantages as well as advantages. Disadvantages include fatigue, covering absentees, overtime, limiting toxic exposure, and possible moonlighting when workers have large blocks of leisure time.
- Make the schedule simple and predictable. People want to be able to plan their personal lives. Make work schedules predictable. Publicly post them in advance so people can plan; 30 days in advance is a good policy.

Source: Konz and Johnson 2000, from Knauth, 1993.

(e.g., talk radio), (3) varying the visual environment (e.g., windows with a view), and (4) varying the climate (change air temperature, velocity).

3. Minimize the fatigue dose.

   The problem is that the fatigue dose becomes too great to overcome easily. Two aspects are intensity and work schedule.

   Reduce high intensity levels with good ergonomic practices; for example, use machines and devices to reduce hold and carry activities. Static work (holding) is especially stressful.

   The effect of fatigue increases exponentially with time. Thus, it is important to get a rest before the fatigue becomes too high. Do not permit workers to skip their scheduled break. For very high-intensity work (e.g., sorting express packages), use part-time workers.

4. Use work breaks.

   The problem with a conventional break is that there is no productivity during the break. A solution is to use a different part of the body to work while resting the fatigued part. If a machine is semiautomatic, the worker may be able to rest during the automatic part of the cycle (machine time). Another alternative is job rotation (worker periodically shifts tasks). Fatigue recovery is best if the alternative work uses a distinctly different part of the body—for example, loading/unloading a truck vs. driving it, word processing vs. answering a telephone.

   Not quite as good, but still beneficial, is alternating similar work, as there would be differences in body posture, force requirements, mental activity, because for example, an assembly team might rotate jobs every 30 minutes. In a warehouse, for half a shift, workers might pick cases from a pallet to a conveyor, and during the second half, they would switch with the people unloading the conveyor to trucks. Job rotation also reduces the feeling of inequity because everyone shares the good and bad jobs. However, it does require cross-trained people (able to do multiple jobs). However, this in turn increases scheduling flexibility.

5. Use frequent short breaks.

   The problem is how to divide break time. The key to the solution is that fatigue recovery is exponential. If recovery is complete in 30 minutes, it takes only 2 minutes to drop from 100% fatigue to 75% fatigue; it takes 21 minutes to drop from 25% fatigue to no fatigue. Thus, give break time in small segments. Some production is lost for each break. Reduce this loss by not turning the machine off and on, taking the break near the machine, and so on.

6. Maximize the recovery rate.

   The problem is recovering as quickly as possible. In technical terms, reduce the fatigue half-life.

   For environmental stressors, reduce contact with the stressor. Use a cool recovery area to recover from heat, a quiet area to recover from noise, no vibration to recover from vibration. For muscle stressors, it helps to have a good circulation system (to be in good shape). Active rest seems better than passive rest. The active rest may be just walking to the coffee area (blood circulation in the legs improves dramatically within <20 steps). Another technique to get active rest is to have the operator do the material handling for the workstation (obtain supplies, dispose of finished components).

7. Increase the recovery/work ratio.

   The problem is insufficient time to recover. The solution is to increase recovery time or decrease work time. For example, if a specific joint is used 8 hr/day, there are 16 hr to recover; 2 hr of recovery/1 hr of work. If the work of the two arms is alternated so that one arm is used 4 hr/day, there are 20 hr to recover; 5 hr of recovery/1 hr of work. Overtime, moonlighting or 12 hr shifts can cause problems. Working 12 hr/day gives 12 hr of recovery, so there is 1 hr of recovery/1 hr of work.

2.6 Error-Reduction Guidelines

Harry and Schroeder (2000) describe the importance from a business viewpoint of joining the "cult of perfectability." Table 12 gives l0 general guidelines to reduce errors. They are organized into three categories: planning, execution, and allowing for error.

1. Get enough information.

   Generating enough information calls for (1) generating/collecting the relevant information and (2) ensuring that the user receives the information.

   Generating the information can be aided by computerization. But the information in the computer needs to be correct! Warnings (e.g., auditory alarm when truck backs up) can generate information; unfortunately, not all warnings are followed.

   Ensure information reception. Just because you put a message in a person's mailbox

TABLE 12 Guidelines for Error Reduction

Planning
1. Get enough information.
2. Ensure information is understood.
3. Have proper equipment/procedures/skill.
4. Don't forget.
5. Simplify the task.

Execution
6. Allow enough time.
7. Have sufficient motivation/attention.

Allow for errors
8. Give immediate feedback on errors.
9. Improve error detectability.
10. Minimize consequences of errors.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

(voicemail, e-mail), doesn't mean the person accesses it! Do you know how to use all the buttons on your telephone?

2. Ensure that information is understood.

Although there is much information available on good ergonomic design of displays (instruments as well as print such as text, figures, and tables), unfortunately, not all designers seem to have paid attention to them. Communication with the general public is especially difficult due to their lack of training on your specific system as well as their diverse ages, vision, language abilities, and so on. Equipment and procedures for the general public need to have usability tests using diverse customers.

When transmitting information to others, a good technique is to have them repeat the information back to you in their words. Technically this information feedback has two steps: forward (you to them) and backward (them to you). As the number of communication nodes increases, the need for feedback increases. Critical messages thus should be written and require confirmation from the user, not only of the receipt of the message but also of the user's understanding of its message.

3. Have proper equipment/procedures/skill.

Four aspects of equipment are design, amount, arrangement, and maintenance.

With proper equipment design, the potential error is eliminated and procedures and skill are not necessary. Don't forget foreseeable use and misuse. Some examples are a ground-fault circuit interrupter to eliminate electrocution; a surge protector for computers; design of connectors so they cannot fit into the wrong plug. Equipment guards not only must be installed, they also must remain installed (and not removed by the operator). In some cases, "antiergonomics" is desirable—making something difficult to use. Lockout/tagout prevents equipment from being activated while it is being maintained.

Equipment amount basically involves duplication. Some examples are duplication of computer files, equipment (four engines on an airplane), and even labor (copilot).

Equipment arrangement involves the concepts of series and parallel systems as well as standby systems. Redundancy can just be duplication of the procedure, but it also can be having some information confirm the remainder—an error-checking code. For example, giving a meeting date as Thursday, February 25, lets the Thursday confirm the 25. A ZIP code confirms the names of city and state on a postal address.

Equipment maintenance typically is assumed as a given. However the Union Carbide incident at Bhopal, India, where over 2500 people were killed, demonstrates what can happen when there is an insufficient maintenance budget, for both spare parts and for labor.

Procedures describe the sequence of body motions necessary to accomplish a task. Skill is the eye/brain/hand coordination to perform the body motion.

Develop standard procedures (protocols). Then users need to be trained on the protocol. The mind seeks patterns and rules and will follow them if they are apparent. Training may require a formal procedure of (1) writing down the training procedure and training aids, (2) designating specific people as trainers, and (3) documenting that a specific person has been trained by a specific trainer on a specific date.

Skill has both mental and physical aspects. Skill or knowledge can be memorized, but an alternative is a job aid (such as information stored in a computer). Practice is important because rookies make more mistakes than old pros; that's why sports teams practice so much.

4. Don't forget.

Two approaches to avoid forgetting are to reduce the need to remember and to use memory aids.

Avoid verbal orders: they leave no reference to refresh the memory. If you receive a verbal order, write it down to create a database for future reference.

A list is a popular memory aid; the many electronic organizers being sold attest to the need for making a list. Memory aids include more than that, however. Forms not only include information but also indicate what information is needed. Patterns aid memory. For example, it is easier to remember a meeting is on the first Friday of each month than to remember the 12 individual days. At McDonald's, orders are filled in a standard sequence; next time you are there, see if you can determine the standard sequence. Other memory aids include the fuel gauge and warning lights in cars and the sticker on the windshield showing when the oil was changed.

5. Simplify the task.

Two ways to simplify tasks are to reduce the number of steps and to improve communication.

Familiar examples of reducing steps are autodialers on telephones and use of address lists for email. Many retailers now have their computers communicate point-of-sale information not only to their own central computer but also to their vendors' computers. The original concept was to speed vendors' manufacturing and shipping, but the retailers soon realized that the elimination of all the purchasing paperwork led to drastic error reductions.

Improve communication by improving controls (how people communicate to machines) and displays (how machines communicate to people); details are covered in a number of ergonomics texts. But even letters and numbers can be confused—especially the number 0 and the letter O and the number 1 and the letter l. Use all letter codes or all numeric codes, not a mixture. If a mixed code is used, omit 0, O, 1 and l. Also, do not use both a dash and an underline.

Guidelines 6 and 7 concern execution of tasks.

6. Allow enough time.

Under time stress, people make errors. Start earlier so there is more margin between the start time and the due time. Another technique to reduce time/person by putting more people on the task; this flexibility requires cross-trained people.

7. Have sufficient motivation/attention.

Motivation is not a replacement for engineering. Gamblers are motivated to win, but wanting to win is not enough. Motivation can be positive (aid performance) or negative (hinder performance).

There are many examples of superiors challenging subordinate's decisions but relatively few of subordinates challenging superior's decisions. It is difficult, though important, to require important decisions to be challenged by or justified to subordinates. Company policies and culture and openness to criticism are important factors. Even requiring superiors to explain decisions to subordinates can be valuable: when superiors try to explain the decision, they may realize it is a bad decision.

Lack of attention can occur from simple things such as distractions, but it can also be caused by people sleeping on the job or being drunk or ill. One solution technique is to have observation by other people (either personally or by TV) so sleeping, heart attacks, and so on can be observed.

Guidelines 8, 9, and 10 discuss allowing for errors.

8. Give immediate feedback on errors.

Two subdivisions of feedback are error detection and reducing delay.

Errors can be detected by people (inspection) or by machines. Machines can passively display the error (needle in the red zone) or actively display the error (audio warning, flashing light) or even make an active correction (turn on furnace when temperature in the room drops). The machine also can fail to respond (computer doesn't work when the wrong command is entered). Machines also can present possible errors for human analysis (spell-check programs). Error messages should be specific and understandable. The message ''Error'' flashing on the screen is not as useful as ''Invalid input—value too high. Permissible range is 18 to 65 years.'' Error messages should not be too irritating or people will turn them off.

Reduce the time (latency) until the error is detected. The more quickly the error is detected, the easier it is to detect why the error occurred. A quick correction may also reduce possible damage. For example, your car may make a noise if you remove the key while the headlights are on; thus, you can turn the lights off before the battery dies.

9. Improve error detectability.

Consider the error as a signal and the background as noise. Improve detectability by having a high signal/noise ratio—a good contrast. Amplify the signal or reduce the noise.

One way to enhance a signal is to match it. For example, if someone gives you a message, repeat it back so he or she can see that you understand the message. A signal also can be increased. For example, putting a switch on can also trigger an indicator light. The signal should not conflict with population stereotypes (expected relationships) and thus be camouflaged. For example, don't put regular coffee in a decaf container; don't put chemicals in a refrigerator that also contains food.

Reduce the noise by improving the background. Traffic lights are made more visible by surrounding the lights with a black background. When talking on the phone, turn down the radio.

10. Minimize consequences of errors.

Design equipment and procedures so they are less sensitive to errors—so they are fail-safe. A well-designed system anticipates possible problems. What if an instrument light fails? the address on a letter is incorrect? the pilot has a heart attack? the air conditioning system fails? the paint drips from the brush? Some solutions are redundant instrument lights, return addresses on letters, copilots, windows that open, painting while using a dropcloth. The consequences of many errors can be reduced by guards (e.g., guards to prevent hands touching moving pulley and belts, gloves to prevent chemical burns, hard hats to prevent head injuries).

Ease of recovery also is important. What can be done if a box jams on a conveyor turn? if a paycheck is lost? if a mistake is made on a computer input? if a car tire has a blowout? Longer time available for recovery helps. For example, a fire door with a 1 hr rating is better than one with a 0.1 hr rating.

3. GATHERING/ORGANIZING INFORMATION

This section will discuss What to study (Pareto), videotaping jobs, searching for solutions, between-operations analysis, and within-operation analysis.

3.1. What to Study (Pareto)

Engineering time is a valuable resource; don't waste time on unimportant problems. To check quickly whether a project is worth considering, calculate (1) savings/year if material cost is cut 10% and (2) savings/year if labor cost is cut 10%.

The concept of the "insignificant many and the mighty few" (Pareto distribution) is shown in Figure 7. Cause (x-axis) and effect (y-axis) are not related linearly. The key concept is that the bulk of the problem (opportunity) is concentrated in a few items. Pareto diagrams are a special form of a histogram of frequency counts; the key is that the categories are put into order with the largest first and smallest last; then a cumulative curve is plotted. Therefore, using the Pareto concept, if your design concerns crime, it should concentrate on the few individuals who cause most of the crimes; if it is to improve quality, it should concentrate on the few components that cause most of the problems. See Table 13. Fight the giants!

3.2. Videotaping Jobs

Videotaping is useful for task analysis and for training. Some tips are:

- Have spare batteries and a battery charger. A charger may take eight hours to recharge a battery.
- Use a camera feature that continuously displays the date on the screen; this establishes without question when the scene was shot. You may wish also to photo a card with the operator's name.
- Plan the location of camera and subject ahead of time. Use a tripod. Multiple views are best. Use some combination of a front view, a side view, a back view, "stepladder" views (a partial-plan view), overall views, and closeup views. Begin a scene with a full view (far shot, wide-angle view) and then zoom as desired.
- If possible, videotape multiple operators doing the same task. This permits comparisons of methods. Small differences can be detected on tape because the tape can be frozen and/or repeated. For example, how are items oriented? Is the sequence of steps the same? Does one operator deviate joints more or less than other operators?
- If the view is perpendicular to the operator front or side, the projected view can be frozen and angles determined on the screen using a protractor. These angles can then be used as input to computer models.

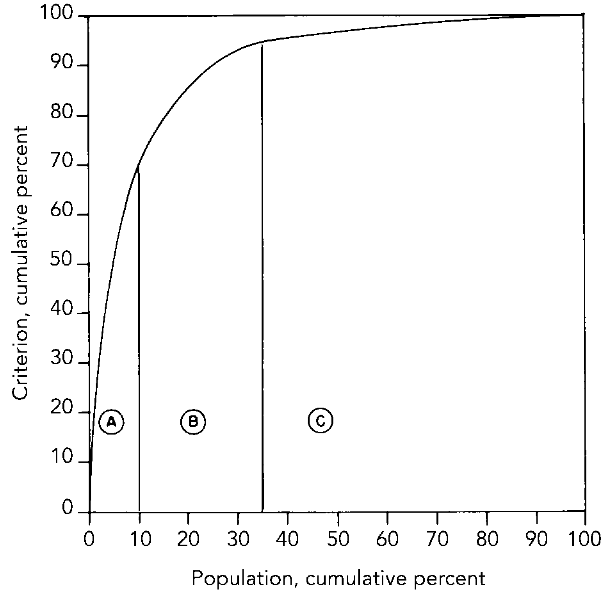


Figure 7 Pareto Distribution. Many populations have a Pareto distribution, in which a small portion of population has a large proportion of the criterion. Inventories, for example, can be classified by the ABC system. "A" items may comprise 10% of the part numbers but 70% of the inventory cost. "B" items may comprise 25% of the part numbers and 25% of the cost; the total of A + B items then comprise 35% of the part numbers and 95% of the cost. "C" items thus comprise the remaining 65% of the part numbers but only 5% of the cost. Concentrate your efforts on the "A" items. Don't use gold cannons to kill fleas! (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

- Take lots of cycles for your stock tape. Each scene should have several cycles. You can always edit, but it is expensive to go back and shoot more tape.
- For analysis of time, there are three alternatives:

1. Put a digital clock in the scene.
2. Use a camera–VCR system with a timecode feature, which will allow you to quickly locate any time on the tape.
3. Count the number of frames (VHS has 30 frames/sec [0.033 sec/frame]).

TABLE 13 Items vs. Opportunities (Pareto Distribution)

Item	Opportunity
A few products	produce most of the direct labor dollars.
products	have most of the storage requirements.
products	produce most of the profit.
operations	produce most of the quality problems.
operations	produce most of the cumulative trauma.
machines	use most of the energy.
time studies	cover most of the direct labor hours.
individuals	commit most of the crimes.
individuals	have most of the money.
individuals	drink most of the beer.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

- During taping, use audio only as a notepad. For management presentations and training, dub in a voice reading a script.
- The VCR should have both a freeze-frame and single-frame advance.
- For task analysis, have the operator as well as the engineer view the tape. The operator can point out to the engineer what is being done and why—perhaps with audio dubbing. The tape also can show the operator some problems and how other operators do the task.
- For a training or management presentation video, you will need to cut and paste your stock tape. For the script, divide a column of paper into two columns labeled audio and visual. In the audio column, enter the words to be read. In the visual column, enter the tape counter reading (start and stop) of the stock videotape. For your final tape, you also will use text shots created in the studio (title page, points you want to emphasize, conclusions, etc.), studio shots of you, and shots of still photos. Then create the final product by blending all the pieces. Perhaps you can even add music!

3.3. Searching for Solutions

Table 14 shows the general procedure for an assessment. Also see Section 4 for the steps of engineering design. The typical sequence is (1) being informed by management that there is a problem, (2), (3), (4) getting information by observation and from the operators and supervisors, and (5) making measurements. Then propose a solution. Rather than depending on inspiration for the solution, most people prefer to follow a more systematic procedure, using tried-and-true techniques established by people who have previously worked on this type of problem.

Searching for an alternative method can be done by following Kipling's ''six honest men'' (who, what, why, when, where, and how), following checklists (see examples in Section 3.5), or using the acronym SEARCH, where:

S = Simplify operations
E = Eliminate unnecessary work and material
A = Alter sequence
R = Requirements
C = Combine operations
H = How often

''Eliminate'' should be done first, then the others in any sequence.

TABLE 14 Procedure for Ergonomics and Productivity Assessment

Source of Information	Method	Data Collected
1. Management	Unstructured interview Collect statistics	Manufacturing measures of productivity, yield throughput, types of defects; job descriptions, injury and absenteeism rates
2. Plant	Ergonomics checklist walk-through	Investigator observes ergonomics and productivity, verified by operator
3. Operators	Unstructured interview Questionnaires Task analyses Videos	Comments on ergonomics, productivity; task analyses; job descriptions
4. First-line supervisors	Unstructured interviews	Current problems in manufacturing process; housekeeping
5. Field measurements	Light, sound meters; tape measures	Ambient environment; workstation dimensions

Reprinted from *International Journal of Industrial Ergonomics,* Vol. 15, M. Helander and G. Burri, ''Cost Effectiveness of Ergonomics and Quality Improvements in Electronics Manufacturing,'' pp. 137–151, copyright © 1995, with permission from Elsevier Science.

If possible, for comparison, obtain measures before and after the change.

E = Eliminate unnecessary work and material. Four subcategories are (1) eliminate unneeded work, (2) eliminate work where costs are greater than benefits, (3) use self-service, and (4) use the exception principle.

Examples of unneeded work are obsolete distribution and mailing lists. Once a year send a letter saying that unless the person completes the form and returns it, that person will no longer receive a copy. An example of costs greater than benefits is the staffing of tool cribs. Self-service at a crib is probably considerably cheaper than having an attendant get the item (consider grocery stores). Another example of self-service is replacing a single mailbox with multiple mailboxes designated ‘‘company—this building,’’ ‘‘company—other buildings,’’ and ‘‘U.S. mail.’’ An example of the exception principle is a reserved parking stall with a strange car in it. Normally the police would give a ticket. Using the exception principle, the police would give a ticket only if the stall owner complained.

S = Simplifying operations is shown by special register keys in fast-food operations. The key indicates a specific item (‘‘Big Mac’’) rather than a price. This reduces pricing errors and improves communication with the cooking area. For e-mail, the return key simplifies finding the return address and entering it.

A = Altering sequence has three subdivisions: (1) simplify other operations, (2) reduce idle/delay time, and (3) reduce material-handling cost.

Modifying when something is done may influence how it is done. For example, machining is easier before a material is hardened than after. On a car assembly line, installing a brake cylinder assembly is easier before the engine is installed. Idle/delay time often can be reduced. For example, in a restaurant, the server can bring coffee when bringing the breakfast menu instead of making two trips. Or the idle time might be used fruitfully by double tooling. Consider having two fixtures on a machine so that while the machine is processing the material on fixture 1, the operator is loading/unloading fixture 2. An example of reducing material-handling costs is using an automatic guided vehicle (bus) instead of moving items on a fork truck (taxi).

R = Requirements has two aspects (1) quality (capability) costs and (2) initial vs. continuing costs.

Costs rise exponentially (rather than linearly) vs. quality. Therefore, do not ‘‘goldplate’’ items. Indirect materials are a fruitful area to investigate because they tend not to be analyzed. For example, one firm compared the varnish applied on motor coils vs the varnish purchased; there was a 10% difference. It was found that the varnish was bought in 55-gallon drums. When the drums were being emptied, they were turned right side up when there was still considerable product inside. The solution was a stand that permitted the varnish to drip out of the drum over a period of hours.

Initial vs. continuing costs means you should focus not just on the initial capital cost of a product but on the life-cycle costs (capital cost plus operating cost plus maintenance cost).

C = Combine operations is really a discussion of general purpose vs. special purpose. For example, do you want to have a maintenance crew with one specialist doing electrical work, another doing plumbing, and another doing carpentry? Or three people, each of whom can do all three types of work? Most firms now are going for the multi-skilled operator because it is difficult to find sufficient work to keep all specialists always busy.

H = How often is a question of the proper frequency. For example, should you pick up the mail once a day or four times a day? Should solution pH be tested once an hour, once a day or once a week?

3.4. Between-Operations Analysis

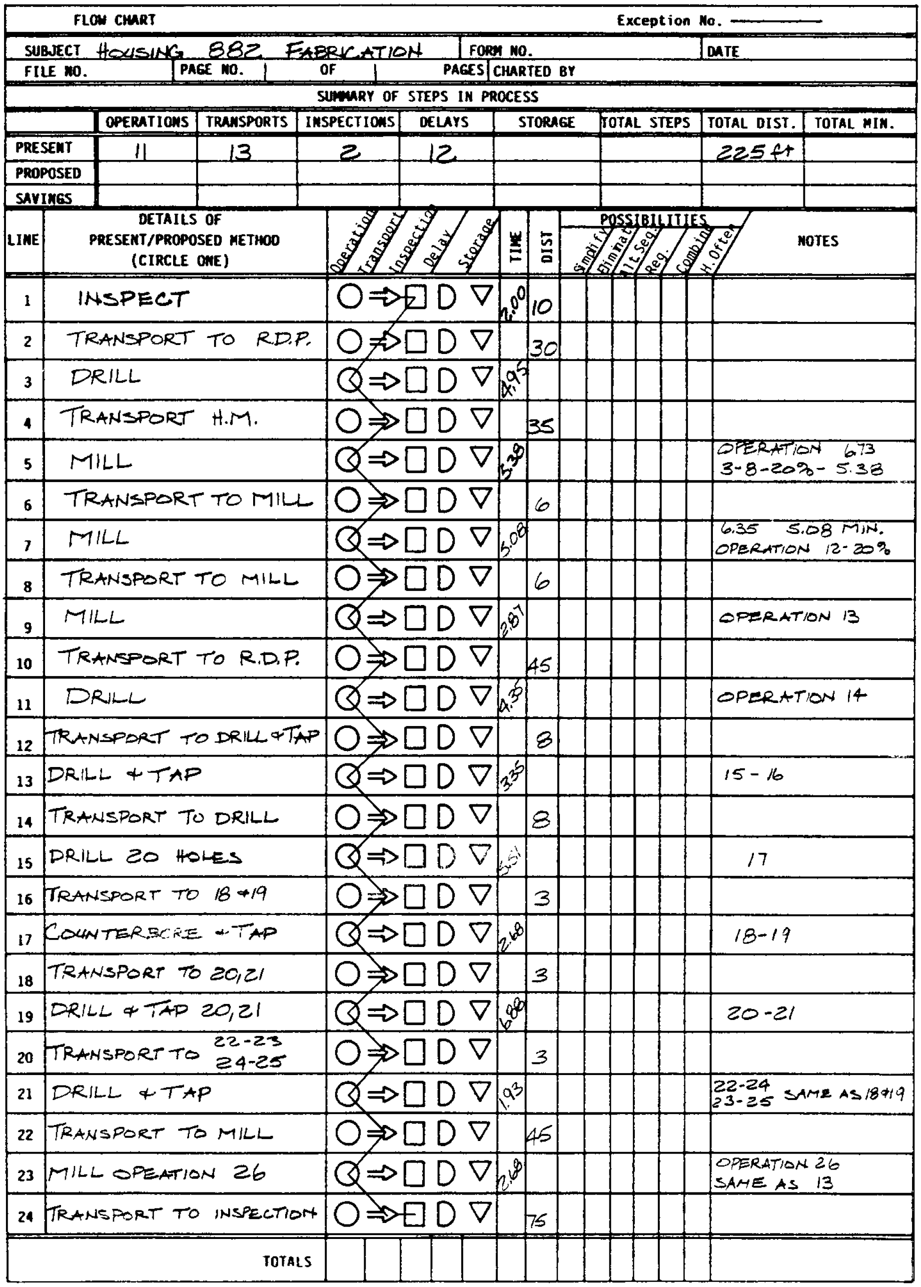
This section will discuss flow diagrams, multiactivity charts, arrangement (layout) of equipment, and balancing flow lines.

3.4.1. Flow Diagrams

Flow diagrams and their associated process charts are a technique for visually organizing and structuring an overview (‘‘mountaintop’’ view) of a between-workstations problem. There are three types: single object, assembly/disassembly, and action–decision.

3.4.1.1. Single Object Figure 8 shows a single-object process chart following a housing in a machine shop; the single object also can be a person. Some examples of following a person are vacuuming an office and unloading a semitrailer.

Figure 9 shows the five standard symbols for process charts. Some people put a number inside each symbol (identifying operation 1, 2, 3) and some don’t. Some people emphasize ‘‘do’’ operations by darkening those circles (but not get-ready and put-away operations) and some don’t. Since a process chart is primarily a communication tool for yourself, take your choice. Most operations have scrap and rework; Figure 10 shows how to chart them. At the end of the chart, summarize the number of operations, moves (including total distance), inspections, storages and delays. Estimate times for storages (which are planned) and delays (which are unplanned storages). Since this is a big-picture analysis, detailed times for operations and inspections normally are not recorded.



FLOW CHART Exception No. ———

SUBJECT HOUSING 882 FABRICATION		FORM NO.		DATE
FILE NO.	PAGE NO.	OF	PAGES	CHARTED BY

SUMMARY OF STEPS IN PROCESS

	OPERATIONS	TRANSPORTS	INSPECTIONS	DELAYS	STORAGE	TOTAL STEPS	TOTAL DIST.	TOTAL MIN.
PRESENT	11	13	2	12			225 ft	
PROPOSED								
SAVINGS								

LINE	DETAILS OF PRESENT/PROPOSED METHOD (CIRCLE ONE)	Operation / Transport / Inspection / Delay / Storage	TIME	DIST	Simplify	Eliminate	Alt.Seq.	Req.	Combine	H.Often	NOTES
1	INSPECT	○ ⇨ [✓ □] D ▽	2.00	10							
2	TRANSPORT TO R.D.P.	○ [✓ ⇨] □ D ▽		30							
3	DRILL	[✓ ○] ⇨ □ D ▽	4.95								
4	TRANSPORT H.M.	○ [✓ ⇨] □ D ▽		35							
5	MILL	[✓ ○] ⇨ □ D ▽	5.38								OPERATION 6 73 3-8-20%- 5.38
6	TRANSPORT TO MILL	○ [✓ ⇨] □ D ▽		6							
7	MILL	[✓ ○] ⇨ □ D ▽	5.08								6.35 5.08 MIN. OPERATION 12-20%
8	TRANSPORT TO MILL	○ [✓ ⇨] □ D ▽		6							
9	MILL	[✓ ○] ⇨ □ D ▽	2.87								OPERATION 13
10	TRANSPORT TO R.D.P.	○ [✓ ⇨] □ D ▽		45							
11	DRILL	[✓ ○] ⇨ □ D ▽	4.35								OPERATION 14
12	TRANSPORT TO DRILL + TAP	○ [✓ ⇨] □ D ▽		8							
13	DRILL + TAP	[✓ ○] ⇨ □ D ▽	3.35								15 - 16
14	TRANSPORT TO DRILL	○ [✓ ⇨] □ D ▽		8							
15	DRILL 20 HOLES	[✓ ○] ⇨ □ D ▽	3.51								17
16	TRANSPORT TO 18 + 19	○ [✓ ⇨] □ D ▽		3							
17	COUNTERBORE + TAP	[✓ ○] ⇨ □ D ▽	2.68								18-19
18	TRANSPORT TO 20,21	○ [✓ ⇨] □ D ▽		3							
19	DRILL + TAP 20,21	[✓ ○] ⇨ □ D ▽	6.88								20-21
20	TRANSPORT TO 22-23 24-25	○ [✓ ⇨] □ D ▽		3							
21	DRILL + TAP	[✓ ○] ⇨ □ D ▽	1.93								22-24 23-25 SAME AS 18+19
22	TRANSPORT TO MILL	○ [✓ ⇨] □ D ▽		45							
23	MILL OPEATION 26	[✓ ○] ⇨ □ D ▽	2.68								OPERATION 26 SAME AS 13
24	TRANSPORT TO INSPECTION	○ [✓ ⇨] □ D ▽		75							
	TOTALS										

Figure 8 Single-Object Process Chart. The chart follows a single object or person—in this example, an object. For good analysis, estimate distances and times. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

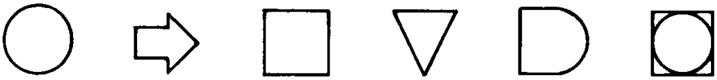


Figure 9 Five Standard Symbols for Process Charts. They are a circle (operation), arrow (move), square (inspect), triangle with point down (storage) and D (delay). A circle inside a square is a combined operation and inspection. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

Next you would consider possible improvements using Kipling's six honest men (who, what, where, why, when and how), SEARCH, and various checklists. This may lead to a proposed method where you would show reductions in number of operations, inspections, storages, distance moved, and so on. Because flow diagrams organize complex information and thus are a useful communication tool, you may want to make a polished copy of the flow diagrams to show to others. Since many potential improvements involve reducing movements between workstations, a map showing the movements is useful; Figure 11 shows a flow diagram of Joe College making a complex job of drinking beer in his apartment. Generally a sketch on cross-section paper is sufficient.

3.4.1.2. Assembly/Disassembly Diagrams Assembly flow diagrams (see Figure 12) tend to point out the problems of disorganized storage and movements. Figure 13 shows a disassembly diagram; another example would be in packing houses.

3.4.1.3. Action–Decision Diagrams In some cases, you may wish to study decision making; then consider an action–decision diagram (see Figure 14) or a decision structure table (see Section 3.5.2).

3.4.2. Multiactivity Charts

The purpose of a multiactivity chart is to improve utilization of multiple related activities. See Figure 15. The two or more activities (column headings) can be people or machines; they can also be parts of a person (left hand vs. right hand) or parts of a machine (cutting head vs. fixture 1 vs. fixture 2). The time axis (drawn to a convenient scale) can be seconds, minutes, or hours.

Example charts and columns might be milling a casting (operator, machine), cashing checks (cashier, customer 1, customer 2), and serving meals (customer, server, cook).

For each column, give cycles/year, cost/minute, and percent idle. Make the idle time distinctive by cross-hatching, shading, or coloring red.

Improve utilization by (1) reducing idle time in a column, (2) shifting idle time from one column to another, or (3) decreasing idle time of an expensive component (such as person) by increasing idle time of a machine. McDonald's uses this third concept when they have two dispensers of Coke in order to reduce the waiting time of the order taker (and thus the customer).

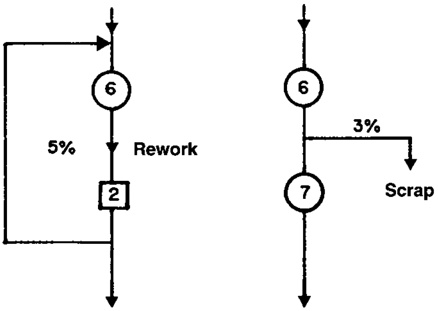


Figure 10 Rework (left) and Scrap (right) for Flow Diagrams. Rework and scrap often are more difficult to handle than good product. Even though rework is often concealed, most items have rework. Material-removal scrap occurs in machining and press operations; scrap units occur everywhere. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

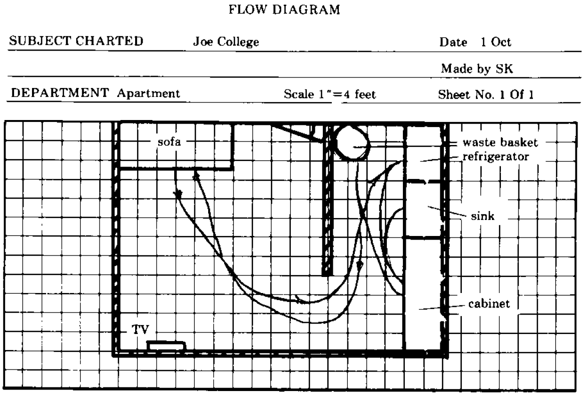

FLOW DIAGRAM
SUBJECT CHARTED Joe College
Date 1 Oct
Made by SK
DEPARTMENT Apartment
Scale 1"=4 feet
Sheet No. 1 Of 1
sofa
waste basket
refrigerator
sink
cabinet
TV

Figure 11 Flow Diagram. This is a map often used with process charts. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

Some time is considered to be free. For example, "inside machine time" describes the situation in which an operator is given tasks while the machine is operating. Since the operator is at the machine anyway, these tasks are "free." A McDonald's example is to grasp a lid while the cup is filling.

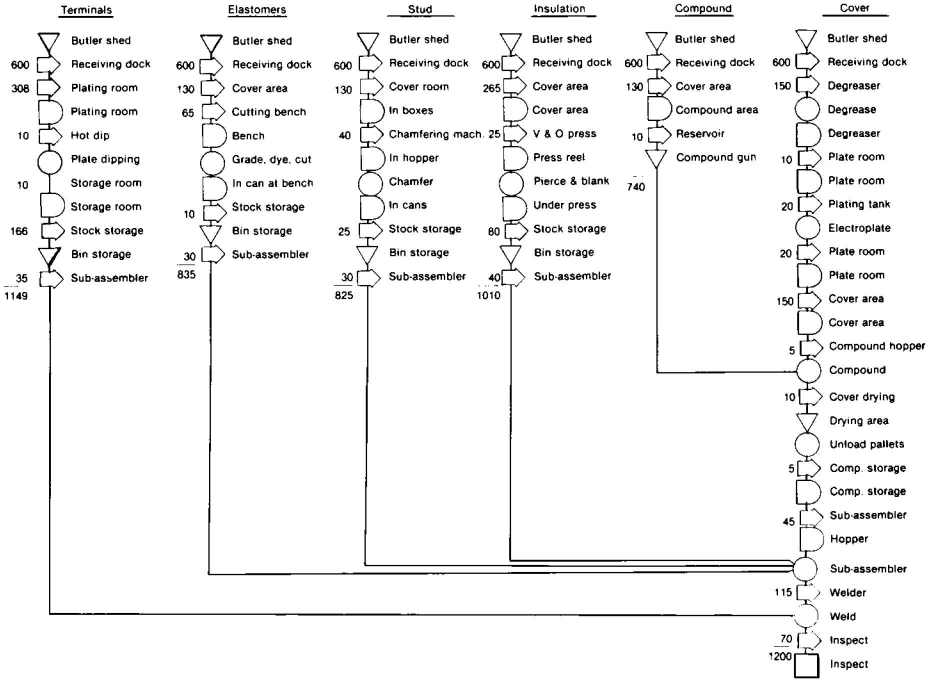

Terminals
Butler shed
600 Receiving dock
308 Plating room
Plating room
10 Hot dip
Plate dipping
10 Storage room
Storage room
166 Stock storage
Bin storage
35 Sub-assembler
1149
Elastomers
Butler shed
600 Receiving dock
130 Cover area
65 Cutting bench
Bench
Grade, dye, cut
In can at bench
10 Stock storage
Bin storage
30 Sub-assembler
835
Stud
Butler shed
600 Receiving dock
130 Cover room
In boxes
40 Chamfering mach.
In hopper
Chamfer
In cans
25 Stock storage
Bin storage
30 Sub-assembler
825
Insulation
Butler shed
600 Receiving dock
265 Cover area
Cover area
25 V & O press
Press reel
Pierce & blank
Under press
80 Stock storage
Bin storage
40 Sub-assembler
1010
Compound
Butler shed
600 Receiving dock
130 Cover area
Compound area
10 Reservoir
Compound gun
740
Cover
Butler shed
600 Receiving dock
150 Degreaser
Degrease
Degreaser
10 Plate room
Plate room
20 Plating tank
Electroplate
20 Plate room
Plate room
150 Cover area
Cover area
5 Compound hopper
Compound
10 Cover drying
Drying area
Unload pallets
5 Comp. storage
Comp. storage
45 Sub-assembler
Hopper
Sub-assembler
115 Welder
Weld
70 Inspect
1200 Inspect

Figure 12 Assembly Process Chart. This type of chart shows relationships among components and helps emphasize storage problems. Each column is an item; the assembly is on the right. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

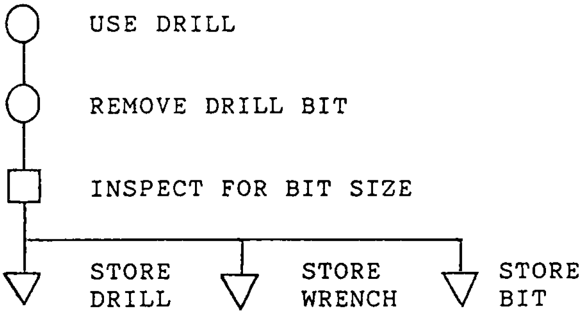
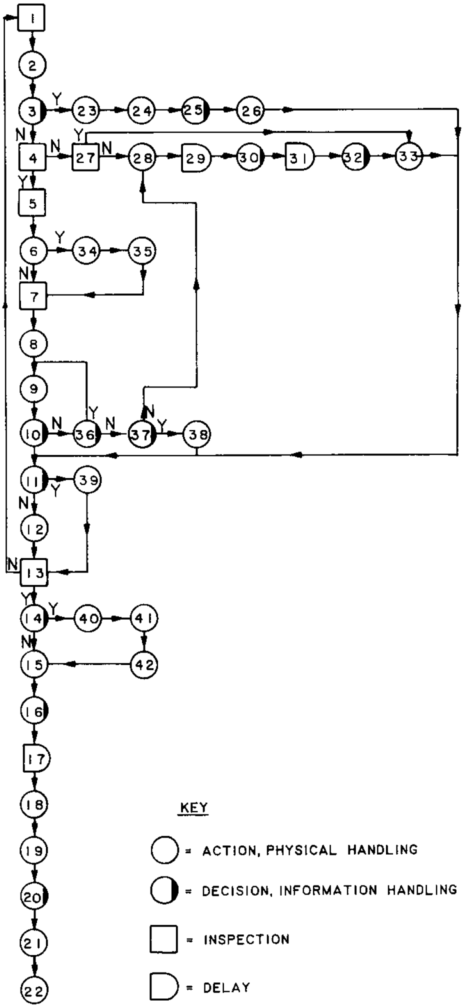


Figure 13 Disassembly Process Chart. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)



1. Search items
2. Pick up item with right hand (both hands if large)
3. Is the item to be weighed? Yes/No
4. Inspect for code Yes/No
5. Inspect for identical items amongst load
6. Are number of similar items greater than 2? Yes/No
7. Inspect for positioning of code
8. Pass item to left hand
9. Pass item over scanner
10. Was scan successful? Yes/No
11. Is the item fragile or vulnerable? Yes/No
12. Place item on the second conveyor
13. Inspect—was this the last item? Yes/No
14. Does customer require a carrier bag? Yes/No
15. Press total
16. Tell customer the cost
17. Wait for payment
18. Receive payment
19. Key in amount received
20. Determine change
21. Cash money, give change
22. Obtain and give receipt
23. Transfer item to left hand
24. Put item on scales
25. Obtain price/kilo
26. Key in price/kilo with right hand
27. Inspect for price mark Yes/No
28. Press intercom button
29. Wait (continue working)
30. When buzzer sounds give the product details
31. Wait (continue working)
32. Receive price information
33. Key in price with right hand
34. Key in number of identical items
35. Place remainder of identical items on second conveyor
36. Decide whether to scan again
37. Read code? Yes/No
38. Key in code
39. Place on ledge in front of customer
40. Reach for carrier bag
41. Give bag to customer
42. Key in price of bag

Figure 14 Action–Decision Flow Diagram. This chart of a seated checkout operator in England shows use of a laser scanner. Note that in Europe the customer bags the groceries. (From Wilson and Grey 1984)

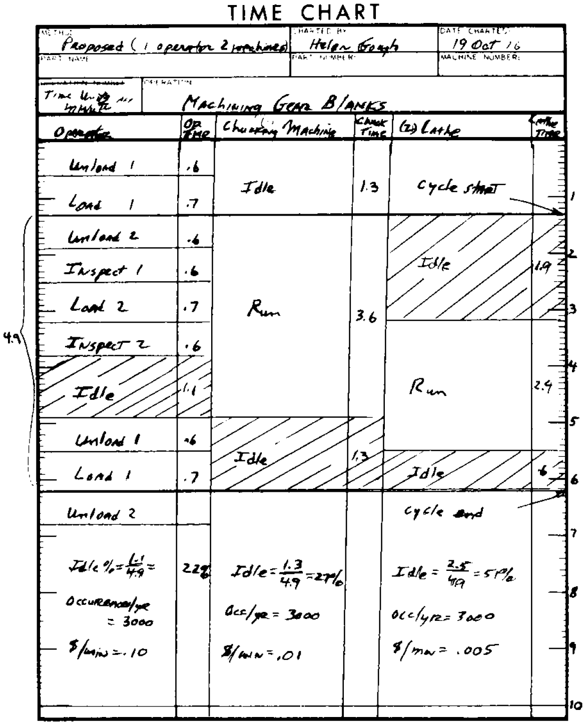

TIME CHART
METHOD
Proposed (1 operator 2 machines)
CHARTED BY
Helen Gough
DATE CHARTED
19 Oct 16
PART NAME
PART NUMBER
MACHINE NUMBER
Time Units are minutes
OPERATION
Machining Gear Blanks
Operator
Op Time
Chucking Machine
Chuck Time
(2) Lathe
Lathe Time
Unload 1
.6
Load 1
.7
Idle
1.3
Cycle start
Unload 2
.6
Inspect 1
.6
Load 2
.7
Inspect 2
.6
Idle
1.1
Run
3.6
Idle
1.9
Run
2.9
Unload 1
.6
Load 1
.7
Idle
1.3
Idle
.6
Unload 2
Cycle end
Idle % = 1.1/4.9 =
22%
Occurrences/yr = 3000
$/min = .10
Idle = 1.3/4.9 = 27%
Occ/yr = 3000
$/min = .01
Idle = 2.5/4.9 = 51%
Occ/yr = 3000
$/min = .005
4.9
1
2
3
4
5
6
7
8
9
10

Figure 15 Multiactivity Chart. The basic concept is to have two or more activities (columns) with a common, *scaled,* time axis. The goal is to improve utilization, so emphasis is placed on idle time. For each column, calculate the percent idle time, the occurrences/year, and the cost/minute. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

As pointed out above in (3), it is possible to have more than one machine/operator—this is double tooling. For example, use two sets of load/unload fixtures on an indexing fixture to reduce idle time while waiting for the machine to cut. Or one person can service two machines (i.e., 1 operator/2 machines or 0.5 operator/machine). It is also possible to have 0.67 or 0.75 operators/machine. Do this by assigning 2 people/3 machines or 3/4, that is, having the workers work as a team, not as individuals. This will require cross-trained operators, which is very useful when someone is absent.

Kitting is the general strategy of gathering components before assembly to minimize search-and-select operations and ensure that there are no missing parts. Subassemblies also may be useful. McDonald's Big Mac has a "special sauce," which is just a mixture of relish and mayonnaise, premixed to ensure consistency.

A disadvantage of the multiactivity chart is that it requires a standardized situation. Nonstandardized situations are difficult to show. For them, use computer simulation.

3.4.3. Arrangement (Layout) of Equipment

This section describes location of one item in an existing network. Arrangement of the entire facility is covered in Chapter 55.

Problem. Table 15 shows some examples of locating one item in an existing network of customers. The item can be a person, a machine, or even a building. The network of customers can be people, machines, or buildings. The criterion minimized is usually distance moved but could be energy lost or time to reach a customer.

Typically an engineer is interested in finding the location that minimizes the distance moved—a minisum problem. An example would be locating a copy machine in an office. An alternative objec-

TABLE 15 Examples of Locating an Item in an Existing Network of Customers with Various Criteria to be Minimized

New Item	Network of Customers	Criterion Minimized
Machine tool	Machine shop	Movement of product
Tool crib	Machine shop	Walking of operators
Time clock	Factory	Walking of operators
Inspection bench	Factory	Movement of product or inspectors
Copy machine	Office	Movement of secretaries
Warehouse or store	Market	Distribution cost
Factory	Warehouses	Distribution cost
Electric substation	Motors	Power loss
Storm warning siren	City	Distance to population
IIE meeting place	Locations of IIE members	Distance traveled
Fire station	City	Time to fire

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

tive is to minimize the maximum distance—a minimax or worst-case problem. An example minimax problem would be to locate an ambulance so everyone could be reached in no more than 15 minutes. There is extensive analytical literature on this "planar single-facility location problem"; see Francis et al. (1992, chap. 4).

The following is not elegant math but "brute force" calculations. The reason is that, using a hand calculator, the engineer can solve all reasonable alternatives in a short time—say 20 minutes. Then, using the calculated optimum for the criterion of distance, the designer should consider other criteria (such as capital cost, maintenance cost) before making a recommended solution.

Solution. The following example considers (1) the item to be located as a machine tool, (2) the network of customers (circles in Figure 16) as other machine tools with which the new tool will exchange product, and (3) the criterion to be minimized as distance moved by the product.

In most real problems, there are only a few possible places to put the new item; the remaining space is already filled with machines, building columns, aisles, and so forth. Assume there are only two feasible locations for the new item—the A and B rectangles in Figure 16.

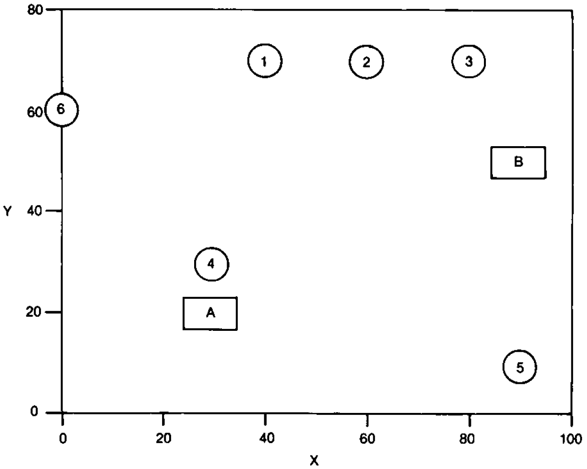


Figure 16 Information for Location of One Item. Customers (in circles) can be served from either location A or B. Table 17 gives the importance of each customer. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

Travel between a customer and A or B can be (1) straight line (e.g., conveyors), (2) rectangular (e.g., fork trucks using aisles), or (3) measured on a map (e.g., fork trucks using one-way aisles, conveyors following nondirect paths). Travel may be a mixture of the three types.

Some customers are more important than others, and thus the distance must be weighted. For a factory, an index could be pallets moved per month. For location of a fire station, the weight of a customer might depend on number of people occupying the site or the fire risk.

The movement cost of locating the new item at a specific feasible location is:

$$\text{MVCOST} = \text{WGTK (DIST)}$$

where MVCOST = index of movement cost for a feasible location
WGTK = weight (importance) of the Kth customer of N customers
DIST = distance moved

For rectangular:

$$\underset{K=1}{\overset{N}{\text{MVCOST}}} = (|X_{ij} - X_k| + |Y_{ij} - Y_k|)$$

For straight line:

$$\underset{K=1}{\overset{N}{\text{MVCOST}}} = \Sigma \sqrt{(X_{i,j} - X_k)^2 + (Y_{i,j} - Y_k)^2}$$

For the two locations given in Table 16, Table 17 shows the MVCOST. Movement cost at B is 67,954/53,581 = 126% of A.

If you wish to know the cost at locations other than A and B, calculate other locations and plot a contour map. A contour map indicates the best location is $X = 42$ and $Y = 40$ with a value of 32,000. Thus, site A is 6,000 from the minimum and site B is 13,000 from the minimum.

The previous example, however, made the gross simplification that movement cost per unit distance is constant. Figure 17 shows a more realistic relationship of cost vs. distance. Most of the cost is fixed (loading/unloading, starting/stopping, paperwork). Cost of moving is very low. Thus:

For rectangular:

$$\text{DIST} = L_k + C_k\,(|X_{i,j} - X_k| + |Y_{i,j} - Y_k|)$$

For straight line:

$$\text{DIST} = L_k + C_k \sqrt{(X_{i,j} - X_k)^2 + (Y_{i,j} - Y_k)^2}$$

where L_k = load + unload cost (including paperwork) per trip between the Kth customer and the feasible location

TABLE 16 Importance of Each Customer
Customers 1 to 6 can be served either from location A ($X = 30$, $Y = 20$) or from location B ($X = 90$, Y = 50). Which location minimizes movement cost?

Customer	Coordinate		Weight or Importance	Movement Type
	X	*Y*		
1	40	70	156	Straight line
2	60	70	179	Straight line
3	80	70	143	Straight line
4	30	30	296	Rectangular
5	90	10	94	Rectangular
6	0	60	225	Rectangular

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

TABLE 17 Cost of Locating a New Machine at Location A or B

		Site A		Site B	
Customer	Weight, Pallets/ Month	Distance, Meters	Cost, m-Pallets/ Month	Distance	Cost, m-Pallets/ Month
1	156	51	7,956	54	8,424
2	179	58	10,382	36	6,444
3	143	71	10,153	22	3,146
4	296	10	2,960	80	23,680
5	94	70	6,580	40	3,760
6	225	70	15,750	100	22,500
			53,781		67,954

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

Since WGTK was pallets/month and DIST was in meters, MVCOST = meter-pallets/month.

$$C_k = \text{cost/unit distance (excluding } L_k)$$

Assume, for customers 1, 2, and 3, that L_k = \$0.50/trip and C_k = \$0.001/m; for customers 4, 5, and 6, L_k = \$1/trip and C_k = \$0.002/m. Then the cost for alternative A = \$854 + \$79.07 = \$933.07 while the cost for B = \$854 + \$117.89 = \$971.89. Thus, B has a movement cost of 104% of A. When making the decision where to locate the new item, use not only the movement cost but also installation cost, capital cost, and maintenance cost. Note that the product (WGTH)(DIST) (that is, the \$854) is independent of the locations; it just adds a constant value to each alternative.

Cost need not be expressed in terms of money. Consider locating a fire station where the customers are parts of the town and the weights are expected number of trips in a 10-year period. Then load might be 1 min to respond to a call, travel 1.5 min/km, and unload 1 min; the criterion is to minimize mean time/call.

The distance cost might rise by a power of 2 (the inverse square law) for problems such as location of a siren or light.

3.4.4. Balancing Flow Lines

The two most common arrangements of equipment are a job shop and a flow line (often an assembly line). The purpose of balancing a flow line is to minimize the total idle time (balance delay time). There are three givens: (1) a table of work elements with their associated times (see Table 18), (2) a precedence diagram showing the element precedence relationships (see Figure 18), and (3) required

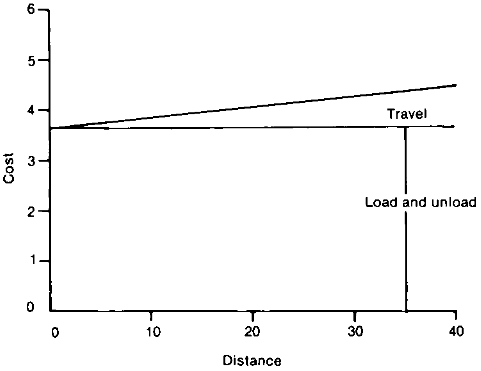


Figure 17 Relationship of Cost vs. Distance. In general, material-handling cost is almost independent of distance. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

TABLE 18 Elements and Work Times for the Assembly-Line Balancing Problem

Element	Work Time/Unit, hr
1	0.0333
2	0.0167
3	0.0117
4	0.0167
5	0.0250
6	0.0167
7	0.0200
8	0.0067
9	0.0333
10	0.0017
	0.1818

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

Each element time is assumed constant. In practice, each element time is a distribution.

units/minute from the line. The three unknowns are (1) the number of stations, (2) the number of workers at each station, and (3) the elements to be done at each station.

1. What is the total number to be made, and in how long a time? For example, 20,000 units could be made in 1000 hr at the rate of 20/hr, 500 hr at 40/hr, or some other combination. Continuous production is only one of many alternatives; a periodic shutdown might be best. Assume we wish to make 20,000 units in 1000 hr at the rate of 20/hr. Then each station will take 1000 hr/20,000 units = 0.05 hr/unit; cycle time is .05 hr.
2. Guess an approximate number of workstations by dividing total work time by cycle time: 0.1818 hr/0.05 hr/station = 3.63 stations. Then use 4 workstations with 1 operator at each.
3. Make a trial solution as in Table 19 and Figure 19. Remembering not to violate precedence, identify each station with a cross-hatched area. Then calculate the idle percentage: 0.0182/(4 × 0.05) = 9.1%.

But this can be improved. Consider Table 20. Here stations 1 and 2 are combined into one superstation with two operators. Elemental time now totals 0.0950. Since there are two operators, the

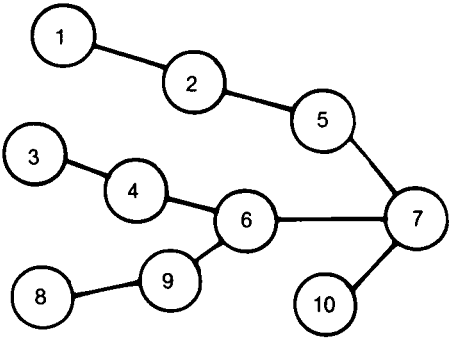


Figure 18 Precedence Diagram Showing the Sequence Required for Assembly. The lines between the circles are not drawn to scale. That is, elements 4 and 9 both must be competed before 6, but 9 could be done before or after 4. Precedence must be observed. Thus, elements 3, 4, and 9 could not be assigned to one station and elements 8 and 6 to another. However, 8, 9, and 10 could be done at one station. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

TABLE 19 Trial Solution for Assembly-Line Balance Problem

Station	Element	Element Time, hr/unit	Station Element Time, hr/unit	Station Idle Time, hr/unit	Cumulative Idle Time, hr/unit
1	1	0.0333			
	2	0.0167	0.0500	0	0
2	8	0.0067			
	9	0.0333	0.0400	0.0100	0.0100
3	3	0.0117			
	4	0.0167			
	6	0.0167	0.0451	0.0049	0.0149
4	5	0.0250			
	7	0.0200			
	10	0.0017	0.0467	0.0033	0.0182
Idle percent = 0.0182/(4 × 0.05)= 9.1%					

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.
Cycle time = 0.0500 hr.

idle time/operator is 0.0025. So far there is no improvement over the solution of Table 19. *But there is idle time at each station.* Thus, the line cycle time can be reduced to 0.0475. The new idle time becomes 0.0083/(4 × 0.0475) = 4.4% instead of 9.1%. The point is that small changes in cycle time may have large benefits.

Small modifications of other "rigid facts" can be beneficial.

Consider element sharing. That is, operators/station need not be 1.0. One possibility is more than 1 operator/station. Two operators/station yields 2.0 operators/station. Three operators/two stations yields 1.5 operators/station. This permits a cycle time that is less than an element time. For example, with 2 operators/station, each would do every other unit.

Get fewer than 1 operator/station by having operators walk between stations or having work done off-line (i.e., using buffers). Also, operators from adjacent stations can share elements. For example, station D does half of element 16 and 17 and station E does half of element 16 and 17.

Remember that cycle times are not fixed. We assumed a cycle time of 0.05 hr (i.e., the line runs 1000 hr or 1000/8 = 125 days). As pointed out above, it is more efficient to have cycle time of

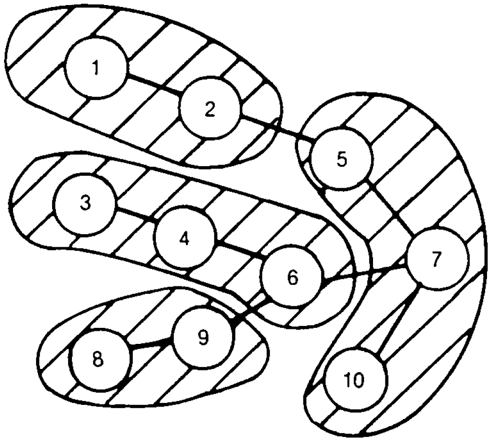


Figure 19 Graphic Solution of Table 19. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission)

TABLE 20 Trial Solution for Assembly-Line Balance Problem

Station	Number of Operators	Element	Element Time, hr	Work Time, hr	Idle Time, hr	Cumulative Idle Time, hr
1	2	1	0.0333			
		2	0.0167			
		3	0.0167			
		4	0.0167			
		6	0.0166	0.0950	0.0000	0.0000
2	1	8	0.0067			
		9	0.0333	0.0400	0.0075	0.0075
3	1	5	0.0250			
		7	0.0200			
		10	0.0017	0.0467	0.0008	0.0083

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.
Cycle time = 0.0475 hr; two operators at station 1.

0.0475 hr (i.e., 20,000 units × 0.0475 = 950 hr = 950/8 = 118.75 days). What is the best combination of setup cost, balance delay cost, and inventory carrying cost?

Elements can be redefined. For example, elements 16, 17, and 18 may be redefined to eliminate element 17 by allocating all of 17 to element 16 or to element 18 or to allocate a portion to element 16 or 18.

Reconsider the allocation of elements to subassemblies and assembly. For example, a nameplate might be added at a subassembly instead of at the assembly station.

3.5. Within-Operation Analysis

This section will discuss fish diagrams, decision structure tables, and checklists.

3.5.1. Fish Diagrams

Fish diagrams graphically depict a multidimensional list. See Figures 20 and 21. Professor Ishikawa developed them while on a quality control project for Kawasaki Steel; they are the "cause" side of cause-and-effect diagrams. The diagram gives an easily understood overview of a problem.

Start with the "effect" (a fishhead), a specific problem. Then add the "cause" (the fish body), composed of the backbone and other bones. A good diagram will have three or more levels of bones (backbone, major bones, minor bones on the major bones). There are various strategies for defining the major bones. Consider the 4 M's: manpower, machines, methods, materials. Or use the 4 P's: policies, procedures, people, plant. Or be creative.

Figures 20 and 21 were used at Bridgestone to reduce the variability of the viscosity of the splicing cement used in radial tires. Figure 21 shows the variability of the four operators before and after the quality circle studied the problem.

A very effective technique is to post the fish diagram on a wall near the problem operation. Then invite everyone to comment on possible problems and solutions; many quality problems can be reduced with better communication.

3.5.2. Decision Structure Tables

Decision structure tables are a print version of "if, then" statements (the "what if" scenario) in computer programs. They also are known as protocols or contingency tables; see Table 21 for an example. They unambiguously describe complex, multivariable, multirule decision systems. A spreadsheet is a computerized decision structure table.

Decision structure tables give better quality decisions due to (1) better decision analysis (higher-quality personnel make the decision, using complex analysis techniques if necessary) and (2) less time pressure at the time the decision is made. They are a game plan worked out in advance, not in the heat of battle. However, in addition, decision structure tables force a good methods analysis because all possibilities are considered systematically. The tables also are good training aids.

3.5.3. Checklists

Checklists allow a novice to use the distilled expertise of others. Table 6 is a checklist to prioritize potential ergonomic problems; it focuses on force, duration, and repetition for various body parts.

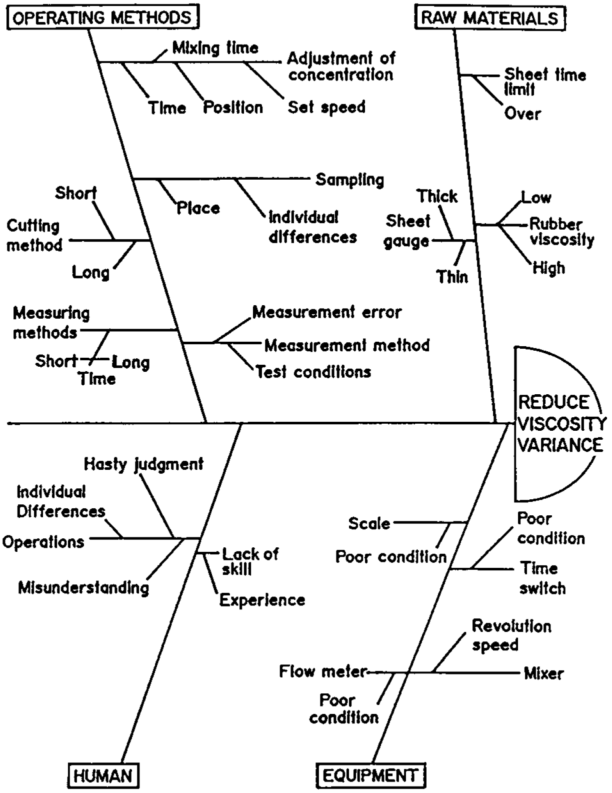

OPERATING METHODS
Mixing time
Adjustment of concentration
Time
Position
Set speed
Sampling
Place
Individual differences
Short
Cutting method
Long
Measuring methods
Short
Long
Time
Measurement error
Measurement method
Test conditions
RAW MATERIALS
Sheet time limit
Over
Thick
Sheet gauge
Thin
Low
Rubber viscosity
High
REDUCE VISCOSITY VARIANCE
Hasty judgment
Individual Differences
Operations
Misunderstanding
Lack of skill
Experience
Scale
Poor condition
Poor condition
Time switch
Revolution speed
Flow meter
Mixer
Poor condition
HUMAN
EQUIPMENT

Figure 20 Fish Diagrams Used at Bridgestone Tire. The goal was to reduce viscosity variance. The four main categories were operating methods, raw materials, humans, and equipment. The variance was reduced when it was discovered that not all operators followed the standard method; those following the standard had *more* variance. See Figure 21. (From R. Cole, *Work, Mobility, and Participation: A Comparative Study of American and Japanese Industry.* Copyright © 1979 The Regents of the University of California. By permission)

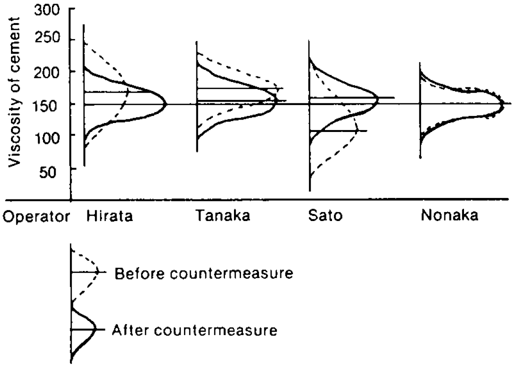


Figure 21 Distribution Curves for Viscosity Variance before and after Quality Circle Project. Reducing variance (noise) is an excellent quality-improvement technique because the effects of various changes are seen much more clearly for processes with little variability. (From R. Cole, *Work, Mobility, and Participation: A Comparative Study of American and Japanese Industry.* Copyright © 1979 The Regents of the University of California. By permission)

TABLE 21 Example of a Decision Structure Table to Select Drill Size (National Special Thread Series) before Tapping or Clearance Drill Size

If Bolt Diameter (in.) Is	And Threads Per Inch Is	Then Drill Size For 75% Thread Is	Then Clearance Drill Bit Size Is
1/16	64	3/64	51
5/64	60	1/16	45
3/32	48	49	40
7/64	48	43	32

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.
Assume you wish a hole large enough so the drill will not touch the bolt threads.

Checklists can also focus more precisely. Table 7 is a checklist for just the upper extremity. Table 8 is a checklist to prioritize potential ergonomic problems due to posture; it focuses on force and duration for neck, trunk, and general body.

4. ENGINEERING DESIGN

After reviewing the job design guidelines from Section 2 and the information from Section 3, you can design alternatives. Remember the five steps of engineering design with the acronym DAMES: Define the problem, Analyze, Make search, Evaluate alternatives, and Specify and sell solution. See Table 22.

1. Define the problem broadly.

Usually the designer is not given the problem but instead is confronted with the current solution. The current solution is not the problem but just one solution of among many possible solutions. The broad, detail-free problem statement should include the number of replications, the criteria, and the schedule. At this stage, putting in too much detail makes you start by defending your concept rather than opening minds (yours and your clients') to new possibilities. At this stage, the number of replications should be quite approximate (within ±500%). Criteria (see Section 2.1) usually are multiple (capital cost, operating cost, quality) rather than just one criterion. The schedule defines priorities and allocation of resources for the project.

2. Analyze in detail.

Now amplify step 1 (defining the problem) with more detail on replications, criteria, and schedule.

- What are the needs of the users of the design (productivity, quality, accuracy, safety, etc.)? See Sections 2.1, 2.2, 2.3, 2.4, 2.5, and 2.6.
- What should the design achieve?
- What are limits (also called constraints and restrictions) to the design?
- What are the characteristics of the population using the design? For example, for designing an office workstation, the users would be adults within certain ranges (age from 18 to 65, weight from 50 to 100 kg, etc.).

The design should consider not only the main activities (the ''do,'' e.g., the assembly) but also the ''get-ready'' and ''put-away'' and support activities such as setup, repairs, maintenance, material handling, utilities, product disposal, and user training.

Since people vary, designers can follow two alternatives: (a) Make the design with fixed characteristics; the users adjust to the device. One example would be an unadjustable chair. Another example would be a machine-paced assembly line. (b) Fit the task to the worker. One example would be an adjustable chair. Another example would be a human-paced assembly line (i.e., with buffers) so all workers could work at individual paces.

3. Make search of the solution space.

Now design a number of alternatives—not just one.

Now use the information you gathered using the techniques of Section 3. Also get design ideas from a number of sources: workers, supervisors, staff people, other engineers, vendors, suppliers, and so on. (Benchmarking is a technique of obtaining ideas from other organizations.) The solution space will be reduced by various economic, political, aesthetic, and legal constraints. Among the feasible solutions (solutions that work), try to select the best one—the optimum solution.

TABLE 22 DAMES: The Five Steps of Engineering Design (Define, Analyze, Make search, Evaluate, Specify and sell)

Step	Comments	Example
Define the problem broadly.	Make statement broad and detail-free. Give criteria, number of replications, schedule	Design, within 5 days, a workstation for assembly of 10,000/yr of unit Y with reasonable quality and low mfg cost.
Analyze in detail.	Identify limits (constraints, restrictions). Include variability in components and users. Make machine adjust to person, not converse.	Obtain specifications of components and assembly. Obtain skills availability of people; obtain capability/availability of equipment. Get restrictions in fabrication and assembly techniques and sequence. Obtain more details on cost accounting, scheduling, and tradeoffs of criteria.
Make search of solution space.	Don't be limited by imagined constraints. Try for optimum solution, not feasible solution. Have more than one solution.	Seek a variety of assembly sequences, layouts, fixtures, units/hr, handtools, etc.
Evaluate alternatives.	Trade off multiple criteria. Calculate benefit/cost.	Alt. A: installed cost $1000; cost/unit = $1.10. Alt. B: installed costs $1200; cost/unit = $1.03.
Specify and sell solution.	Specify solution in detail. Sell solution. Accept a partial solution rather than nothing. Follow up to see that design is implemented and that design reduces the problem.	Recommend Alt. B. Install Alt. B1, a modification of B suggested by the supervisor.

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

A problem is the tendency of designers to be satisfiers rather than optimizers. That is, designers tend to stop as soon as they have one feasible solution. For example, when designing an assembly line, the designer may stop as soon as there is a solution. For an optimum solution, there must be a number of alternatives to select from. Alternatives tend to suggest further alternatives, so stopping too soon can limit solution quality and acceptance.

4. Evaluate alternatives.

To get better data for your evaluation, consider trying out your alternatives with mockups, dry runs, pilot experiments, and simulations.

You will need to trade off multiple criteria—usually without any satisfactory tradeoff values. For example, one design of an assembly line may require 0.11 min/unit while another design may require 0.10 min/unit; however, the first design gives more job satisfaction to the workers. How do you quantify job satisfaction? Even if you can put a numerical value on it, how many ‘‘satisfaction units’’ equal a 10% increase in assembly labor cost?

Consider a numerical ranking, using a equal interval scale for each criterion. (Method A requires 1.1 min/unit, while method B requires 1.0 min/unit; method A requires 50 m^2 of floor space, while method B requires 40 m^2.) However, managers generally want to combine the criteria. Table 23 shows one approach. After completing the evaluation, have the affected people sign off on the evaluation form. Then go back and select features from the alternatives to get an improved set of designs.

5. Specify and sell solution.

Your abstract concept must be translated into nuts and bolts—detailed specifications. Then you must convince the decision makers to accept the proposal. A key to acceptance is input

TABLE 23 Step 4 Is to Evaluate the Alternatives

Criterion	Weight	Present		Alternative 1		Alternative 2	
Minimum investment	6	A	24	B	18	A	24
Ease of supervision	10	C	20	C	20	B	30
Ease of operation	8	C	16	C	16	C	16
Ease of expansion and contraction	2	C	4	C	4	B	6
Total points			64		58		76
Relative merit (100% is best)			84%		76%		100%

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

The criteria and weights will depend upon specific management goals. Grade each alternative (A = Excellent = 4; B = Good= 3; C = Average = 2; D = Fair = 1; F = Bad = 0). Then calculate the alternative's gradepoint (grade × weight). Defining the best as 100%, calculate the percent for each alternative.

from others (especially users and decision makers) early in the design stages (steps 1, 2, and 3). At the meeting where the final proposal is presented, there should be no surprises; the participants should feel you are presenting material they have already seen, commented on, and approved. That is, the selling is done early, not late; if they did not buy your approach then, you have modified it until it is acceptable. One common modification is partial change instead of full change; a test market approach instead of immediate national rollout; change of some of the machines in a department instead of all the machines; a partial loaf rather than a full loaf.

The installation needs to be planned in detail (who does what when). This is a relatively straightforward, though time-consuming, process. The installation plan typically is approved at a second meeting. It also should have preapproval by everyone before the decision meeting. During the installation itself, be flexible—improvements may become apparent from the suggestions of supervisors, operators, skilled-trades workers, and so on.

Finally, document the results. A documentary video and photos give opportunities to praise all the contributors (and may even reduce resistance to your next project).

REFERENCES

Cole, R. (1979), *Work, Mobility and Participation: A Comparative Study of American and Japanese Industry,* University of California Press, Berkeley.

Corlett, N., and Bishop, R. (1976), "A Technique for Assessing Postural Discomfort," *Ergonomics,* Vol. 19, pp. 175–82.

Helander, M., and Burri, G. (1995), "Cost Effectiveness of Ergonomics and Quality Improvements in Electronics Manufacturing," *International Journal of Industrial Ergonomics,* Vol. 15, pp. 137–151.

Keyserling, M., Brouwer, M., and Silverstein, B. (1992), "A Checklist for Evaluating Ergonomic Risk Factors Resulting from Awkward Postures of the Legs, Trunk and Neck," *International Journal of Industrial Ergonomics,* Vol. 9, pp. 283–301.

Knauth, P. (1993), "The Design of Shift Systems," *Ergonomics,* Vol. 36, Nos. 1–3, pp. 15–28.

Konz, S., and Goel, S. (1969), "The Shape of the Normal Work Area in the Horizontal Plane," *AIIE Transactions,* Vol. 1, No. 4, pp. 359–370.

Konz, S., and Johnson, S. (2000), *Work Design: Industrial Ergonomics,* 5th Ed., Holcomb Hathaway, Scottsdale, AZ.

Lifshitz, Y., and Armstrong, T. J. (1986), "A Design Checklist for Control and Prediction of Cumulative Trauma Disorder in Hand Intensive Manual Jobs," in *Proceedings of the Human Factors Society 30th Annual Meeting*.

Putz-Anderson, V., Ed. (1988), *Cumulative Trauma Disorders,* Taylor & Francis, London.

Rodgers, S. A. (1992), "A Functional Job Analysis Technique," *Occupational Medicine: State of the Art Reviews,* Vol. 7, No. 4, pp. 679–711.

Wilson, J., and Grey, S. (1984), "Reach Requirements and Job Attitudes at Laser-Scanner Checkout Stations," *Ergonomics,* Vol. 27, No. 12, pp. 1247–1266.

ADDITIONAL READING

Francis, R., McGinnis, L., and White J., *Facility Layout and Location: An Analytical Approach,* Prentice Hall, Englewood Cliffs, NJ, 1992.

Harry, M., and Schroeder, R., *Six Sigma,* Doubleday, New York, 2000.

CHAPTER 53

Time Standards*

STEPHAN KONZ
Kansas State University

1. WHY DETERMINE TIME/JOB? 1392
- 1.1. Cost Allocation 1392
- 1.2. Production and Inventory Control 1392
- 1.3. Evaluation of Alternatives 1392
- 1.4. Acceptable Day's Work 1392
- 1.5. Incentive Pay 1392

2. ESTABLISHING TIME STANDARDS 1392
- 2.1. Nonengineered (Type 2) Estimates 1392
  - 2.1.1. Historical Records 1392
  - 2.1.2. Ask Expert 1393
  - 2.1.3. Time Logs 1393
  - 2.1.4. Work (Occurrence) Sampling 1393
- 2.2. Engineering (Type 1) Estimates 1393
  - 2.2.1. Stopwatch Time Study 1393
  - 2.2.2. Standard Data 1393

3. ADJUSTMENTS TO TIME: ALLOWANCES 1394
- 3.1. Three Levels of Time 1394
- 3.2. Personal Allowances 1394
- 3.3. Fatigue Allowances 1394
  - 3.3.1. Physical: Physical Fatigue 1395
  - 3.3.2. Physical: Short Cycle 1395
  - 3.3.3. Physical: Static Load (Body Posture) 1395
  - 3.3.4. Physical: Restrictive Clothing 1395
  - 3.3.5. Mental: Concentration/Anxiety 1395
  - 3.3.6. Mental: Monotony 1395
  - 3.3.7. Environmental: Climate 1395
  - 3.3.8. Environmental: Dust, Dirt, and Fumes 1395
  - 3.3.9. Environmental: Noise and Vibration 1398
  - 3.3.10. Environmental: Eye Strain 1398
  - 3.3.11. Overview of Fatigue Allowances 1398
- 3.4. Delay Allowances 1398

4. ADJUSTMENTS TO TIME: LEARNING 1400
- 4.1. Learning 1400
  - 4.1.1. Individual Learning 1400
  - 4.1.2. Organization Learning (Manufacturing Progress) 1400
  - 4.1.3. Quantifying Improvement 1400
  - 4.1.4. Typical Rates for Organization Progress 1404
  - 4.1.5. Typical Values for Learning 1405
  - 4.1.6. Example Applications of Learning 1405

5. DOCUMENTING, USING, AND MAINTAINING STANDARDS 1406
- 5.1. Documenting Standards 1406
- 5.2. Using Standards 1406
  - 5.2.1. Reports 1406
  - 5.2.2. Consequences of Not Making Standard 1406
- 5.3. Maintaining Standards (Auditing) 1407

REFERENCES 1407

---

*This chapter is a concise version of the material in Konz and Johnson (2000).

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. WHY DETERMINE TIME/JOB?

It is useful to know the direct labor cost/unit, especially when the job is repetitive. Five typical applications are:

1. Cost allocation
2. Production and inventory control
3. Evaluation of alternatives
4. Acceptable day's work
5. Incentive pay

1.1. Cost Allocation

To determine cost/unit, you need the direct material cost, the direct labor cost, and various miscellaneous costs (called overhead or burden). Direct labor cost is (direct labor time)(wage cost/hr). So you need to determine how long the job takes. But, in addition, overhead costs usually are allocated as a percentage of direct labor (e.g., overhead is 300% of direct labor cost). So again you need direct labor time. Without good estimates of the cost of production to compare vs. selling price, you don't know your profit/unit (it may even be negative!). The goal is to improve and control costs through better information.

1.2. Production and Inventory Control

Without time/unit, you can not schedule or staff (i.e., use management information systems). How many people should be assigned to the job? When should production start in order to make the due date and thus avoid stockouts?

1.3. Evaluation of Alternatives

Without time/unit, you can not compare alternatives. Should a mechanic repair a part or replace it with a new one? Is it worthwhile to use a robot that takes 10 seconds to do a task?

1.4. Acceptable Day's Work

Sam picked 1600 items from the warehouse today—is that good or bad? Supervisors would like to be able to compare actual performance to expected performance. Many applications of standards to repetitive work have shown improvement in output of 30% or more when measured daywork systems are installed in place of nonengineered standards. Output increases about 10% more when a group incentive payment is used and 20% more when an individual incentive is used.

1.5. Incentive Pay

A minority of firms use the pay-by-results (the carrot) approach. If you produce 1% more, you get paid 1% more. This works for the firm because even though direct labor cost/unit stays constant, overhead costs do not increase and thus total cost/unit decreases.

2. ESTABLISHING TIME STANDARDS

There are two basic strategies: nonengineered (subjective) standards (''did take'' times) and engineered (objective) standards (''should take'' times). The techniques to use depend upon the cost of obtaining the information and the benefits of using the information.

2.1. Nonengineered (Type 2) Estimates

''Quick and dirty'' information can be obtained at low cost. But using ''dirty'' information increases the risk of errors in decisions. Since nonengineered standards are not preceded by methods or quality analysis, they are ''did take'' times, not ''should take'' times.

There are four approaches: historical records, ask expert, time logs, and work (occurrence) sampling.

2.1.1. Historical Records

Standards from historical records tend to be very ''dirty'' (although cheap). For example, in the warehouse, how many cases can be picked per hour? From shipping records, determine the number of cases shipped in January, February, and March. From personnel, determine the number of employees in shipping in each month. Divide total cases/total hours to get cases/hr. Ignore changes in product output, product mix, absenteeism, delays, and so on.

2.1.2. Ask Expert

Here you ask a knowledgeable expert how long a job will take. For example, ask the maintenance supervisor how long it will take to paint a room. Ask the sales supervisor how many customers can be contacted per week. A serious problem is that the expert may have an interest in the answer. For example, a "hungry" maintenance supervisor wants work for his group and so quotes a shorter painting time; a sales supervisor may be able to hire more staff (and thus increase her prestige and power) by giving a low estimate of customers/sales representative.

2.1.3. Time Logs

It may be that a job is "cost plus" and so the only problem is how many hours to charge to a customer. For example, an engineer might write down, for Monday, 4.0 hr for project A, 2.5 hr for B, and 3.5 hr for C. Obviously there are many potential errors here (especially if work is done on project D, for which there no longer is any budget).

2.1.4. Work (Occurrence) Sampling

This technique is described in more detail in Chapter 54. It is especially useful when a variety of jobs are done intermittently (e.g., as in maintenance or office work). Assume that during a three-week period a maintainer spends 30% of the time doing carpentry, 40% painting, and 30% miscellaneous; this means 120 hr × 0.4 = 48 hr for painting. During the three-week period, 10,000 ft^2 of wall were painted or 10,000/48 = 208 ft^2/hr. Note that the work method used, work rate, delays, production schedule, and so on are not questioned.

2.2. Engineering (Type 1) Estimates

Engineered estimates of time must be preceded by a methods and quality analysis; the result is a "should take" time, not a "did take" time. MIL-STD-1567A requires for all individual type 1 standards (assuming the basic standards system is in place):

1. Documentation that the method was analyzed before time was determined
2. A record of the method or standard practice followed when the time standard was developed
3. A record of rating (if time study was used)
4. A record of the observed times (if time study) or predetermined time values were used
5. A record of the computations of standard time, including allowances

You would not expect a time standard established in 1960 to be valid today; things are different now, we think. But what differs? Thus, step 2 (record of method followed) is essential information for maintaining standards.

There are two basic ways of determining time/job: stopwatch time study and standard data.

2.2.1. Stopwatch Time Study

Stopwatch time study, which is described in detail in Chapter 54, requires an operator to do the operation; thus it cannot be done ahead of production. In general, it requires the operator to do the operation over and over rather than doing different tasks intermittently (such as might be done in office or maintenance work). Before the timing is done, the method must be analyzed.

In repetitive work, where detailed methods analysis is desired, a videotape of the task can be made; the analyst can study the tape instead of a live operator.

Because of learning, do not do time studies on operators who are early on the learning curve. If the study must be done early, label it temporary and restudy it in, say 30 days.

2.2.2. Standard Data

Standard data can be at the micro level or the macro level (see Chapter 54). In this approach, the analyst visualizes what the job entails (a danger is that the analyst may not think of some of the steps [elements] needed to complete the job). After determining the method, the analyst uses a table or formula to determine the amount of time for each element. The database elements are expressed in normal time (i.e., rating is included), so no additional rating is required. Then normal time is increased with allowances to obtain standard time.

Compared with time study, the standard data method has three advantages: (1) cost of determining a standard is low (assuming you have a database with standard times); (2) consistency is high because everyone using the database should get the same times; and (3) ahead-of-production standards are helpful in many planning activities. But among these three roses are two thorns: (1) you may not have the money to build the database (the databases are built from stopwatch studies and predetermined times); and (2) the analyst must imagine the work method; even experienced analysts may overlook some details or low-frequency elements.

3. ADJUSTMENTS TO TIME: ALLOWANCES

3.1. Three Levels of Time

Time is reported at three levels:

1. *Observed time:* The raw (unadjusted) time taken by the worker. It does not have any rating, allowance or learning adjustment.
2. *Normal time:*

$$\text{Normal time} = (\text{observed time})(\text{rating})$$

   The observer estimates the pace of the worker in relation to normal pace. Normal time is the time an experienced operator takes when working at a 100% pace. See Chapter 54 for more details.
3. *Standard time:* For allowances expressed as a percent of shift time:

$$\text{Standard time} = \text{normal time}/(1 - \text{allowances})$$

   For allowances expressed as a percent of work time:

$$\text{Standard time} = \text{normal time } (1 + \text{allowances})$$

   It is a policy decision by the firm whether to give allowances as a percent of shift or work time. Normal time needs to be increased from standard time by personal, fatigue, and delay allowances.

3.2. Personal Allowances

Personal allowances are given for such things as blowing your nose, going to the toilet, getting a drink of water, smoking, and so on. They do not vary with the task but are the same for all tasks in the firm. There is no scientific or engineering basis for the percent to give. Values of 5% (24 minutes in a 480-minute day) seem to be typical.

Most firms have standardized break periods (coffee breaks)—for example, 10 minutes in the first part of the shift and the same in the second part. It is not clear whether most firms consider this time as part of the personal allowance or in addition to it.

The midshift meal break (lunch) is another question. This 20–60-minute break obviously permits the worker to attend to personal needs and recover from fatigue. Yet lunch usually is not considered as part of allowances—even if the lunch period is paid.

Some firms give an additional break if work is over 8 hours. For example, if a shift is over 10 hours, there is an additional break of 10 minutes after the 9th hour.

In addition, some firms give an additional allowance to all workers for cleanup (either of the person or the machine), putting on and taking off of protective clothing, or travel. In mines, the travel allowance is called portal-to-portal pay; pay begins when the worker crosses the mine portal, even though the worker will not arrive at the working surface until some time later.

3.3. Fatigue Allowances

The rationale of fatigue allowances is to compensate the person for the time lost due to fatigue. In contrast to personal allowances, which are given to everyone, fatigue allowances are given only for cause—for fatigue. No fatigue? Then no fatigue allowance!

Another challenge is the concept of machine time. With the increasing capabilities of servomechanisms and computers, many machines operate semiautomatically (operator is required only to load/unload the machine) or automatically (machine loads, processes, and unloads). During the machine time of the work cycle, the operator may be able to drink coffee (personal allowance) or talk to the supervisor (delay allowance) or recover from fatigue. Thus, as a general principle, give a fatigue allowance only for the portion of the work cycle outside the machine time.

The following will discuss the fatigue allowances developed by the International Labor Organization (ILO 1992). They were supplied by a British consulting firm. Use of the ILO values is complex. Remembering that fatigue allowances are given for work time only (not machine time), sum the applicable fatigue allowance points. Then, using Table 1, convert points to percent time. For a more detailed discussion of allowances, see Konz and Johnson (2000, chap. 32).

The fatigue factors are grouped into three categories: physical, mental, and environmental.

3.3.1. Physical: Physical Fatigue

Table 2 shows how the ILO makes a distinction among carrying loads, lifting loads, and force applied. In the NIOSH lifting guideline, the lift origin and destination, frequency of move, angle, and container are considered as well as load.

3.3.2. Physical: Short Cycle

Table 3 gives the fatigue allowance to allow time for the muscles to recover.

3.3.3. Physical: Static Load (Body Posture)

Table 4 gives the allowance for poor posture.

3.3.4. Physical: Restrictive Clothing

Table 5 gives the allowance for restrictive clothing.

3.3.5. Mental: Concentration/Anxiety

Table 6 gives the allowance for concentration/anxiety.

3.3.6. Mental: Monotony

Table 7 gives the allowance for monotony. In the author's opinion, allowances for monotony, boredom, lack of a feeling of accomplishment, and the like are questionable. These factors are unlikely to cause fatigue and thus increase time/cycle. These factors primarily reflect unpleasantness and thus should be reflected in the wage rate/hr rather than the time/unit.

3.3.7. Environmental: Climate

Table 8 gives the allowance for climate.

3.3.8. Environmental: Dust, Dirt, and Fumes

Table 9 gives the allowance for dust, dirt and fumes.

TABLE 1 Conversion from Points Allowance to Percent Allowance for ILO

Points	0	1	2	3	4	5	6	7	8	9
0	10	10	10	10	10	10	10	11	11	11
10	11	11	11	11	11	12	12	12	12	12
20	13	13	13	13	14	14	14	14	15	15
30	15	16	16	16	17	17	17	18	18	18
40	19	19	20	20	21	21	22	22	23	23
50	24	24	25	26	26	27	27	28	28	29
60	30	30	31	32	32	33	34	34	35	36
70	37	37	38	39	40	40	41	42	43	44
80	45	46	47	48	48	49	50	51	52	53
90	54	55	56	57	58	59	60	61	62	63
100	64	65	66	68	69	70	71	72	73	74
110	75	77	78	79	80	82	83	84	85	87
120	88	89	91	92	93	95	96	97	99	100
130	101	103	105	106	107	109	110	112	113	115
140	116	118	119	121	122	123	125	126	128	130

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

The second column (0) gives the 10s, and the remaining columns give the units. Thus, 30 points (0 column) = 15%; 31 points (1 column) = 16%; 34 points = 17%. The percent allowance is for manual work time (not machine time) and includes 5% personal time for coffee breaks.

TABLE 2 Carrying, Lifting, and Body Force Allowances

Weight or Force, kg	Push Points	Carry Points	Lift Points
1	0	0	0
2	5	5	10
3	8	9	15
4	10	13	18
5	12	15	21
6	14	17	23
7	15	20	26
8	17	21	29
9	19	24	32
10	20	26	34
11	21	29	37
12	23	31	40
13	25	33	44
14	26	34	46
15	27	36	50
16	28	39	50
17	30	40	53
18	32	42	56
19	33	44	58
20	34	46	60

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

The ILO tables go to 64 kg. Push includes foot pedal push and carry on the back. Carry includes hand carry and swinging arm movements. Weight is averaged over time. A 15 kg load lifted for 33% of a cycle is 5 kg.

TABLE 3 Short-Cycle Allowances

Points	Cycle Time, min
1	0.16–0.17
2	0.15
3	0.13–0.14
4	0.12
5	0.10–0.11
6	0.08–0.09
7	0.07
8	0.06
9	0.05
10	Less than 0.05

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

TABLE 4 Posture Allowances

Points	Activity
0	Sitting easily
2	Sitting awkwardly or mixed sitting and standing
4	Standing or walking freely
5	Ascending or descending stairs, unladen
6	Standing with a load; walking with a load
8	Climbing up or down ladders; some bending, lifting, stretching, or throwing
10	Awkward lifting; shoveling ballast to container
12	Constant bending, lifting, stretching, or throwing
16	Coal mining with pickaxes; lying in low seam

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

TABLE 5 Restrictive Clothing Allowances

Points	Clothing
1	Thin rubber (surgeon's) gloves
2	Household rubber gloves; rubber boots
3	Grinder's goggles
5	Industrial rubber or leather gloves
8	Face mask (e.g., for paint spraying)
15	Asbestos suit or tarpaulin coat
20	Restrictive protective clothing and respirator

ILO (1979) considers clothing weight in relation to effort and movement. Also consider whether it affects ventilation and breathing.

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

TABLE 6 Concentration/Anxiety Allowances

Allowance Points	Degree
0	Routine, simple assembly; shoveling ballast
1	Routine packing, washing vehicles, wheeling trolley down clear gangway
2	Feed press tool (hand clear of press); topping up battery
3	Painting walls
4	Assembling small and simple batches (performed without much thinking); sewing machine work (automatically guided)
5	Assembling warehouse orders by trolley; simple inspection
6	Load/unload press tool; hand feed into machine; spraypainting metalwork
7	Adding up figures; inspecting detailed components
8	Buffing and polishing
10	Guiding work by hand on sewing machine; packing assorted chocolates (memorizing patterns and selecting accordingly); assembly work too complex to become automatic; welding parts held in jig
15	Driving a bus in heavy traffic or fog; marking out in detail with high accuracy

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

ILO (1979) considers what would happen if the operator were to relax attention, responsibility, need for exact timing, and accuracy or precision required.

TABLE 7 Monotony Allowances

Allowance Points	Degree
0	Two people on jobbing work
3	Cleaning own shoes for 0.5 hr on one's own
5	Operator on repetitive work; operator working alone on nonrepetitive work
6	Routine inspection
8	Adding similar columns of figures
11	One operator working alone on highly repetitive work

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

ILO (1979) considers the degree of mental stimulation and whether there is companionship, competitive spirit, music, and so on.

TABLE 8 Climate Allowances

Points for Temperature/Humidity

Humidity, %	Temperature, °C Up to 24	 24–32	 Over 32
Up to 75	0	6–9	12–16
76–85	1–3	8–12	15–26
Over 85	4–6	12–17	20–36

Points Wet

Points	
0	Normal factory operations
1	Outdoor workers (e.g., postman)
2	Working continuously in the damp
4	Rubbing down walls with wet pumice block
5	Continuous handling of wet articles
10	Laundry washhouse, wet work, steamy, floor running with water, hands wet

Points Ventilation

Points	
0	Offices; factories with office-type conditions
1	Workshop with reasonable ventilation but some drafts
3	Drafty workshops
14	Working in sewer

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

ILO (1979) considers temperature/humidity, wet, and ventilation. For temperature/humidity, use the average environmental temperature. For wet, consider the cumulative effect over a long period. For ventilation, consider quality/freshness of air and its circulation by air conditioning or natural movement.

3.3.9. Environmental: Noise and Vibration

Table 10 gives the allowance for noise and vibration.

3.3.10. Environmental: Eye Strain

Table 11 gives the allowance for eye strain.

3.3.11. Overview of Fatigue Allowances

In general, the fatigue allowances seem to have inadequate range. In addition, note from Table 1 that points are not converted one for one to percent. A person with 0 points for fatigue gets a 10% fatigue allowance. A person with 30 points gets a 15% fatigue allowance—an increase of only 5%.

In addition, neither the length of the workday nor the number of days/week is specified. Presumably it is 8 hr/day and 5 days/week. The author does not recommend changing the allowance for working a shorter or longer time period. Any adjustment should be in the discipline level (see Section 5.2.2).

3.4. Delay Allowances

Delay allowances should vary with the task but not the operator. They compensate for machine breakdowns, interrupted material flow, conversations with supervisors, machine maintenance and cleaning, and so on. If the delay is long (e.g., 30 minutes), the operator clocks out (records the start and stop time of the delay on a form) and works on something else during the clocked-out time. Delays usually permit the operator to take some personal time and reduce fatigue; that is, they also serve as personal allowances and fatigue allowances.

How do you set a delay allowance? One possibility is to record the delays during a work sampling study or time study. For example, if there were 4 minutes of delay during 100 minutes of time study, then 4% could be used for the delay allowance.

Errors in delay allowances can occur from poor sampling or changing conditions.

To obtain a valid sample of delays, the sample must represent the total shift, not just the middle of the shift. That is, the delays must be observed at the start and stop of the shift and just before and after lunch and coffee breaks, in addition to the middle of the shift. Also observe delays on the second and third shifts.

TABLE 9 Dust, Dirt, and Fumes Allowances

Points	Dust
0	Office, normal light assembly, press shop
1	Grinding or buffing with good extraction
2	Sawing wood
4	Emptying ashes
6	Finishing weld
10	Running coke from hoppers into skips or trucks
11	Unloading cement
12	Demolishing building
Points	**Dirt**
0	Office work, normal assembly operations
1	Office duplicators
2	Dustman (garbage collector)
4	Stripping internal combustion engine
5	Working under old motor vehicle
7	Unloading bags of cement
10	Coal miner; chimneysweep with brushes
Points	**Fumes**
0	Lathe tuning with coolants
1	Emulsion paint, gas cutting, soldering with resin
5	Motor vehicle exhaust in small commercial garage
6	Cellulose painting
10	Molder procuring metal and filling mold

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

For dust, consider both volume and nature of the dust. The dirt allowance covers ‘‘washing time’’ where this is paid for (e.g., 3 min for washing). Do not allow both time and points. For fumes, consider the nature and concentration; whether toxic or injurious to the health; irritating to eyes, nose, throat, or skin; odor.

TABLE 10 Noise and Vibration Allowances

Points	Noise Category
0	Working in a quiet office, no distracting noise; light assembly work
1	Work in a city office with continual traffic noise outside
2	Light machine shop; office or assembly shop where noise is a distraction
4	Woodworking machine shop
5	Operating steam hammer in forge
9	Riveting in a shipyard
10	Road drilling
Points	**Vibration Category**
1	Shoveling light materials
2	Power sewing machine; power press or guillotine if operator is holding the material; cross-cut sawing
4	Shoveling ballast; portable power drill operated by one hand
6	Pickaxing
8	Power drill (2 hands)
15	Road drill on concrete

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

ILO (1979) considers whether the noise affects concentration, is a steady hum or a background noise, is regular or occurs unexpectedly, is irritating or soothing. Consider the impact of the vibration on the body, limbs, or hands and the addition to mental effort as a result, or to a series of jars or shocks.

TABLE 11 Eye Strain Allowances

Points	Eye Strain Category
0	Normal factory work
2	Inspection of easily visible faults; sorting distinctively colored articles by color; factory work in poor lighting
4	Intermittent inspection for detailed faults; grading apples
8	Reading a newspaper in a bus
10	Continuous visual inspection (cloth from a loom)
14	Engraving using an eyeglass

From International Labour Office, *Introduction to Work Study,* 4th (Rev.) Ed., pp. 491–498. Copyright © International Labour Organization 1992.

ILO (1979) considers the lighting conditions, glare, flicker, illumination, color, and closeness of work and for how long strain is endured.

Conditions change over time. A reasonable procedure is to give delay allowances an expiration date, for example, two years after being set. After two years, they must be redetermined.

4. ADJUSTMENTS TO TIME: LEARNING

4.1. Learning

Failure to adjust standard time for learning is the primary cause of incorrect times. Learning occurs both in the individual and in the organization.

4.1.1. Individual Learning

Individual learning is improvement in time/unit even though neither the product design nor the tools and equipment change. The improvement is due to better eye–hand coordination, fewer mistakes, and reduced decision time.

4.1.2. Organization Learning (Manufacturing Progress)

This is improvement with changing product design, changing tools and equipment, and changing work methods; it also includes individual learning. Often it is called manufacturing progress.

Consider the server Maureen serving breakfast. During the individual learning period, she learned where the coffeepot and cups were, the prices of each product, and so on. The amount of time she took declined to a plateau. Then management set a policy to serve coffee in cups without saucers and furnish cream in sealed, one-serving containers so the container need not be carried upright. These changes in product design reduced time for the task. Other possible changes might include a coffeepot at each end of the counter. A different coffeepot might have a better handle so less care is needed to prevent burns. The organization might decide to have the server leave the bill when the last food item is served. Organization progress comes from three factors: operator learning with existing technology, new technology, and substitution of capital for labor.

Point 1 was just discussed. Examples of new technology are the subsurface bulblike nose on the front of tankers (which increased tanker speed at very low cost) and solid-state electronics. Moore's law states that the number of transistors on a given chip size (roughly a gauge of chip performance) doubles every 1.5–2 years. Some example numbers are 3,500 transistors/chip in 1972, 134,000 in 1982, 3,100,000 in 1993, and 7,500,000 in 1997.

Use of two coffee pots by Maureen is an example of substituting capital for labor. Another example is the use of the computer in the office, permitting automation of many office functions. The ratio of capital/labor also can be improved by economies of scale. This occurs when equipment with twice the capacity costs less than twice as much. Then capital cost/unit is reduced and fewer work hours are needed/unit of output.

4.1.3. Quantifying Improvement

"Practice makes perfect" has been known for a long time. Wright (1936) took a key step when he published manufacturing progress curves for the aircraft industry. Wright made two major contributions. First, he quantified the amount of manufacturing progress for a specific product. The equation was of the form Cost = a (number of airplanes)^b; (see Figure 1). But the second step was probably even more important: he made the data a straight line (by putting the curve in the axis!) (see Figure 2). That is, the data is on a log–log scale.

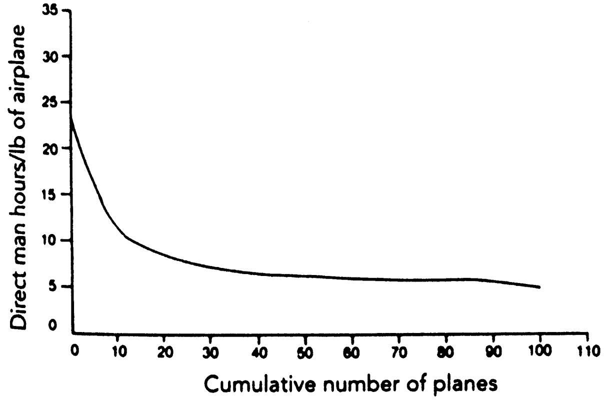


Figure 1 Practice Makes Perfect. As more and more units are produced, the fixed cost is divided by more and more units, so fixed cost/unit declines. In addition, variable cost/unit declines because fewer mistakes are made, less time is spent looking up instructions, better tooling is used, and so on. The variable cost data usually can be fitted with an equation of the form $y = ax^b$. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.)

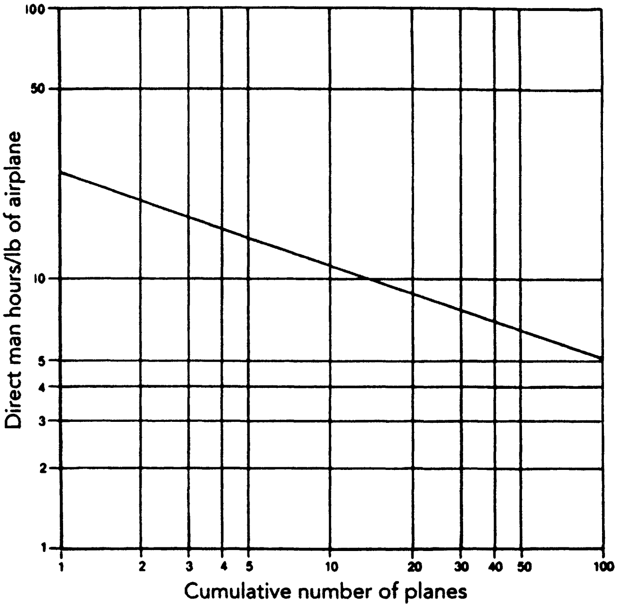


Figure 2 Log–Log Scale. Supervisors like straight lines. Plotting $y = ax^b$ on log–log paper gives a straight line. The key piece of information supervisors desire is the rate of improvement—the slope of the line. The convention is to refer to reduction with doubled quantities. If quantity $x_1 = 8$, the quantity $x_2 = 16$. Then if cost at x_1 is $y_1 = 100$ and cost at x_2 is $y_2 = 80$, this is an 80% curve. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.)

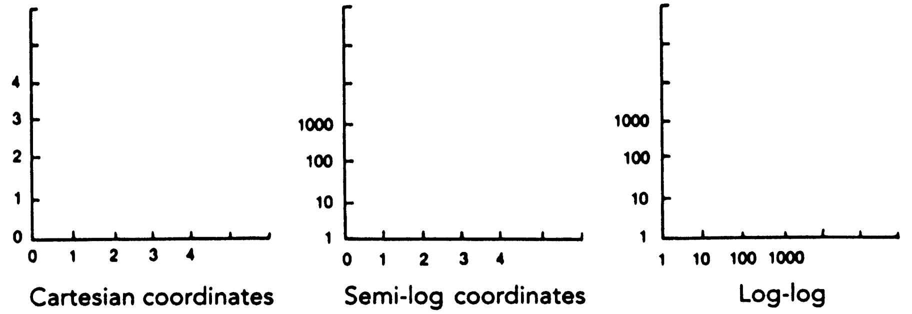


Figure 3 Cartesian Coordinates, Semi-Log Coordinates, and Log–Log Coordinates. Cartesian coordinates have equal distances for equal numerical differences; that is, the linear difference from 1 to 3 is the same as from 8 to 10. On a log scale, the same distance represents a constant *ratio*; that is, the distance from 2 to 4 is the same as from 30 to 60 or 1000 to 2000. Semi-log paper has one axis Cartesian and one axis log. Log–log (double log) paper has a log scale on both axes. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.)

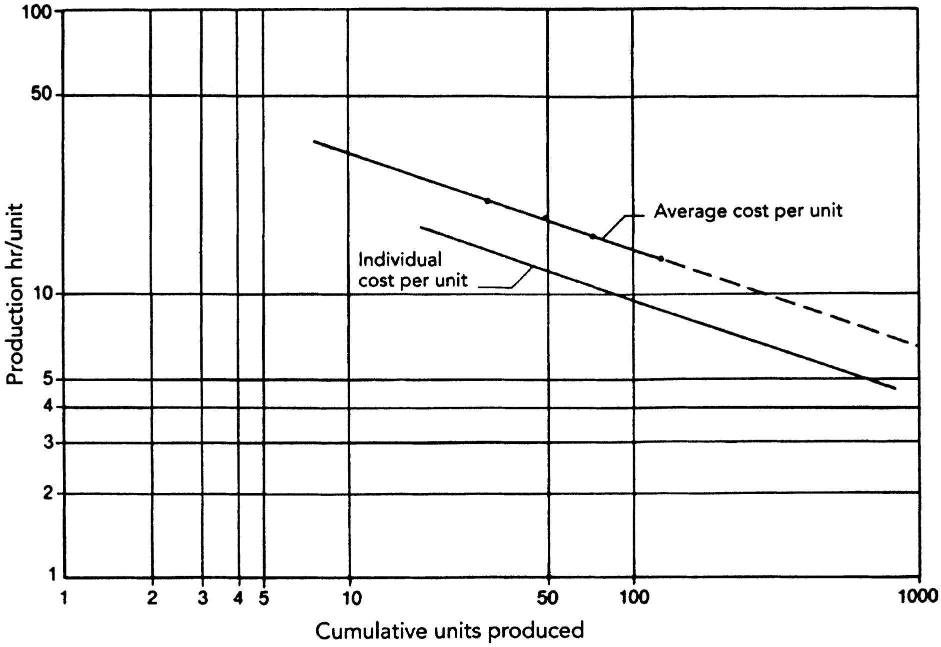


Figure 4 Average Cost/Unit from Table 12. Table 12 gives a 79% curve. Cost/unit is the cost of the *n*th unit; average cost/unit is the sum of the unit costs/*n*. Cost/unit can be estimated by multiplying average cost/unit by the factor from Table 12. The average cost of the first 20 units is estimated as 25.9 from the fitted line; the cost of the 20th unit is 25.9(0.658) = 17.0 hr. (From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.)

TABLE 12 Factors for Various Improvement Curves

Improvement Curve, % Between Doubled Quantities	Learning Factor, b, For Curve $y = ax^b$	Multiplier to Determine Unit Cost if Average Cost Is Known	Multiplier to Determine Average Cost if Unit Cost Is Known
70	−0.515	0.485	2.06
72	−0.474	0.524	1.91
74	−0.434	0.565	1.77
76	−0.396	0.606	1.65
78	−0.358	0.641	1.56
80	−0.322	0.676	1.48
82	−0.286	0.709	1.41
84	−0.252	0.746	1.34
85	−0.234	0.763	1.31
86	−0.218	0.781	1.28
88	−0.184	0.813	1.23
90	−0.152	0.847	1.18
92	−0.120	0.877	1.14
94	−0.089	0.909	1.10
95	−0.074	0.926	1.08
96	−0.059	0.943	1.06
98	−0.029	0.971	1.03

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

The multipliers in columns 3 and 4 are large-quantity approximations. For example, for a 90% curve, the table value in column 4 is 1.18. A more precise value at a quantity of 10 = 1.07, at 50 = 1.13, and at 100 = 1.17. A more precise value for an 85% curve at a quantity of 100 = 1.29; a more precise value for a 95% curve at a quantity of 100 = 1.077.

TABLE 13 Time and Completed Units as They Might Be Reported for a Product

Month	Units Completed (Pass Final Inspection)	Month's Direct Labor Hours Charged To Project	Cumulative Units Completed	Cumulative Work Hours Charged to Project	Average Work hr/unit
March	14	410	14	410	29.3
April	9	191	23	601	26.1
May	16	244	39	845	21.7
June	21	284	60	1129	18.8
June	24	238	84	1367	16.3
August	43	401	127	1708	13.4

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

On a log scale, the physical distance between doubled quantities is constant (i.e., 8 to 16 is the same distance as 16 to 32 or 25 to 50) (see Figure 4). Wright gave the new cost as a percent of the original cost when the production quantity had doubled. If cost at unit 10 was 100 hr and cost at unit 20 was 85 hr, then this was an 85/100 = 85% curve. Since the curve was a straight line, it was easy to calculate the cost of the 15th or the 50th unit. If you wish to solve the $Y = ax^b$ equation instead of using a graph (calculators are much better now than in 1935), see Table 12.

For example, assume the time for a (i.e., cycle 1)= 10 min and there is a 90% curve (i.e., $b = -0.152$), then the time for the 50th unit is: $Y = 10\ (50)^{-0.152} = 10/(50)^{0.152} = 5.52$ min.

Table 13 shows how data might be obtained for fitting a curve. During the month of March, various people wrote down on charge slips a total of 410 hours against this project charge number.

The average work hours/unit during March then becomes 29.3. The average *x*-coordinate is (1 + 14)/2 = 7.5. Because the curve shape is changing so rapidly in the early units, some authors recommend plotting the first lot at the 1/3 point[(1 + 14)/3] and points for all subsequent lots at the midpoint.

During April, 9 units passed final inspection and 191 hours were charged against the project. Cumulative hours of 601 divided by cumulative completed output of 23 gives average hr/unit of 26.1. The 26.1 is plotted at (15 + 23)/2 = 19. As you can see from the example data, there are many possible errors in the data, so a curve more complex than a straight line on log–log paper is not justified. Figure 4 shows the resulting curve.

Although average cost/unit is what is usually used, you may wish to calculate cost at a specific unit. Conversely, the data may be for specific units and you may want average cost. Table 12 gives the multiplier for various slopes. The multipliers are based on the fact that the average cost curve and the unit cost curve are parallel after an initial transient. Initial transient usually is 20 units, although it could be as few as 3. The multiplier for a 79% slope is (0.641 + 0.676)/2 = 0.658. Thus, if we wish to estimate the cost of the 20th unit, it is (24.9 hr)(0.658) = 16.4 hr.

Cost/unit is especially useful in scheduling. For example, if 50 units are scheduled for September, then work-hr/unit (for a 79% curve) at unit 127 = (13.4)(0.656) = 8.8 and at unit 177 = 7.8. Therefore between 390 and 440 hours should be scheduled.

Looking at Figure 4, you can see that the extrapolated line predicts cost/unit at 200 to be 11.4 hr, at 500 to be 8.3, and at 1,000 to be 6.6. If we add more cycles on the paper, the line eventually reaches a cost of zero at cumulative production of 200,000 units. Can cost go to zero? Can a tree grow to the sky? No.

The log–log plot increases understanding of improvement, but it also deceives. Note that cost/unit for unit 20 was 24.9 hr. When output was doubled to 40 units, cost dropped to 19.7; doubling to 80 dropped cost to 15.5; doubling to 160 dropped cost to 12.1; doubling to 320 dropped cost to 9.6; doubling to 640 dropped cost to 7.6. Now consider the improvement for each doubling. For the first doubling from 20 to 40 units, cost dropped 5.20 hr or 0.260 hr/unit of extra experience. For the next doubling from 40 to 80, cost dropped 4.2 hr or 0.105 hr/unit of extra experience. For the doubling from 320 to 640, cost dropped 2.0 hr or 0.006 hr/unit of extra experience. In summary, the more experience, the more difficult it is to show additional improvement.

Yet the figures would predict zero cost at 200,000 units, and products just aren't made in zero time. One explanation is that total output of the product, in its present design, is stopped before 200,000 units are produced. In other words, if we no longer produce Model T's and start to produce Model A's, we start on a new improvement curve at zero experience. A second explanation is that the effect of improvement in hours is masked by changes in labor wages/hr. The Model T Ford had a manufacturing progress rate of 86%. In 1910, when 12,300 Model T Fords had been built, the price was $950. When it went out of production in 1926 after a cumulative output of 15,000,000, the price was $270; $200 in constant prices plus inflation of $70.

The third explanation is that straight lines on log–log paper are not perfect fits over large ranges of cycles. If output is going to go to 1,000,000 cumulative units over a 10-year period, you really shouldn't expect to predict the cost of the 1,000,000th unit (which will be built 10 years from the start) from the data of the first 6 months. There is too much change in economic conditions, managers, unions, technology and other factors. Anyone who expects the future to be perfectly predicted by a formula has not yet lost money in the stock market.

4.1.4. Typical Values for Organization Progress

The rate of improvement depends on the amount that can be learned. The more that can be learned, the more will be learned. The amount that can be learned depends upon two factors: (1) amount of previous experience with the product and (2) extent of mechanization. Table 14 gives manufacturing progress as a function of the manual/machine ratio. Allemang (1977) estimates percent progress from

TABLE 14 Prediction of Manufacturing Progress

Percent of Task Time		
Manual	Machine	Manufacturing Progress, %
25	75	90
50	50	85
75	25	80

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

product design stability, a product characteristics table (complexity, accessability, close tolerances, test specifications, and delicate parts), parts shortage, and operator learning.

Konz and Johnson (2000) have two detailed tables giving about 75 manufacturing progress rates reported in the literature.

4.1.5. Typical Values for Learning

Assume learning has two components: (1) cognitive learning and (2) motor learning (Dar-El et al. 1995a, b). Cognitive learning has a greater improvement (say 70% curve), while motor learning is slower (say 90% curve). For a task with both types, initially the cognitive dominates and then the motor learning dominates. Use values of 70% for ‘‘pure cognitive,’’ 72.5% for ‘‘high cognitive,’’ 77.5% for ‘‘more cognitive than motor,’’ 82.5% for ‘‘more motor than cognitive,’’ and 90% for ‘‘pure motor.’’ Konz and Johnson (2000) give a table of 43 tasks for which learning curves have been reported.

The improvement takes place through reduction of fumbles and delays rather than greater movement speed. Stationary motions such as position and grasp improve the most while reach and move improve little. It is reduced information-processing time rather than faster hand speed that affects the reduction.

The range of times and the minimum time of elements show little change with practice. The reduction is due to a shift in the distribution of times; the shorter times are achieved more often and the slower times less often—‘‘going slowly less often’’ (Salvendy and Seymour 1973).

The initial time for a cognitive task might be 13–15 times the standard time; the initial time for a manual task might be 2.5 times the standard time.

4.1.6. Example Applications of Learning

Table 15 shows the effect of learning/manufacturing progress on time standards. The fact that labor hr/unit declines as output increases makes computations using the applications of standard time more complicated. Ah, for the simple life!

4.1.6.1. Cost Allocation Knowing what your costs are is especially important if you have a make–buy decision or are bidding on new contracts. If a component is used on more than one product (standardization), it can progress much faster on the curve since its sales come from multiple sources. Manufacturing progress also means that standard costs quickly become obsolete.

Note that small lots (say due to a customer emergency) can have very high costs. For example, if a standard lot size is 100 and labor cost is 1 hr/unit and there is a 95% curve, a lot of 6 would have a labor cost about 23% higher (1.23 hr/unit). Consider charging more for special orders!

4.1.6.2. Scheduling Knowing how many people are needed and when is obviously an important decision. Also, learning/manufacturing progress calculations will emphasize the penalties of small lots.

4.1.6.3. Evaluation of Alternatives When alternatives are being compared, a pilot project might be run. Data might be obtained for 50–100 cycles. Note that the times after the pilot study should be substantially shorter due to learning. In addition, the learning/manufacturing progress rate for alternatives A and B might differ so that what initially is best will not be best in the long run.

TABLE 15 Demonstration of the Learning Effect on Time Standards

Learning Curve, %	Time/Unit at			Percent of Standard at		
	2X	4X	32X	2X	4X	32X
98	0.98	0.96	0.90	102	104	111
95	0.95	0.90	0.77	105	111	129
90	0.90	0.81	0.59	111	124	169
85	0.85	0.72	0.44	118	138	225

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

Even a small learning rate can have a major effect on performance. X = experience level of the operator when time study was taken, for example, 50, 100, or 500 cycles. The table gives time/unit based on a time standard of 1.0 min/unit; therefore, if actual time standard were 5.0 min/unit, then time/unit at 98% and 2X would be 0.98 (5.0) = 4.9.

4.1.6.4. Acceptable Day's Work Assume a time standard y is set at an experience level x. Assume further, for ease of understanding, that $y = 1.0$ min and $x = 100$ units. A time study technician, Bill, made a time study of the first 100 units produced by Sally, calculated the average time, and then left. Sally continued working. Let's assume a 95% rate is appropriate. Sally completes the 200th piece shortly before lunch and the 400th by the end of the shift. The average time/unit for the first day is about 0.9 min (111% of standard). The second day, Sally completes the 800th unit early in the afternoon. She will complete the 3200th unit by the end of the week. The average time of 0.77 min/unit during the week yields 129% of standard!

Point. If none of the operators in your plant ever turns in a time that improves as they gain experience, do you have stupid operators or stupid supervisors?

Next point. The magnitude of the learning effect dwarfs potential errors in rating. You might be off 5% or 10% in rating but this effect is trivial compared versus the errors in using a time standard without considering learning/mfg. progress.

Third point. Any time standard that does not consider learning/manufacturing progress will become less and less accurate with the passage of time.

Final point. Since learning and manufacturing progress are occurring, output and number of work hours should not both be constant. Either the same number of people should produce more or fewer people can produce a constant output.

5. DOCUMENTING, USING, AND MAINTAINING STANDARDS

5.1. Documenting Standards

Standards are part of a goal-setting system, and control is essential to any goal-setting system.

The more detailed the data, the more detailed the possible analysis; you can always consolidate data after they is gathered, but you can't break the data down if they are consolidated before they are gathered. Computerization permits detailed recording and thus analysis of downtime, machine breakdown, setup time, and so on. With bar coding of parts and computer terminals at workstations, it is feasible to record times for each individual part. For example, for product Y, operator 24 completed operation 7 on part 1 at 10:05, part 2 at 10.08, and so on. More commonly, however, you would just record that for product Y, operator 24 started operation 7 at 10:00 and completed the 25 units at 11:50. Least useful would be recording, for product Y, that all operations on 25 units were completed on Tuesday. Companies tend to develop elaborate codes for types of downtime, quality problems, and so on. Be careful about downtime reporting: it is easy to abuse.

5.2. Using Standards

5.2.1. Reports

Variance is the difference between standard performance and actual performance. Generally, attention is focused on large negative variances so that impediments to productivity can be corrected. Performance should be fed back to both workers and management at least weekly. Daily reports highlight delays and production problems; monthly reports smooth the fluctuations and show long-term trends.

5.2.2. Consequences of Not Making Standard

If the standard is used to determine an acceptable day's work or pay incentive wages, the question arises, What if a person doesn't produce at a standard rate?

For a typical low-task standard such as methods time measurement (MTM), over 99% of the population should be able to achieve standard, especially after allowances are added. The relevant question is not what workers are able to do but what they actually do.

Comparisons of performance vs. standard should be over a longer time period (such as a week) rather than a short period (such as a day).

The first possibility to consider is learning. As noted above, for cognitive work, the first cycle time may be as much as 13 times MTM standard and, for motor work, 2.5 times MTM standard. If a new operator's performance is plotted vs. the typical learning curve for that job, you can see whether the operator is making satisfactory progress. For example, Jane might be expected to achieve 50% of standard the first week, 80% the second week, 90% the third week, 95% the fourth week, and 100% the fifth week. Lack of satisfactory progress implies a need for training (i.e., the operator may not be using a good method).

If Jane is a permanent, experienced worker, the below-standard performance could be considered "excused" or "nonexcused." Excused failure is for temporary situations—bad parts from the supplier, back injuries, pregnancy for females, and so forth. For example, "Employees returning to work from Worker's Compensation due to a loss-of-time accident in excess of 30 days will be given consideration based upon the medical circumstances of each individual case."

Nonexcused performances are those for which the worker is considered capable of achieving the standard but did not make it.

Table 16 shows some example penalties. Most firms have a "forget" feature; for example, one month of acceptable performance drops you down a step. The use of an established discipline procedure allows workers to self-select themselves for a job. The firm can have only minimal preemployment screening and thus not be subject to discrimination charges.

Standards are based on an eight-hour day, but people work longer and shorter shifts. Because most standards tend to be loose, people can pace themselves and output/hr tends to be constant over the shift. I do not recommend changing the standard for shifts other than eight hours.

The level at which discipline takes place is negotiable between the firm and the union. For example, it might be 95% of standard—that is, as long as workers perform above 95% of standard, they are considered satisfactory. However, this tends to get overall performance slightly higher—say 98%. The best long-range strategy probably is to set discipline at 100% of standard. Anything less will give a long-term loss in production, especially if a measured daywork system is used instead of incentives.

Firms can use several strategies to improve the group's performance and reduce output restrictions. Basically, they allow the employees as well as the firm to benefit from output over 100%.

The primary technique is to give money for output over 100%. A 1% increase in pay for a 1% increase in output is the prevalent system.

Another alternative is to give the worker time off for output over 100%. For example, allow individuals to "bank" weekly hours earned over 100%. Most people will soon run up a positive balance to use as "insurance." This can be combined with a plan in which all hours in the bank over (say) 20 hr are given as scheduled paid time off. This tends to drop absenteeism because workers can use the paid time off for personal reasons.

5.3. Maintaining Standards (Auditing)

A standard can *restrict* productivity if it has not been updated because workers will produce only to the obsolete standard and not to their capabilities.

What to audit? To keep the work-measurement system up to date, accurate, and useful, MIL-STD-1567A says the audit should determine (1) the validity of the prescribed coverage, (2) the percentage of type I and II coverage, (3) use of labor standards, (4) accuracy of reporting, (5) attainment of goals, and (6) results of corrective actions regarding variance analysis.

How often to audit? Auditing should be on a periodic schedule. A good procedure is to set an expiration date on each standard at the time it is set; MIL-STD-1567 says annually. A rule of thumb is to have it expire at 24 months if the application is <50 hr/year, at 12 months if between 50 and 600 hr/year, and at 6 months if over 600 hr/year. Then, when the standard expires, and if it is still an active job, an audit is made. If it is not active, the standard will be converted from permanent to temporary. Then, if the job is resumed, the temporary can be used for a short period (e.g., 30 days) until a new permanent standard is set. An advantage of a known expiration date is that if a standard is audited (and perhaps tightened), the operator will not feel picked on.

If the resources for auditing are not sufficient for doing all the audits required, use the Pareto principle. Audit the "mighty few" and don't audit the "insignificant many." However, when the standard on one of the insignificant many passes the expiration date, convert the permanent standard to temporary.

TABLE 16 Example Discipline Levels for Not Producing Enough

Step	Description
0	Normal operator, acceptable performance
1	Oral warning
2	Oral warning; detailed review of method with supervisor or trainer
3	Written warning; additional training
4	Written warning; some loss of pay
5	Written warning; larger loss of pay
6	Discharge from job

From *Work Design: Industrial Ergonomics,* 5th Ed., by S. Konz and S. Johnson. Copyright © 2000 by Holcomb Hathaway, Pub., Scottsdale, AZ. Reprinted with permission.

A published set of rules ensures that everyone is treated fairly. Most organizations have a similar set of rules for tardiness and absenteeism.

REFERENCES

Allemang, R. (1977), "New Technique Could Replace Learning Curves," *Industrial Engineering,* Vol. 9, No. 8, pp. 22–25.

Dar-El, E., Ayas, K., and Gilad, I. (1995a), "A Dual-Phase Model for the Individual Learning Process in Industrial Tasks," *IIE Transactions,* Vol. 27, pp. 265–271.

Dar-El, E., Ayas, K. and Gilad, I. (1995b), "Predicting Performance Times for Long Cycle Tasks," *IIE Transactions*, Vol. 27, pp. 272–281.

International Labour Office (1992), *Introduction to Work Study*, 4th (Rev.) Ed., International Labour Office, Geneva.

Konz, S., and Johnson, S. (2000), *Work Design: Industrial Ergonomics,* 5th Ed., Holcomb Hathaway, Scottsdale, AZ.

Salvendy, G., and Seymour, W. (1973), *Prediction and Performance of Industrial Work Performance*, John Wiley & Sons, New York, p. 17.

Wright, T. (1936), "Factors Affecting the Cost of Airplanes," *Journal of Aeronautical Sciences*, Vol. 3, February, pp. 122–128.

CHAPTER 54

Work Measurement: Principles and Techniques

AURA CASTILLO MATIAS
University of the Philippines Diliman

1. WORK MEASUREMENT: AN OVERVIEW 1410
- 1.1. Basic Procedure of Work Measurement 1410
- 1.2. Work Measurement Techniques 1411

2. TIME STUDY 1411
- 2.1. A Fair Day's Work 1411
- 2.2. Time Study Equipment 1411
  - 2.2.1. Time-Recording Equipment 1411
  - 2.2.2. Time Study Board 1414
  - 2.2.3. Time Study Forms 1414
- 2.3. Requirements for Effective Time Study 1414
- 2.4. Conducting the Study 1417
  - 2.4.1. Choosing an Operator 1417
  - 2.4.2. Breaking the Job into Elements 1418
  - 2.4.3. Number of Cycles to Study 1419
  - 2.4.4. Principal Methods of Timing 1420
  - 2.4.5. Recording Difficulties Encountered 1420
- 2.5. Rating the Operator's Performance 1422
  - 2.5.1. Scales of Rating 1423
  - 2.5.2. Selecting a Rating System 1425
- 2.6. Allowances 1426
- 2.7. Calculating the Standard Time 1426
  - 2.7.1. Temporary Standards 1427
  - 2.7.2. Setup Standards 1427

3. PREDETERMINED TIME STANDARDS 1427
- 3.1. Methods–Time Measurement (MTM) 1429
  - 3.1.1. MTM-1 1429
  - 3.1.2. MTM-2 1429
  - 3.1.3. MTM-3 1435
  - 3.1.4. MTM-V 1436
  - 3.1.5. MTM-C 1436
  - 3.1.6. MTM-M 1438
  - 3.1.7. Specialized MTM Systems 1438
- 3.2. Maynard Operations Sequence Technique (MOST) 1439
- 3.3. Macromotion Analyses 1441
- 3.4. Guidelines for System Selection 1442

4. STANDARD DATA 1443
- 4.1. Developing Standard Time Data 1447
- 4.2. Uses of Standard Data 1448

5. WORK SAMPLING 1448
- 5.1. Definitions and Objectives of Work Sampling Studies 1448
- 5.2. Work Sampling Methodology 1449
- 5.3. Work Sampling Study Plans 1451
  - 5.3.1. Determining the Observations Needed 1451
  - 5.3.2. Determining Observation Frequency 1453
  - 5.3.3. Observations and Data Recording 1456
  - 5.3.4. Using Control Charts 1457

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

5.4. Establishing Standard Times 1457
5.5. Computerized Work Sampling 1458
6. MEASUREMENT OF INDIRECT LABOR OPERATIONS 1458
6.1. Indirect and Expense Work Standards 1459
6.2. Advantages of Indirect Work Standards 1460
7. SELECTED SOFTWARE 1462
REFERENCES 1462
ADDITIONAL READING 1462

1. WORK MEASUREMENT: AN OVERVIEW

Work measurement, as the name suggests, provides management with a means of measuring the time taken in the performance of an operation or series of operations. Work measurement is the application of techniques designed to establish standard times for a qualified worker to carry out a specified job at a defined level of performance. With today's increasing global competition among producers of products or providers of service, there has been an increasing effort to establish standards based on facts and scientific methods rather than the use of estimates based on judgment or experience (Niebel and Freivalds 1999). Sound standards have many applications that can mean the difference between the success or failure of a business. Accurately established time standards make it possible to produce more within a given plant, thus increasing the efficiency of the equipment and the optimal utilization of personnel. Poorly established standards, although better than no standards at all, lead to high costs, low productivity, and labor unrest.

Common uses of work measurement include the following:

- To compare the efficiency of alternative methods. Other conditions being equal, the method that takes the least time will be the best method.
- To balance the work of members of teams, in association with multiple activity charts, so that, as nearly as possible, each member has tasks taking an equal time to perform.
- To determine, in association with man and machine multiple activity charts, the number of machines an operator can run.
- To provide information on which the planning and scheduling of production can be based, including plant and labor requirements for carrying out the program of work, the utilization of available machine and labor capacity, and delivery promises.
- To provide information for labor cost control and to enable standard costs to be fixed and maintained.
- To provide information on which incentive plans can be based.

Time standards are always based on a specific method and working conditions. As time passes, improvements may be introduced, whether by the supervisor, the methods engineer, management, or the operator. Regardless of who introduced the method changes, they will directly affect the time standard being applied. This should signal the need for the method to be restudied and reevaluated. The time study analyst should be advised of such changes so that the portion of the operation that is affected by the change can be reviewed and the time standard recalculated.

To ensure that the methods applied when the time study was conducted are still being employed, regular audits should be scheduled by the time study analysts. The more frequently a standard is applied, the more frequent the audits should be. Time study audits do not necessarily lead to conducting a time study of the entire method. If, however, a major modification in the method is observed (e.g., change in tools used, different sequence in operations, change in materials processed, change in process or approach), then a new detailed study must be performed.

1.1. Basic Procedure of Work Measurement

In general, the basic procedure of work measurement uses the following steps:

1. *Select* the work to be studied.
2. *Record* all the relevant data relating to the circumstances in which the work is being done, the methods, and the elements of activity in them.
3. *Examine* the recorded data and the detailed breakdown critically to ensure that the most effective method and motions are being used and that unproductive and foreign elements are separated from productive elements.

4. *Measure* the quantity of work involved in each element in terms of time using the appropriate work-measurement technique.
5. *Calculate* the standard time for the operation, which in the case of stopwatch time study will include time allowances to cover relaxation, personal needs, etc.
6. *Define* precisely the series of activities and method of operation for which the time has been compiled and issue the time as standard for the activities and methods specified.

1.2. Work Measurement Techniques

Operations managers and engineering staff need to be aware of the many different work measurement techniques that are available so that they can make appropriate choices as operating conditions change. Successful installation of any of the work-measurement techniques needs the full support of management to commit the time and financial resources necessary on a continuing basis. The objective of this chapter is to discuss these different alternatives, compare their advantages and disadvantages, and offer some guidelines for proper selection and application.

There are four types of work measurement techniques: time study, predetermined time standards (PTS), standard data, and work sampling. All of these techniques are based on facts and consider each detail of work. The first three techniques measure the normal time required to perform the entire work cycle. Work sampling measures the proportion of time of various work activities that constitute a job. Table 1 describes these techniques and where each is appropriately used. Table 2 provides a further comparison among work measurement techniques in terms of their advantages and disadvantages.

2. TIME STUDY

Time study is a work measurement technique for recording the times and rates of working for the elements of a specified job carried out under specified conditions and for analyzing the data so as to obtain the time necessary for carrying out the job at a defined level of performance. The objective of time study is to determine reliable time standards for all work, both direct and indirect, with due allowance for fatigue and for personal and unavoidable delays, for the efficient and effective management of operations.

The establishment of reliable and accurate time standards makes it possible for companies to define their capacity or output, thus increasing the efficient use of equipment and obtaining optimum utilization of personnel. Because time is a common measure for all jobs, time standards can be used to investigate the difference between actual and standard performance and take appropriate action where necessary. It can also be used to facilitate job design as a basis for comparing various methods of doing the job, purchasing the most productive new equipment, introducing sound production controls, designing an efficient workplace layout, and balancing between work schedules and available manpower. Other benefits of establishing reliable time standards include budgetary control, development of incentive plans, and ensuring that quality specifications are met.

2.1. A Fair Day's Work

Time study is often referred to, among industry practitioners, as a method of determining a "fair day's work." In general, that means fair to both the company and the employee. Employees are expected to give the full day's work that they get paid for, with reasonable allowances for personal delays, unavoidable delays, and fatigue. Employees are expected to perform the prescribed method at a pace, neither fast nor slow, that may be considered representative of a full day's output by a well-trained, experienced, and cooperative operator.

2.2. Time Study Equipment

The equipment needed to develop reliable standards is minimal, easy to use and often inexpensive. Only four basic items are needed: an accurate and reliable stopwatch, a time study board, a well-designed time study form, and a calculator to compute the recorded observations. Videotape equipment can also be very useful.

2.2.1. Time-Recording Equipment

Several types of time-recording equipment are available today. The most common are decimal minute stopwatches, computer-assisted electronic stopwatches, and videotape cameras.

Two types of stopwatches are used today, the mechanical decimal minute watch and the electronic stopwatch. The mechanical decimal minute watch has 100 divisions on its face, each division equal to 0.01 min. One minute requires one revolution of the long hand. The small dial on the watch face has 30 divisions, each representing 1 min. For every full revolution of the long hand, the small hand moves 1 division, or 1 min.

Electronic stopwatches provide resolution to 0.001 sec and an accuracy of ±0.002%. They are lightweight and have a digital display. They permit the timing of any number of individual elements

TABLE 1 Techniques of Work Measurement

Method	Definition and Where Used
Time study	Technique for recording the times of performing a certain job or its elements carried out under specified conditions and for analyzing the data so as to obtain the time necessary for an operator to carry it out at a defined rate of performance. • Where there are repetitive work cycles of short to long duration OR • Where wide variety of dissimilar work is performed OR • Where process control elements constitute a part of the cycle
Predetermined time standards (PTS)	Technique whereby times established for basic human motions (classified according to the nature of the motion and the conditions under which it is made) are used to build up the time for a job at a defined level of performance. • Where work is predominantly operator controlled OR • Where there are repetitive cycles of short to medium duration OR • Where it is necessary to plan work methods in advance of production OR • Where there has been controversy over time study results OR • Where there has been controversy over consistency of existing standards
Standard data	Technique that refers to all the tabulated elemental standards, curves, alignment charts, and tables that are compiled from time studies and predetermined time standards (PTS) to allow the measurement of a specific job without the use of a timing device. Formula construction represents a simplification of standard data. It involves design of algebraic expression or a system of curves that establishes a time standard in advance of production by substitution of known values peculiar to the job for the variable elements. • Where there are similar work of short to long duration OR • Where there has been controversy over time study results OR • Where there has been controversy over consistency of existing standards
Work sampling	Technique used to investigate the proportions of total time devoted to the various activities that constitute a job or work situation. • Where there are considerable differences in work content from cycle to cycle, as in shipping, materials handling, and clerical activities, OR • Where activity studies are needed to show machine or space utilization, or the percentage of time spent on various activities OR • Where there is an objection to stopwatch time studies

while also measuring the total elapsed time. Thus, they can be conveniently used for both continuous and snapback timing (see Section 2.4.4). Electronic watches have become increasingly more affordable than mechanical stopwatches, and it is expected that mechanical watches will quickly disappear from use.

Three types of computer-assisted electronic stopwatches are now on the market: the DataMyte 1010, the COMPU-RATE, and the OS-3 Plus Event Recorder. The DataMyte 1010 all-solid-state battery-operated data collector, developed by the Electro/General Corporation in 1971, is a practical alternative to both mechanical and electronic stopwatches. Observed data are keyed in and recorded in a solid-state memory in computer language. Elapsed time readings are recorded and computed automatically. Data recordings may be directly downloaded from the DataMyte to most personal computers through an output cable. The instrument is self-contained and can be carried around the workplace. The rechargeable battery power provides approximately 12 hr of continuous operation. It takes 40–50% less time to conduct time studies with the DataMyte and a computer than with a stopwatch and a hand calculator.

TABLE 2 Comparison of Different Work-Measurement Techniques

Time Study	Predetermined Time Standards	Standard Data	Work Sampling
Advantages:			
1. Enables analysts to observe the complete cycle, providing an opportunity to suggest and initiate methods improvement. 2. The only method that actually measures and records the actual time taken by an operator. 3. More likely to provide coverage of those elements that occur less than one per cycle. 4. Quickly provides accurate values for machine-controlled elements 5. Relatively simple to learn and explain.	1. Forces detailed and accurate descriptions of the workplace layout; motion patterns; and shape, size, and fit of components and tools. 2. Encourages work simplification to reduce standard times. 3. Eliminates performance rating. 4. Permits establishing methods and standards in advance of actual production. 5. Permits easy and accurate adjustments of time standards to accommodate minor changes in method 6. Provides more consistent standards.	1. Averages performance rating. 2. Establishes consistent standards. 3. Permits establishing methods and standards in advance of actual production. 4. Allows standards to be established rapidly and inexpensively. 5. Permits easy adjustment of time standards to accommodate minor changes in method.	1. Eliminates tension caused by constant observation of the worker. 2. Does not require continuous observations over a long period of time. 3. Represents typical or average conditions over a period of time where conditions change from hour to hour or day to day. 4. Permits simultaneous development of standards from different operations. 5. Ideally suited to studies of machine utilization, activity analyses, and delays. 6. Can be used with performance rating to determine standard times.
Disadvantages:			
1. Requires the performance rating of a worker's skill and effort. 2. Requires continuous observation of the worker over repeated work cycles. 3. May not provide accurate evaluation of noncyclic elements. 4. Requires work to be performed by an operator.	1. Depends on complete and accurate descriptions of the required methods for the accuracy of the time standard. 2. Requires more time for the training of competent analysts. 3. More difficult to explain to workers and supervisors. 4. May require more work-hours to establish standards for long-cycle operations. 5. Must use stopwatch or standard data for process-controlled and machine-controlled elements.	1. Involves significant costs in setting up the standard data. Once these costs are incurred, additional costs are needed to maintain the system. 2. May not accommodate small variations in method. 3. May require more skilled analysts for complex formulas. 4. More difficult to explain to workers and supervisors. 5. May result to significant inaccuracies if extended beyond the scope of the data used in their development.	1. Assumes that the worker uses an acceptable and standard method. 2. Requires that the observer to identify and classify a wide range of work activities and delays. 3. Should be confined to constant populations. 4. Makes it more difficult to apply a correct performance rating factor than does time study.

The COMPU-RATE, developed by Faehr Electronic Timers, Inc., is a portable device using batteries, which provide about 120 hr of running time. Manual entries are required only for the element column and the top four lines of the form. This allows the analyst to concentrate on observing the work and the operator performance. Time can be in thousandths of a minute or one hundred-thousandths of an hour. The COMPU-RATE software system computes the mean and median element values, adjustment of mean times to normal time after inputting the performance rating, and allowed times in minutes and/or hours per piece and pieces per hour. Errors can be corrected through an edit function.

The GageTalker Corporation (formerly Observational Systems) markets the OS-3 Plus Event Recorder. This versatile recorder is useful for setting and updating standard times, machine downtime studies, and work sampling. The device allows the analyst to select the time units appropriate for the study—0.001 min, 0.0001 hr, or 0.1 sec. The total time, frequency, mean times, performance ratings, normal times, allowances, standard times, and pieces per hour may be printed through printer interface devices. In addition, the standard deviation, the maximum and minimum element values and frequencies are also available.

Videotape cameras are an excellent means for recording operator's methods and elapsed time to facilitate performance rating. By recording exact details of the method used, analysts can study the operation frame by frame and assign normal time values. Projecting the tape at the same speed at which the pictures were taken and then rating the performance of the operator can also be used. For example, a half minute to deal 52 cards into a four-hand bridge deck is considered by many to be typical of normal performance. An operator carrying a 9 kg (20 lb) load a distance of 7.6 m (25 ft) may be expected to take 0.095 min when working at a normal pace (Niebel 1992).

Observing the videotape is a fair and accurate way to rate performance because all the facts are documented. Also, the videotape can be used to uncover potential methods improvements that otherwise would have not been observed using a stopwatch procedure. Videotapes are also excellent training material for novice time study analysts, particularly in acquiring consistency in conducting performance rating.

2.2.2. Time Study Board

It is often convenient for time study analysts to have a suitable board to hold the time study form and the stopwatch. The board should be light but made of sufficiently hard material. Suitable materials include ¼ in. plywood and smooth plastic. The board should have both arm and body contacts for comfortable fit and ease of writing while being held. A clip holds the time study form. The stopwatch should be mounted in the upper right-hand corner of the board for right-handed observers. Standing in the proper position, the analyst should be able to look over the top of the watch to the workstation and follow the operator's movements while keeping both the watch and time study form in the immediate field of vision.

2.2.3. Time Study Forms

It is important that a well-designed form be used for recording elapsed time and working up the study. The time study form is used to record all the details of the time study. It should be designed so that the analyst can conveniently record watch readings, foreign elements (see Section 2.4.2), and rating factors and calculate the allowed time. It should also provide space to record all pertinent information concerning the operator being observed, method studied, tools and equipment used, department where the operation is being performed, and prevailing working conditions. Figures 1 and 2 illustrates a time study form (Niebel 1992) that is sufficiently flexible to be used for practically any type of operation. Various elements of the operation are recorded horizontally across the top of the sheet, and the cycles studied are entered vertically row-wise. The four columns under each element are R for ratings, W for watch reading, OT for observed time, and NT for normal time. At the back of the time study form, a sketch of the workstation layout is drawn and a detailed description is given of the work elements, tools and equipment utilized, and work conditions. Computed results of the time study can also be summarized next to each work element defined.

2.3. Requirements for Effective Time Study

Certain fundamental requirements must be met before the time study can be undertaken. First, all details of the method to be followed and working conditions must be standardized at all points where it is to be used before the operation is studied. Furthermore, the operator should be thoroughly acquainted with the prescribed method to be followed before the study begins. Without standardization, the time standards will have little value and will continually be a source of distrust, grievances, and internal friction. Reasons for selecting a particular job may be any of the following:

1. The job in question is a new one, not previously carried out (new product, component, operation, or set of activities)
2. A change in material or method of working has been made and a new time standard is required.

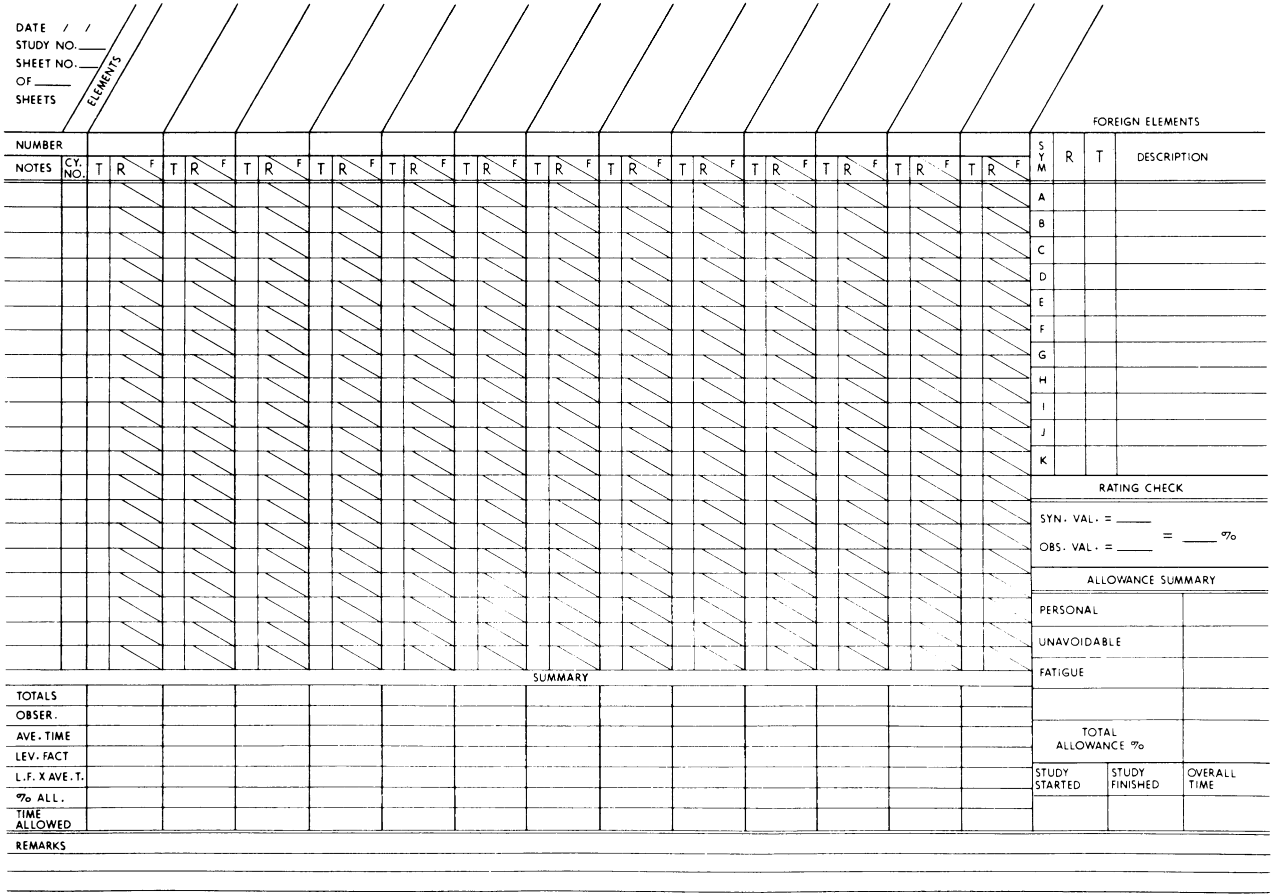
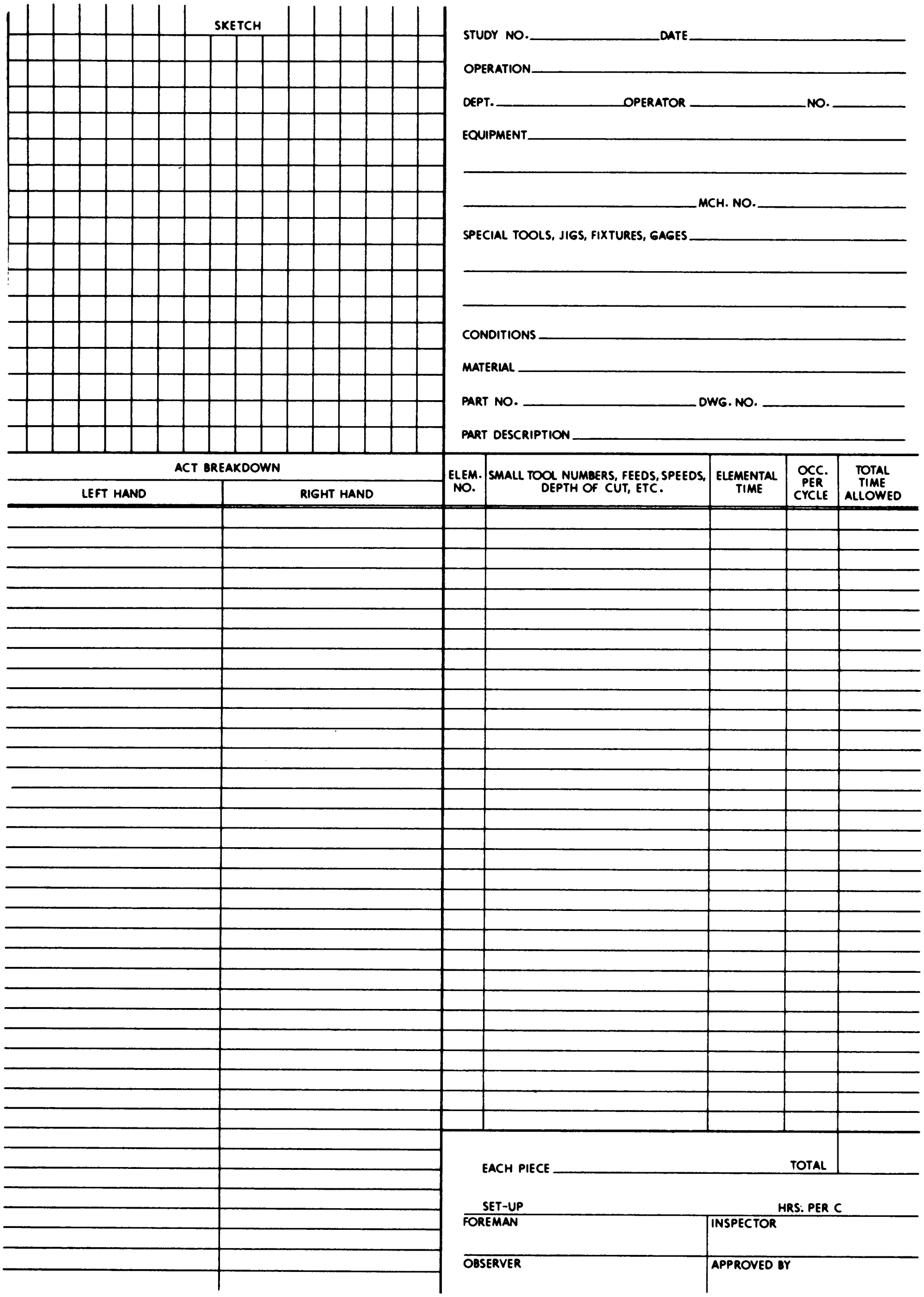

DATE / /
STUDY NO.____
SHEET NO.____
OF ____
SHEETS
ELEMENTS
NUMBER
NOTES
CY. NO.
T
R
F
FOREIGN ELEMENTS
SYM
R
T
DESCRIPTION
A
B
C
D
E
F
G
H
I
J
K
RATING CHECK
SYN. VAL. = ____
OBS. VAL. = ____
= ____ %
ALLOWANCE SUMMARY
PERSONAL
UNAVOIDABLE
FATIGUE
TOTAL ALLOWANCE %
STUDY STARTED
STUDY FINISHED
OVERALL TIME
SUMMARY
TOTALS
OBSER.
AVE. TIME
LEV. FACT
L.F. X AVE. T.
% ALL.
TIME ALLOWED
REMARKS

Figure 1. Front of Time Study Form. (From Niebel 1992)



SKETCH

STUDY NO. ____________ DATE ____________

OPERATION ____________

DEPT. ____________ OPERATOR ____________ NO. ____________

EQUIPMENT ____________

____________ MCH. NO. ____________

SPECIAL TOOLS, JIGS, FIXTURES, GAGES ____________

CONDITIONS ____________

MATERIAL ____________

PART NO. ____________ DWG. NO. ____________

PART DESCRIPTION ____________

ACT BREAKDOWN		ELEM. NO.	SMALL TOOL NUMBERS, FEEDS, SPEEDS, DEPTH OF CUT, ETC.	ELEMENTAL TIME	OCC. PER CYCLE	TOTAL TIME ALLOWED
LEFT HAND	RIGHT HAND					

EACH PIECE ____________ TOTAL

SET-UP ____________ HRS. PER C

FOREMAN ____________ INSPECTOR

OBSERVER ____________ APPROVED BY

Figure 2 Back of Time Study Form. (From Niebel 1992)

3. A complaint has been received from a worker or worker's representative about the time standard for an operation.
4. A particular operation appears to be a bottleneck holding up subsequent operations and possibly (through accumulations of work in process behind it) previous operations.
5. Standard times are required before an incentive plan is introduced.
6. A piece of equipment appears to be idle for an excessive time or its output is low, and it therefore becomes necessary to investigate the method of its use.
7. The job needs studying as a preliminary to making a methods study or to compare the efficiency of two proposed methods.
8. The cost of a particular job appears to be excessive.

Second, analysts should inform the union steward, the department supervisor, and the operator that the job is to be studied. Each of these parties can then make specific preparations to allow a smooth and coordinated study. The operator should verify that he or she is performing the correct method and should become acquainted with all details of that operation. The supervisor should check the method in advance to ensure that feeds, speeds, cutting tools, lubricants, and so forth conform to standard practice (as established by the methods department) and that sufficient material is available on hand. If several operators are available for the study, the time study analyst and the supervisor should select the operator that represents the average or above-average performance within the group (Niebel 1992). This type of operator will give the most satisfactory results than either a low-skilled or highly superior operator (see Section 2.4.1 for choosing an operator). The union steward should then make sure that only trained, competent operators are selected for the study, explain to the operator why the study is being undertaken, and answer any pertinent questions raised by the operator on the study.

Because of many human interests and reactions associated with the time study technique, a full understanding among the supervisor, employee, union steward, and time study analyst is essential. Since time standards directly affect both the employee and company financially, time study results must be completely dependable and accurate. Inaccuracies and poor judgment may result in loss of confidence by the operator and the union, which may ultimately undo harmonious labor relations between the union and management. To achieve and maintain good human relations, the time study analyst should ensure that the correct method is being used, accurately record the times taken, honestly evaluate the performance of the operator, and conduct himself or herself in a manner that will gain and hold the respect and confidence of both labor and management.

2.4. Conducting the Study

To ensure success, the actual conduct of a time study must be systematic and apply a scientific approach. The following elements should be included:

1. Obtaining and recording all the information available about the job, the operator, and surrounding conditions that is likely to affect the carrying out of the work
2. Recording a complete description of the method, breaking down the operation into elements (see Section 2.4.2)
3. Examining the detailed breakdown to ensure that the most effective method and motions are being used and determining the number of cycles to study
4. Measuring with a timing device (usually a stopwatch) and recording the time taken by the operative to perform each element of the operation
5. At the same time, assessing the effective speed of work of the operator relative to the observer's concept of the rate corresponding to standard rating
6. Extending the observed times to basic times
7. Determining the allowances for the operation
8. Determining the standard time for the operation

2.4.1. Choosing an Operator

After reviewing the job in operation, both the supervisor and the analyst should agree that the job is ready to be studied. If more than one person is performing the job, several criteria should be considered when selecting a qualified operator for the study. In general, a qualified operator is one who has acquired the skill, knowledge, and other attributes to carry out the work to satisfactory standards of quantity, physical specifications, and safety. Normal pace is neither fast nor slow and gives due consideration to the physical, mental, and visual requirements of a specific job.

The qualified operator usually performs the work consistently and systematically, thereby making it easier for the analyst to apply a correct performance factor. The operator should be completely trained in the method and should demonstrate motivation in doing a good job. The operator should have confidence in both time study methods and the analyst, be cooperative, and willingly follow suggestions made by the supervisor or the analyst.

In cases where only one operator performs the operation, the analyst needs to be especially careful in establishing the performance rating. It is possible that the operator may be performing either fast or slow.

Once the qualified operator has been selected, the analyst should approach the operator in a friendly manner and demonstrate understanding of the operation to be studied to obtain full cooperation. The operator should have the opportunity to ask questions and be encouraged to offer suggestions in improving the job. The analyst, in return, should answer all queries frankly and patiently and willingly receive the operator's suggestions. Winning the respect and goodwill of the operators helps establish a fair and acceptable time standard.

2.4.2. Breaking the Job into Elements

For ease of measurement, the operation should be divided into groups of motions known as elements. An element is a distinct part of a specified job selected for convenience of observation. It is a division of work that can be easily measured with a stopwatch and that has readily identifiable breakpoints. To divide the operation into individual elements, the analyst should carefully observe the operator over several cycles. A work cycle is the sequence of elements that are required to perform a job or yield a unit of production. The sequence may include elements that do not occur every cycle.

It is desirable to determine, prior to conducting the time study, the elements into which the operation is divided. Breaking down the operation into elements improves method description and provides good internal consistency during the time study. Elements also make it possible to reuse the data and permit different ratings for the work cycle.

Elements are classified into the following types:

1. Repetitive element: an element that occurs in every work cycle of the job
2. Occasional element: an element that does not occur in every work cycle of the job but that may occur at regular or irregular intervals
3. Constant element: an element for which the basic time remains constant whenever it is performed (e.g., stop machine)
4. Variable element: an element for which the basic time varies in relation to some characteristics of the product, equipment, or process (e.g., walk "X" meters)
5. Manual element: an element performed by a worker
6. Machine element: an element automatically performed by a power-driven machine or process
7. Governing element: an element occupying a longer time than that of any other element being performed concurrently
8. Foreign element: an element observed during a study that after analysis is not found to be a necessary part of the job

Each element should be recorded in proper sequence, including a definition of its breakpoints or terminal points. The analyst should follow some basic guidelines when dividing the job into elements:

1. Ascertain that all elements being performed are necessary. If some are unnecessary and the objective is to come up with a standard time, the time study should be discontinued and a method study should be conducted.
2. Elements should be easily identifiable (with clear start and break-off points) so that once established, they can be repeatedly recognized. Sounds (such as machine motor starts and stops) and change in motion of hand (putting down tool, reaching for material) are good breakpoints.
3. Elements should be as short in duration as possible to be timed accurately. The minimum practical units for timing are generally considered to be 0.04 min or 2.4 sec (Niebel and Freivalds 1999). For less-trained observers, it may be 0.07–0.10 minutes. Very short cycles should, if possible, be next to long elements.
4. Elements should be chosen so that they represent naturally unified and recognizably distinct segments of the operation. To illustrate, reaching for a tool can be detailed as reaching, grasping, moving, and positioning. This can be better treated as a whole and describe as "obtaining and positioning of wrench."
5. Irregular elements should be separated from regular elements.
6. Machine-paced elements should be separated from operator-controlled elements; the division helps recognize true delays.

7. Constant elements should be separated from variable elements.
8. Foreign or accidental elements should be identified.
9. End points should be defined (break-off points).

2.4.3. Number of Cycles to Study

Since time study is a sampling procedure, it is important that an adequately sized sample of data be collected to ensure that the resulting standard is reasonably accurate. From an economic standpoint, the activity of the job and the duration of the cycle influence the number of cycles that can be studied. The General Electric Company has established Table 3 as an approximate guide to the number of cycles to observe.

A more accurate sample size can be established using statistical methods. It is known that averages of samples $\bar{x}$ drawn from a normal distribution of observations are distributed normally about the population mean μ. The variance about the population mean μ equals σ^2/n, where n equals the sample size and σ^2 equals the population variance. Normal curve theory leads to the following confidence interval:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

The preceding equation assumes that the standard deviation of the population is known. In general, this is not true, but the population standard deviation may be estimated by the sample standard deviation s, where:

$$s = \sqrt{\frac{\Sigma^2 x_i^2}{n - 1} - \frac{(\Sigma x_i)^2}{n(n - 1)}}$$

However, since pilot time studies involve only small samples ($n < 30$) of a population; a t-distribution must be used. The confidence interval equation is then:

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

The ± term can be considered an error term expressed as a fraction of $\bar{x}$:

$$k\bar{x} = \frac{ts}{\sqrt{n}}$$

where k = an acceptable percentage of $\bar{x}$. If we let N be the number of observations for the actual time study, solving for N yields:

TABLE 3 Recommended Number of Observation Cycles

Cycle Time, min	Recommended Number of Cycles
0.01	200
0.25	100
0.50	60
0.75	40
1.00	30
2.00	20
2.00–5.00	15
5.00–10.00	10
10.00–20.00	8
20.00–40.00	5
40.00–above	3

From B. Niebel and A. Freivalds, *Methods, Standards, and Work Design,* 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.

$$N = \left\{ \frac{st}{k\bar{x}} \right\}^2$$

For example, a pilot study of 30 readings for a given element showed that $\bar{x} = 0.25$ and $s = 0.08$. A 5% probability of error for 29 DOF (30 − 1 DOF) yields $t = 2.045$ (see Table 4 for values of t). Solving for N yields (rounded up):

$$N = \left[\frac{(0.08)(2.045)}{(0.05)(0.25)} \right]^2 = 171.3 \approx 172 \text{ observations}$$

Thus, a total of 172 observations need to be taken for the pilot sample $\bar{x}$ to be within ±5% of the population mean μ.

2.4.4. *Principal Methods of Timing*

There are two techniques of recording the elemental times during a time study: the continuous method and the snapback method. Each technique has its advantages and disadvantages, which are described as follows.

2.4.4.1. Continuous Timing The watch runs continuously throughout the study. It is started at the beginning of the first element of the first cycle to be timed and is not stopped until the whole study is completed. At the end of each element, the watch reading is recorded. The purpose of this procedure is to ensure that all the time during which the job is observed is recorded in the study.

The continuous method is often preferred over the snapback method for several reasons. The principal advantage is that the resulting study presents a complete record of the entire observation period. Consequently, this type of study is more appealing to the operator and the union because no time is left out of the study, and that all delays and foreign elements have been recorded. The continuous method is also more suited for measuring and recording short elements. Since no time is lost in snapping the hand back to zero, accurate values are obtained on successive short elements.

More clerical work is involved when using the continuous method. The individual element time values are obtained on successive subtractions after the study is completed to determine the elapsed elemental times.

2.4.4.2. Snapback Timing The hands of the stopwatch are returned to zero at the end of each element and are allowed to start immediately. The time for each element is obtained directly. The mechanism of the watch is never stopped and the hand immediately starts to record the time of the next element. It requires less clerical work than the continuous method because time recorded is already time of elements. Elements observed out of order by the operator can also be readily recorded without special notation.

Among the disadvantages of the snapback method are:

1. If the stopwatch is analog, time is lost in snapping back to zero and cumulative error is introduced into the study. However, this error becomes insignificant when using a digital stopwatch because only the delay is in the display but internally it does not lag.
2. Short elements are difficult to time (0.04 min and less).
3. No verification of overall time with the sum of elemental watch readings is given.
4. A record of complete study is not always given in the snapback method. Delays and foreign elements may not be recorded.

2.4.5. *Recording Difficulties Encountered*

During the course of the study, the analyst may occasionally encounter variations in the sequence of elements originally established. First, the analyst may miss reading an element. Second, the analyst may observe elements performed out of sequence. A third variation is the introduction of foreign elements during the course of the study.

2.4.5.1. Missed Readings When missing a reading, analysts should in no circumstance approximate or endeavor to record the missed value. Put a mark M on the W column (referring to Figure 1) or the area where the time should have been logged. Occasionally, the operator omits an element. This is handled by drawing a horizontal line through the space in the W column. This occurrence should happen infrequently; it usually indicates an inexperienced operator or a lack of a standard method. Should elements be omitted repeatedly, the analyst should stop the study and investigate the necessity of performing the omitted elements in cooperation with the supervisor and the operator.

TABLE 4 Percentage Points of the *t*-Distribution

	Probability $(P)^a$												
n	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	363.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.449	5.405
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	3.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.265	0.391	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.265	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.265	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.265	0.390	0.531	0684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.265	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.265	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.265	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.265	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.656
30	0.127	0.265	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

From R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research,* 6th Ed., Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

[a] Probabilities refer to the sum of the two tail areas; for single tail, divide the probability by 2.

2.4.5.2. Variations from Expected Sequence of Work This happens frequently when observing new and inexperienced workers or on a long-cycle job made of many elements. If using snapback timing, just note the jump from one element to another by putting arrows. For continuous timing:

1. Draw a horizontal line in the box where the time is to be logged.
2. Below the line, log the time operator began the element, and above, record the time it ended.
3. Continue logging in this format until the next element observed is the normal element in the sequence.

2.4.5.3. Foreign Elements Foreign elements are unavoidable delays such as talk with the supervisor, tool breakage, material replenishment, etc. or personal delays such as drinking, resting, etc. A separate area should be assigned where details of the delay can be recorded (letter code A, B, C . . . , start and end time, description). Put the code in the box of the element done upon resumption.

2.4.5.4. Short Foreign Elements Occasionally, foreign elements can be of a short duration that makes it impossible to time. Delays like dropping and retrieving of tools, wiping of brows, or short answers to supervisor's inquiries are example of such situations. In these situations, continue timing the elements but encircle the time and put in the remarks column the foreign element encountered.

2.5. Rating the Operator's Performance

The purpose of performance rating is to determine, from the time actually taken by the operator being observed, the standard time that can be maintained by the average qualified worker and that can be used as a realistic basis for planning, control, and incentive plans. The speed of accomplishment must be related to an idea of a normal speed for the same type of work. This is an important reason for doing a proper method study on a job before attempting to set a time standard. It enables the analyst to gain a clear understanding of the nature of the work and often enables him or her to eliminate excessive effort or judgment and so bring the rating process nearer to a simple assessment of speed.

Standard performance is the rate of output that qualified workers will naturally achieve without undue fatigue or overexertion during a typical working day or shift, provided that they know and adhere to the specified method and are motivated to apply themselves to their work. This performance is denoted as 100 on the standard rating and performance scales. Depending on the skill and effort of the operator, it may be necessary to adjust upwards to normal the time of a good operator and adjust downwards to normal the time of a poor operator.

In principle, performance rating adjusts the mean observed time (OT) for each element performed during the study to the normal time (NT) that would be required by a qualified operator to perform the same work:

$$\mathrm{NT} = (\mathrm{OT})\left(\frac{R}{100}\right)$$

where R is expressed as a percentage. If the analyst decides that the operation being observed is being performed with less effective speed than the concept of standard, the analyst will use a factor of less than 100 (e.g., 75 or 90). If, on the other hand, the analyst decides that the effective rate of working is above standard, it has a factor greater than 100 (e.g., 115 or 120).

Unfortunately, there is no universally accepted method of performance rating, nor is there a universal concept of normal performance. The majority of rating systems used are largely dependent on the judgment of the time study analyst. It is primarily for this reason that the analyst reflects high personal qualifications.

To define normal performance, it is desirable for a company to identify benchmark examples so that various time study analysts can develop consistency in performance rating. The benchmark examples should be supplemented by a clear description of the characteristics of an employee carrying out a normal performance.

To do a fair job of rating, the time study analyst must evaluate the job in terms of the factors that affect the rate of working. Factors outside and within the worker's control include:

- *Outside the worker's control:*
  1. Variations in the quality or other characteristics of the material used, although they may be within the prescribed tolerance limits
  2. Changes in the operating efficiency of equipment within the useful life
  3. Minor and unavoidable changes in methods or conditions of operations
  4. Variations in the mental attention necessary for the performance of certain elements
  5. Changes in climatic and other surrounding conditions such as light and temperature.

- *Within the worker's control:*
    1. Acceptable variations in the quality of the product
    2. Variations due to his or her ability
    3. Variations due to attitude of mind, especially the operator's attitude to the organization

2.5.1. *Scales of Rating*

For a comparison between the observed rate of working and the standard rate to be made effectively, it is necessary to have a numerical scale against which to make the assessment. The rating can then be used as a factor by which the observed time can be multiplied to give the basic time, which is the time it would take the qualified worker, motivated to apply himself or herself, to carry out the element at standard rating.

There are several scales of rating in use, the most common of which are the following:

2.5.1.1. Westinghouse System Developed by Westinghouse Electric Corporation, this is one of the oldest rating systems. It has had wide application, especially on short-cycle, repetitive operation where performance rating of the entire study takes place (Niebel 1992). This method considers four factors in evaluating performance: skill, effort, environmental condition, and consistency. The overall performance factor rating is obtained by adding the sum of the equivalent numerical values of the four factors to unity. It must be noted that the performance factor is applicable only for manually performed elements. All machine-controlled elements are rated 100. Table 5 summarizes the ratings for each factor.

In 1949, Westinghouse Electric Corporation developed a new rating method called the performance rating plan (Niebel and Freivalds 1999). In addition to using the operator-related physical attributes, the company attempted to evaluate the relationship between those physical attributes and the basic divisions of work.

The characteristics and attributes that the Westinghouse performance-rating plan considers are dexterity, effectiveness, and physical application. These three major classifications are assigned nine attributes that carry numerical weights: three are related to dexterity, four to effectiveness, and two to physical application. Table 6 provides values for each of these attributes at various levels of performance.

Westinghouse rating procedures are appropriate for either cycle rating or overall study rating. Both of the Westinghouse rating techniques demand considerable training to differentiate the levels of each attribute.

2.5.1.2. Synthetic Rating Morrow (1946) developed a procedure known as the synthetic method to eliminate judgment of the time study observer, therefore giving more consistent results. The procedure determines a performance factor for representative effort elements of the work cycle by comparing actual elemental observed times and fundamental motion data (see Section 3). This factor would then be applied to the remainder of the manually controlled elements comprising the study. The ratio is given by:

$$P = \frac{\text{fundamental motion time}}{\text{observed mean element time}}$$

For example, the observed average time of work element 2 is 0.08 min and the corresponding fundamental motion time is 0.096. In addition, the observed average time of work element 5 is 0.22 with corresponding fundamental motion time of 0.278. The performance factor of element 2 and element 5 are computed as

$$P_2 = \frac{0.096}{0.08} = 1.2 \qquad P_5 = \frac{0.278}{0.22} = 1.26$$

The mean of the performance factors of elements 2 and 5 is

$$\overline{P} = \frac{1.2 + 1.26}{2} = 1.23 \text{ or } 123\%$$

This rating factor is then used for all other effort elements of the method.

2.5.1.3. Speed Rating Today, speed rating is probably the most widely used rating system (Niebel 1992). Speed rating is a performance evaluation that considers only the rate of accomplishment of the work per unit time. In this method, observers measure the effectiveness of operator against concept of a normal operator and then assign a percentage to indicate the ratio of the observed

TABLE 5 The Westinghouse Rating System

Skill ratings		
+0.15	A1	Superskill
+0.13	A2	Superskill
+0.11	B1	Excellent
+0.08	B2	Excellent
+0.06	C1	Good
+0.03	C2	Good
0.00	D	Average
−0.05	E1	Fair
−0.10	E2	Fair
−0.16	F1	Poor
−0.22	F2	Poor
Effort ratings		
+0.13	A1	Excessive
+0.12	A2	Excessive
+0.10	B1	Excellent
+0.08	B2	Excellent
+0.05	C1	Good
+0.02	C2	Good
0.00	D	Average
−0.04	E1	Fair
−0.08	E2	Fair
−0.12	F1	Poor
−0.17	F2	Poor
Environmental condition ratings		
+0.06	A	Ideal
+0.04	B	Excellent
+0.02	C	Good
0.00	D	Average
−0.03	E	Fair
−0.07	F	Poor
Consistency ratings		
+0.04	A	Perfect
+0.03	B	Excellent
+0.01	C	Good
0.00	D	Average
−0.02	E	Fair
−0.04	F	Poor

From B. Niebel and A. Freivalds, *Methods, Standards, and Work Design,* 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.

performance to standard performance. A rating of 100% is considered normal. A rating of 125% means a slow operator and a rating of 90% means a fast operator.

To be consistently accurate in using speed rating, the analyst must be familiar with the work being studied. The analyst first appraises the operator as whether performance is above or below the concept of normal rate. Then the precise position in the rating scale is determined. Benchmarks of various activities can be used for comparing with the performance being observed. It is recommended that analysts undergo a comprehensive training program prior to conducting independent studies (Niebel 1992).

2.5.1.4 Objective Rating The objective rating method developed by Mundel and Danner (1994) eliminates the difficulty of establishing normal speed criteria for every type of work by establishing a single work assignment to which the pace of all other jobs are compared. After the pace is judged, a second factor indicating its relative difficulty is assigned. The performance rating is expressed as the product of the pace rating factor and the job difficulty adjustment factor.

Factors influencing the difficulty adjustment that have been assigned numerical values for a range of degrees are: (1) amount of body used, (2) foot pedals, (3) bimanualness, (4) eye–hand coordination,

TABLE 6 The Westinghouse Performance-Rating Plan

	Performance Level Values				
Attribute	+ Above		0 Expected	− Below	
Dexterity					
Displayed ability in use of equipment and tools and in assembly of parts	+0.06	+0.03	0	−0.02	−0.04
Certainty of movement	+0.06	+0.03	0	−0.02	−0.04
Coordination and rhythm		+0.02	0	+0.02	
Effectiveness					
Displayed ability to continually replace and retrieve tools	+0.06	+0.03	0	−0.02	−0.04
Displayed ability to facilitate, eliminate, combine, or shorten motions	+0.06	+0.03	0	−0.04	−0.08
Displayed ability to use both hands with equal ease	+0.06	+0.03	0	−0.04	−0.08
Displayed ability to confine efforts to necessary work			0	−0.04	−0.08
Physical Application					
Work pace	+0.06	+0.03	0	−0.04	−0.08
Attentiveness			0	−0.02	−0.04

From Niebel 1992.

(5) handling or sensory requirements, and (6) weight handled or resistance encountered. The sum of the numerical values for each of the factors constitute the difficulty factor. Hence performance rating would be

$$P = \text{(pace rating)(difficulty adjustment)}$$

Tables of percentage values for the effects of various difficulties in operation performed can be found in Mundel and Danner (1994).

2.5.2. Selecting a Rating System

From a practical-standpoint, the performance-rating technique that is easiest to understand, apply, and explain is speed rating when augmented by standard data benchmarks (see Section 4). As is true of all procedures requiring the exercise of judgment, the simpler and more concise the technique, the easier it is to use and, in general, the more valid the results. Five criteria are used to ensure the success of the speed rating procedure (Niebel 1992):

1. *Experience by the time study analyst in the class of work being performed:* The analyst should be sufficiently familiar with the details of the work being observed as well as have experience as an observer.
2. *Use of standard data benchmarks on at least two of the elements performed:* The accuracy of the analyst's ratings can be validated using standard data elements. Standard data are a helpful guide to establishing performance factors.
3. *Regular training in speed rating:* It is important that time study analysts receive regular training in performance rating. Training may involve observation of films or videotape illustrating a variety of operations characteristic of the company's operations.
4. *Selection of an operator who has adequate experience:* As much as possible, the analyst should select an operator who has had sufficient experience on the job. It is desirable to select a cooperative operator who in the past has performed regularly at normal or above normal pace.
5. *Use of the mean value of three or more independent studies:* It is good practice to conduct more than one study before arriving at the standard time. These can be made on the same operator at different times, or on different operators. The total error due to both the performance rating and the measurement of the elemental times is reduced when the averages of several independent studies are used in computing standards.

2.6. Allowances

The fundamental purpose of all allowances is to add enough time to normal production time to enable the average worker to meet the standard when performing at standard performance. Even when the most practical, economical, and effective method has been developed, the job will still require the expenditure of human effort, and some allowance must therefore be made for recovery from fatigue and for unavoidable delays. Allowance must also be made to enable a worker to attend to his personal needs. Other special allowances may also have to be added to the basic time in order to arrive at a fair standard.

Fatigue and personal needs allowances are in addition to the basic time intended to provide the worker with the opportunity to recover from the physiological and psychological effects of carrying out specified work under specified conditions and to allow attention to personal needs. The amount of allowance will depend on the nature of the job, the work environment, and individual characteristics of the operator (e.g., age, physical condition, and working habits). Usually a 5% allowance for personal delays is appropriate in the majority of work environments nowadays. It is also considered good practice to provide an allowance for fatigue on the effort elements of the time study.

Special allowances include many different factors related to the process, equipment, materials, and so on and are further classified into unavoidable delays and policy allowances. Unavoidable delays are a small allowance of time that may be included in a standard time to meet legitimate and expected items of work or delays, such as due to power outages, defective materials, waiting lines, late deliveries, and other events beyond the control of the operator. Precise measurement of such occurrences is uneconomical because of their infrequent or irregular occurrence.

Policy allowance (ILO 1979) is an increment, other than bonus increment, applied to standard time (or to some constituent part of it) to provide a satisfactory level of earnings for a specified level of performance under exceptional circumstances (e.g. new employees, differently abled, workers on light duty, elderly). Special allowances may also be given for any activities that are not normally part of the operation cycle but that are essential to the satisfactory performance of the work. Such allowances may be permanent or temporary and are typically decided by management with possible union negotiations (Niebel and Freivalds 1999). Policy allowances should be used with utmost caution and only in clearly defined circumstances. The usual reason for making a policy allowance is to line up standard times with requirements of wage agreements between employers and trade unions.

Whenever possible, these allowances should be determined by a time study. Two methods are frequently used for developing standard allowance data. A production study, which requires analysts to study two to three operations over a long period, records the duration and reason for each idle interval. After a reasonably representative sample is established, the percent allowance for each applicable characteristic is determined. The other method involves work sampling studies. This method requires taking a large number of random observations. (See Chapter 53 for more information on allowances.)

The allowance is typically given as a percentage and is used as a multiplier, so that normal time (NT) can be readily adjusted to the standard time (ST):

$$\text{ST} = \text{NT} + \text{NT(allowance)} = \text{NT}(1 + \text{allowance})$$

The total allowance is determined by adding individual allowance percentages applicable to a job. For example, total allowance is computed as the sum of personal needs (5%), fatigue (4%), and unavoidable delays (1%), equal to 10%. Normal time would then be multiplied by 1.1 to determine standard time.

2.7. Calculating the Standard Time

The standard time for the operation under study is obtained by the sum of the elemental standard times. Standard times may be expressed in minutes per piece or hours per hundred pieces for oper-

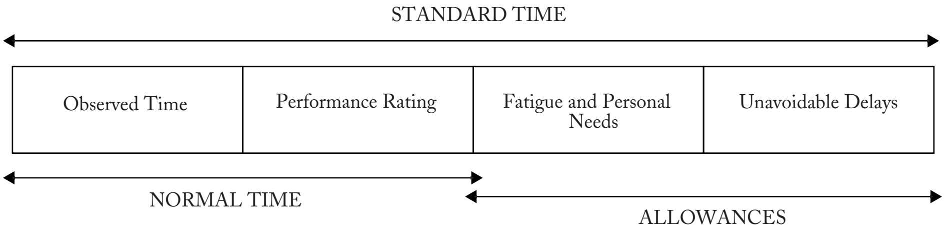


Figure 3 Breakdown of Standard Time.

ations with short cycles. In this form, it is convenient to compute the operator's efficiency and daily earnings if an incentive plan is applied. The percent efficiency of the operator can be expressed as:

$$E = 100\left(\frac{H_e}{H_c}\right)$$

where E = percent efficiency
H_e = earned standard hours
H_c = clock hours on the job

For example, the work standard for an operation is 11.5 min/piece. In an 8-hour shift, an operator produced 50 pieces in a given working day. The standard hours earned would be:

$$H_e = \frac{50 \text{ pieces} \times 11.50 \text{ min/piece}}{60 \text{ min/hr}} = 9.58 \text{ hr}$$

The operator's efficiency would be computed as:

$$\text{Efficiency} = \frac{100 \times 9.58 \text{ hr}}{8 \text{ hr}} = 119.7\%$$

2.7.1. Temporary Standards

In many cases, time study analysts establish a standard on a relatively new operation wherein there is insufficient volume for the operator to reach his long-term efficiency. Employees require time to become proficient in any new or different operation. If the analyst rates the operator based on the usual concept of output (i.e., rating the operator below 100), the resulting standard may be unduly tight and make it difficult to achieve any incentive earnings. On the other hand, if a liberal standard is established, this may cause an increase in labor expense and other related production costs.

The most satisfactory solution to such situations would be the establishment of temporary standards. The analyst establishes the standard based on the difficulty of the job and the number of pieces to be produced. Then, by using a learning curve (see Chapter 53 on learning curves) for the work, as well as existing standard data, an equitable temporary standard for the work can be established. When released to the production floor, the standard should be clearly identified as temporary and applicable to only to a fixed quantity or fixed duration. Upon expiration, temporary standards should be immediately replaced by permanent standards.

2.7.2. Setup Standards

Setup standards are for those elements that involve all events that take place between completion of the previous job and the start of the present job. Setup standards also include teardown or put-away elements, such as clocking in on the job, getting tools from the tool crib, getting drawings from the dispatcher, setting up the machine, removing tools from the machine, returning tools to the tool crib, and clocking out from the job. The analysts need to use the same care and precision in studying the setup time because only one cycle, if not a few, can be observed and recorded in one day. Because setup elements are of long duration, there is a reasonable amount of time to break the job down, record the time, and evaluate the performance as the operator proceeds from one work element to the next.

3. PREDETERMINED TIME STANDARDS

Predetermined time standards (PTS) are techniques that aim at defining the time needed for the performance of various operations by derivation from preset standards of time for various motions and not by direct observation and measurement. PTS is a work measurement technique whereby times established for basic human motions (classified according to the nature of the motion and the conditions under which it is made) are used to build up the time for a job at defined levels of performance. They are derived as a result of studying a large sample of diversified operations.

There are a number of different PTS systems, and it will be useful to present the main ways in which the systems vary. The differences are with respect to the levels and scope of application of data, motion classification, and time units. Essentially, these predetermined time systems are sets of motion–time tables with explanatory rules and instructions on the use of the motion–time values. Most companies require considerable training in the practical application of these techniques to earn certification before analysts are allowed to apply the Work-Factor, MTM, or MOST systems. Figure 4 illustrates the derivation of all of these predetermined time systems.

No two PTS systems have the same set of time values. This is partly due to the fact that different systems have different motion classes and the time data therefore refer to different things. Variations

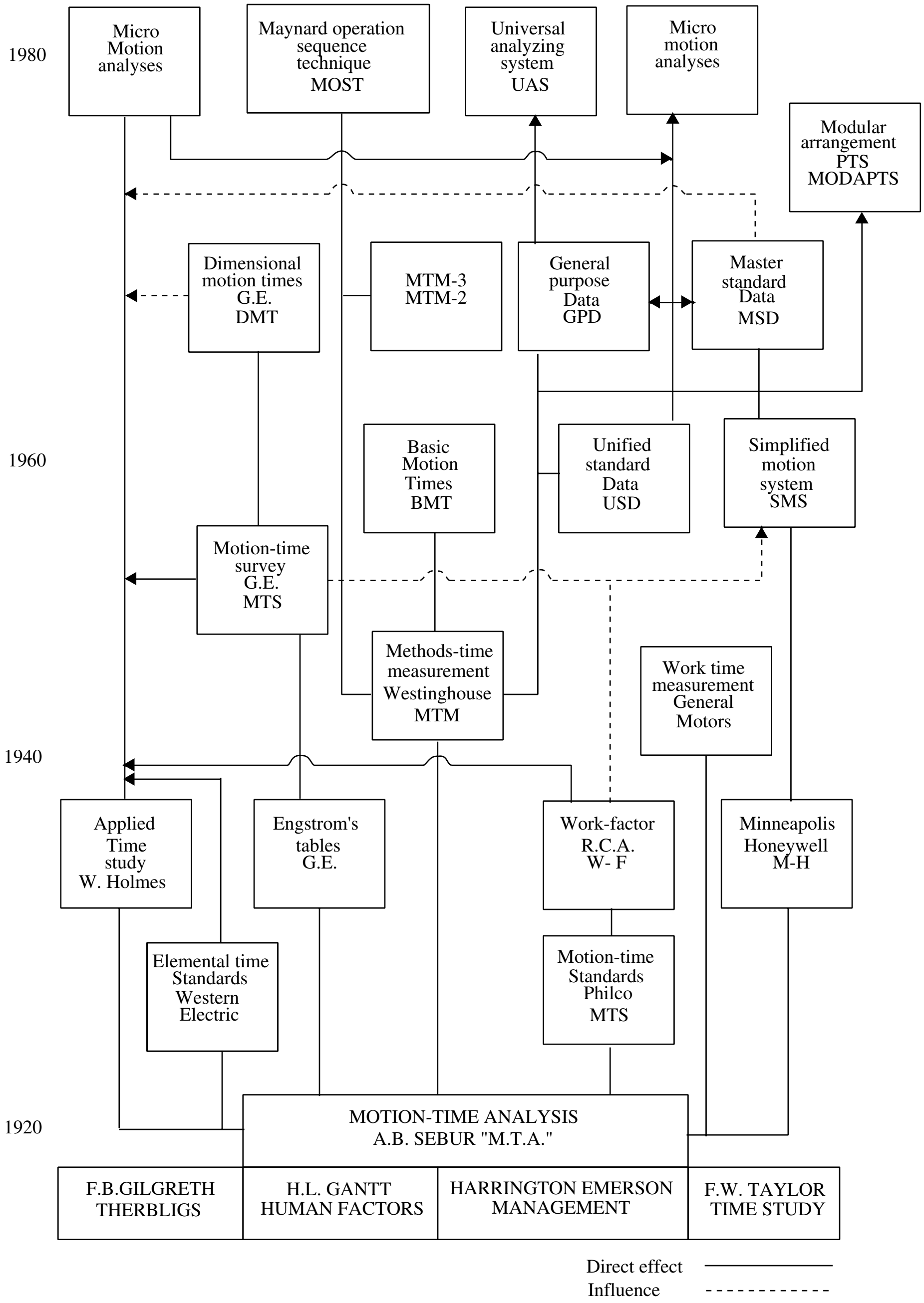

1980
1960
1940
1920
Micro Motion analyses
Maynard operation sequence technique MOST
Universal analyzing system UAS
Micro motion analyses
Modular arrangement PTS MODAPTS
Dimensional motion times G.E. DMT
MTM-3 MTM-2
General purpose Data GPD
Master standard Data MSD
Basic Motion Times BMT
Unified standard Data USD
Simplified motion system SMS
Motion-time survey G.E. MTS
Methods-time measurement Westinghouse MTM
Work time measurement General Motors
Applied Time study W. Holmes
Engstrom's tables G.E.
Work-factor R.C.A. W- F
Minneapolis Honeywell M-H
Elemental time Standards Western Electric
Motion-time Standards Philco MTS
MOTION-TIME ANALYSIS A.B. SEBUR "M.T.A."
F.B.GILGRETH THERBLIGS
H.L. GANTT HUMAN FACTORS
HARRINGTON EMERSON MANAGEMENT
F.W. TAYLOR TIME STUDY
Direct effect
Influence

Figure 4 Family Tree of Predetermined Times. (From Sellie 1992)

in time units are due to differences in the methods adopted for standardizing the motion times, the choice of the basic unit (seconds, minutes, hours, or fractions of a minute), and the practice of adding contingency allowances or not.

The scope of application of a PTS system can be universal or generic, functional, or specific. A universal system is one that is designed for body members in general. Its application is not restricted to any specific type of work. The motion descriptions only identify the body member being used. Examples of generic terms are REACH, TRANSPORT, GRASP, WALK. Examples of universal systems are MTM-1 to MTM-3, Work-Factor, MOST, and MODAPTS.

A functional system defines motion element times for a particular type of activity, such as clerical work (MTM-C), uses of microscopes (MTM-M), and so on. The element names indicate the function for which the system was developed. For example, FILE is a common element name in office work environments.

A specific system is one where the motion–time tables were constructed for specific operations or work areas. Examples are standard motion–time tables developed for electronic tests (MTM-TE), to measure one-of-a-kind and small-lot production (MTM-MEK).

The approach in applying PTS systems uses the following general procedure:

1. Summarize all left- and right-hand motions required to perform the job properly (e.g., SIMO chart).
2. Determine governing or limiting elements for elements done simultaneously.
3. Eliminate or delete nonlimiting elements.
4. Summarize only the limiting or governing elements.
5. Determine from the PTS table the basic time to perform the elements.
6. Add up the basic elemental times of limiting elements obtained from time tables.

3.1. Methods–Time Measurement (MTM)

Methods–Time Measurement systems were first developed by Harold B. Maynard in 1946 at the Westinghouse Electric Corporation in collaboration with Gustave J. Stegemerten and John L. Schwab. Maynard was commissioned by Westinghouse to develop a system for describing and evaluating operations methods.

MTM became the first widely used predetermined time system. In the United States, it is administered, advanced, and controlled by the MTM Association for Standards and Research. This nonprofit association is one of 12 associations comprising the International MTM Directorate. The success of the MTM systems is attributed to the active committee structure made up of members of the association.

The original MTM system is now known as MTM-1. Subsequent modifications were later developed to provide easier and quicker systems by condensing and reducing the number of motion options and time values. MTM-2 and MTM-3 are examples of second-level and third-level MTM data. In addition, the MTM family of systems include MTM-V, MTM-C, MTM-M, MTM-TE, MTM-MEK, and MTM-UAS.

3.1.1. MTM-1

MTM-1 data are the result of frame-by-frame analyses of motion picture films of different types of work. The film was analyzed for the motion content of the operation in question, rated by the Westinghouse technique, evaluated to determine the degree of difficulty caused by the characteristics of the motion and frame counts measured to yield normal motion times. The first set of predetermined time standards defined time values for the seven fundamental motions of: REACH, MOVE, TURN, GRASP, POSITION, DISENGAGE, and RELEASE.

Further analyses categorized some fundamental motions into different distinct cases. Five cases of REACH, 3 cases of MOVE, 2 cases of RELEASE, and 18 cases of POSITION were established based on factors that affect motion times, such as distance, weight of the object, and type of motion.

Table 7 summarizes the MTM-1 values. Time values are expressed in terms of a new time unit known as the time-measurement unit (TMU), and assigned a unit value of 0.00001 hr, equal to 1 TMU. The tabulated values do not include any allowances. Proponents of the MTM-1 system state that fatigue allowance in not needed in most applications because the time values are based on a work rate that can be sustained at steady state for the working life of a healthy employee.

In using this system, first all left-hand and right-hand motions required to perform a job properly are summarized and tabulated. Then the rated times in TMU for each motion are determined from the MTM-1 tables. The time required for a normal performance of the task is obtained by adding only the limiting motions (the longer time between two simultaneous motions predominates). The nonlimiting motion values are then either circled or deleted. Figure 5 illustrates the use of MTM-1 in analyzing a simple operation.

3.1.2. MTM-2

MTM-2 is a system of synthesized MTM data and is the second general level of MTM data. It consists of single basic MTM motions and certain combinations of basic MTM motions. MTM-2

TABLE 7 Summary of MTM-1 Data

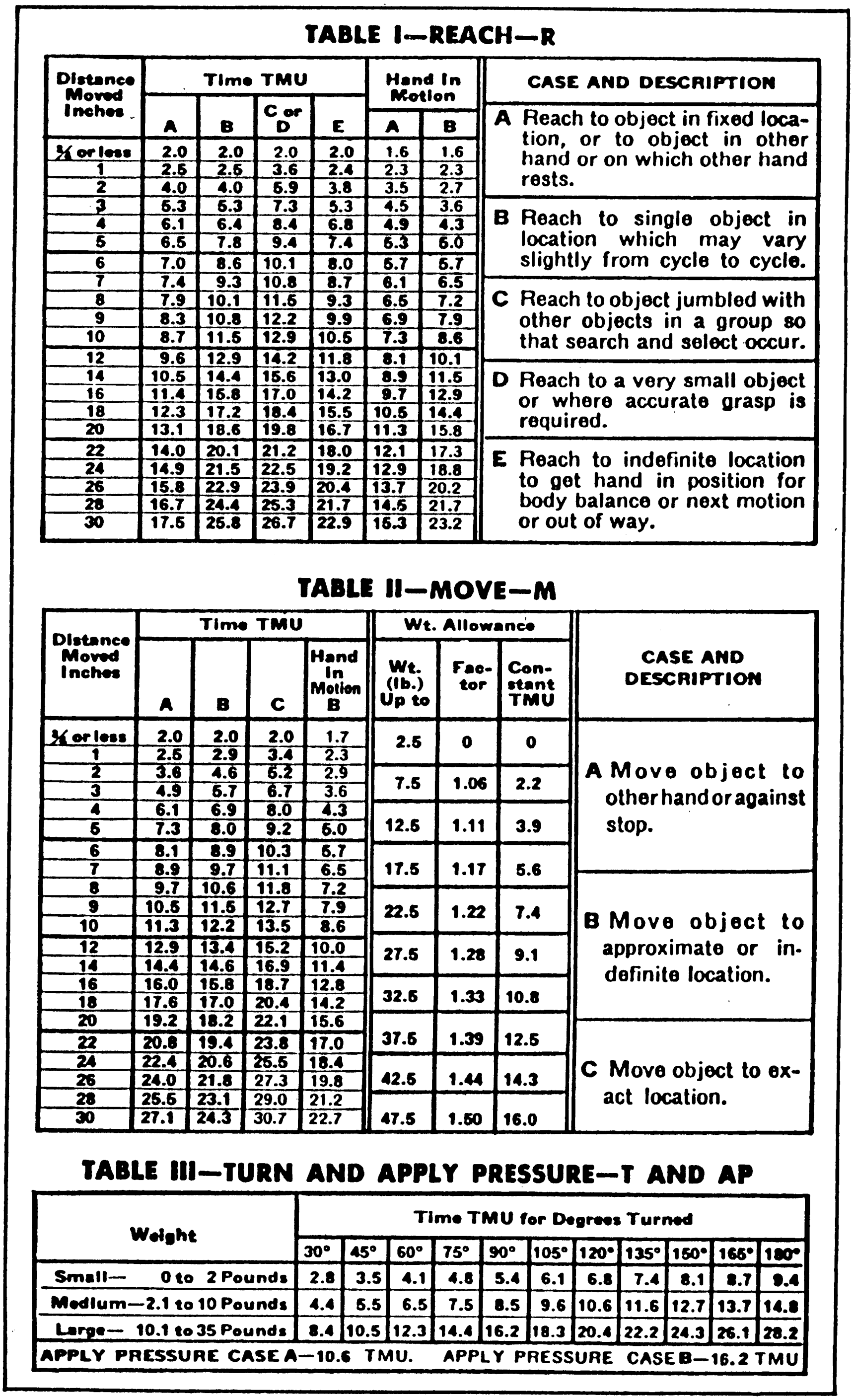


TABLE I—REACH—R

Distance Moved Inches	Time TMU A	Time TMU B	Time TMU C or D	Time TMU E	Hand In Motion A	Hand In Motion B
¾ or less	2.0	2.0	2.0	2.0	1.6	1.6
1	2.5	2.5	3.6	2.4	2.3	2.3
2	4.0	4.0	5.9	3.8	3.5	2.7
3	5.3	5.3	7.3	5.3	4.5	3.6
4	6.1	6.4	8.4	6.8	4.9	4.3
5	6.5	7.8	9.4	7.4	5.3	5.0
6	7.0	8.6	10.1	8.0	5.7	5.7
7	7.4	9.3	10.8	8.7	6.1	6.5
8	7.9	10.1	11.5	9.3	6.5	7.2
9	8.3	10.8	12.2	9.9	6.9	7.9
10	8.7	11.5	12.9	10.5	7.3	8.6
12	9.6	12.9	14.2	11.8	8.1	10.1
14	10.5	14.4	15.6	13.0	8.9	11.5
16	11.4	15.8	17.0	14.2	9.7	12.9
18	12.3	17.2	18.4	15.5	10.5	14.4
20	13.1	18.6	19.8	16.7	11.3	15.8
22	14.0	20.1	21.2	18.0	12.1	17.3
24	14.9	21.5	22.5	19.2	12.9	18.8
26	15.8	22.9	23.9	20.4	13.7	20.2
28	16.7	24.4	25.3	21.7	14.5	21.7
30	17.5	25.8	26.7	22.9	15.3	23.2

CASE AND DESCRIPTION	
A	Reach to object in fixed location, or to object in other hand or on which other hand rests.
B	Reach to single object in location which may vary slightly from cycle to cycle.
C	Reach to object jumbled with other objects in a group so that search and select occur.
D	Reach to a very small object or where accurate grasp is required.
E	Reach to indefinite location to get hand in position for body balance or next motion or out of way.

TABLE II—MOVE—M

Distance Moved Inches	Time TMU A	Time TMU B	Time TMU C	Hand In Motion B
¾ or less	2.0	2.0	2.0	1.7
1	2.5	2.9	3.4	2.3
2	3.6	4.6	5.2	2.9
3	4.9	5.7	6.7	3.6
4	6.1	6.9	8.0	4.3
5	7.3	8.0	9.2	5.0
6	8.1	8.9	10.3	5.7
7	8.9	9.7	11.1	6.5
8	9.7	10.6	11.8	7.2
9	10.5	11.5	12.7	7.9
10	11.3	12.2	13.5	8.6
12	12.9	13.4	15.2	10.0
14	14.4	14.6	16.9	11.4
16	16.0	15.8	18.7	12.8
18	17.6	17.0	20.4	14.2
20	19.2	18.2	22.1	15.6
22	20.8	19.4	23.8	17.0
24	22.4	20.6	25.5	18.4
26	24.0	21.8	27.3	19.8
28	25.5	23.1	29.0	21.2
30	27.1	24.3	30.7	22.7

Wt. Allowance Wt. (lb.) Up to	Factor	Constant TMU
2.5	0	0
7.5	1.06	2.2
12.5	1.11	3.9
17.5	1.17	5.6
22.5	1.22	7.4
27.5	1.28	9.1
32.5	1.33	10.8
37.5	1.39	12.5
42.5	1.44	14.3
47.5	1.50	16.0

CASE AND DESCRIPTION	
A	Move object to other hand or against stop.
B	Move object to approximate or indefinite location.
C	Move object to exact location.

TABLE III—TURN AND APPLY PRESSURE—T AND AP

Weight	Time TMU for Degrees Turned 30°	45°	60°	75°	90°	105°	120°	135°	150°	165°	180°
Small— 0 to 2 Pounds	2.8	3.5	4.1	4.8	5.4	6.1	6.8	7.4	8.1	8.7	9.4
Medium—2.1 to 10 Pounds	4.4	5.5	6.5	7.5	8.5	9.6	10.6	11.6	12.7	13.7	14.8
Large— 10.1 to 35 Pounds	8.4	10.5	12.3	14.4	16.2	18.3	20.4	22.2	24.3	26.1	28.2

APPLY PRESSURE CASE A—10.6 TMU. APPLY PRESSURE CASE B—16.2 TMU

TABLE 7 (*Continued*)

TABLE IV—GRASP—G

Case	Time TMU	DESCRIPTION
1A	2.0	Pick Up Grasp—Small, medium or large object by itself, easily grasped.
1B	3.5	Very small object or object lying close against a flat surface.
1C1	7.3	Interference with grasp on bottom and one side of nearly cylindrical object. Diameter larger than ½".
1C2	8.7	Interference with grasp on bottom and one side of nearly cylindrical object. Diameter ¼" to ½".
1C3	10.8	Interference with grasp on bottom and one side of nearly cylindrical object. Diameter less than ¼".
2	5.6	Regrasp.
3	5.6	Transfer Grasp.
4A	7.3	Object jumbled with other objects so search and select occur. Larger than 1" x 1" x 1".
4B	9.1	Object jumbled with other objects so search and select occur. ¼" x ¼" x ⅛" to 1" x 1" x 1".
4C	12.9	Object jumbled with other objects so search and select occur. Smaller than ¼" x ¼" x ⅛".
5	0	Contact, sliding or hook grasp.

TABLE V—POSITION*—P

CLASS OF FIT		Symmetry	Easy To Handle	Difficult To Handle
1—Loose	No pressure required	S	5.6	11.2
		SS	9.1	14.7
		NS	10.4	16.0
2—Close	Light pressure required	S	16.2	21.8
		SS	19.7	25.3
		NS	21.0	26.6
3—Exact	Heavy pressure required.	S	43.0	48.6
		SS	46.5	52.1
		NS	47.8	53.4

*Distance moved to engage—1" or less.

TABLE VI—RELEASE—RL

Case	Time TMU	DESCRIPTION
1	2.0	Normal release performed by opening fingers as independent motion.
2	0	Contact Release.

TABLE VII—DISENGAGE—D

CLASS OF FIT	Easy to Handle	Difficult to Handle
1—Loose—Very slight effort, blends with subsequent move.	4.0	5.7
2—Close — Normal effort, slight recoil.	7.5	11.8
3—Tight — Considerable effort, hand recoils markedly.	22.9	34.7

TABLE VIII—EYE TRAVEL TIME AND EYE FOCUS—ET AND EF

Eye Travel Time $= 15.2 \times \frac{T}{D}$ TMU, with a maximum value of 20 TMU.

where T = the distance between points from and to which the eye travels.
D = the perpendicular distance from the eye to the line of travel T.

Eye Focus Time = 7.3 TMU.

TABLE 7 (*Continued*)

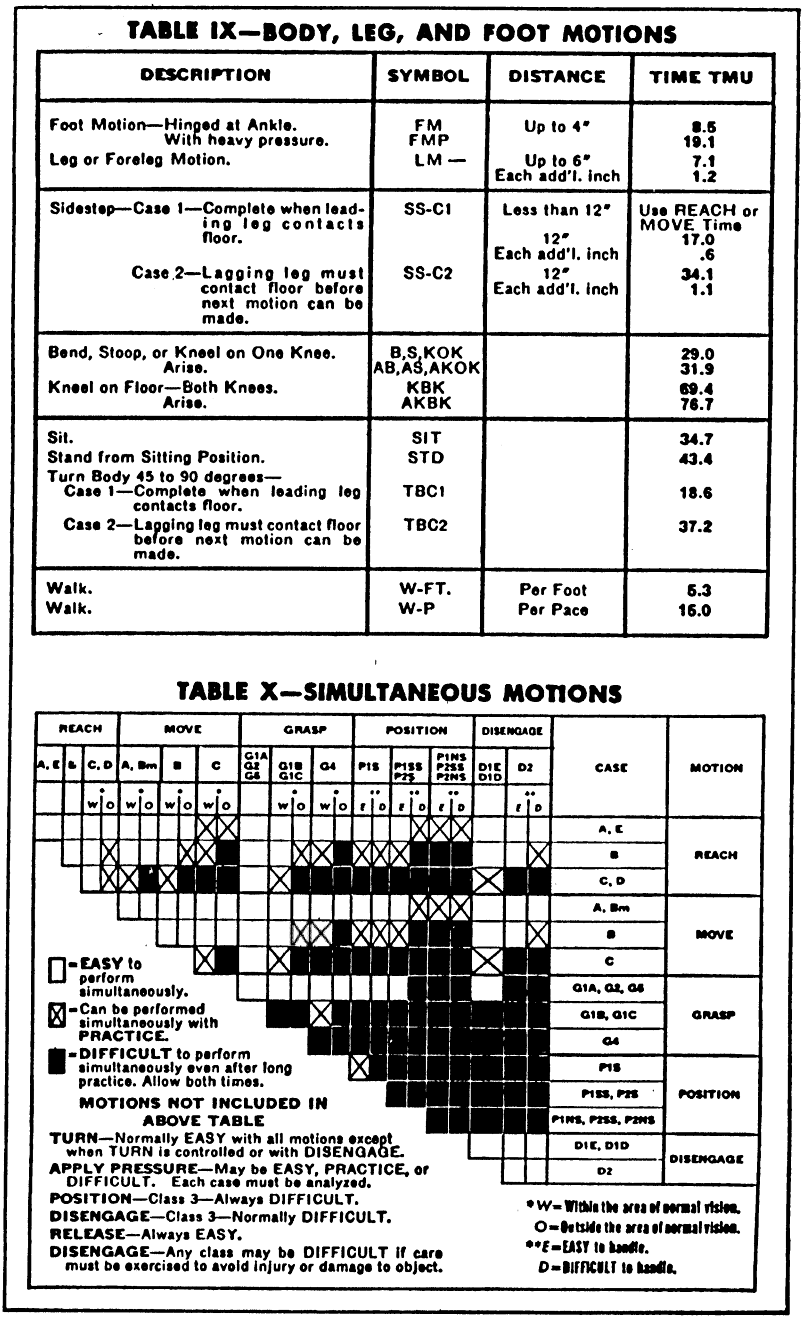


TABLE IX—BODY, LEG, AND FOOT MOTIONS

DESCRIPTION	SYMBOL	DISTANCE	TIME TMU
Foot Motion—Hinged at Ankle.	FM	Up to 4"	8.5
With heavy pressure.	FMP		19.1
Leg or Foreleg Motion.	LM —	Up to 6"	7.1
		Each add'l. inch	1.2
Sidestep—Case 1—Complete when leading leg contacts floor.	SS-C1	Less than 12"	Use REACH or MOVE Time
		12"	17.0
		Each add'l. inch	.6
Case 2—Lagging leg must contact floor before next motion can be made.	SS-C2	12"	34.1
		Each add'l. inch	1.1
Bend, Stoop, or Kneel on One Knee.	B,S,KOK		29.0
Arise.	AB,AS,AKOK		31.9
Kneel on Floor—Both Knees.	KBK		69.4
Arise.	AKBK		76.7
Sit.	SIT		34.7
Stand from Sitting Position.	STD		43.4
Turn Body 45 to 90 degrees—			
Case 1—Complete when leading leg contacts floor.	TBC1		18.6
Case 2—Lagging leg must contact floor before next motion can be made.	TBC2		37.2
Walk.	W-FT.	Per Foot	5.3
Walk.	W-P	Per Pace	15.0

TABLE X—SIMULTANEOUS MOTIONS

REACH				MOVE						GRASP					POSITION						DISENGAGE			CASE	MOTION
A, E	B	C, D		A, Bm		B		C		G1A G2 G5	G1B G1C		G4		P1S		P1SS P2S		P1NS P2SS P2NS		D1E D1D	D2			
		W*	O	W*	O	W*	O	W*	O		W*	O	W*	O	E**	D	E**	D	E**	D		E**	D		
☐	☐	☐	☐	☐	☐	☐	☐	⊠	⊠	☐	☐	☐	☐	☐	☐	☐	☐	⊠	⊠	⊠	☐	☐	☐	A, E	REACH
	☐	☐	⊠	☐	☐	☐	⊠	⊠	■	☐	☐	⊠	⊠	■	⊠	⊠	⊠	■	■	■	☐	☐	⊠	B	
		☐	⊠	⊠	■	⊠	■	■	■	☐	⊠	■	■	■	■	■	■	■	■	■	⊠	■	■	C, D	
				☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	⊠	⊠	⊠	☐	☐	☐	A, Bm	MOVE
						☐	☐	☐	☐	☐	☐	⊠	⊠	■	⊠	⊠	⊠	■	■	■	☐	☐	⊠	B	
								⊠	■	☐	⊠	■	■	■	■	■	■	■	■	■	⊠	■	■	C	
										☐	☐	☐	☐	☐	☐	☐	☐	■	■	■	☐	■	■	G1A, G2, G5	GRASP
											■	■	⊠	■	■	■	■	■	■	■	■	■	■	G1B, G1C	
													■	■	■	■	■	■	■	■	■	■	■	G4	
															⊠	■	■	■	■	■	■	■	■	P1S	POSITION
																	■	■	■	■	■	■	■	P1SS, P2S	
																			■	■	■	■	■	P1NS, P2SS, P2NS	
																					☐	☐	☐	D1E, D1D	DISENGAGE
																						☐	☐	D2	

☐ -**EASY** to perform simultaneously.

⊠ -Can be performed simultaneously with **PRACTICE.**

■ -**DIFFICULT** to perform simultaneously even after long practice. Allow both times.

MOTIONS NOT INCLUDED IN ABOVE TABLE

TURN—Normally EASY with all motions except when TURN is controlled or with DISENGAGE.

APPLY PRESSURE—May be EASY, PRACTICE, or DIFFICULT. Each case must be analyzed.

POSITION—Class 3—Always DIFFICULT.

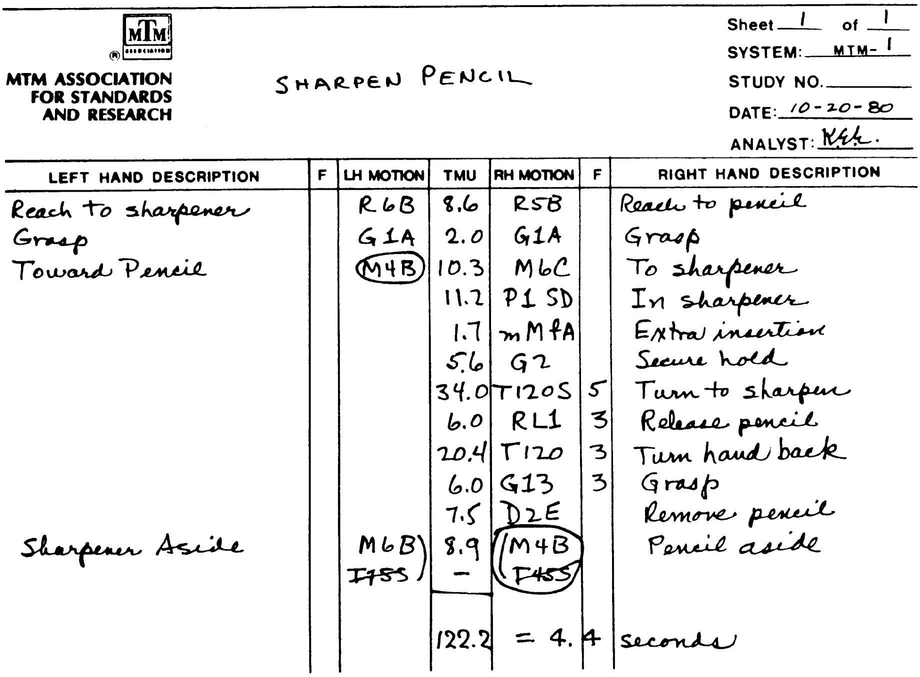
DISENGAGE—Class 3—Normally DIFFICULT.

RELEASE—Always EASY.

DISENGAGE—Any class may be DIFFICULT if care must be exercised to avoid injury or damage to object.

*W = Within the area of normal vision.
O = Outside the area of normal vision.
**E = EASY to handle.
D = DIFFICULT to handle.

Courtesy: MTM Association.



MTM ASSOCIATION
FOR STANDARDS
AND RESEARCH

SHARPEN PENCIL

Sheet 1 of 1
SYSTEM: MTM-1
STUDY NO.
DATE: 10-20-80
ANALYST: KEK.

LEFT HAND DESCRIPTION	F	LH MOTION	TMU	RH MOTION	F	RIGHT HAND DESCRIPTION
Reach to sharpener		R6B	8.6	R5B		Reach to pencil
Grasp		G1A	2.0	G1A		Grasp
Toward Pencil		M4B	10.3	M6C		To sharpener
			11.2	P1SD		In sharpener
			1.7	mMfA		Extra insertion
			5.6	G2		Secure hold
			34.0	T120S	5	Turn to sharpen
			6.0	RL1	3	Release pencil
			20.4	T120	3	Turn hand back
			6.0	G1B	3	Grasp
			7.5	D2E		Remove pencil
Sharpener Aside		M6B	8.9	M4B		Pencil aside
		~~T45S~~	–	~~T45S~~		
			122.2	= 4.	4	seconds

Figure 5 An MTM-1 Analysis of Sharpening Pencil with Hand-Held Sharpener. (From Sellie 1992)

recognizes 11 classes of actions. The motion times range from 3 to 61 TMU. The 11 categories and their symbols are:

Category	Code(s)
GET	GA, GB, GC
PUT	PA, PB, PC
GET WEIGHT	GW
PUT WEIGHT	PW
REGRASP	R
APPLY PRESSURE	A
EYE ACTION	E
FOOT ACTION	F
STEP	S
BEND & ARISE	B
CRANK	C

GET and PUT categories are affected by the distance traveled by the knuckle at the base of the index finger for hand motions and the path traveled by the fingertips, if only the fingers move. Five distance classes are defined (see Table 8).

GET is a composite of REACH, GRASP, and RELEASE. The time required to perform a GET motion is affected by the case involved (distinguished by the grasping action involved), the distance reached, and the weight of the object handled. The three cases A, B, and C of GET are judged using the decision model in Figure 6.

PUT is considered a combination of MOVE and POSITION. It involves moving an object to a destination with the hand or fingers. It starts with the object grasped and under control at the initial place and includes all transporting and correcting motions necessary to place the object. The motion

TABLE 8 Summary of MTM-2 Data

				MTM-2 (TMU)			
Range	Code	GA	GB	GC	PA	PB	PC
Up to 2 in. (5 cm)	−2	3	7	14	3	10	21
Over 2–6 in. (15 cm)	−6	6	10	19	6	15	26
Over 6–12 in. (30 cm)	−12	9	14	23	11	19	30
Over 12–18 in. (45 cm)	−18	13	18	27	15	24	36
Over 18 in. (45 cm)	−32	17	23	32	20	30	41
MTM Association ®	GW 1–per 2 lb (1 kg)				PW 1–per 10 lb (4.5 kg)		
	A	R	E	C	S	F	B
	14	6	7	15	18	9	61

Courtesy: MTM Association.

ends with the object still under control at the intended place. Similarly, the time required to perform a PUT motion is affected by distance and weight variables. Three cases of PUT are likewise distinguished based on the number of correcting motions required. A correcting motion is an unintentional stop, hesitation, or change in direction at the terminal point. Identification of the cases of PUT can be made using the decision model shown in Figure 7. When there is engagement of parts following a correction, an additional PUT motion is allowed when the distance exceeds 2.5 cm (1 in.).

A further consideration is that a PUT motion can be accomplished either as an insertion or an alignment. An insertion involves placing one object into another while an alignment involves orienting a part on a surface. Table 9 assists the analyst in better identifying the appropriate case.

GET WEIGHT (GW) is the action required for the hand and arm to take up the weight of the object. It occurs after the fingers have closed on the object in the preceding GET motion and is accomplished before any actual movement takes place. The time value for a GW is 1 TMU per kilogram (2.2 lb). For instance, a 4 kg (8.8 lb) load handled by both hands will be assigned a time value of 2 TMU since the effective weight per hand will be 2 kg (4.4 lb). When the weight is less than 1 kg (2.2 lb) per hand, no GW is assigned.

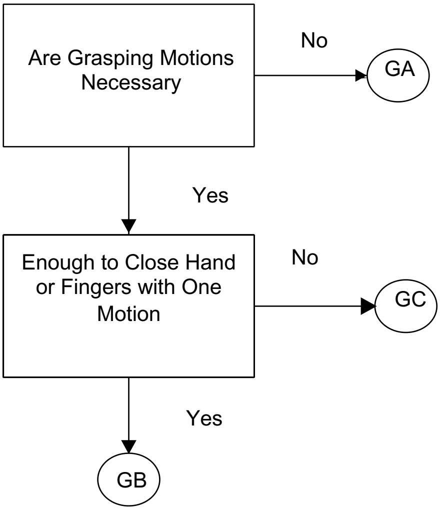

Are Grasping Motions Necessary
No
GA
Yes
Enough to Close Hand or Fingers with One Motion
No
GC
Yes
GB

Figure 6 Algorithm for Determining Case of GET. (MTM Association)

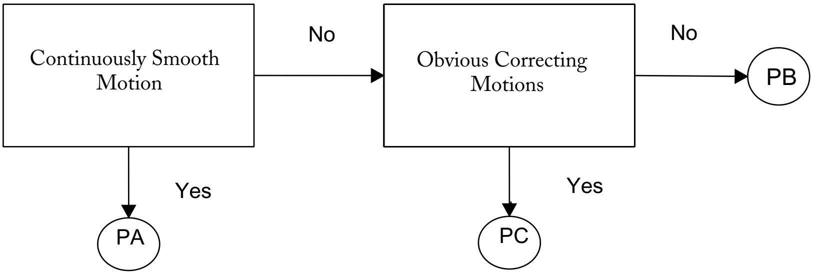


Figure 7 Algorithm for Determining Case of PUT. (MTM Association)

PUT WEIGHT (PW) is an addition to a PUT motion due to the weight of the object moved. Additions are 1 TMU per 5 kg (11 lb) of effective weight, up to a maximum of 20 kg (44 lb).

The category REGRASP (R) is a hand action with the purpose of changing the grasp on an object with the hand maintaining control. It is assigned a time of 6 TMU.

APPLY PRESSURE (A) applies to the action of exerting muscular force on an object to achieve control, restrain, or overcome resistance to motion. It can be performed by any body member and the object is not displaced more than 6.4 mm (0.25 in.) during the action. It is assigned a time of 14 TMU. An example of A is the final tightening action made with a screwdriver.

EYE ACTION (E) is an action that involves either recognizing a readily distinguishable characteristic of an object or shifting the aim of the axis of vision to a new viewing area. A single eye focus is defined as moving beyond a 10 cm (3.94 in.) diameter circle at a typical viewing distance of 40 cm (15.75 in.). Recognition time considered is sufficient only for simple binary decisions. The estimated value of E is 7 TMU. The value of E is allowed only when E is independent of hand or body motions.

FOOT movements are 9 TMU and STEP movements are 18 TMU. The time for STEP movement is based on an 85 cm (34 in.) pace. The decision model shown in Figure 8 helps differentiate a STEP movement from a FOOT movement.

BEND & ARISE (B) occurs when the body changes its vertical position. Examples include sitting down, standing up, and kneeling. B is assigned a time value of 61 TMU.

CRANK (C) is the motion of moving an object in a circular path of more than half a revolution with the hand or finger. For less than half a revolution, a PUT is used instead. Two variables are considered in applying the C motion: the number of revolutions and the weight of the object. A time of 15 TMU is allotted for each complete revolution. Where weight is significant, PW is applied to each revolution. The number of revolutions should be rounded to the nearest whole number.

In performing MTM-2 analysis, the principle of limiting motion or combined motions applied in MTM-1 is also used. That is, for two simultaneous motions, the longer time predominates.

3.1.3. MTM-3

The third-level MTM system is MTM-3. This level was developed to supplement MTM-1 and MTM-2. The system is helpful in work situations where saving time takes preference over accuracy. The accuracy of MTM-3 is within ±5%, with a 95% confidence level compared to MTM-1 analyses for cycles of approximately 4 min (Niebel and Freivalds 1999). MTM-3, however, cannot be applied to operations that require eye focus or eye travel time because the MTM-3 data do not include these motions.

TABLE 9 Comparison of Insertion and Alignment PUT Cases (mm/in.)

	PA	PB	PC
Insertion	Clearance > 10.2 mm (0.4 in.)	Clearance < 10.2 mm (0.4 in.)	Tight fit
Alignment	Tolerance >6.3 mm (0.25 in.)	1.6 mm (0.06 in.) < Tolerance < 6.3 mm (0.25 in.)	Tolerance < 1.6 mm (0.06 in.)

From B. Niebel and A. Freivalds, *Methods, Standards, and Work Design,* 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.

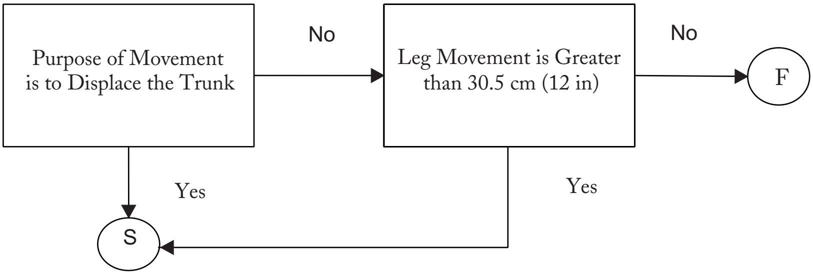

Purpose of Movement is to Displace the Trunk
No
Leg Movement is Greater than 30.5 cm (12 in)
No
F
Yes
Yes
S

Figure 8 Algorithm for Differentiating between STEP (S) and FOOT (F) Motion. (MTM Association)

The MTM-3 consists of only four categories of manual motions:

1. *Handle:* getting control of an object with the hand or fingers and placing the object in a new location
2. *Transport:* moving an object to a new location with the hand or fingers
3. *Step and foot motions:* step movement based on an 85 cm (34 in.) pace and foot movements
4. *Bend and arise:* a motion where the body changes its vertical position

Table 10 presents a summary of MTM-3 data. It consists of 10 time values, ranging from 7–61 TMU.

3.1.4. MTM-V

This functional PTS system was developed by Svenka MTM Grupen, the Swedish MTM Association, for use in metal-cutting operations. The MTM-V system is of the fourth level and contains time values for handling and adjusting workpieces of any weight and size, including machine tool setup, attaching crane hooks, and other mechanical handling equipment. The system does not include process-controlled activities. The system's 12 elements are of two types: those that can be accomplished by the hand and finger alone and those that require the use of a hand tool to accomplish the objective. This technique is said to be about 23 times faster to apply than MTM-1 (Niebel and Freivalds 1999).

3.1.5. MTM-C

MTM-C is a functional work-measurement system used to establish time standards for clerical-related work tasks at two levels of job description, precision, and speed of analysis. This system was developed by a consortium of banking and service industries.

Level 1 data categories cover nine areas. A six-digit numeric coding system (similar to MTM-V) is used to provide a detailed description of the operation being studied, each of which is documented with a specific MTM-1 motion pattern. Level 2 data are directly traceable to Level 1 and to MTM-1, which covers the same activities at a combined motion level. Distance ranges are reduced to one, and elements use simplified alphanumeric and mnemonic codes.

Calculating standard time using MTM-C takes less time to perform than with MTM-1 because the number of elements used to describe the clerical operation is reduced significantly. MTM-C Level

TABLE 10 Summary of MTM-3 Data

MTM-3 (TMU)					
		Handle		Transport	
Cm	In.	HA	HB	TA	TB
15.24	6	18	34	7	21
15.24	6	34	48	16	29
		SF 18		B 61	

Courtesy: MTM Association.

TABLE 11 MTM-C Level-1 Elements

MTM-C OPERATION ANALYSIS

VALIDATION
Sheet of

MTM ASSOCIATION
FOR STANDARDS
AND RESEARCH

MTM-C LEVEL 1
Replace page in 3-ring binder

DEPARTMENT: Clerical
ANALYST: CNR
DATE: 11/77

No.	Description	Reference	Element TMU	Occurrence per Cycle	TMU per Cycle
1	OPEN BINDER				
	Get binder from shelf	113 520	21	1	21
	Aside to desk	123 002	22	1	22
	Get cover	112 520	14	1	14
	Open cover	212 100	15	1	15
2	LOCATE CORRECT PAGE				
	Read on first page	510 000	7	2	14
	Locate approximate	451 120	16	3	48
	Identify page number	440 630	22	3	66
	Locate correct page	450 130	18	4	72
	Identify pages	440 630	22	3	66
3	REPLACE PAGES				
	Get binder rings	112 520	14	1	14
	Open rings	210 400	21	1	21
	Get old sheet	111 100	10	1	10
	Aside sheet to basket	123 002	22	1	22
	Get new sheet	111 100	10	1	10
	Insert sheet in binder	462 104	64	1	64
	Get rings	112 520	14	1	14
	Close rings	222 400	21	1	21
4	CLOSE COVER AND ASIDE BINDER				
	Get cover	111 520	8	1	8
	Close cover	222 100	13	1	13
	Get binder	112 520	14	1	14
	Aside binder to shelf	123 002	22	1	2
	TOTAL TMU PER CYCLE				571
	ALLOWANCES ______%				
	STANDARD HOURS PER ____ UNIT				
	UNITS PER HOUR				

From B. Niebel and A. Freivalds, *Methods, Standards, and Work Design,* 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.

TABLE 12 MTM-C Level-2 Elements

MTM-C OPERATION ANALYSIS					VALIDATION Sheet of
MTM ASSOCIATION FOR STANDARDS AND RESEARCH	MTM-C LEVEL 2 Replace page in 3-ring binder				
DEPARTMENT: Clerical	ANALYST: CNR				DATE: 2/77
No.	Description	Reference	Element TMU	Occurrence per Cycle	TMU per Cycle
	Get and aside binder	G5A2	29	1	29
	Open cover	O1	29	1	29
	Read first page	RN2	14	1	17
	Locate pages	LC12	129	1	129
	Identify pages	130	22	6	132
	Open rings	O4	35	1	35
	Remove sheet	G1A2	32	1	32
	New sheet on rings	H114	84	1	84
	Close rings	C4	35	1	35
	Close cover	C1	27	1	27
	Aside binder	G5A2	29	1	29
				TOTAL TMU PER CYCLE	575

From B. Niebel and A. Freivalds, *Methods, Standards, and Work Design,* 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.

1 can be calculated faster than MTM-2 and MTM-C Level 2 is faster than MTM-3. Comparing the standards for replacing a page in a three-ring binder, developed first using MTM-1, then using MTM-C Level 1 (see Table 11), and finally using MTM-C Level 2 (see Table 12), exhibits how closely these three standards agree (see Table 13).

3.1.6. MTM-M

MTM-M is a functional, basic-level system specifically designed for evaluating work using a stereoscopic microscope. The data used were original data developed through the efforts of the United States/Canada MTM Association. In general, MTM-M is a higher-level system, similar to MTM-2.

This system is contained in four major tables and one subtable. Four variables when selecting the appropriate data must be considered: (1) type of tool; (2) condition of the tool; (3) terminating characteristic of the motion; and (4) distance/tolerance ratio. Additional factors that have impact on motion performance include the tool load state (empty or loaded), microscopic power, distance moved, positioning tolerance, purpose of the motion, and simultaneous motions.

3.1.7. Specialized MTM Systems

There are three other specialized MTM systems: MTM-TE, MTM-MEK, and MTM-UAS. MTM-TE was developed for electronic tests and uses two levels of data derived from MTM-1. Level 1 includes

TABLE 13 Comparison of MTM-1, MTM-C (1), and MTM-C (2)

Techniques	Number of Elements	Standard (TMU)
MTM-1	57	577.8
MTM-C Level 1	21	577
MTM-C Level 2	11	575

From B. Niebel and A. Freivalds, *Methods, Standards, and Work Design,* 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.

the elements get, move, body motions, identify, adjust, and miscellaneous data. Level 2 includes get and place, read and identify, adjust, body motions, and writing.

The second specialized system, MTM-MEK, was designed to measure one-of-a-kind and small-lot production. It also is a two-level system developed from MTM-1 and can analyze all manual activities as long as certain requirements are met. MTM-MEK can be used provided that the operation is not highly repetitive; the method used to perform a given operation typically varies from cycle to cycle; the task is complex and requires employee training; and the workplace, tools, and equipment used are universal in nature. The data in MTM-MEK consist of 51 time values classified into the following eight categories: get and place, handle tool, place, operate, motion cycles, fasten or loose, body motions, and visual control. In addition, standard data for a wide range of assembly tasks in one-of-a-kind and small-lot production are also available. These data consist of 290 time values in the following categories: fasten, clamp and unclamp, clean and/or apply lubricant/adhesive, assemble standard parts, inspect and measure, mark, and transport.

The third specialized system, MTM-UAS, is a third-level system. It was developed to provide a process description as well as to determine the allowed times for activities related to batch production. MTM-UAS can be applied to batch production activities that have the following characteristics: similar tasks, customized workplace, good levels of work organization, detailed instructions, and well-trained operators. The system consists of 77 time values in seven of the eight categories used in MTM-MEK: get and place, place, handle tool, operate, motion cycles, body motions, and visual control. MTM-UAS is about eight times faster to apply than MTM-1. At cycle times greater than 4.6 min, the standard calculated using MTM-UAS is within $\pm 5\%$ of that calculated from MTM-1, with a 95% confidence level (Niebel and Freivalds 1999).

3.2. Maynard Operations Sequence Technique (MOST)

The Swedish Division of H.B. Maynard and Company, Inc. developed Maynard Operation Sequence Technique (MOST) from 1967 to 1972. An outgrowth of MTM, MOST is a simplified system developed by Zandin (1980) as a result of an extensive review of MTM data. With MOST, analysts can establish standards at least five times faster than with MTM-1 without compromising accuracy.

The MOST system is based on the structure and theory of MTM-1 and MTM-2, and its systems can be applied to direct productive work as well as material handling, distribution, maintenance, and clerical activities. It is applicable for any cycle length and repetitiveness for as long as there are variations in the motion pattern from one cycle to another.

MOST utilizes larger blocks of fundamental motions than MTM-2. In contrast to MTM-2, which is built around 37 time values for describing manual work, MOST utilizes only 16 time fragments. MOST identifies three basic sequence models: general move, controlled move, and tool use.

- *General move:* The general move sequence is defined as the spatial free movement of an object through the air. This can account for as much as 35% of the work of a machine operator and even more for an assembly worker. A specific move sequence consists of three phases, each with a subset of parameters: get (A, B, G), put (A, B, P), and return (A). Thus, this activity is represented by the following sequence of seven letters or subactivities:

$$A\ B\ G\ A\ B\ P\ A$$

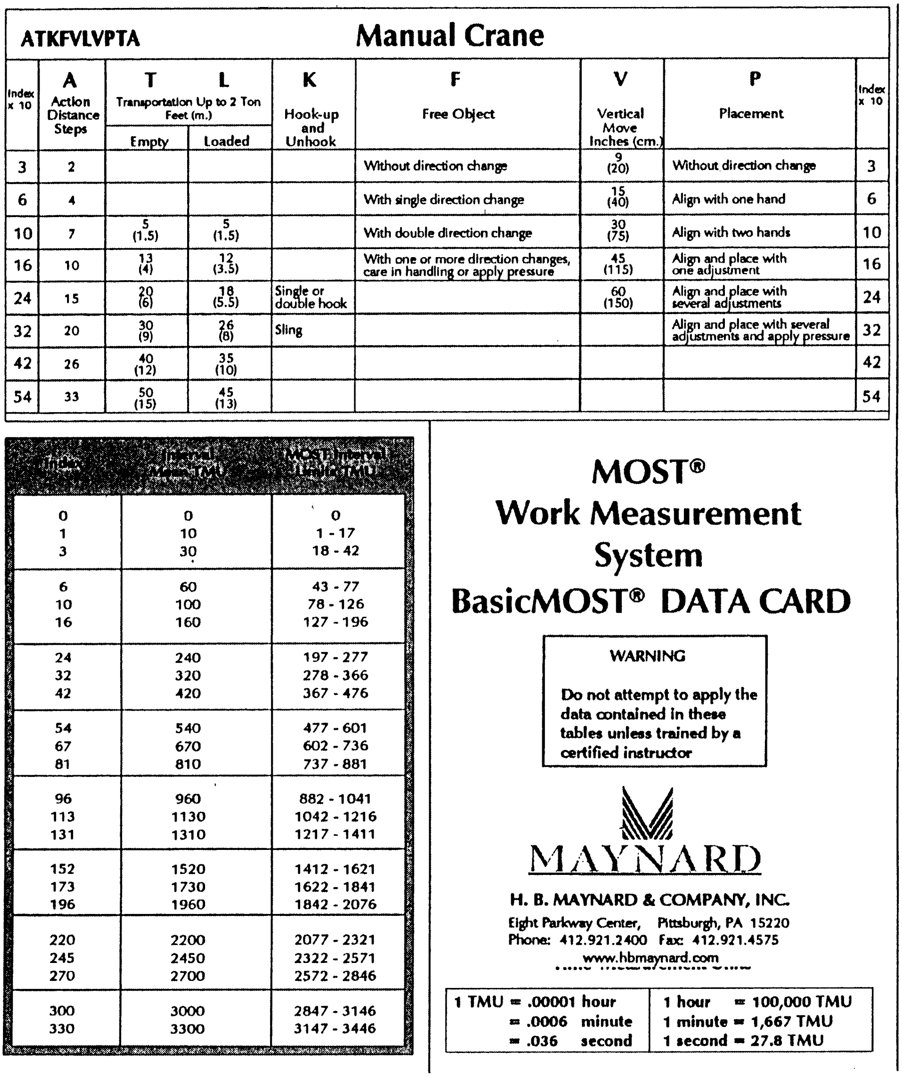
where A = action distance (primarily horizontal distance of hand or body motions)
B = bend (mainly vertical body motions)
G = grasp or gain control
P = position or place

The variations for each subactivity are indicated by a time-related index number to the applicable parameter. MOST uses index numbers of 0, 1, 3, 6, 10, and 16, corresponding to the relative difficulty of the parameter as shown in Table 14. Simply adding the index numbers and multiplying the sum by a scaled factor of 10 to yield the appropriate TMUs obtains the time value for the sequence. For instance, a move sequence is indicated by the index figure:

$$A_1B_0G_1A_1B_0P_1A_1$$

where A_1 = reach to washer with 12.7 cm (5 in.) travel
B_0 = no body motion
G_1 = grasp washer
A_1 = place washer with 12.7 cm (5 in.) travel
B_0 = no body motion
P_1 = place washer with a loose fit
A_1 = return to original position with 12.7 cm (5 in.) travel

TABLE 14 Basic MOST Data Card



ATKFVLVPTA **Manual Crane**

Index x 10	A Action Distance Steps	T Transportation Up to 2 Ton Feet (m.)	L	K Hook-up and Unhook	F Free Object	V Vertical Move Inches (cm.)	P Placement	Index x 10
		Empty	Loaded					
3	2				Without direction change	9 (20)	Without direction change	3
6	4				With single direction change	15 (40)	Align with one hand	6
10	7	5 (1.5)	5 (1.5)		With double direction change	30 (75)	Align with two hands	10
16	10	13 (4)	12 (3.5)		With one or more direction changes, care in handling or apply pressure	45 (115)	Align and place with one adjustment	16
24	15	20 (6)	18 (5.5)	Single or double hook		60 (150)	Align and place with several adjustments	24
32	20	30 (9)	26 (8)	Sling			Align and place with several adjustments and apply pressure	32
42	26	40 (12)	35 (10)					42
54	33	50 (15)	45 (13)					54

Index	Interval Mean TMU	MOST Interval Limits TMU
0	0	0
1	10	1 - 17
3	30	18 - 42
6	60	43 - 77
10	100	78 - 126
16	160	127 - 196
24	240	197 - 277
32	320	278 - 366
42	420	367 - 476
54	540	477 - 601
67	670	602 - 736
81	810	737 - 881
96	960	882 - 1041
113	1130	1042 - 1216
131	1310	1217 - 1411
152	1520	1412 - 1621
173	1730	1622 - 1841
196	1960	1842 - 2076
220	2200	2077 - 2321
245	2450	2322 - 2571
270	2700	2572 - 2846
300	3000	2847 - 3146
330	3300	3147 - 3446

MOST®
Work Measurement
System
BasicMOST® DATA CARD

WARNING

Do not attempt to apply the data contained in these tables unless trained by a certified instructor

MAYNARD

H. B. MAYNARD & COMPANY, INC.
Eight Parkway Center, Pittsburgh, PA 15220
Phone: 412.921.2400 Fax: 412.921.4575
www.hbmaynard.com

1 TMU = .00001 hour	**1 hour = 100,000 TMU**
= .0006 minute	**1 minute = 1,667 TMU**
= .036 second	**1 second = 27.8 TMU**

The standard time for this sequence is (1 + 0 + 1 + 1 + 0 + 1 + 1) × 10 = 50 TMUs.

- *Controlled move:* The controlled move sequence describe the movement of an object when it either remains in contact with a surface or remains attached to another object during the movement. It covers manual operations such as cranking, pulling a starting lever, turning a steering wheel, and engaging a starting switch. In controlled move sequences, the following parameters are considered: the previously defined action distance (A), body motion (B), gain control (G); and the new parameters move controlled (M), process time (X), and align (I). The controlled move parameters are further defined in Table 14.
- *Tool use:* The tool-use sequence covers the use of common hand tools. Cutting, gauging, fastening, and writing with tools are all covered by this sequence. The tool-use sequence embraces a combination of general move and controlled move activities. Additional parameters unique to

TABLE 14 (*Continued*)

ABG Get ABP Put A Return General Move

Index x 10	A Action Distance	B Body Motion	G Gain Control	P Placement	Index x 10
0	≤ 2 in. (5 cm.)	No Body Motion	No Gain Control Hold	No Placement Hold Toss	0
1	Within Reach		Grasp Light Object Grasp Light Objects Simo	Lay Aside Loose Fit	1
3	1 - 2 Steps	Sit without adjustments Stand without adjustments Bend and Arise 50% occ.	Get Non-simo Get Heavy/Bulky Get Blind Get Obstructed Free Interlocked Disengage Collect	Loose Fit Blind Place with Adjustments Place with Light Pressure Place with Double Placement	3
6	3 - 4 Steps	Bend and Arise		Position with Care Position with Precision Position Blind Position Obstructed Position with Heavy Pressure Position with Intermediate Moves	6
10	5 - 7 Steps	Sit Stand			10
16	8 - 10 Steps	Bend and Sit Climb on Climb off Stand and Bend Through Door			16

A Action Distance Extended Values

Index	Steps	Distance (ft.)	Distance (m.)
24	11-15	38	12
32	16-20	50	15
42	21-26	65	20
54	27-33	83	25
67	34-40	100	30
81	41-49	123	38
96	50-57	143	44
113	58-67	168	51
131	68-78	195	59
152	79-90	225	69
173	91-102	255	78
196	103-115	288	88
220	116-128	320	98
245	129-142	355	108
270	143-158	395	120
300	159-174	435	133
330	175-191	478	146

ABG Get MXI Move/Actuate A Return Controlled Move

Index x 10	M Move Controlled Push/Pull/Pivot	Crank	X Process Time Seconds	Minutes	Hours	I Alignment	Index x 10
0	No Action	No Action	No Process Time			No Alignment	0
1	Push/Pull/Pivot ≤ 12 in. (30 cm.) Push Button Push or Pull Switch Rotate Knob		.5 sec.	.01 min.	.0001hr.	Align to 1 Point	1
3	Push/Pull/Pivot > 12 in. (30 cm.) Push/Pull with Resistance Seat Unseat Push/Pull with High Control Push/Pull 2 Stages ≤ 12 in.(30 cm.) Push/Pull 2 Stages ≤ 24 in. Total	1 Rev.	1.5 sec.	.02 min.	.0004 hr.	Align to 2 Points ≤ 4 in. (10 cm.)	3
6	Push/Pull 2 Stages > 12 in. (30 cm.) Push/Pull 2 Stages > 24 in. Total Push with 1 - 2 Steps	2 - 3 Revs.	2.5 sec.	.04 min.	.0007 hr.	Align to 2 Points > 4 in. (10 cm.)	6
10	Push/Pull 3 - 4 Stages Push with 3 - 5 Steps	4 - 6 Revs.	4.5 sec.	.07 min.	.0012 hr.		10
16	Push with 6 - 9 Steps	7 - 11 Revs.	7.0 sec.	.11 min.	.0019 hr.	Align with Precision	16

M Push or Pull Extended Values

Index	Steps
24	10-13
32	14-17
42	18-22
54	23-28
67	29-34

Crank Extended Values

Index	Revs.
24	12-16
32	17-21
42	22-28
54	29-36

X Process Time Extended Values

Index	Seconds	Minutes	Hours
24	9.5	.16	.0027
32	13.0	.21	.0036
42	17.0	.28	.0047
54	21.5	.36	.0060
67	26.0	.44	.0073
81	31.5	.52	.0088
96	37.0	.62	.0104
113	43.5	.72	.0121
131	50.5	.84	.0141
152	58.0	.97	.0162
173	66.0	1.10	.0184
196	74.5	1.24	.0207
220	83.5	1.39	.0232
245	92.5	1.54	.0257
270	102.0	1.70	.0284
300	113.0	1.88	.0314
330	124.0	2.06	.0344

this activity are: fasten (F), loosen (L), cut (C), surface treat (S), record (R), think (T), and measure (M). These are further defined also in Table 14.

Two adaptations to the MOST work-measurement system are also in use: mini MOST and maxi MOST. Mini MOST measures identical, short-cycle operations, while maxi MOST measures long-cycle operations with significant variations in actual method from one cycle to another.

MOST is now available in a computer-aided format. This computerized version allows the analyst to retrieve suboperation data and perform the numerical operations to determine the standard of performance for the input characteristics of the method being analyzed. Another relatively new computerized version, ErgoMOST, enables the analyst to examine ergonomic problems in the workplace. ErgoMOST uses a biomechanical model to calculate stresses of push/pull and lifts, highlight awkward postures and repetitive body movements, and quantify the relative risk of the job using ergonomic stress indices. The computer software provides a utility function for identifying ergonomic recommendations and generating reports. The advantages of using a computerized system vs. manual applications are faster speed at which application can be made, minimized errors, and facilitation of numerical operations.

3.3. Macromotion Analyses

Standards International, Inc. has developed two specialized predetermined time systems: MICRO Motion Analyses and MACRO Motion Analyses. MICRO Motion Analyses is used for precise methods specifications and time standards, while MACRO Motion Analyses is for general-purpose data. They were developed to provide improvements over MTM and Work-Factor with much input from several of Standards International's clientele. Specifically, MTM and Work-Factor systems were not found to be adequate for some special types of motions, which entailed describing these motions and assigning appropriate time values using the individual user's judgement. Also, some analysts found difficulty in using the tables. These systems have been proven to be valid in thousands of applications,

TABLE 14 (*Continued*)

ABG Get Tool **ABP** Put Tool ***** Use Tool **ABP** Aside Tool **A** Return **Tool Use**

	F L Fasten or Loosen											
	Finger Action	Wrist Action				Arm Action					Tool Action	
Index x 10	Spins	Turns	Strokes	Cranks	Taps	Turns		Strokes	Cranks	Strikes	Screw Dia.	Index x 10
	Fingers, Screw-driver	Hand, Screw-driver, Ratchet, T-Wrench	Wrench, Allen key	Wrench, Allen key, Ratchet	Hand, Hammer	Ratchet	T-Wrench 2-Hands	Wrench, Allen key	Wrench, Allen key, Ratchet	Hand, Hammer	Power Wrench	
1	1	-	-	-	1	-	-	-	-	-	-	1
3	2	1	1	1	3	1	-	1	-	1	1/4" (6mm)	3
6	3	3	2	3	6	2	1	-	1	3	1" (25mm)	6
10	8	5	3	5	10	4	-	2	2	5		10
16	16	9	5	8	16	6	3	3	3	8		16
24	25	13	8	11	23	9	6	4	5	12		24
32	35	17	10		30	12	8	6		16		32
42	47	23	13		39	15	11	8		21		42
54	61	29	17		50	20	15	10		27		54

P Tool Placement

Tool	Index
Hammer	0 (1)
Fingers or Hand	1 (3 or 6)
Knife	1 (3)
Scissors	1 (3)
Pliers	1 (3)
Writing Instrument	1
Measuring Device	1
Surface Treating Device	1
Screwdriver	3
Ratchet	3
T-Wrench	3
Fixed End Wrench	3
Allen Wrench	3
Power Wrench	3
Adjustable Wrench	6

I Alignment of Machining Tools

Index	Align to
3	Workpiece
6	Scale Mark
10	Indicator Dial

Alignment of Nontypical Objects

Index	Positioning Method
0	Against stop(s)
3	1 adjustment to stop
6	2 adjustments to stop(s) 1 adjustment to 2 stops
10	3 adjustments to stop(s) 2-3 adjustments to linemark

Nontypical Object Characteristics
Flat, Large, Flimsy, Sharp, Difficult to Handle

ABG Get Tool **ABP** Put Tool ***** Use Tool **ABP** Aside Tool **A** Return **Tool Use**

	C Cut				S Surface Treat			M Measure	R Record			T Think			
Index x 10	Twist/ Bend	Cutoff	Cut	Slice	Air-Clean	Brush-Clean	Wipe	Measure	Write		Mark	Inspect	Read		Index x 10
	Pliers		Scissors	Knife	Nozzle	Brush	Cloth	Measuring Device	Pencil		Marker	Eyes, Fingers	Eyes		
		Wire	Cut(s)	Slice(s)	sq. ft. (0,1 m.^{2})	sq. ft. (0,1 m.^{2})	sq. ft. (0,1 m.^{2})	in.(cm.) ft.(m.)	Digits	Words	Digits	Points	Digits, Single Words	Text of Words	
1	Grip		1	-	-	-	-		1	-	Check Mark	1	1	3	1
3		Soft	2	1	-	-	1/2		2	-	1 Scribe Line	3	3 Gauge	8	3
6	Twist Bend-Loop	Medium	4	-	1 Spot Point Cavity	1 Small Object	-		4	1	2	5 Touch for Heat	6 Scale Value Date or Time	15	6
10		Hard	7	3	-	-	1	Profile-Gauge	6	-	3	9 Feel for Defect	12 Vernier-Scale	24	10
16	Bend Cotter Pin		11	4	3	2	2	Fixed Scale Caliper 12 in. (30 cm.)	9 Signature or Date	2	5		Table Value	38	16
24			15	6	4	3	-	Feeler-Gauge	13	3	7			54	24
32			20	9	7	5	5	Steel-Tape 6 ft. (2 m.) Depth Micrometer	18	4	10			72	32
42			27	11	10	7	7	OD-Micrometer 4 in. (10 cm.)	23	5	13			94	42
54			33					ID-Micrometer 4 in. (10 cm.)	29	7	16			119	54

Reprinted with permission of H. B. Maynard and Company. © 1998 H. B. Maynard & Company, Inc. MOST®, Most is a registered trademark of H. B. Maynard and Company, Inc.

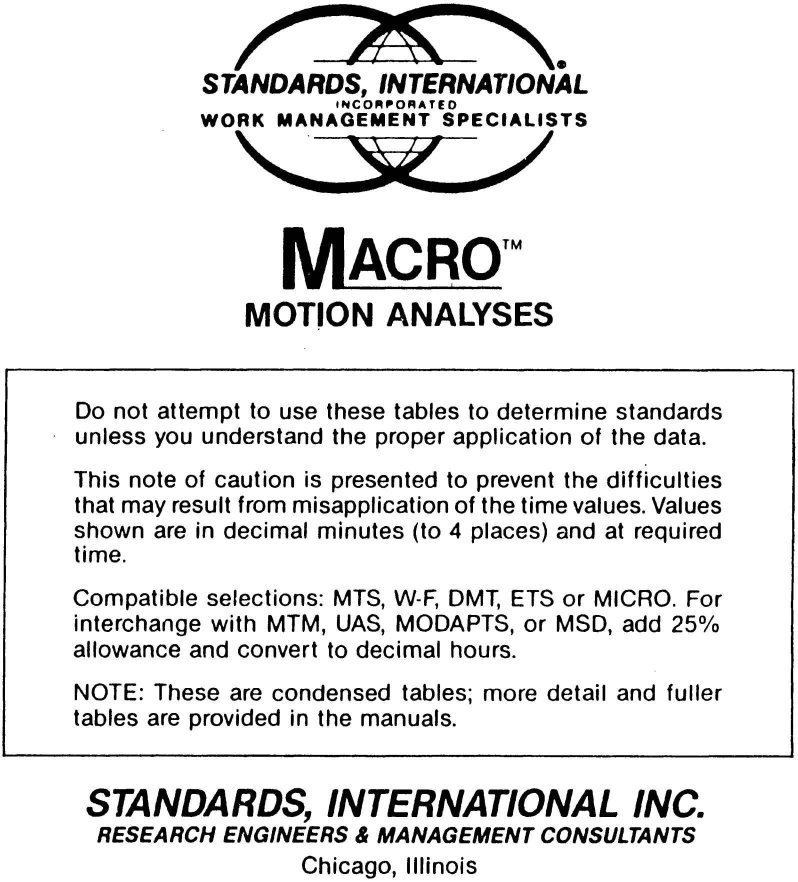
having been used by Standards International for several years in conjunction with their consulting activities. The MACRO Motion tables are provided in Table 15.

3.4. Guidelines for System Selection

The choice of PTS system to use is generally characterized by the following:

1. *Terminology:* PTS terminology is primarily verbs that are used to denote the action taken by body members. Each system defines its own set of action words for the body motions covered by its time values. A term used in one system may or may not mean the same in another system. Therefore, in order to properly use any PTS system, one needs to have a clear understanding of the rules in the use of the system.
2. *Work pace:* Performance rating is not necessary when using predetermined motion times because motion–time values have been established in advance. For example, MTM data were leveled to represent a 100% or "normal" performance level.
3. *Methods description required:* The extent of methods description provided by the analyst is usually determined by the detail required by the system. Although a more detailed description can be provided, it is typical for the analyst to provide only the minimum required. Condensed systems tend to have a lesser degree of methods description than basic-level systems. The more methods description is required, the longer the time needed for analysis.
4. *Accuracy desired:* Some operations need very precise time values for manual operations. Precise measurement is possible with several of the detailed PTS systems where the work motions are well structured and predictable. On other operations, the work motions are unpredictable and it is impractical to make a detailed analyses. Tables are available that provide quick approximate time values.

TABLE 15 MACRO Motion Analysis



STANDARDS, INTERNATIONAL
INCORPORATED
WORK MANAGEMENT SPECIALISTS

MACRO™
MOTION ANALYSES

Do not attempt to use these tables to determine standards unless you understand the proper application of the data.

This note of caution is presented to prevent the difficulties that may result from misapplication of the time values. Values shown are in decimal minutes (to 4 places) and at required time.

Compatible selections: MTS, W-F, DMT, ETS or MICRO. For interchange with MTM, UAS, MODAPTS, or MSD, add 25% allowance and convert to decimal hours.

NOTE: These are condensed tables; more detail and fuller tables are provided in the manuals.

STANDARDS, INTERNATIONAL INC.
RESEARCH ENGINEERS & MANAGEMENT CONSULTANTS
Chicago, Illinois

5. *Speed and ease of application:* The length of the individual motion time or element time shown on the PTS tables determines the speed of application. The shorter the average element time, the greater the number of elements that need to be recorded and the longer it will take to perform the analysis. The ease of application is influenced by the clarity of the terminology and rules for the motions to be used. If the motion–time rules are unclear or inexact, a great deal of time may be lost in selecting the right time value from the PTS tables and the chances of errors will increase.

Exact understanding and correct application of predetermined motion times require careful training in the system employed. Guidance from a thoroughly trained practitioner in the system is preferable. Other major considerations include the type of work to be measured and the purpose of the measurement. Analysts must pick the system to fit the work and the need.

The ultimate guideline on the system to be used depends on management's objectives and beliefs. If management requires accurate performance measures, the system selected should be one that will give accurate values for the type of work to be measured. On the other hand, if management desires only approximate results, it is unnecessary to spend much time and effort in developing precise measurements.

4. STANDARD DATA

Many operations in a given system have several common elements. The element "walking," for example, is a component of many different jobs. When these jobs are timed, the same common

TABLE 15 (*Continued*)

MACRO™ MOTION ANALYSES

Frequently used General Purpose Data Elements recommended for:

1. Evaluation of manual motions (and costs) by methods personnel and production supervisors.
2. Development of standards for short run, long cycle operations by methods/standards personnel.

OBTAIN AND PLACE			DIST. RANGE IN INCHES			
WT.	CONDITIONS OF OBTAIN	PLACE ACCURACY	CODE	6"	18"	30"
2 LBS. OR LESS	EASY GRASP	Approximate	OEA	90	150	190
		Close	OEC	110	170	210
		Tight	OET	130	190	230
	DIFFICULT GRASP	Approximate	ODA	100	160	200
		Close	ODC	120	180	220
		Tight	ODT	140	200	240
	GRASP HANDFUL	Approximate	OHA	150	190	230
OVER 2 LBS. THRU 18 LBS.		Approximate	OWHA	160	230	270
		Close	OWHC	170	250	290
		Tight	OWHT	180	270	310
OVER 18 LBS. THRU 48 LBS.		Approximate	OW2HA	190	290	320
		Close	OW2HC	200	310	340
		Tight	OW2HT	210	330	360

PLACE ONLY	CODE	6"	18"	30"
Approximate	PA	40	60	80
Close	PC	60	80	100
Tight	PT	70	100	130
Add for: Weights: OVER 2 LBS THRU 24 LBS. 40; OVER 24 LBS. 60				

element is timed again and again. The function of the analyst would therefore be made much easier if the analyst had a set of data from which he or she could readily derive standard times for these common work elements without necessarily going into the process of timing each one. If, for instance, a standard time would be derived for the particular element "walking" and could be read directly from a table, this would not only reduce effort and cost but also lead to greater consistency in time estimations.

It is however, difficult to visualize a situation where all the possible elements making up a job could be timed and stored for future retrieval. We may therefore conclude that in practice it is better to restrict the number of jobs for which standard data are derived.

TABLE 15 (*Continued*)

ASSEMBLE NON-ROUND OBJECT	CODE	UP TO & INCL. 18 LBS.	OVER 18 LBS. THRU 48 LBS.
UP TO & INCL. 1/2" CLEARANCE	ANT	90	120
OVER 1/2" CLEARANCE	ANC	50	70

ASSEMBLE	CODE	PLUG		
		UP TO & INCL. 1/4"	OVER 1/4" THRU 1/2"	OVER 1/2" THRU 1"
Loose Fit	ASYL	60	50	40
Normal Fit	ASYN	80	70	60
Add for: Simo (S) 15 Temporary Blind (TB) 15		Apply Pressure (AP) 40 Regrasp & Apply Pressure (RAP) 60		

CIRCULAR MOTIONS	CODE	DIAMETER		
		3" & UNDER	OVER 3" THRU 12"	OVER 12"
Revolution without Deceleration	CR	50	80	90
Revolution to General Location	CRG	60	110	120
Revolution to Exact Location	CRE	80	140	150

MOTION-PATTERNS

WALKING						(Measured at Toe or Heel)
NUMBER OF STEPS	OPEN/BASIC (WO)		CONFINED (WC)		RESTRICTED (WR)	
	DIST. FT.	TIME	DIST. FT.	TIME	DIST. FT.	TIME
1 (1 Leg)	2.5'	120	2.5'	130	2'	140
1 (Both Legs)	2.5'	200	2.5'	220	2'	240
2	5'	260	5'	280	4'	300
5	13.5'	500	13.5'	550	11'	600
10	26'	900	26'	1000	21'	1100
Each Add'l Pace — Add	2.5'	80	2.5'	90	2'	100
Add for Turn Over 120°	—	100	—	100	—	100
Stairs/Up (SU) 130 /Down (SD) 100		Bend (B) 160 Arise (AB) 130		Sit (SI) 230 Stand (ST) 280		

The reliability of the data can be increased if as many common elements as possible that are performed in the same way are grouped together from analysis and if a sufficient amount of accumulated or collected data on each element has been analyzed by a trained analyst. Making sure that all the factors affecting a certain element have been taken into consideration can further increase reliability.

Standard data refers to all the tabulated elemental standards, curves, alignment charts, and tables that are compiled to allow the measurement of a specific job without the use of a timing device.

TABLE 15 (*Continued*)

VISUAL & MENTAL PROCESSES	CODE	TIME
READ — Per Word	ER	50
INSPECT — Per Criteria — Per Occurrence	EI	50
RECALL, DECIDE, REACT or CALCULATE — Per Criteria — Per Occurrence	MA	50
MOVE HEAD TO MICROSCOPE & SEE	HM	130
WRITE — Per Character	WC	90

MOTION-PATTERNS

TOOL HANDLING

TYPE OF TOOL	MOTION PATTERN	DISTANCE 9"	18"	30"	36"
Screw Driver, Mallet	OX-RH-E	210	270	340	360
Spin-Tites, Knives	OX-RH-G	180	230	290	310
File (w/Rd. Handle)					
Pliers, Wire Strippers	OX-DH-E	240	300	360	380
Scissors, Cutters	OX-DH-G	200	260	310	330
Pencils, Brushes	OX-TH-E	210	270	330	350
	OX-TH-G	170	220	280	300
Soldering Iron: —	OX-SI-E	310	380	440	470
in Holster	OX-SI-G	280	340	390	410
Air Tool: —	OX-ATB-E	260	340	400	430
on Bench	OX-ATB-G	240	300	360	380
Air Tool: —	OX-ATH-E	370	450	520	540
in Holster	OX-ATH-G	340	410	470	490
Air Tool: —	OX-ATS-E	170	220	260	270
Suspended	OX-ATS-G	140	170	210	220

Motion Pattern E: Pick up tool; move directly to work; place tool aside.
Motion Pattern G: Pick up tool; move to general area (move to assemble, etc., must be added); place tool aside.

RELATED ELEMENTS	CODE	TIME
TWEEZERS GRASP	GT	140
PALM & UNPALM TOOL	PXT	30
START THREADS	STT	90
THREAD ON OR OFF — Per Twist	TO	30
USE: SCREWDRIVER — Per Twist	SDT	70
NUT RUNNER — Per Twist	NRT	70
BOX OR END WRENCH — Per Turn	BEW	180
RATCHET — Per Turn	RW	60
HAMMER — Per Hit	HB	90
TIGHTEN OR LOOSEN w/HAND TOOL	ASYX	100

© 1989. All rights reserved. Permission courtesy of Clifford N. Sellie, CEO, Standards, International.

These are element time standards, from time studies, that have been proven to be accurate and reliable. There are three levels of detail: MOTION, ELEMENTAL, and TASKS. The more refined the standard data element, the broader is its range of application. Although it has the widest usage, it takes longer to develop.

Motion-level systems are typified by MTM, MODAPTS, or other PTS systems. It involves times of the smallest component range from about 0.01 to 1 sec (e.g., reach, grasp, move, position, release). Elemental level system components vary from about 1 to 1000 sec. Components come from either PTS systems or time study (e.g., get equipment, polish shoes, put equipment away). Task-level component times range upward. Components come from elemental combinations, from time study, from occurrence sampling, or even from employee activity logs (e.g., loading of truck, driving truck 200 km [124 mi], unloading truck).

Components can either be constant or variable. It is critical in developing standard data to model variable components. Example of variable elements are include in Table 16, illustrating the levels of detail for standard time systems (Konz 1990).

The advantages of the standard data approach include the following:

1. Provides a catalog of facts, therefore saving time, effort, and money
2. Permits analysts to concentrate on methods improvement
3. Provides a quick and reliable tool for methods and standards evaluation
4. Allows standards to be determined prior to production
5. Results in more consistent standards among jobs, departments, and plants
6. Eliminates the subjective rating of the operator’s speed performance

4.1. Developing Standard Time Data

The general approach in developing standard time data is as follows:

1. *Decide on the coverage:* The coverage should be restricted to one or more departments or work areas or to a limited range of processes within the system in which several similar elements, performed by the same method, are involved in carrying out the jobs.
2. *Break the jobs into elements:* In this case, try to identify as many elements as possible that are common to the various jobs.

TABLE 16 Levels of Detail for Standard Time Systems

Motions (Typical time range from 0.1 to 1 sec)			
Element	Code	Time, TMU	
Reach	R10C	12.9	
Grasp	G4B	9.1	1 sec = 27.8 TMU
Move	M10B	12.2	
Position	P1SE	5.6	1 TMU = 0.036 sec
Release, etc.	RL1	2.0	
Constant elements (Typical time range 1 sec to 1000 sec)			
Element		Time, sec	
Assemble bracket to unit		8	
Get equipment		90	
Polish a shoe		130	
Put equipment away		47	
Load carton on pallet		15	
Pack box		40	
Variable elements (Typical time range 1 sec to 1000 sec)			
Element	Parameters		
Walk	F(distance traveled, load carried)		
Assemble bracket	F(number of holes, type of fastener)		
Pack box	F(type of component packed, number packed)		
Repair TV set	F(manufacturing of set, type of defect)		

From S. Konz, *Work Design: Industrial Ergonomics,* 3d Ed. Copyright © 1990 Publishing Horizons, Inc. Reprinted by permission.

3. *Decide on the type of reading:* Decide whether to use readings based on stopwatch time studies or derived from PTS systems such as MTM.
4. *Determine the relevant factors:* There are factors that are likely to affect the time for each element. Classify them into major and minor factors.
5. To validate the standard time data, measure the time taken to perform the activity from actual observation.

4.2. Uses of Standard Data

For easy reference, standard data should be classified as either constant or variable data elements. Constant standard data elements should be tabulated. Variable data can either be tabulated or expressed as a curve or equation through formula construction.

Formula construction represents a simplification of standard data. It involves the design of an algebraic expression or a system of curves that establishes a time standard in advance of production by substitution of known values peculiar to the job for the variable element.

In this age of computers, standard data can be stored, retrieved, and accumulated. Several software systems have fundamental motion databases. The software can select, retrieve, and modify the appropriate motion or elements to generate a time standard. When properly applied, standard data allow the establishment of accurate time standards prior to actual production quickly and consistently. This is especially useful for estimating the cost of new products or work, preparing quotations and for subcontracting purposes and for establishing standards for indirect labor. In addition, this technique provides management with fair and proven satisfactory standards that help alleviate labor–management conflicts.

5. WORK SAMPLING

Among work measurement techniques, work sampling is one of the simplest available to time study analysts, yet it can be adapted to analyze complex and sophisticated patterns of work. Work sampling is a method of finding the proportions of total time devoted to various activities that constitute a job or work situation by random sampling. It is called by other names, including activity sampling, ratio-delay study, random observation method, snap-reading method, and observation ratio study.

The results of work sampling are effective for determining machine and personnel utilization, allowances applicable to the job and production standards. They are also particularly useful for analyzing nonrepetitive or irregularly occurring activities where no complete methods or frequency descriptions are available. Although the same information can be obtained using time study procedures, work sampling often provides this information faster and at considerably less cost.

Work sampling studies normally extend over a long period (two- to four-week period) to enable a comparatively large number of observations at random intervals to be taken. The percentage of time that a process is in a certain activity is obtained by the ratio of observations of the given activity to the total number of observations. Consider an analyst who takes 100 observations of the shop floor at random intervals over several weeks. It was observed that a certain machine was idle in 30 instances for miscellaneous reasons. The estimated idle time of this machine would then be 30% of the working day.

L. H. C. Tippett first applied work sampling in the British textile industry during the early 1930s. The technique was later introduced in the United States by Morrow (1941) under the name ‘‘ratio-delay study.’’ The technique did not gain wide acceptance at the outset. It gained attention when the name ‘‘work sampling’’ was coined by Brisley (1952) and Waddell (1952) at the time when increased attention was being paid to indirect labor. Also, the increasing number of able practitioners brought more discriminating use of the technique.

The accuracy of the data determined using work sampling depends mainly on the number of observations made, the length of the period under which the study was undertaken, and the conditions under which the study was conducted. Reliability and accuracy of information can only be obtained by using a sufficiently large sample size and a sampling period representing typical conditions.

5.1. Definitions and Objectives of Work Sampling Studies

Work is defined in work sampling studies as the specific activity being studied. It is further broken down into mutually exclusive categories of work. Exhaustive descriptions of the activity are defined so that an observation can be readily classified as belonging to one and only one category. Work sampling studies are not necessarily confined to studying work environments but can also be applied to other activities or events. They can be used to estimate, for example, the percentage of TV commercials that are about beer, percentage of people wearing glasses, and the like.

The theory of work sampling is based on the fundamental law of probability. That is, an event can either be present or absent at any given instant. This concept is presented in the following expression:

$$(p + q)^n = 1$$

where p = probability of a single occurrence
$q = (1 - p)$ = probability of an absence of an occurrence
n = number of observations

If the above expression is expanded according to the binomial theorem, the distribution of the probabilities obtained from the expansion is known as the binomial distribution. According to elementary statistics, as n becomes larger, the binomial distribution approaches the normal distribution. Since work sampling studies involve large sample sizes, the normal distribution is a satisfactory approximation of the binomial distribution. Rather than use the binomial distribution, it is more convenient to use the distribution of a proportion, with a mean p and a standard deviation of $\sqrt{pq/n}$ as the approximately normally distributed random variable.

The immediate goal of work sampling is to produce an estimate of the proportion of total time of each category of work a job has been subdivided. To establish a good estimate, two qualities must be taken into consideration: absence of bias and low variance.

A biased estimate is one that is apt to be too high or too low because of the procedure used to determine it. For example, if one work category is "clean up" and the analyst always leaves before the end of the shift, the estimate for this category will probably be too low because the observer may miss some occurrences that take place after he or she leaves the work area. Similarly, if workers are able to anticipate the times when observations are taken, the results of the study may be biased by the workers themselves. It is expected that workers will engage in productive activity as soon as the analyst is seen taking observations.

Work sampling studies are undertaken for various purposes, including obtaining general information, justifying proposed changes, and setting standards. Implicitly, obtaining general information allows management to achieve cost reduction through improved utilization of equipment or manpower and the identification of areas that need additional facilities. General information studies are also used to establish and monitor standards for indirect labor, in which case pace rating is done at the time of observation and some unit (or units) of output must be collected.

Work sampling studies are also frequently used to substantiate a subjective opinion. For example, the cost of installing a new materials-handling system may need to be justified with a reputable estimate of the amount of time currently being spent on materials-handling activities. Although the design of such a special-purpose study is easier than of a general information study, problems may arise with the objectivity of the analysts. The attempt to prove that the suggestion is justified may introduce bias into the estimates.

Stopwatch time studies and predetermined time standards are generally superior to work sampling for setting time standards on any well-defined job. However, work sampling can be used if other systems are not appropriate, such as for determining the unavoidable delay allowances. Because work sampling studies extend over several weeks, many different causes of delay can be captured and hence the information obtained can be very useful in establishing allowances.

Work sampling is frequently used to set job standards for jobs having irregular components that vary in the amount of time devoted to any one unit of output. For example, time for reconditioning and repairs on machines varies depending on the severity of the damage or cause of the breakdown. When a large sample population of jobs, representative of the population of future jobs, is observed, the established standard is expected to fit the mean of that population.

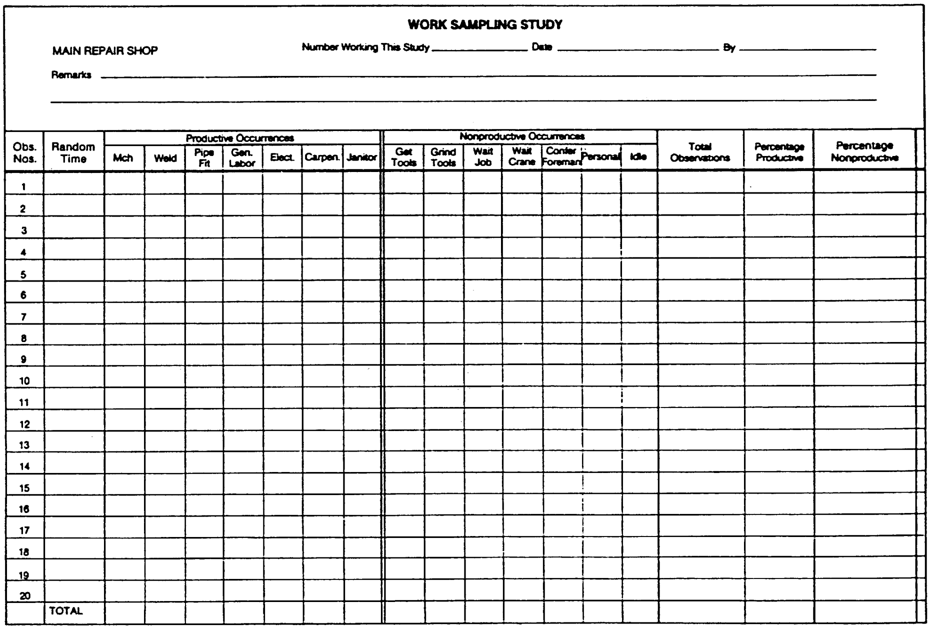
5.2. Work Sampling Methodology

Before any actual observations for work sampling study can be made, detailed planning must be done. Because work sampling is a flexible tool adaptable to both very simple and very complex studies, it is impossible to provide a single step-by-step procedure for designing and conducting a study that will fit all uses of work sampling. Hence, the following list of steps is presented as an example procedure rather than a comprehensive guide.

1. *Determine the scope of the study:* It is important to decide on the objective of the work sampling. Select the area to be studied and determine the method of measuring output. Make sure that the work sampling technique is indeed appropriate for the concern area.
2. *Brief personnel involved in the study:* Before beginning a work sampling program, the analyst must explain and solicit the cooperation of all the members of the organization who will be affected by the results of the study. Labor–management disputes are minimized by a careful explanation of the study's objective(s), methodology and deliverables and the possible effects the study will have on the members and their work.
3. *Determine the general method to be used to record data:* In general, this is easily determined by the available equipment. For example, electronic work-measurement machines with op-

tional work sampling software (see Section 2.2.1) can be extremely helpful in scheduling random observations, performance rating individual readings, and producing summary statistics. On the other hand, the purchase of a new computer system for a specific study can cause unreasonable delay.

4. *Classify the activity into work categories:* It is essential from the outset that the analyst be clear in his or her own mind about what he or she wants to achieve and why. Avoid ambiguity when classifying activities. Categories should be consistent with the end use of the study, initially recognizable by sight. Categories should be carefully defined so as to be mutually exclusive or nonoverlapping.
5. *Design the appropriate forms to be used for recording data:* An observation form to record the data to be gathered during the work sampling study should be designed. Because each work sampling study is unique from the standpoint of the total observations needed, the random times that the observations are made, and the information being sought, a standard form is not usually accepted. The best form is tailored to the study objectives. Figure 9 is an example of a work sampling study form.
6. *Carefully plan the frequency and interval of sampling:*
   - **(a)** *Conduct preliminary sampling:* Make a small random observation (about 30 samples) to generate the approximate values of p and q for key categories of the activity.
   - **(b)** *Determine confidence level and margin of error preferred for the study:* Confidence level is commonly pegged at 95%, 99%, and 99.9% with equivalent Z-values of 1.96, 2.56 and 3.3 respectively.
   - **(c)** *Determine the sample size* n *and the time period of the study:* Determine the approximate number of observations that will be required and divide this by the number of days of the study and by the number of observations per round. This should give you the number of observation rounds per day that are needed. If this figure is unreasonably large, then either the number of days or the desired precision must be evaluated. It is better if the study period is long. This would better show normal fluctuations in production or work (e.g., if you need to make 300 observations, you should not make all 300 observations in one day but spread it over several observation days.).
   - **(d)** *Select the necessary random times:* Using tables, calculators, or computers, establish the necessary random times when observations are to be taken. Although it is not necessary



WORK SAMPLING STUDY

MAIN REPAIR SHOP Number Working This Study ________ Date ________ By ________

Remarks ________

Obs. Nos.	Random Time	Productive Occurrences							Nonproductive Occurrences							Total Observations	Percentage Productive	Percentage Nonproductive
		Mch	Weld	Pipe Fit	Gen. Labor	Elect.	Carpen.	Janitor	Get Tools	Grind Tools	Wait Job	Wait Crane	Confer Foreman	Personal	Idle			
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
	TOTAL																	

Figure 9 Work Sampling Study Form. (From B. Niebel and A. Freivalds, Methods, Standards, and Work Design, 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.)

to generate all predetermined times of observations for the entire study, the times for the first few days should be established to make the sampling method clear to the analyst. For example, random number generators built into most scientific calculators can be used to identify random observation times during which observations are taken.

7. *Take a short test study:* The first day's data, and possibly the second or third if they are recognized to contain bias, will need to be discarded (Pape 1992). Individual circumstances such as familiarity of the workforce with the work sampling and the cost of data collection will determine the length of the test study. Familiarity with the activity and the work sampling scheme should be established by the analyst.
8. *Make and record observations:* Adhere to the work sampling scheme. Gather and record data so that it can be used in multiple ways. Take note of other identifiers or qualifiers of work categories such as day of the week, location, shift, and the like. As the study progresses, control charts (see Section 5.3.4) can be set up to plot the daily estimates of event/activity occurrences to monitor one ore more of the p values for key categories.
9. *Process the results and reevaluate the precision of estimates:* The analysis of results can be calculated readily in the record sheet. It is possible to find out the percentage of effective time compared with that of delays, analyze the reasons for ineffective time, and ascertain the percentage of time spent by a worker, groups of workers, or a machine in a given work element. As the study progresses, confidence intervals can be calculated periodically. Results of this reevaluation may permit terminating the study early or may suggest that it be extended beyond the original planned length.
10. *Write up and file results:* The form of the report will depend on the objectives of the study. The report should contain the estimates of the proportion of time allocated to the various categories, the precision of these estimates, and the computed standard times. Reports should be retained for future reference in similar or follow-up studies.

5.3. Work Sampling Study Plans

Detailed planning must be done prior to taking actual observations. The plans start with a preliminary estimate of the activities on which information is sought. This estimate can frequently be made from historical data. However, if a reasonable estimate cannot be made, preliminary sampling of the work area for two or three days can be used as basis for these estimates.

Once preliminary estimates have been made, the analyst can determine the desired accuracy of the results expressed as a tolerance, or limit error, within a stated confidence level. Next, the analyst must estimate the number of observations to be made and determine the frequency of observations. Finally, the work sampling form on which to tabulate the data is designed, as well as the control charts used in conjunction with the study.

5.3.1. Determining the Observations Needed

To determine the number of observations needed or sample size, the analyst must set the desired accuracy of the results. The more observations made, the more valid the final results will be. Three thousand observations give considerably more reliable results than 300. However, because of the cost of obtaining so many observations and the marginal improvement in accuracy, 300 observations may be ample.

The size of the sample is therefore important, and we can express our confidence in whether or not the sample is representative by using a desired confidence level. It is customary to estimate precision by assuming a normal distribution for each of the p's and claiming $(1 - \alpha)$ confidence interval that

$$p - [Z_{\alpha/2}\,(\sigma_p)] < p' < p + [Z_{\alpha/2}\,(\sigma_p)]$$

The quantity σ_p is the standard deviation of the estimate p. The true variance of p depends on the manner in which observations are made. Hence,

$$p - Z_{\alpha/2}\sqrt{\frac{pq}{n}} < p' < p + Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

The quantity $Z_{\alpha/2}$ is a constant associated with the normal distribution such that the probability of any one normally distributed quantity being more than $Z_{\alpha/2}$ from its true value is less than α. The 95% confidence obtainable by using 1.96 and the 90% confidence obtainable by using 1.645 are the most frequently used in practice (see Table 17).

For example, a preliminary study estimates that a machine is in set up 25 times out of 100 observations. The interval with approximately a 95% of containing p' is then

TABLE 17 Confidence Levels and $Z_{\alpha/2}$ Values

Desired Confidence Level $1 - \alpha$	$Z_{\alpha/2}$
0.90	1.645
0.95	1.96
0.98	2.326
0.99	2.576

$$0.25 - 1.96\sqrt{\frac{(0.25)(0.75)}{100}} < p' < 0.25 + 1.96\sqrt{\frac{(0.25)(0.75)}{100}}$$

$$0.165 < p' < 0.335$$

There are two methods of determining the sample size: the statistical method and the nomogram method:

5.3.1.1. Statistical Method The formula used for this method is,

$$n = \frac{p(1 - p)(Z_{\alpha/2})^2}{e^2}$$

where p = percentage occurrence of the element being sought, expressed as a decimal
n = total number of random observations upon which p is based
e = precision (accuracy desired)
$Z_{\alpha/2}$ = constant associated with the normal distribution

Before we can use this formula, however, we need to have at least an idea of the value of p. The first step is therefore to carry out a number of random observations in the working area. Let us assume that some 100 random observations were carried out as a preliminary study and that these showed the machine to be operational 90% of the time. We thus use this approximate value as the estimate of p.

Choosing a confidence level of 95% with a 2% precision, the sample size is computed as:

$$n = \frac{(0.9)(0.1)(1.96)^2}{(0.02)^2} = 865 \text{ observations}$$

If the analyst does not have the time or capability to collect 865 observations but can only collect 500 data points, the above equation can be inverted to solve for the resulting error limit:

$$e = Z_{\alpha/2}\sqrt{\frac{p(1 - p)}{n}}$$

$$e = 1.96\sqrt{\frac{0.9(0.1)}{500}} = 0.0263$$

With 500 observations, the accuracy of the study is reduced to ±2.63%. Thus, there is a direct tradeoff between the error or accuracy of the study and the number of observations collected.

Alternatively, some practitioners find it easier to specify the *relative* precision of estimate desired in a study. Instead of setting $p \pm e$, a relative precision range of $p \pm Rp$. The proportion R is termed the relative precision in contrast to the precision e, which, for clarity, can be:

$$n_R = \frac{(1 - p)(Z_{\alpha/2})^2}{pR^2}$$

referred to as the absolute precision. Since $R = e/p$, substituting in the sample size formula yields:

$$R = Z_{\alpha/2}\sqrt{\frac{(1-p)}{pn}}$$

If a precision limit within 10% of e is desired (R = 0.10), the required sample size for the above example is:

$$n_R = \frac{(0.1)(1.96)^2}{(0.9)(0.10)^2} = 43 \text{ observations}$$

The determination of the sample size can be improved as data are gathered, and p can be estimated from the initial observations taken. After a study has been take for I days, the new sample size is determined by,

$$N(I) = \frac{p(1-p)(Z_{\alpha/2})^2}{e^2}$$

But n observations have already been obtained. Hence the additional number of observations necessary to produce the absolute precision e will be:

$$N(I) - n = n\left\{\left(\frac{p(1-p)}{n}\right)\frac{Z_{\alpha/2}{}^2}{e^2} - 1\right\} = n\left(V\frac{Z_{\alpha/2}{}^2}{e^2} - 1\right)$$

If observations are taken at an average rate of n/I observations per day, the additional days required to produce the absolute precision e will be equal to:

$$\text{Additional days} = I\left(V\frac{(Z_{\alpha/2})^2}{e^2} - 1\right)$$

For example, 5 days of data have produce 125 observations of a particular category during 500 total worker observations so that $p = 125/500 = 0.25$. If a precision of 0.02 with 95% confidence is desired, a total of 1801 (i.e., $n = [(0.25)(0.75)(1.96)^2]/(0.02)^2)$ observations will be needed. Since 500 observations have already been taken, an additional 1301 observations will still be needed. Because it has taken 5 days to get the first 500 observations (or 100 observations/day), it is reasonable to expect that the study will require an additional 1301/100 = 13.01 days.

The above equation yields this figure immediately as:

$$5\left(\frac{(0.25)(0.75)(1.96)^2}{500(0.02)^2} - 1\right) = 13.01$$

Software for determining the observations required for work sampling study is readily available today. These programs perform all the statistical calculations required to determine sample sizes and confidence intervals. For example, calculations for 90%, 95%, and 99% confidence intervals for a sample can be calculated. They can also provide the number of samples necessary to achieve the desired confidence for a specified degree of accuracy.

5.3.1.2. Nomogram Method An easier way to determine sample size is to read off the number of observations needed directly from a nomogram developed by Moskowitz (1965) such as the one shown in Figure 10. To use the nomogram, we draw a line from the *element percentage occurrence* ordinate p (e.g., $p = 20\%$) to intercept the *precision interval* (accuracy required) ordinate D (e.g., $D = 0.04$ or 4%) and extend it until it meets the vertical line in the center of the exhibit. This intersection identifies the pivot point. Connecting that pivot point by a new line with the desired confidence level on the short scale, for example 95%, and extending this new line to the right-most scale yields the necessary sample size, which for this example is $n = 384$ observations. By a similar procedure, the nomograph can also be used to solve for precision or for confidence level, given the other three quantities.

5.3.2. Determining Observation Frequency

Two important principles are important in selecting observation times: randomization and stratification. Randomization is used to reduce the possible bias of periodic observations times introduced by worker anticipation of observation times. Observations must be taken at random times within a day and the total number of observations stratified by the number of days of the study period. Stratification is undertaken to reduce the variance of the estimates. It is also applied to obtain representativeness

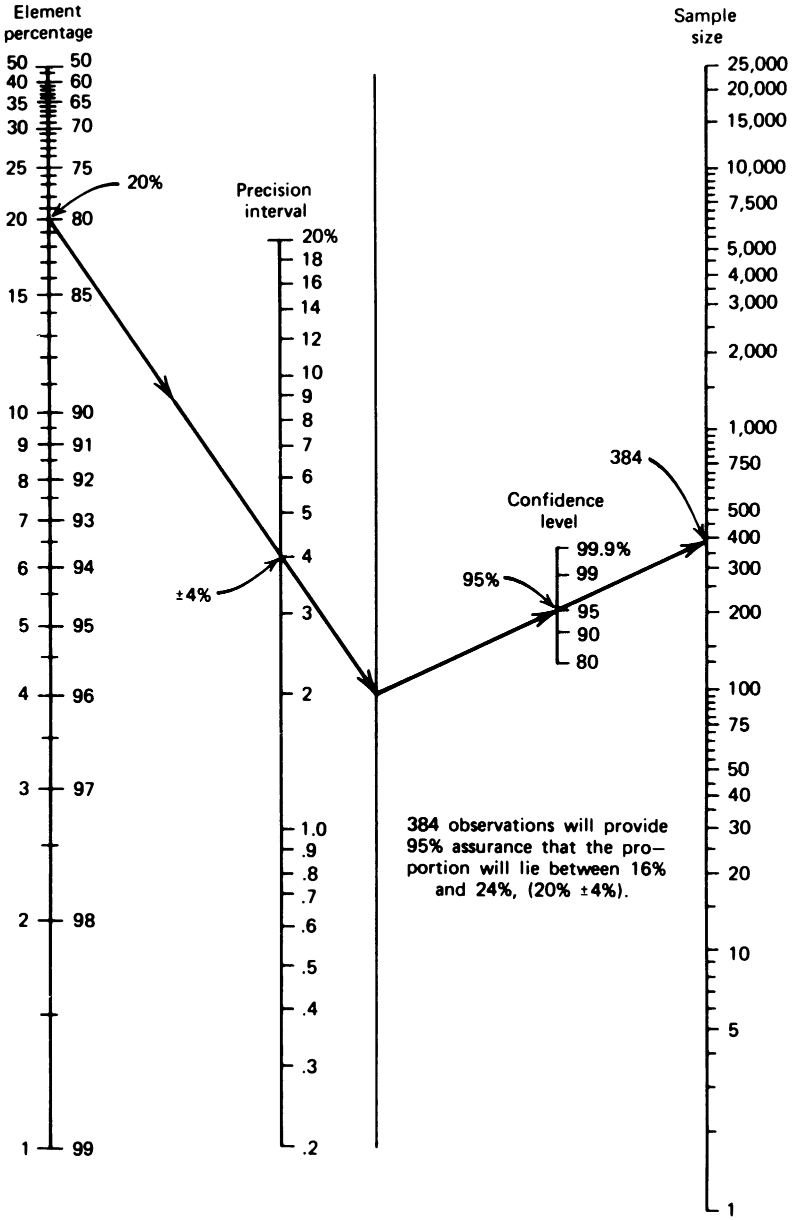


Figure 10 Nomograph. (Reprinted by permission from Moskowitz 1965)

of the sample based on various attributes. Spreading out the observations will give better estimates and a more representative result. It also allows the use of the data in multiple ways, such as identifying seasonal activities and reasons for unavoidable delays.

The frequency of the observations depends, for the most part, on the sample size and the time available to develop the data. If 3600 observations need to be completed in 20 calendar days, the analyst will need to obtain approximately 3600/20 = 180 observations per day. Other factors that determine observation frequency include:

- Nature of work being observed
- Length of work cycle: short work (small assembly) or long work (major machining process)
- Pattern of work: repetitive or irregular; peak and slump season; cyclical work
- Number of analysts (or other constraining resources) available

Determining a target length for a work sampling study requires rational judgment on the part of the analyst. The study should be long enough to include normal fluctuations in production. The longer the overall study, the better the chance of observing average conditions. However, extending the study over an infinitely long time span may prove useless and costly. Usually, work sampling studies are made over a period ranging from two to four weeks.

After determining the number of observations per day, the analyst must select the actual time needed to record the observations. To obtain a representative sample, observations should be taken at all times of the day. There are many ways of randomizing the occurrence of the observations. One method is to use tables of random numbers published in many handbooks and textbooks or the simple random number generators programmed in many hand-held scientific calculators.

Any type of watch can be used to identify the predetermined times of observations. Computers can also be used to determine the schedule of daily observations. Another alternative to help analysts decide when to make daily observations is a random reminder. A pocket-sized instrument beeps at random times, letting analysts know when to make the next observation. Electronic clipboards with optional work sampling software can also be extremely helpful. For example, the OS-3 Plus recorder (see Section 2.2) is available with a work sampling program that permits the scheduling of random observations and the performance rating of individual readings. It also produces a summary statistics and formatted printed reports.

In one approach, the analyst may select numbers from the statistical table of random numbers, ranging from 1 to 480. If each number carries a value, in minutes, the numbers selected can then set the time, in minutes, from the beginning of the day to the time for taking the observations. For example, the random numbers 25 and 152 would mean that the analyst should make a series of observations 25 minutes and 152 minutes after the beginning of the shift. If the day begins at 8 a.m., then at 8:25 a.m., an inspection of the work area would begin, followed by an observation made at 10:32 a.m.

Another approach considers four adjacent digits in the random number table. The first digit is the day identifier, with numbers 1 to 5 identifying the workday Monday through Friday. The second digit is the hour identifier, with numbers 0 to 8 added to the starting time of work (e.g., 7:00 a.m.). The third and fourth digits are the minutes identifiers, with numbers between 0 and 60 acceptable.

Computers can also be used to determine the schedule of daily observations. For example, work sampling programs for the DataMyte collecter described in Section 2.2 can print random time schedules.

Moder and Kahn (1980) propose a decision process for selecting a sampling procedure for each day of a work sampling study (see Figure 11). They assume that observations will be stratified by days under all plans. They define four sampling strategies: systematic random sampling (SyRS), stratified continuous random sampling (StCRS), stratified noncontinuous random sampling (StNCRS), and restricted random sampling (RRS).

Systematic random sampling is practiced if, when r observations per worker are to be taken during a day of t min duration, a time is selected at random during the first t/r min of the day. Subsequent observations are made at intervals of exactly t/r min.

In stratified, continuous random sampling, the use of the word "continuous" implies that the observer is assigned 100% to work sampling and thus works "continuously" at work sampling. The observations are still snap observations, and no continuous time study is implied. Randomization is achieved by (1) random selection of the starting point of each observation round (numbers can be assigned to each station and a number selected at random), (2) random selection of several different routes (assign numbers to various feasible routes), and (3) coin flip determination of the direction of route will be taken (clockwise or counterclockwise). The degree of randomization needed to deter worker bias can be decided by those who are familiar with the job process being studied.

Although Moder and Kahn (1980) use the notation "stratified, noncontinuous random sampling" for the procedure when a single random time is selected during each t/r min interval during the day, it could also apply to the practice of selecting 3 random times in each hour when 24 times are desired during an 8 hr day. If observation rounds are being made, routes can be randomized as with StCRS.

In restricted random sampling, "random sampling" implies selection of r times at random from the total time period of duration t. When an observer is making a full observation round and two times are selected to close together to allow completion of the earlier round, the restriction implies that the second time *selected* is rejected and replaced with another random time. Since this procedure is adopted at the beginning, no bias is induced by the method, and variance may be slightly decreased.

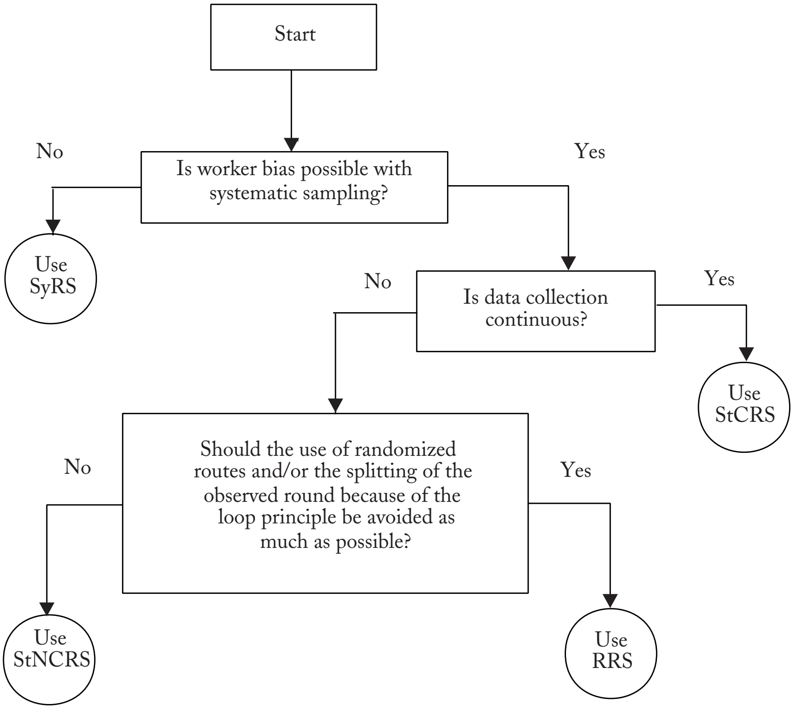

Start
Is worker bias possible with systematic sampling?
No
Yes
Use SyRS
Is data collection continuous?
No
Yes
Use StCRS
Should the use of randomized routes and/or the splitting of the observed round because of the loop principle be avoided as much as possible?
No
Yes
Use StNCRS
Use RRS

Figure 11 Decision Tree for the Selection of a Sampling Procedure for Work Sampling Studies. (From Pape 1992)

5.3.2.1. Loop Principle When observation rounds of approximately 5 min or more are being made, some bias will be incurred if the loop principle is not recognized. A time selected by any of the preceding procedures is customarily taken to be the starting point of an observation round to be made during a specific period (a day for RRS and SyRS, t/r min or hr for StNCRS). However, to keep the round within the specific period, that round cannot be observed during the first few minutes of the period. Thus, biases can be induced when end and beginning activity differs from midperiod activity.

To reduce this source of bias, consider the last observation round of the period as extending into the first few minutes of the same day. That is, if a round takes 10 min, and if the last observation time of the day is 3 min before closing, the last 70% of the round is conducted during the first 7 min of the workday and before the "first" round of the day begins.

If a work sampling study is being conducted of only a few workers or machines, the loop principle can usually be ignored.

5.3.3. Observations and Data Recording

The methods of collecting work sampling data vary with the size and the purpose of the study. Work sampling data can be collected by one of three methods: self-observation, observation by a trained analyst or the supervisor, and use of video camera equipment.

5.3.3.1. Self-observation If the purpose of a work sampling study can be achieved simply by having the workers who are being studied become aware of their own work, self-observation can be effective. The time for an observation can be signaled by a auditory signal (buzzer or beeper). The operator records on a simple ledger what he or she was doing opposite the number of the observation at the time the signal was heard. Although operators are expected to know exactly the work categories of their jobs, estimates produced by this method are subject to a great deal of bias. Instead, this technique is used more effectively among semiprofessional personnel, who are more conscientious about the effectiveness of their time usage.

5.3.3.2. Observation by a Trained Analyst or Supervisor In making the observations, the analyst must anticipate the expected recording. The observation should be made at the same point a given distance from the work area. If the operator is observed to be idle, the analyst must determine the reason for the idleness.

Recording consists of simply making a stroke in the form of the appropriate work element at the proper and predetermined time. No stopwatches are used. The analyst should avoid recording what has just happened or what will be happening. Rather, the analyst should record what is actually happening at the exact moment of the observation.

5.3.3.3. Observations from a Video Camera Unbiased work sampling studies involving only people can be performed using a video camera (Niebel and Freivalds 1999). The arrival of an observer at the work area may influence the activity of the operator. The operator becomes productively engaged as soon as the analyst is seen approaching the work area. However, it is essential that the analyst reviewing the film be familiar with the process being studied. Despite the advantage of replay, it is possible that in some instances some observations can be ambiguous or omitted.

5.3.4. Using Control Charts

Control chart techniques used in statistical quality control activities can be applied to work sampling studies. In particular, the p-chart can be used by time study analysts because work sampling studies deal with percentages or proportions (see Chapter 69 for more information on control charts). A control chart similar to Figure 12 (Niebel and Freivalds 1999) can be constructed by plotting the p' values for each day of observation.

In statistical process control, p-charts are used to indicate whether or not the process is control. In general, analysts use the $\pm 3\sigma$ limits as control limits on the p-chart. In a similar manner, time study analysts can use this chart to indicate whether work sampling data fall within the $\pm 3\alpha$ limits of the estimate p. In other words, if a sample has a value that falls outside the $\pm 3\sigma$ limits, the sample is said to be out of control.

As in statistical control work, points that fall outside the upper or lower control limits may be of statistical significance. For example, two successive out-of-control points may indicate that the population being observed has changed. That is, conditions in the work environment may have changed or process/work improvement been realized. Control charts can show the progressive improvement work, thus triggering the need to change the standards applied.

5.4. Establishing Standard Times

The computational requirements of a work sampling study are very simple. First, the observed time (OT) for a given work category or element is calculated from the working time divided by the number of units produced during that time:

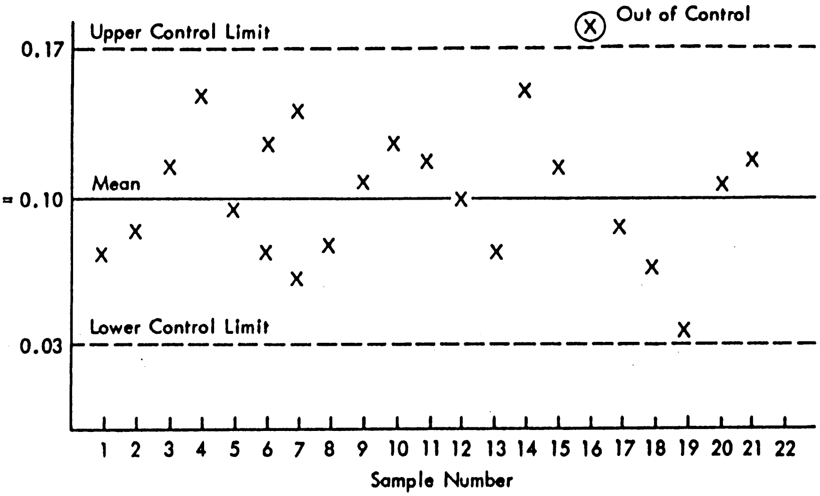


Figure 12 Sample Control Chart. (From B. Niebel and A. Freivalds, Methods, Standards, and Work Design, 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.)

$$\text{OT} = \frac{Tn_i}{Pn}$$

where T = total time
n_i = number of occurrences for element I
n = total number of observations
P = total production for period studied

The normal time (NT) is obtained by multiplying the observed time by the average rating:

$$\text{NT} = (\text{OT})\,\frac{R}{100}$$

where R = average rating = $\sum \frac{R}{n_i}$

Finally, the standard time (ST) is found by adding allowances to the normal time:

$$\text{ST} = \text{NT}\,(1 + \text{allowance})$$

To illustrate, a drill press operator was observed to have drilled 350 units during an 8 hr workshift. The work sampling study recorded that this represented 92% of her total working day (working + idle) with an average rating of 110%. Applying an allowance of 15%, the standard time for the drilling process is obtained by:

$$\text{ST} = \text{OT}\left(\frac{R}{100}\right)(1 + \text{allowance})$$

$$\text{ST} = \left[\frac{480}{350} \times 0.92\right]\left[\frac{110}{100}\right](1.15) = 1.596 \text{ min}$$

5.5. Computerized Work Sampling

With the aid of a computer, the total cost of work sampling studies can be reduced by approximately 35% because of the high percentage of clerical effort relative to actual observation time (Niebel and Freivalds 1999). The majority of the clerical effort involved in summarizing work sampling data is significantly reduced. Computer software can be applied to determine the sample size and sampling scheme (number of daily observations required, number of trips and the time of day for each trip, etc.), calculate percentages and accuracy, plot data on control charts, and summarize daily and cumulative results.

One such computer software is the mechanized activity sampling technique (MAST), developed to automate the clerical work and mathematical calculations involved in work sampling studies. MAST also assists the analyst in computing the element, developing performance ratings, checking statistical accuracies, preparing and maintaining control charts, and extrapolating the data into equivalent staffing needs and/or machines and annual costs.

Users of MAST claim the following benefits of the automated system (Niebel and Freivalds 1999):

1. Reduction of clerical routines
2. Results of the study are achieved more rapidly and presented in a professional format
3. Reduction in the cost of conducting the work sampling study
4. Improved accuracy in computations
5. Fewer errors committed by analysts
6. Incentive to make greater use of the work sampling technique

6. MEASUREMENT OF INDIRECT LABOR OPERATIONS

Most of the advances in work measurement have dealt mainly in the measurement of direct labor. The emphasis during the first half of the 20th century was on controlling costs due to direct labor. Because the indirect components of cost were highly variable and difficult to trace, the quantitative skills needed to analyze many problems associated with the indirect components were not available.

Employees classified under indirect labor include shipping and receiving, trucking, inventory, inspection, material handling, toolroom, and janitorial and maintenance. Expense labor positions are found in office operations such as clerical, accounting, sales, management, and engineering.

The rapid growth in the numbers of office workers, maintenance workers, and other indirect and expense employees is due to several factors. First, the increased mechanization of industry and the complete automation of many processes, including the use of robots, have decreased the need for craftsmen and operators. This trend toward increased mechanization has resulted in a huge demand for electronics specialists, electricians, technicians, and other service personnel. The design of complicated machines and controls has also resulted in greater demand for engineers, designers and draftsmen.

Second, the tremendous increase in paperwork brought about by legislation is responsible to a large extent for an increasing need for clerical personnel.

Third, office and maintenance work has not been subjected to the methods study and technical advances that have been applied so effectively to direct labor in industrial processes. With a large share of most payrolls earmarked for indirect and expense labor, progressive management is beginning to realize the opportunities for the application of methods and standards in this area.

6.1. Indirect and Expense Work Standards

The tools used for establishing time standards for indirect and expense work are the same as those used for direct work: time study, PTS systems, and standard data and work sampling. However, because of the high degree of variability characteristic of most indirect and expense work, time would not permit using stopwatch time studies for each and every standard developed. Standards will usually be established using a combination of techniques depending on the nature of the work. Other methods such as queuing theory and Monte Carlo simulation (see Chapter 100 for more on queuing models) are frequently used to determine delays due to waiting time in service facilities such as the stockroom or tool crib.

As for other work, methods analysis should precede work measurement in all indirect work operations. After completion of a thorough methods program, development of standards can be performed. Once standards have been developed for most of the common elements used in indirect work, time standards for specific tasks can be calculated quickly and economically.

All indirect and expense work is considered a combination of four divisions: (1) direct work, (2) transportation, (3) indirect work, and (4) unnecessary work and delays.

Direct work is that segment of the operation that discernibly advances the progress of work. For example, in operating a fax machine, the direct work elements may include insert document into paper feed, locate and dial telephone number, press send button. Such direct work can be measured using conventional techniques such as stopwatch time study, standard data, and fundamental motion data. Fundamental motion data systems such as Work-Factor, MTM-2, and MOST have been widely applied in establishing standards.

Transportation is the work performed by movements during the course of the job or from one task to another. Typical transportation elements include walk, carry load, ride elevator, push cart, and ride on motor truck. Transportation elements can be measured using standard data and work sampling.

The indirect work portion of indirect or expense labor is activities that cannot be evaluated by physical evidence in the completed job. It can be further classified into three categories: (1) tooling, (2) material, and (3) planning.

Tooling work elements include the acquisition, disposition, and maintenance of all tools needed to perform an operation. Typical elements under this category include getting and checking tools and equipment, cleaning tools, repairing and calibrating tools, and returning tools to the tool crib. Tooling elements can be measured using work sampling or standard data.

Material work elements involve acquiring and checking the material used in an operation and disposing of scrap. Examples of material work elements are getting materials, making minor repairs to materials, and picking up and disposing of scrap. Like tooling elements, material work elements can be measured. Use historical data to determine their frequency.

The planning elements represent the most difficult area in which to establish standards. Consulting with the supervisor, planning work program, inspecting, checking, and testing are common examples. In this case, work sampling is the most practical technique to provide a basis for determining the time required to perform the planning elements.

Unnecessary work and delay encountered in indirect and expense work are mainly due to queues. Workers need to stand in line at the tool crib, the stockroom, or the fax machine, photocopy machine, or some other equipment. Through the application of queuing theory, analysts may be able to determine the waiting time incurred in such activities as well as determine the optimum number of service facilities to improve service quality.

Where maintenance and other indirect operations are numerous and diversified, efforts have been made to reduce the number of time standards for indirect operations through universal indirect labor standards (UILS). The principle behind UILS is the assignment of the major proportion of indirect operations to appropriate groups (Niebel and Freivalds 1999). Each group has its own standard, which is the average time for all indirect operations assigned to the group. For example, the following indirect operations replacing defective part, replacing limit switch, and repairing door may represent group A. The standard time for any indirect operation performed in group A may be set at 45 minutes.

This time represents the mean ($\bar{x}$) of all jobs within the group and the dispersion of the jobs within the group for $\pm 2\sigma$ is some predetermined percentage of $\bar{x}$.

Three principal steps in introducing a universal indirect labor system, called time slotting, are:

1. Determine the number of standards (groups or slots) to do a satisfactory job. (20 slots should be used when the range is up to 40 hours).
2. Determine the numerical standard representative of each group of operations contained in each slot.
3. Assigning the standard to the appropriate slot of indirect labor work as it occurs.

The first step is to determine good benchmark standards, based on measurements of an adequate sample of the indirect labor for which the UILS system is being developed. A relatively large number of standards (200 or more) that is representative of the entire population of indirect needs to be established. Competent analysts can develop these measured benchmarks using work measurement techniques.

After the benchmarks are established, they are then arranged in decreasing order (from shortest to longest time). If there are 20 slots, and if a uniform distribution is used, the time standard for the first slot (UILS one) is computed by calculating the mean of the first 10 benchmark standards. Similarly, the value of UILS two is taken by the mean of benchmark standards 11–20. Succeeding UILS slots will be calculated on the successive 10 benchmark standards. Hence, the UILS 20 would be equal to the mean of benchmark standards 191 through 200. Engineers have used this procedure extensively in the development of UILS.

Another approach to designing a UILS system is cluster analysis (Knott 1992). Here, given a distribution of job times, the time boundaries and slot times can be determined to give a priori criterion. This might be, for example, absolute percent error calculated over a time period of one week.

$$\text{Absolute percent error} = \frac{\text{actual standard time} - \text{universal standard time}}{\text{actual standard time}} \times 100$$

More reliable UILS result from using the normal distribution rather than the uniform distribution. For 20 slots, the 200 standards would not be assigned as 10 per slot. Rather, the standard normal variable would be divided into 20 equal intervals. For example, the standard normal variable may have a truncated range of:

$$-3.0 \le z \le 3.0$$

which accounts for 99.87% of the area under the normal curve. The range of each interval would be 0.3. The benchmark standards used in the compilation of the mean of each of the 20 slots would equal:

$$\frac{P(z \in \text{interval})(200)}{0.9987}$$

Slot number 1 and 20 would have:

$$\frac{P(-3.0 \le z \le -2.7)(200)}{0.9987} = \frac{P(2.7 \le z \le 3.0)(200)}{0.9987}$$

Investigations have shown that the distribution of job times in practice is not normal but has a “positive skew” (Knott 1992). A typical distribution of job times is shown in Figure 13. A simple slotting scale has been constructed into this distribution.

To establish standards on indirect and expense work, Table 18 is presented as a guide for choosing the appropriate method.

6.2. Advantages of Indirect Work Standards

Applying standards on indirect work offer distinct advantages to both the employer and the employee. Some of these advantages are:

1. Installation of standards leads to many operating improvements.
2. Establishment of standards results in better performance.
3. Indirect labor costs are related to the workload, regardless of fluctuations in the overall workload.

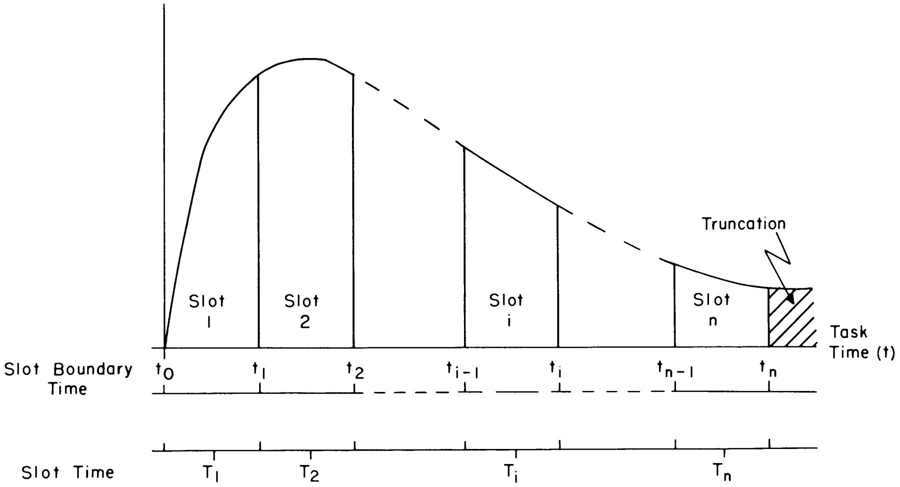


Figure 13 Elements of a Time-Slotting Scale. (From Knott 1992)

4. Labor loads can be budgeted.
5. The efficiency of various indirect labor departments can be determined.
6. The costs of such items as specific repairs, reports, and documents are allocated.
7. System improvements can be evaluated prior to installation.
8. The establishment of incentive wage-payment plans on indirect work is allowed.
9. Accurate planning and scheduling of indirect labor leads to timely performance.
10. Employees require less supervision with the establishment of standard measures of performance.

TABLE 18 Guide to Establishing Indirect Labor and Expense Standard

Indirect and Expense-Type Work	Recommended Method of Establishing Standards
Routine maintenance. Work standards 0.5 to 3 hr	Standard data, MTM-2, MTM-3, Work-Factor, MOST, macromotion analyses
Complicated maintenance, standards 3–40 hr	Slotting based on universal indirect labor (UIL) standards
Shipping and receiving	Standard data, MTM-2, MTM-3, Work-Factor, MOST, macromotion analyses
Tool room	Slotting based on UIL standards
Inspection	Standard data, MTM-2, Work-Factor, MOST, macromotion analyses
Tool design	Slotting based on UIL standards
Buying	Standards based on historical records, analysis, and work sampling
Accounting	Standards based on historical records, analysis, and work sampling
Plant engineering	Standards based on historical records, analysis, and work sampling
Clerical	Standard data, MTM-2, Work-Factor, MOST, micro- and macromotion analyses
Janitorial	Standard data, slotting based on UIL standards
General management	Standards based on historical records, analysis, and work sampling

From B. Niebel and A. Freivalds, *Methods, Standards, and Work Design,* 10th Ed. © 1999 McGraw-Hill Companies, Inc. Reprinted by permission.

Since indirect labor operations are difficult to standardize by virtue of their nonrepetitive characteristic, they are infrequently subjected to methods analysis. Consequently, this area usually offers a greater potential for cost reduction and increasing profitability through methods and time study than shop-floor operations. Methods improvement, along with employee training, makes it possible and practical to establish standards on indirect labor operations.

7. SELECTED SOFTWARE

ErgoMOST (1997), H.B. Maynard and Co., Pittsburgh

MAST (1998), Applied Research Laboratories Division of Bausch & Lomb, Inc., Rochester, NY

MOST (1997), H.B. Maynard and Co., Pittsburgh

MTM Link (1998), MTM Association, Des Plaines, IL

REFERENCES

Brisley, C. L. (1952), "How You Can Put Work Sampling to Work," *Factory Management and Maintenance,* Vol. 110, No. 7, pp. 84–89.

Fisher, R. A., and Yates, F. (1963), *Statistical Tables for Biological, Agricultural, and Medical Research,* 6th Ed., Oliver & Boyd, Edinburgh.

International Labour Office (1979), *Introduction to Work Study,* 4th Ed., International Labour Office, Geneva.

Knott, K. (1992), "Indirect Operations: Measurement and Control," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1723–1754.

Konz, S. (1990), *Work Design: Industrial Ergonomics,* 3rd Ed., Publishing Horizons, Scottsdale, AZ, p. 474.

Moder, J. J., and Kahn, H. D. (1980), "Selection of Work Sampling Observation Times: Part 1—Stratified Sampling," *AIIE Transactions,* Vol. 12, No. 1, pp. 23–31.

Morrow, R. L. (1941), "Ratio Delay Study," *Mechanical Engineering,* Vol. 63, No. 4, pp. 302–303.

Morrow, R. L. (1946), *Time Study and Motion Economy,* Ronald Press, New York.

Moskowitz, A. D. (1965), "A Monograph for Work Sampling," *Work Study and Management Services,* Vol. 9, pp. 349–350.

Mundel, M. E., and Danner, D. L. (1994), *Motion and Time Study: Improving Productivity,* 7th Ed., Prentice Hall, Englewood Cliffs, NJ.

Niebel, B. (1992), "Time Study," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1599–1638.

Niebel, B., and Freivalds, A. (1999), *Methods, Standards and Work Design,* 10th Ed., McGraw-Hill, New York.

Pape, E. S. (1992), "Work Sampling," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1699–1721.

Sellie, C. N. (1992), "Predetermined Motion–Time Systems and the Development and Use of Standard Data," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1639–1698.

Tippett, L. H. D. (1935), "A Snap-Reading Method of Making Time Studies of Machines and Operatives in Factory Surveys," *Journal of the Textile Institute Transactions I,* Vol. 26, February, pp. 51–55.

Waddell, H. L. (1952), "Work Sampling—A New Tool to Help Cut Costs, Boost Productivity, Make Decisions," *Factory Management and Maintenance,* Vol. 110, No. 7, p. 83.

Zandin, K. B. (1980), *MOST Work Measurement Systems,* Marcel Dekker, New York.

ADDITIONAL READING

Panico, J. A., "Work Standards: Establishment, Documentation, Usage and Maintenance," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, 1992, pp. 1549–1574.

Quick, J. H., Duncan, J. H., and Malcolm, J. A., *Work-Factor Time Standards,* McGraw-Hill, New York, 1962.

IV.D

Systems and Facilities Design

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 55

Facilities Size, Location, and Layout

JAMES A. TOMPKINS
Tompkins Associates

1. **INTRODUCTION** **1465**
2. **PROCEDURE—THE MACRO ANALYSIS (LSMP)** **1467**
    - 2.1. Supply Chain Needs 1467
    - 2.2. Customer Satisfaction Standards 1468
    - 2.3. The Strategic Master Plan and Establishing a Baseline 1469
    - 2.4. Network Analysis 1470
        - 2.4.1. Strategic Distribution Network Planning 1472
        - 2.4.2. Do Not Underestimate the Importance of Distribution 1475
    - 2.5. Team Selection 1475
        - 2.5.1. Real Estate Brokers 1476
        - 2.5.2. Government Agencies 1476
        - 2.5.3. Utilities 1476
        - 2.5.4. Consultants 1476
3. **PROCEDURE—THE MICRO ANALYSIS (SITE SELECTION)** **1476**
    - 3.1. Community Selection 1476
    - 3.2. A Site-Selection Checklist 1477
        - 3.2.1. Comparing Specific Sites Based on Organization-Specific Criteria 1486
        - 3.2.2. Evaluation of Addressable Locations 1486
    - 3.3. Environmental Factors 1489
    - 3.4. Free Trade Zones 1489
    - 3.5. Site Visitation 1490
    - 3.6. Finalizing the Process 1490
4. **MOVING FROM SITE SELECTION TO CONSTRUCTION** **1490**
    - 4.1. Methods of Project Delivery 1491
        - 4.1.1. Design-Bid-Build (Traditional Method) 1492
        - 4.1.2. Construction Management (CM Method) 1493
        - 4.1.3. Design–Build 1494
        - 4.1.4. Team Design/Construct 1495
    - 4.2. Selecting an Architect 1496
    - 4.3. Selecting a Contractor 1499
    - 4.4. Project Management 1499
    - 4.5. Summary Points for Construction 1501
5. **CONCLUSION** **1501**

ADDITIONAL READING **1501**

1. INTRODUCTION

Site selection for a new factory or distribution center is a complicated and arduous process and must be based on the strategic vision of the organization, the requirements of the supply chain, and the needs of the customer. Determined both quantitatively and qualitatively, proper site selection usually involves upper management, since the level of success of the new facility will have a major impact on the bottom line of the organization.

The purposes for site selection are relocation, expansion, and/or decentralization. Motivations for selecting a site can vary depending on the purpose (e.g., the urgency of the matter may eliminate some complex quantitative assessment or some site research), but we will treat the process in the aggregate, touching on generic steps taken to make a good decision.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

Many stakeholders will give different reasons for selecting a site for a new industrial facility. When the word gets out that your organization is in the market, several individuals and companies eager to help you find the right location will be in contact. The Chamber of Commerce, realtors, local and state governments, and developers will offer substantial incentive packages to locate in their location.

The objective of all site selections is to turn a property or an existing facility into a weapon of competitive advantage. As Figure 1 indicates, an organization must transition from a big-picture analysis of its strategies and mission (macro analysis), to an assessment of addressable locations (micro analysis), to the construction phase. Though construction can happen without the comprehensive analysis discussed in this chapter, competitive advantage is rare when the upfront evaluation work is not done completely.

Among the more common mistakes made by corporations when selecting a site are:

- *Proceeding with a site search without a plan for the new facility:* Do not start looking until you know what you are looking for. You must establish criteria for the site's requirements. This will tell you what size warehouse your organization needs, what its footprint should look like, what the column spacing should be, dock and road requirements, etc. From this information, the site selection process can be narrowed to those properties that will accommodate the footprint of the original building design, as well as any planned expansions. Do not try to fit a round peg into a square hole. Immediately eliminate sites that do not accommodate the facility as designed.
- *Allowing premature publicity:* There are advantages and disadvantages to keeping your search for a site confidential as long as possible. Announcing your plan to the public may result in free publicity, support from local communities, and a head start on employee recruiting. On the other hand, announcing your intentions may drive up land prices, open the door to an onslaught of people and organizations hoping to influence your decisions, and feed the rumor mill in a way that might be detrimental to the company.
- *Failure to use a good criteria checklist:* The only way to find a site that fits your specific needs is through the use of a detailed customized criteria checklist. If you are using a one-page document, you've left out a lot of important details. To be an effective tool, the checklist must be as comprehensive as you can make it, and should include both long-and short-term economical, community, and quality of life factors.
- *Failure to align project plans with future requirements and technological trends:* Do not build for today or tomorrow; build for 2, 5, or even 10 years from now. How will the new facility accommodate increased sales, need for crossdocking, and different product lines? These are questions to be answered before you move, rather than after occupying the building.
- *Failure to accurately estimate the true cost of doing business at each proposed location:* The lowest cost site may not be the most economical place to do business. Your comparison of sites must include a thorough and detailed analysis of projected production costs. Do not let yourself be surprised by the cost of services and utilities like water, electricity, waste disposal, local special taxes, or site security. These factors can have a dramatic long-term cost impact on the bottom line.
- *Giving consideration to intangibles at the wrong time:* The purpose of the macro analysis is to precisely determine how best to utilize the site for competitive advantage. Allowing personal preferences about location and style only serves to shift the emphasis of the process from strategic and organizational requirements.
- *Failure to use consultants to supplement staff skills:* Collecting data on possible sites, accurately projecting operating costs, and evaluating incentive packages is a time-consuming process that cannot be rushed. Few owners have the time or in-house resources to do this job correctly. Facilities-planning consultants can collect the data and provide an unbiased source of advice on each facet of the site-selection process. With a wide breadth of knowledge and accessibility to time-saving planning tools, consultants make for an efficient and effective means of moving the process forward.

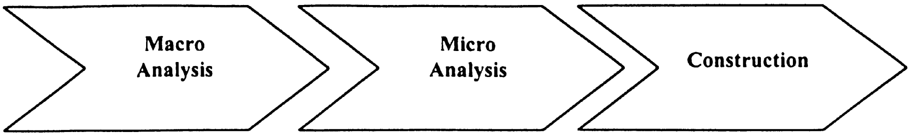


Figure 1 The Progression from Concept to Reality.

2. PROCEDURE—THE MACRO ANALYSIS (LSMP)

As Figure 2 illustrates, the macro analysis methodology is comprised of five parts:

- Organizations must envision themselves as an element of a supply chain that requires continuous improvements among all of its links, as opposed to the siloism of the past. Future competititve advantage hinges on examining the continuous improvement process as it relates to the aggregate supply chain.
- Customer satisfaction is the understanding that customers' perceptions and expectations rather than the organization's idea of what the customer wants, are key to profit maximization.
- Network analysis is the determination of the distribution plan that will provide the customer with the right goods in the right quantity at the right time and place while minimizing distribution costs through the correct balance of warehouses and transportation costs.
- The strategic master plan is the expression of future space, labor, and equipment requirements in order to analyze and justify alternative plans.
- Team selection is the establishment of a cross-functional group of people, internal and external to the organization, who bring their talents to the process.

2.1. Supply Chain Needs

The vision for the future of the site and for the aggregate supply chain must address the concepts of change and integration. Customer requirements, commerce structures, and market demands fluctuate faster than organizations can adapt to them. By understanding the impacts of change and integration and designing facilities around these concepts, organizations can improve chances for competitive advantage.

- *Total integration:* An ultimate customer focus where material and information flow will be designed into the system and the supply chain will be fully optimized.
- *Blurred boundaries:* Traditional customer/supplier and manufacturing/warehousing boundaries will be shifted to simplify, add value, and increase responsiveness.
- *Consolidation:* Efficient and effective transportation infrastructures will allow for high levels of customer satisfaction from fewer points along and throughout the supply chain.
- *Reliability:* Robust, redundant, and fault-tolerant systems will enable organizations to increase up time.
- *Maintainability:* A 24/7 schedule will mandate increased predictive maintenance and self-assessment.

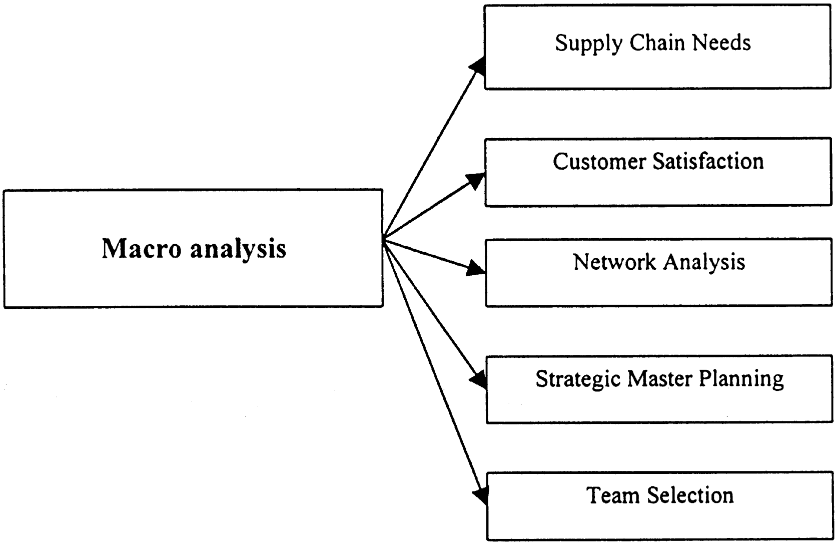


Figure 2 The Macro Analysis.

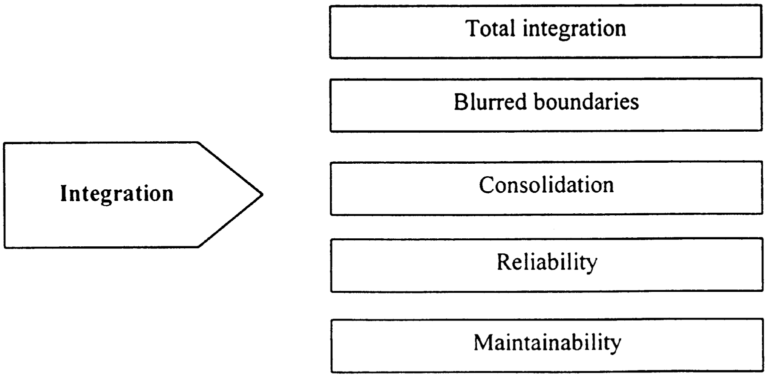

Integration
Total integration
Blurred boundaries
Consolidation
Reliability
Maintainability

Figure 3 The Challenges of Integration.

- *Flexibility:* Addressing change in product variety and rate of new product introductions through soft, friendly systems that handle products of different sizes and weights.
- *Modularity:* Accommodating the change in product volumes through systems that operate at a variety of rates.
- *Upgradeability:* Gracefully incorporating technology, process, and methodology changes into current systems.
- *Adaptability:* Responding to systems requirements on the fly.
- *Selective operability:* Operating in segments without degradation of the overall supply chain.
- *Supportability:* Maintaining the system while automated and nonautomated elements are brought online.

2.2. Customer Satisfaction Standards

Customer requirements are changing constantly, and the one-size-fits-all philosophy is obsolete. The growth of e-commerce has affected the size (decreasing) and frequency (increasing) of orders. In the end, customer perception of quality, which may or may not be an accurate depiction of organizational quality, drives the success of the business. The needs of the customer must be the focus.

Customer satisfaction is a means by which companies attempt to differentiate their products, keep customers loyal, improve profits, increase sales, and thus become the supplier of choice. Customer satisfaction is not based on what the supplier does; rather, it is based on what customers think the supplier does. Since customer satisfaction is an ongoing process of meeting and exceeding expectations, organizations must embrace continuous improvement and always look toward the consumer for affirmation.

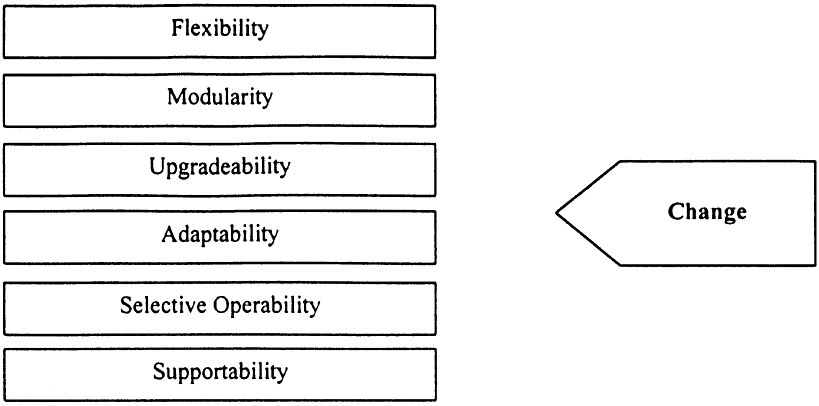

Flexibility
Modularity
Upgradeability
Adaptability
Selective Operability
Supportability
Change

Figure 4 The Challenges of Change.

How does customer satisfaction factor into the macro analysis stage of site selection? Beyond product quality and price, it impacts such processes and methodologies as:

- On-time delivery: From where is the customer order originating? Is the network configuration such that the customer is receiving a shipment as quickly as possible?
- High fill-rate percentage: Is the site large enough to accommodate a distribution/warehousing/manufacturing center that promptly produces, stores, and ships merchandise?
- Ability to adapt: Does the site have the potential to adapt to peak/seasonal shifts?

Customer satisfaction is achieved through an understanding of requirements and expectations. Organizations should ask themselves the following questions when defining their strategic master plan:

- What are the pipeline requirements and expectations?
- What is the organization's impact on achieving customer satisfaction in the supply chain?
- How do the customers perceive current operations, processes, and resources?
- Are there gaps between what is currently possible and what is currently being done for customer satisfaction?
- What service offerings can be created based on pipeline requirements?
- How will the organization measure, track, and improve performance?

2.3. The Strategic Master Plan and Establishing a Baseline

The strategic master plan (SMP) is a seven-step process from which a baseline, or foundation on which costs are analyzed, is developed. Without a baseline, site selection is based on ''crystal ball'' methodology, and crystal balls are known to give wrong answers from time to time.

Step 1: *Document the existing operation:*
- What are your operations costing you now?
- What are your current throughput and storage requirements?
- How much safety stocks does your operation keep now?
- What are the documented standard operating procedures? How do they compare with what actually happens in the warehouse?
- What policies are in place to govern the acquisition, the utilization, and the disposal of resources?
- Establish a baseline against which recommendations for improvements can be measured.

Step 2: *Determine facility requirements for a specified planning horizon:*
- What will your operations cost in the next three years? Five years?
- What will be your processing, material handling, and storage requirements in the next three years? Five years?
- Is there sufficient safety stock to meet the forecasted demand?
- What are your firm's objectives? What resources are needed to meet those objectives?
- How do your firm's objectives take into account the dynamic, global business world?
- What is the impact of customer ordering changes on order picking requirements?

Step 3: *Identify deficiencies in current operations:*
- Are your firm's customers requiring faster delivery, more variety, and more adaptability?
- Does your firm succeed in delivering on-time?
- Are existing facilities, methods, equipment, and/or labor the most efficient and effective means for handling capacity requirements and forecasts?

Step 4: *Identify alternative plans:*
- What alternate sites and operation methods can be considered? What are the quantitative and qualitative issues connected with these alternatives?
- How do these alternate sites and methods reduce or eliminate the deficiencies in the current operation?

Step 5: *Evaluate alternative plans:*

- What are the after-tax costs of the alternatives? The returns on investment?

Step 6: *Select and specify the recommended plan:*

- What are the space, equipment, personnel, and standard operating procedure requirements of the facility over the planning horizon?

Step 7: *Update the SMP:*

- How will changes in the business climate affect the first six steps of the SMP?

Note: The process of developing a SMP is continuous; completion of the first six steps is no cause for celebration, since your business continues to evolve. Recognize that a static plan is as good as no plan at all.

2.4. Network Analysis

Basically, a distribution network is a series of nodes and transportation links. Distribution networks can range from direct shipments from the source to demand points for job shop items to complex multisite networks. The design of a distribution network is dependent on factors such as the type of products, range and volume of products, geographic spread of service area, the level of service required, and the number and type of customers. However, since distribution is a dynamic environment, it is challenged by business issues such as the global marketplace, the level of government involvement, the environment, and energy. At the same time, the customer requirements of increased pace, variety, and adaptability while reducing costs must be understood. Of course, these issues impact the internal pressures of distribution requirements to centralize, utilize third parties, improve information systems, increase productivity, and more fully utilize people. Therefore, the only way to enhance distribution excellence is to pursue the integration of distribution by applying strategic planning.

Strategic planning is the process of deciding on objectives of the firm; changes in the objectives; resources to attain these objectives; and policies to govern the acquisition, use, and disposition of resources. The objective of strategic planning is to define the overall approach to stocking points, transportation, inventory management, customer service, and information systems and the way they relate in order to provide the maximum return on investment.

Strategic planning is an offensive tool designed to guard against a predictable change in requirements, the timing of which can be anticipated. Strategic planning is directed at forecasting future needs far enough in advance of the actual requirements to allow sufficient lead time to meet those needs efficiently. Granted, forecasting with a long planning horizon is a risky business and distribution plans based on such forecasts often prove unworkable. Nevertheless, the forecast is the best available information concerning the future, and it is foolish not to use that information to one's advantage. In fact, the only way to survive the rapidly changing distribution environment today is to have good strategic plans that address the future needs of distribution and the factors influencing distribution. These factors are:

- **Global marketplace:** The global marketplace is a distribution issue. In fact, in today's world there is no choice but to understand the global strategy implications on all distribution decisions. As shifts occur in the world's trading patterns, this changes the distribution requirements, alters the location and number of warehouses, increases pipeline inventories, and creates new transportation opportunities and problems.
- **Government involvement:** A global trend is for governments to deregulate many activities, most notably transportation. It is important that distribution professionals understand that just as government involvement has an impact on distribution, distribution leadership has an obligation to have an impact on government on behalf of distribution.
- **Reverse distribution:** An issue that is closely tied to the issue of government involvement is the issue of reverse distribution. Reverse distribution is the task of recovering packaging and shipping materials and backhauling them to a central collection point for recycling. Handling the mechanics of reverse distribution will require significant attention by distribution professionals. Not only will they need to understand a diverse set of state and federal laws, but they will have to deal with backhauls, handling the waste packaging in their warehouses, and the customer satisfaction issue of recycling.
- **Off-highway vehicles:** The EPA is pushing to regulate off-highway vehicles; this effort will include lift trucks and will further push warehouses in the direction of electric vehicles. The internal combustion lift trucks that will be sold in the future will need to meet much stricter emission standards, but in many applications these vehicles will be replaced by electric vehicles.

- **Energy:** Another issue, like the environment, that has not been a major topic of consideration by distribution professionals is the issue of energy. Nevertheless, the cost of energy is a major concern to transportation companies. In the United States, 60% of all energy consumption is for transportation. Although these costs tend to be buried in the overall cost of transportation, any significant shift in the cost of energy could have an impact on the costs of transportation and therefore on distribution. It is therefore important that, at least as a sensitivity issue, the issue of energy costs be viewed in making all distribution decisions.
- **Pace:** There exists an accelerating rate of change in all aspects of human endeavor: social, political, economic, technological, ecological and psychological. It is not surprising, then, that the reduction of lead times, shorter product lives, and increases in inventory turnover are resulting in significant increases in the pace of change in distribution. Distribution must be more responsive because the demands being placed upon it by customers.
- **Variety:** The variety of tasks to be handled by distribution will continue to increase. Special packaging, unitizing, pricing, labeling, kitting, and delivery requirements will become the norm. Distribution will be required to perform operations that traditionally have been viewed as manufacturing operations. Systems and procedures will be put in place to handle information consistent with the desires of the customers.
- **Flexibility:** The most important aspect of flexible distribution is *versatility*—in equipment, systems, and workers. The design, specification, and implementation of versatile equipment is required to achieve flexible distribution. Warehouse storage rack and material-handling equipment, as well as transportation equipment, should be selected with sufficient versatility to handle today's distribution requirements and, when justifiable, future requirements. Similarly, versatile systems have an impact on adapting customer labeling, automatic identification, communications, and documentation requirements. We never want to find ourselves saying to a customer, "I am sorry, our system does not allow us to accommodate your request." Lastly, we must have multiskilled personnel to achieve flexible distribution. Overly restrictive work rules, excessive job classifications and labor grades, and insufficient training have often resulted in a lack of flexibility in distribution. Multiskilling eliminates barriers between tasks, and workers can better understand the implications of their performance. Throughout distribution organizations, there is a need to destroy the traditional barriers between tasks.
- **Modularity:** The three most important aspects of modular distribution are modular distribution assets, modular work assignments, and time modularity. The issue of modular distribution assets has to do with the expansion and contraction of warehouse space and the increase or decrease of transportation equipment. Similarly, for transportation equipment, purchase and lease decisions, as well as contract terms, should be evaluated while considering both the long-term and short-term fluctuations in traffic. The challenge of modular work assignments has to do with the daily balance of work within a warehouse. Once people have been given multiple skills, it is important to be certain that people are assigned in such a way to allow for a continuous flow of materials through distribution. Lastly, to provide modular distribution is the issue of time modularity. Creativity in employee work schedules can have a significant impact on an operation's output. Many distribution operations have been significantly improved by adjusting work schedules so that there is a balance between the staff on hand and the tasks to be performed. Not addressing the issue of time modularity often results in distribution operations having very low productivity.
- **Price:** A prerequisite for the success of free enterprise is efficient, effective, and low-cost distribution. Although the cost of distribution is less than 10% of the price that the customer must pay, it is of the utmost importance to the customer that even this price be reduced. As a percentage of Gross National Product, distribution costs are down from a high of almost 15% to 11%, and as a percentage of Gross Domestic Product, they are down from a high of almost 18% to below 12%. Thus, it is very important that the cost of distribution be even further reduced.
- **Centralization:** There will be fewer large centralized warehouses in the future to replace the more numerous, smaller, decentralized warehouses of the past. There will be fewer managers and administrative people involved with distribution as integrated distribution is pursued and distribution staffs are centralized. Along with the centralization of warehouses and staffs will come the centralization of order entry, customer service, and data processing. The increased responsiveness of transportation at lower costs, the focus on the total cost of distribution, the realities of customer satisfaction, pace, variety, and adaptability—all are pointed toward the centralization theme. The trend toward centralized distribution will result in higher inventory turnover, which will in turn lead to new opportunities for automation and sophisticated information systems.

- **Third-party logistics:** Third-party logistics (3PL) is the utilization of an outside firm to perform some or all of the distribution functions presently performed internally. As companies better understand integrated distribution and as distribution leadership better understands the costs of distribution, there will be an increasing trend toward the outsourcing of portions of the distribution function.
- **Information systems:** Information technology is impacting everything from business to education to entertainment. It is not surprising, therefore, that information technology is having, and will continue to have, a major impact on distribution. It has become clear that all distribution documentation must be electronically transmitted and not mailed. All distribution paperwork needs to be scrutinized and eliminated whenever possible. It is important for distribution leadership to realize that paperwork means delays, errors, additional work, and therefore wasted time and money. Distribution information systems must be real-time and paperless and standardized throughout the distribution supply chain.
- **Productivity:** Accountability for performance in distribution must be increased. Distribution management must establish standards, identify opportunities for improvement, measure performance, and take action to ensure continuous distribution improvement. The entire distribution function must realize that productivity must be increased. The option of maintaining the status quo is totally unacceptable. The improvement of distribution productivity includes labor productivity, but it goes well beyond labor productivity.
- **People:** Customers drive the business of distribution, but performance depends upon distribution people. Customer satisfaction results from contact with distribution people, and so an important, ongoing distribution issue remains in people. In the past, distribution people were narrowly focused, having a specialized skill or technical strength. These distribution people do not conform to today's distribution needs. The people needed in distribution today must adopt a broader view of distribution, a more integrated understanding of distribution, a team-based, participative organization culture, and a total dedication to the supply chain and to customer satisfaction.

2.4.1. Strategic Distribution Network Planning

Distribution network planning is one of the main areas to which strategic planning is applied. A strategic distribution network plan is developed to meet a specific set of requirements over a given planning horizon. A good plan will determine the best network that will provide the customer with the right goods, in the right quantity, at the right place, at the right time, and minimize the total distribution cost. As the number of warehouses increases, delivery costs decrease and warehouse costs increase. The opposite is also true: as the number of warehouses decreases, the delivery cost increases. Therefore, to minimize total distribution cost it is important to find the best balance of these costs.

The objective of strategic distribution network planning is to determine a plan that indicates the most economical way to ship and receive product while maintaining or increasing customer satisfaction, or simply put, to maximize profits and optimize service. Strategic distribution network planning typically answers the following:

1. How many distribution centers should exist?
2. Where should the distribution center(s) be located?
3. How much inventory should be stocked at each distribution center?
4. What customers should be serviced by each distribution center?
5. How should the customers order from the distribution center?
6. How should the distribution centers order from vendors?
7. How frequently should shipments be made to each customer?
8. What should the service levels be?
9. What transportation methods should be utilized?

Planning a distribution network is a sequential process that continually needs updating. Some companies run into the pitfall of performing steps 3 through 6 before collecting and understanding the most important steps, which are 1 and 2. The answer to distribution network planning is only as good as the data put into the analysis. The steps taken in distribution design are listed below:

1. Document distribution network.
2. Identify delivery requirements.
3. Establish database.
4. Develop alternative networks.
5. Model annual operating costs.

6. Evaluate alternatives.
7. Specify the plan.

2.4.1.1. Document Distribution Network The steps for documenting the distribution network, identifying delivery requirements, and establishing the database can be done simultaneously. The main goal of these steps is to gain an understanding of the current system and define the requirements of the future system. In order to document the existing systems, information must be collected on the distribution centers and the transportation system. In gathering information on the distribution centers, it is critical to collect from all existing sites considered, since the study could result in making recommendations on closing, moving or expanding the facilities. The following information needs to be collected for each site:

- *Space utilizations:* Determine the utilization of the distribution center. This will allow you to determine the amount of physical inventory space that will be required if this facility is to be closed when the analysis is complete. It also identifies how much more inventory can be combined into this location.
- *Layout and equipment:* List the equipment and layout of each facility. If you have a list of equipment available, it will be easier to determine the investment requirements of a new or expanded facility.
- *Warehouse operating procedure:* Understand the order picking and shipping procedures. If there are two product lines in one location, are they picked and shipped together? Understand the differences in operating methods between facilities. This may explain why one facility achieves a higher throughput efficiency per person. Understand how replenishment orders are placed or pushed to the distribution center.
- *Staffing levels:* Document levels by position. Understand which jobs could be consolidated. Collect labor rates by level, including fringe benefits.
- *Receiving and shipping volumes:* Understand the number of incoming and outgoing trucks and the number of docks. This will be important if the facility is required to increase throughput.
- *Building characteristics:* Collect building characteristics such as clear height, lighting levels, column spacing, etc. Collect this for the same reason as layout information, but keep in mind to review expansion capabilities.
- *Access to location:* Review the access to main highways. Determine whether this will have an effect on freight cost.
- *Annual operating cost:* Collect lease cost, taxes, insurance, maintenance, energy cost, and other facility cost.
- *Inventory:* Collect information on inventory turns and levels, fill rates, safety stock levels, and ABC analysis. By having this information, the savings of consolidating facilities can be determined. Also collect which, and how much, stock is slow moving or seasonal to help determine if it should be centralized in one location or whether public warehouse space should be used. Get future inventory goals.
- *Performance reporting:* Understand the performance measures for service requirements, order completeness, shipping accuracy, etc.

The following information should be collected for the transportation system:

- *Freight classes and discounts:* Collect the freight classes and rates used. In addition to freight classes, get the discounts by carrier or location. It is also important to understand where the discounts apply (under which parameters, i.e. routes, minimum weights).
- *Transportation operating procedures:* Understand how a certain mode of transport is selected and how a carrier is selected.
- *Delivery requirements:* What are the delivery requirements (days of delivery) to the customer in days, and how is carrier performance measured? Is order completeness measured?
- *Replenishment weight/cube:* At what weight is a trailer cubed out? Get this information from each replenishment point and for a typical load of general merchandise.

At the end of the site visits, a project team meeting should be held that summarizes the data collected and the assessment of each site. This assessment will give the team insight into its operation, and more than likely they will discover information unknown to management that will be useful in developing alternatives.

To document the future distribution network requirements, it is not only important to understand the factors influencing distribution but also to understand the marketing strategies and sales forecast. The following list identifies questions that should be answered by marketing and sales:

- Are there any new products coming out? From where are they sourced? What is the target market area (geographically)?
- What are the ordering parameters right now? For example, what is the minimum order size? Are they changing any terms of order (i.e., charging for expedite service)?
- What is the direction of the market? (Packaging changes, wholesalers, mass merchants having more volume.)
- Are sales increasing each year?
- Are customer shifts becoming apparent? Are fewer customers handling more volume?
- Have geographic shifts emerged? Have sales increased by geographic regions?

2.4.1.2. Identify Delivery Requirements One of the key data requirements in analyzing a distribution network is that of the delivery requirements (time order placement to receipt of the shipment). If the requirements are not identifiable, a customer satisfaction gap analysis must be undertaken. The gap analysis is a series of questions directed at internal staff and customers. The purpose is to identify discrepancies between customer perception of satisfaction and satisfaction requirements. At some point, the sales sharply decline because competition exceeds both your delivery and your cost (assuming equal product quality). The key is to find the best customer satisfaction that maximizes profits.

2.4.1.3. Establish Database The database of orders that are to be modeled can be established while the existing network is being documented. This information should include ship-to locations, weight of the shipments, products ordered, and the quantity ordered. Once the data are established, the next step should be to validate the data. In order to ensure that the information was transferred properly, print out a few records of invoices and compare these to hard copies. Also, it is a good idea to prepare a summary report (sales, cases sold, weight shipped) for a sanity check to ensure all the data in the files were transferred. Once the data are valid, various analyses such as ABC analysis by picks, location (geographical), volumes, and product volumes by regions of the country should be run. These reports should be used to help determine alternatives.

2.4.1.4. Develop Alternatives Once the data have been collected, the next step consists of developing alternative site locations and operating methods. The input used to determine alternatives consists of site visits, future requirements, database analysis, and customer service surveys. The methods used for the selection of each site will vary.

Sites are not the only option to consider as alternatives. Operating methods must also be considered. Consideration must also be given to criteria such as consolidating vendor shipments, centralizing slow-moving items in one place, keeping company divisions separate, and direct shipment by vendors. Once alternative sites are determined, data must be collected on freight rates, warehouse cost, and labor cost for the alternative sites.

2.4.1.5. Model Annual Operating Cost Modeling software doesn't guarantee the right answer. Modeling should only be used as a tool to aid in the decision process. The real value in distribution planning is the knowledge gained from understanding the working of a company's distribution system, knowledge on distribution planning, and the imagination to use the model in ways to really benefit the distribution network. Alternatives can be close in cost but have a wide range in number of facilities; therefore, it is important to have some other criteria to judge the modeling runs, such as:

1. *Central administrative costs and order processing cost:* Typically, these costs increase with the number of warehouses. It takes more effort to coordinate and manage a larger network of facilities.
2. *Cycle and safety stock carrying costs:* More warehouses means more total system inventory. Inventory theory supports that safety stocks will increase with the number of facilities.
3. *Customer order size effects:* Customers who are close to a warehouse generally tend to order more frequently and in smaller quantity than customers who are farther away. This implies that delivery costs tend to increase on a $/cwt basis as the number of facilities increases.
4. *Interwarehouse transfer cost:* The more distribution centers there are, the greater the coordination problems and the more likely the tendency to transfer inventory between facilities due to imbalanced inventory availability.
5. *Negotiated reduction in warehousing and delivery costs:* The fewer facilities, the greater their individual volume, and hence the more opportunity there is to negotiate more favorable arrangements for warehousing and delivery service.

No matter what modeling method is used, the overall approach should closely resemble the following steps:

- *Validate the existing network:* Run computer model to simulate the existing cost. Compare this cost to actual cost.
- *Run alternative networks:* Once the model is valid, alternative networks should be run for present volumes and forecasted volumes.
- *Summarize runs and rank:* Create a table to summarize cost by alternative. The table should list distribution center cost individually.
- *Summarize all annual costs and satisfaction factors:* Create a table that indicates by alternative all the cost and service factors.
- *Perform a sensitivity analysis:* Sensitivity analysis is based on the idea of setting up runs that fluctuate some components of the data. This could be a cost that is uncertain or has potential to change. By modifying this single parameter, the effect on the run can be determined.
- *Determine all investment costs associated with each alternative:* Such as cost of new warehouse equipment required to save space, expansion, and construction cost or any building modifications such as adding dock doors. This information will be of use in the next step.

2.4.1.6. Evaluate Alternatives The economic analysis compares the recommended network plan to all alternatives. To do this analysis, you must determine all the investments and savings associated with each alternative. Costs such as new warehouse equipment, construction cost, and any building modification should be included. Additionally, the following information must be identified: personnel relocation, severance, stock relocation, computer relocation, taxes, equipment relocation, and the sale of existing land and buildings.

The result of this step should be a return on investment of each alternative compared to the baseline. Once this step is completed, a sensitivity analysis that fluctuates various costs and savings to see which alternatives are the most stable should be performed. To round out the analysis, a qualitative analysis should be performed, looking at such factors as customer service and ease of implementation. Once a conclusion has been reached, a time-phased implementation schedule should be drawn up listing the major steps involved in transferring the distribution network from the existing system to the future system.

2.4.1.7. Specify the Plan The final step in the distribution network planning process is selling the results to top management. This must be expressed so that management can understand the impact of the strategy on the total business. This communication should express not only the finances relating to transportation and warehouse costs, but overall sales and customer satisfaction.

2.4.2. Do Not Underestimate the Importance of Distribution

Distribution is the management of inventory to achieve customer satisfaction. Today, many companies have realized that distribution is a major frontier for both customer satisfaction enhancement and cost reduction. It is important to remember that a good strategic distribution network plan is a requirement of success and that it should not be composed simply of ideas, thoughts, or possibilities whose validity has not been researched. The distribution network plan is based upon a set of premises concerning future sales volumes, inventory levels, transportation cost, and warehouse cost. Requirements should be defined, analyzed, and evaluated and should result in the development of a specific set of strategic requirements. A good distribution network plan is action oriented and time phased and keeps the ultimate customer's requirements at the forefront at all times.

2.5. Team Selection

Once the need to identify a site is realized, an in-house selection team should be established. Facility management can make the site-selection process easier by also establishing alliances with external resources such as:

- Brokers
- Economic developers
- Government agencies
- Utilities
- Consultants

Once a prime geographical area for the new facility has been settled on and management has approved the SMP, the job of selecting the best community and site begins. This is the most difficult and time-consuming part of the process; support from outside sources will help the in-house team narrow down the list of potential candidates.

2.5.1. Real Estate Brokers

Real estate brokers are typically tied into a multiple-listing service that lists all available property in an area. Many agencies also employ state-of-the-art technology so prospective buyers can see many views of a particular site without leaving the broker's office. Be careful to select a broker comfortable with industrial site searches; many advertise commercial expertise, but industrial requirements are different from generic commercial ones, especially insofar as environmental regulations. Remember that brokers are compensated only for successful transactions, and hard-sell approaches may or may not create pressure for the selection team.

2.5.2. Government Agencies

State and local government development agencies, as well as chambers of commerce, are reliable information sources. Economic developers are sometimes hired employees of a city or county to promote area growth; they can also be commercial real estate brokers. A good economic developer will be able to save you time by showing you only properties that meet your needs; they will also steer you away from properties that may have watershed or zoning restrictions that would prohibit locating your facility. An economic developer most likely will have important political connections that can help cut through "red tape" and therefore speed up the process of site selection. Since government staff is motivated to attract new industry to its area, make certain that the area is compatible with company objectives before relying on this information source. A plus is that economic developers are usually aware of all incentive possibilities.

2.5.3. Utilities

Once a general area has been chosen, gas and electric companies can provide useful information on specific sites. Utilities are unbiased sources of advice and often work with brokers on specific land and building details.

2.5.4. Consultants

Facilities planning consultants provide an unbiased source of advice on each facet of the site selection process. With a wide breadth of knowledge and accessibility to time-saving planning tools, consultants make for an efficient and effective means of establishing important quantitative information such as the baseline and for "flushing out" all possible and feasible alternatives.

3. PROCEDURE—THE MICRO ANALYSIS (SITE SELECTION)

Once the "big picture" has been analyzed and a macro analysis developed, the microlevel work, in which a specific site must be chosen, begins. This is illustrated in Figure 5.

3.1. Community Selection

Once the general area for site selection has been determined by means of the macro analysis, the selection team should identify specific communities within that area for serious consideration. Since each community will interpret and administer legislation and government mandated/funded programs differently, the site-selection team must take special care to assess each community against company criteria and objectives. The team must therefore view and evaluate communities as they are likely to exist when the proposed facility comes on-line.

A checklist is an ideal method for measuring community attitudes and trends. The following presents some general subjects against which a community might be evaluated:

1. Are attitudes of government favorable to industry and progress?
2. How acceptable are the educational and training systems, from daycare through university?
3. Are municipal services operating on a level acceptable for community progress?
4. How is the quality of life as far as shopping, entertainment, and medical facilities?
5. Are there accessible support services for industry, such as maintenance and machine shops?
6. What are the residential neighborhoods like?
7. Can a labor force be built without compromising skill and productivity requirements?
8. Are construction and contracting services sufficient to build a facility?
9. What economic incentives are being offered?

One important economic factor is the local tax on inventories, primarily because this cost can vary significantly from one location to another. States or other taxing authorities have used the presence of inventory in storage as a basis for levying franchise, income, or other taxes on the owner of the property. Rates of taxation within states, or even within counties or towns, can also show

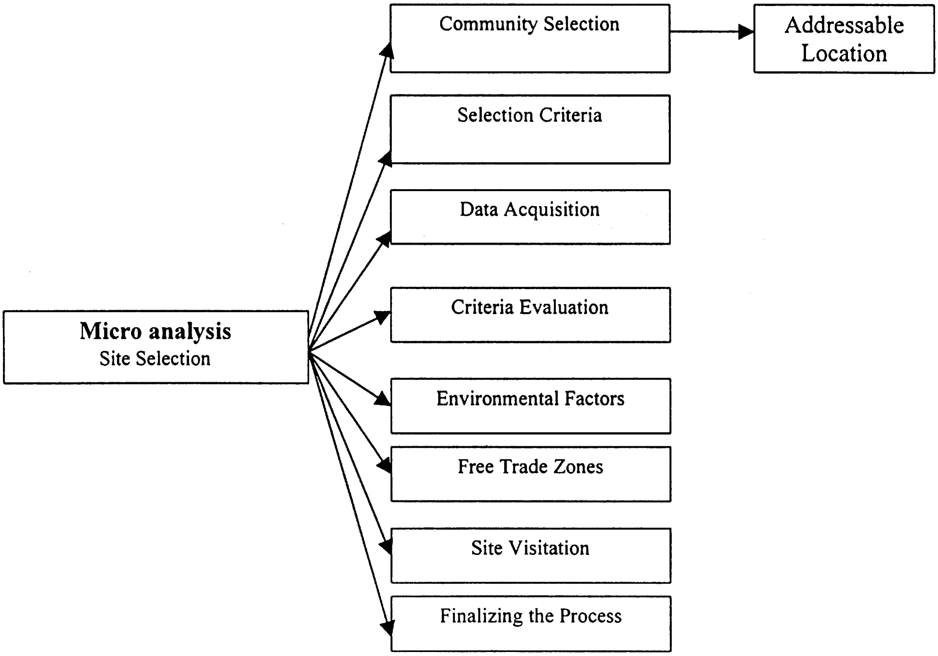

Micro analysis
Site Selection
Community Selection
Addressable Location
Selection Criteria
Data Acquisition
Criteria Evaluation
Environmental Factors
Free Trade Zones
Site Visitation
Finalizing the Process

Figure 5 The Micro Analysis.

significant differences. Because state and local tax situations change frequently, expert advice should be sought when comparing tax policies of different communities.

3.2. A Site-Selection Checklist

One must utilize a comprehensive site selection checklist as part of an assessment of a particular location. A sample checklist is provided below. Checklists such as the one below are useful when establishing contacts in prospective site areas. It is not recommended that site-selection teams solicit information "cold"; in other words, observation, rather than someone else's perception, will provide an accurate depiction of the site and its advantages and disadvantages.

Site Factors

1. General Information
   - Site location
     City ______________
     County ______________
     State ______________
   - Total acreage
   - Approximate cost per acre
   - Approximate dimension of site
     Width ______________
     Length ______________
2. Zoning
   - Current ______Residential ______Light ______Medium ______Heavy Industry ______Commercial ______Other
   - Can zoning be changed? ______Yes ______No
   - Check which, if any, is required: ______Rezoning ______Variance ______Special Exception
   - Probability for success: ______Excellent ______Good ______Fair ______Poor

- Applicable zoning regulations (attach copy)
- Parking/loading regulations
- Open space requirements

  Office/portion

  Maximum building allowed

  Warehouse/DC portion

  Percent of lot occupancy allowed

  Setbacks if required
- On-site waste treatment required
- Height restrictions
- Noise limits
- Odor limits
- Are neighboring uses compatible with proposed use? ______Yes ______No
- Can a good title be secured? ______Yes ______No
- Can protective easements, protective covenants, or mineral rights be tolerated? ______Yes ______No
- Is expansion allowed? ______Yes ______No

  If yes, how much?

3. Topography

- Grade of slope ______level ______ rolling ______ mostly level ______ steep

  Lowest elevation ______

  Highest elevation ______
- Drainage ______Excellent ______Good ______Fair ______Poor
- Are there any: ______Marshlands ______Ponds ______Streams ______Brooks ______Ditches ______Lakes
- Are they: ______On site ______Adjacent to Site ______Bordering Site
- What is the 100-year flood plan?
- Is any part of the site subject to flooding?
- What is the groundwater level? ______feet
- Describe surface soil:
- Does site have any fill? ______Yes ______No
- Soil percolation rate: ______Good ______ Fair
- Load-bearing capacity of soil: ______ lb/ft^2
- Is site wooded? ______Yes ______No

  How much?
- Cost of removal
- Cost of grading
- Storm water discharged: ______ storm sewer ______ other
- Roof drainage discharged: ______ storm sewer ______ other

  Collection required ______Yes ______No

4. Landscaping requirements

- Building and parking lot
- Access road
- Loading zones
- Site buffer

5. Access to site

Is site visible from the highway? ______Yes ______No

Describe access including distance from site to:

  Interstate highways

  Major local roads

  Central business district

Rail
Water
Airport
Describe availability of public transportation
Will access road need to be built? ______Yes ______No
If yes, who will build?
Who will maintain?
Cost of maintenance?
Is rail extended to site? ______Yes ______No
Name of railroad(s)
If no, how far?
Cost of extension to site
Who will maintain extension?

6. Sanitary sewage

- Is sanitary sewage on site? ______Yes ______No
- Reserve capacity in treatment plant ______GPD
- Tap charges
- Special regulations
- Anticipated long-range plans for permanent disposal of sewage

7. Water

- Is water line on site? ______Yes ______No
- Location of water main
- Size of main
- Static pressure ______PSI
- Residual pressure at 1000 GPM flow ______PSI
- Hardness of water
- Is supply adequate? ______Yes ______No
- Capacity of water plant ______ gallons
- Are fire hydrants metered? ______Yes ______No
- If water and sewer not on site:

What is distance to nearest line
Water
Sewer
Line size
Water
Sewer
Cost to extend line
Water
Sewer

8. Sprinklers

- What type of sprinkler system does code permit?
- Is there sufficient water pressure for sprinkler system? ______Yes ______No
- Is water for sprinkler metered? ______Yes ______No
- Is separate water supply required for sprinkler system? ______Yes ______No
- Where can sprinkler drainage be discharged?

9. Electric power

- Is adequate electric power available to site? ______Yes ______No
  Capacity available to site
- Describe high-voltage lines on site
- Service is ______ underground ______overhead
- Is submetering permitted? ______Yes ______No

- Indicate if reduced rates are available for:
  Heat pumps ______Yes ______No
  Electric heating ______Yes ______No
  Insulation ______Yes ______No
- Rates

10. Gas
- Type of gas available ______natural ______LP
- Capacity
- Line size ______inches
- Pressure of gas ______ PSI
- Is submetering permitted? ______Yes ______No
  If not, cost of extension
- Rates

11. Other utilities
- Coal
  Source of supply
  Reserves
  Quality
  Cost per million BTU delivered
  Method of delivery
- Oil
  Source of supply
  Volume
  Quality
  Cost per million BTU delivered
  Method of delivery

12. Taxes
- Date of most recent appraisal
- Real estate tax history, last five years
- History of tax assessments, last five years
- Proposed
  Increases
  Assessments
  Tax rates
- Are abatement programs in effect? ______Yes ______No
  If yes, describe
- Is site in an Enterprise Zone? ______Yes ______No
- Are industrial revenue bonds available? ______Yes ______No
- Services provided for taxes paid

Community Factors

1. Labor history
- Does labor force have deep community roots?
- Do most workers own their own homes?
- Is labor force largely transient?
- Can you determine prospects of future favor tranquillity as evidence by labor turnover or absenteeism?
- Has labor group maintained a good reputation for accepting technological change?
- Do employees have a good reputation for housekeeping practices and care of equipment?
- Labor availability survey
- Population at last census
- Population density per square mile

  - Percent agriculture
  - Total employed in manufacturing
  - Total employed in nonmanufacturing
  - Percent men in labor force
  - County-wide potential employment
  - Unemployed available workers
  - Shift willingness
  - Distribution of available labor
    Skilled
    Semiskilled
    Unskilled
  - For women: average family income and whether basic need exists for supplemental income
  - Do farm areas serve as good labor pool?
  - Is there high degree of farm mechanization?
  - Does community have increasing supply of women seeking industrial jobs?
  - Can you complement rather than compete with existing industry?
  - Will seasonal jobs in nearby resort areas affect labor availability?
  - Is community subject to other seasonal labor variations?
  - Does adequate labor pool exist within reasonable radius?
  - Are young people taking jobs elsewhere?
  - Would better opportunities keep young people at home?
  - Influence of local industry on labor
  - Principle community factors
  - Wage rates, by skill
  - Working hours
  - Shift patterns
  - Hourly or piece rates
  - Fringe benefits
  - Degree of competition for skills
  - Pattern of productivity
  - Seniority provisions
  - Layoff provisions
  - Grievance
  - Presence of any unusual or radical tendencies
  - Does industrial accident rate for community compare favorably with national averages?
  - Will you be direct (or indirect) competition with an industrial pace-setter?

2. Maturity of citizens
  - Do local civic and religious leaders have enlightened and progressive attitude toward business and industry?
  - Do people of community display political awareness?
  - How many voters went to the polls in the last municipal election?
  - How many voters went to the polls in the last national election?
  - Do local people understand how business operates in the American economy?
  - Are there community educational programs directed at young people?
  - Do social and economic backgrounds of community point toward maturity?
  - Is standard of living at or above normal average?

3. Management potential
  - Can prospective workers be expected to grow into added responsibilities?
  - Can you translate evaluation into estimates for potential supervisors and executives?
  - Can you expect to recruit certain management echelons locally?
  - Are specialized skills available such as specific and technical manpower?
  - Have local people responded well to in-plant training?

4. Water pollution
- Will you have waste disposal problems?
- Can streams nearby accommodate waste water?
- Will good business practice plus local or state ordinances call for waste treatment?

5. Transportation
- Rail

  On a rate-blanketing basis, are rates to principle markets satisfactory?

  Has pattern of differential freight rate increases been relatively favorable for your proposed area?

  Area?

  Are there amply freight forwarders for LCL

  Does railroad give transit or stop-off privileges for partial unloading and loading enroute?

  Are there adequate truck handling facilities at freight terminals?

  Is pick-up and deliver service available?

  Which of these principal rail considerations are important?

  Branch or mail line

  Freight schedules

  Switching per day

  Yard limits

  Direction of turnout to private siding from yard

  Orientation of site to roadbed

  Relative elevation of site and roadbed

  Potential construction difficulties such as culvert, fill, bridge, cut

  Does prospective rail carrier favor the use of technologically improved equipment to meet shipper needs?
- Truck

  Are there state laws re: truck size and weight restrictions Is site near a trucking gateway in order to reduce in-transit times?

  Are state gasoline taxes in line with alternate sites?

  Is the pattern of recent truck freight rate increased reasonable?

  Will the new Federal Highway Program help solve trucking problems?

  Is there good access to bridges and culverts?

  Which of these factors are important?

  Natural traffic flow

  Specific routes

  Schedules

  Rates

  Transfers

  Common, contract, or private carrier
- Air

  Is site near a good airport?

  Are rates and schedules or scheduled airlines satisfactory for hire shipment?

  Are there good air-freight forwarders nearby?

  Is airport service convenient for transport of personnel?

  Is there helicopter shuttle service?
- Ocean

  Is proximity to inland water transport important?

  Is proximity to overseas shipping important?

  Does area have alert and progressive port authority?

  Are port facilities closed in winter?

  Is ample lighterage available?

  Is access to port convenient and economical?

  Are water transport rates and schedules competitive?

- Other
  Is railway express service available?
  Are pipelines used as common carriers?
  Does community have desirable level of passenger transportation facilities?
  Are there toll roads or bridges?
  Do winter conditions adversely affect transport?
  Does community have public or private warehouse available to help out witht short-range inventory storage problems?

6. Raw material supply
- Are raw material sources reliable? Close enough?
- Are the raw materials committed to others?
- Are terms of sale and delivery right?
- Are multiple supply areas available?

7. Residential housing
- Are there plentiful rental properties?
- Are houses available in several costs brackets?
- Does extent of home ownership among hourly employees indicate stability and community pride?
- Are residential property values increasing?
- Are attractive suburbs within convenient distance of selected community?
- Is community saddled with submarginal or slum areas? Is rehabilitation in progress?

8. Education
- Assess the number and sufficiency of
  Public schools
  Vocational colleges, trade schools, and apprenticeships
  Foremanship courses
  Adult education, degree programs
- Is school growth keeping up with community growth?
- What is the overall education picture?
  Expense to public
  Teacher salaries
  PTA enthusiasm
  Building program

9. Health and welfare
- Are there satisfactory
  Hospitals
  General practitioners
  Dentists
  Clinics
  Nurses
  Public health facilities
- Do hospitals have adequate ratings by State Board of Health?
- How large an area is served by hospitals?
- Are Blue Cross and allied plans available?
- Does community have a workable disaster plan?
- Are there reasonable state industrial safety and health laws?
- Does the community have adequate and well-enforced sanitary laws?

10. General community aspects
- How sufficient is the community recreation system insofar as:
  Family recreation (e.g. parks and playgrounds)
  Outdoor activities (e.g. golf and tennis)
  Libraries

    Civic attractions (e.g. museums)
    Fraternal organizations
- Sufficiency of public buildings
    Gyms
    Churches
    Auditoriums
- Is the physical center of town attractive?
- Are there good hotels, motels, and restaurants?
- Is banking adequate?
- Are shopping and commercial districts well laid out?

11. Commercial services

- Evaluate quantity and quality of commercial services typically required by industry
    Major repair shops
    Industrial distributors
    Lumber and allied materials
    Stationery
    Local trucking
    Air freight services
    Blueprint services
    HVAC repair
    Testing labs
    Electric motor maintenance
    Lubricants
    Engineering department supplies
    Food and sundry vending
    Railway express
    Postal service
    Industrial repair
    Janitorial service
- Evaluate quantity and quality of construction services and facilities in or near the community
    Architects
    Prime contractors
    Mechanics
    Engineers
    Subcontractors
    Electricians
    Piping
    Construction labor
    Special equipment
    Plasterer
    Painting
    Paving
    Carpentry
    Rigger
    Mason
    Tiling
    Landscape

12. Police

- Does police department have high standards of personnel, equipment, training, and morale?
- Is police patrol provided for industrial properties?
- Are private watchmen or uniformed detective services available?

- Is incidence of crime as low or lower than in surrounding area?
- Does community have disproportionate number of bars?
- Is judiciary system well organized?

13. Fire department
- Does fire department have high standards of personnel, equipment, training, and morale?
- Is community fire insurance classification up near the top?
- Is site within fire hydrant limits?
- Are adjacent communities near enough to send apparatus in case of serious fire?

14. Infrastructure
- Does quality of construction and maintenance indicate an efficient highway department?
- Are roads kept free of ice and snow?
- Is there a satisfactory highway improvement program in place?
- How adequate is garbage collection?
- Does the sewage department have realistic plans for expansion and improvement?

15. Planning and zoning
- Is the city planning commission active and progressive?
- Are smoke, noise, and odors controlled?
- Can facilities expect protection from undesirables?
- Do building inspectors have a reputation for honesty and integrity?

16. Community financial picture
- Does community indebtedness present a healthy picture?
- Is community taxation well balanced between residential, commercial, and industrial sources?
- Is pattern of community expenditures well balanced between needs and income?
- Is total community tax picture in line with services received?
- Are community tax inducements offered to prospective industries?

17. Community business climate
- Is attitude of local officials sympathetic and enthusiastic toward existing and new industry?
- Is record of local government good insofar as honesty, efficiency, and principles?
- Does community have business-sponsored civic organizations dedicated to improving business climate? Have results been achieved?
- Is community industrially well diversified?
- Have any manufacturers migrated from the community recently?
- Are the industries dynamic and growing?

18. Community employer evaluation
- Have most employers demonstrated enlightened management policies?
- Have employers kept pace with rising wage standards voluntarily?
- How do you rate employee/employer communications within the manufacturing community?

19. State taxes
- What is the existing gross debt of the state?
- Do state corporate taxes compare favorably with those of competition elsewhere?
- Is there a state individual income tax?
- Does the state levy property taxes?
- Is there a state sales tax?
- Does the state grant permission to deduct Federal Income Tax?

20. State business climate
- Are state legislative, executive, and judiciary branches performing as well as counterparts in other states?
- Are state salaries attractive enough to get and keep good people?
- Are state wage and hour laws fairly written and administered?
- Is state workman’s compensation satisfactory?
- Is state unemployment compensation equitable?

- Does state have an active and progressive development commission?
- What are state laws regarding:
  Unreasonable union acts?
  Secondary boycotts?
  Illegal strikes and picketing?
- Have other industries been asked whether there are hidden restrictive laws?

3.2.1. Comparing Specific Sites Based on Organization-Specific Criteria

The following example shows how specific sites can be compared. The matrix established is based on weighted criteria deemed important to the client; both qualitative and quantitative factors are included in the assessment. Information is gathered from a variety of sources, including:

- U.S. government agencies
- State economic developers
- County development agencies
- City governments
- Trade groups
- Commercial real estate agents

3.2.2. Evaluation of Addressable Locations

Step 1 is the establishment of criteria and the respective weights each criterion holds. Company X has determined that the following general categories are significant in the determination of the best site for their new warehouse. Further specification of each criterion is noted in step 2.

• Customer service	15%
• Cost	15%
• Infrastructure	15%
• Suppliers	15%
• Labor	20%
• Community	5%
• Incentives	13%
• Climate	2%
	100%

Step 2 is the information-gathering stage. Information is taken from government and community sources.

Criteria	Location X—Greenfield	Location Y—Greenfield
	Customer Satisfaction	
1. Number of days to market		
Boston	4	3
Chicago	3	3
Miami	3	3
New York	4	2
Dallas	3	3
2. Distance to market (mi)		
Boston	971	1118
Chicago	411	542
Miami	839	681
New York	986	899
Dallas	776	894
3. Carriers available	2 terminals, 21 carriers	1 terminal, 9 carriers
4. Parcel service	All	All
5. Interstate access	18 miles to interstate	6 miles to interstate
6. Freight rates—outbound	\$2.09M	\$1.98M
7. Freight rates—inbound	\$570K	\$680K

Criteria	Location X—Greenfield	Location Y—Greenfield
	Cost	
1. Land cost		
Per acre	$20K	–[a]
Total land cost	$440K	–[a]
2. Construction cost		
Distribution	$30/ft^2	$28/ft^2
Manufacturing	$45/ft^2	$40/ft^2
3. Incentives	LEDA	KEA Land $75,000

[a] Financial incentive offered by economic developers was the ‘‘donation’’ of a Greenfield site.

Criteria	Location X—Greenfield	Location Y—Greenfield
	Infrastructure	
1. Fire main supply	2400 GPM	1286 GPM
2. Fire department rating	4	3
3. Power source	LA Utility	KY Authority
4. Sewer	X	Y
5. Service roads	45 ft.	30 ft.
6. Park lighting	Installed	Add
7. Security	Fence optional	Fence optional
8. Data transmission lines	56 KB $941/mon	56 KB $856/mon
9. Telephone lines	Fiberoptic	Fiberoptic
10. Commercial airport (distance)	96 mi	30 mi
11. Local airport runway length	5000 ft	5000 ft
	Suppliers	
1. Tool and die	3	4
2. Corrugated	1	2
3. Material-handling vendors	2	2
4. Pallets	2	2
	Labor	
1. Major employers	XeronTech—300 Tachion Inc.—220	CirrinePlex—500 Handle Tire—350 Goodson Snacks—110
2. Labor Pool	87,400	15,540
Employed	81,500	13,396
Unemployed	5,900	2,144
Rate of unemployment	7%	14%
3. New employers	XeronTech—300 Duplexus Packaging—86	Goodson Snacks—110
4. Industry leaving	None	Nechcon Laminate—200 Abernathy Textiles—640
5. Right to work	No	Yes
6. Union activity	ACTW LIU UFCW	IAM BCTW UGWA
	Community	
1. Local government	Mayor 5 Council members	Mayor Alderman
2. Population	11,295	7,056
3. Schools		
Enrollment	7,000	4,500
Student:teacher ratio	17:1	18:1
Post-secondary schools	(1) Community college (1) Technical school (1) 25,000-student university	(2) Community colleges (1) 4,500-student liberal arts college

Criteria	Location X—Greenfield	Location Y—Greenfield
	Community	
4. Hospitals		
Beds	227	91
Doctors	90	16
Dentists	22	2
5. Recreation		
Parks	10	4
Golf	3	1
Country clubs	2	1
Theaters	4	1
Hotel rooms	350	164
6. Housing		
<$125,000	166	45
>$125,000	71	23
7. Closest metropolitan area	75 mi	25 mi
	Financial Incentives	
1. Economic developer	LEDA	KEA
2. Fire protection	$60,000	$230,000
3. Site preparation	$75,000	$75,000
4. Tax abatements	$150,000	$300,000
5. Other		Land—$300,000
	Climate	
1. Average annual temperature (F)	57.6	60
2. Average annual rainfall (in.)	47	52
3. Average annual snowfall (in.)	12	9
4. Prevailing winds	NE	S

Step 3 is the development of a matrix for each constituent. Each constituent of each criterion is further weighted within its category and ranked on a 0–10 scale, with 10 representing an "ideal." One example is provided in Chart 2. The process will be conducted for all of the criteria.

		Location X—Greenfield		Location Y—Greenfield	
Criterion	**Weight**	**Value**	**Weighted**	**Value**	**Weighted**
		Customer Service			
Number of days to primary market	25%	5	1.25	5	1.25
Number of days to secondary markets	10%	4	0.40	6	0.60
Carriers available	15%	7	1.05	5	0.75
Parcel service	15%	10	1.50	10	1.50
Interstate access	15%	5	0.75	5	0.75
Freight rates (outbound)	10%	5	0.50	5	0.50
Freight rates (inbound)	10%	6	0.60	5	0.50
Totals	100%		6.05		5.85

Step 4 is the development of a matrix showing each general category, its values and weights. A final recommendation is based on this matrix.

		Location X—Greenfield		Location Y—Greenfield	
Criterion	**Weight**	**Value**	**Weighted**	**Value**	**Weighted**
Customer service	15%	6.05	0.91	5.85	0.88
Cost	15%	3.50	0.53	9.60	1.44
Infrastructure	15%	7.67	1.15	5.00	0.75
Suppliers	15%	5.61	0.84	4.87	0.73
Labor	20%	6.22	1.24	7.15	1.43
Community	5%	6.45	0.32	4.33	0.22
Incentives	13%	4.00	0.52	9.75	1.27
Climate	2%	5.22	0.10	5.85	0.12
Totals	100%		5.61		6.84

The methodology by which information is gathered is to establish an open, fruitful dialogue with these sources; this can be through a phone call, fax transmission, e-mail, or postal correspondence. It is vital to the success of any site-selection process that the information be shared on an immediate basis and follow-up (including questions) be timely and constructive.

The deliverable of this comprehensive study of sites is the recommendation of the best fit for expansion or location/relocation.

3.3. Environmental Factors

Even ‘‘clean’’ facilities need to be cognizant of the many regulations protecting the environment. Site-selection teams should be aware that cases can be cited of companies buying sites at low prices but having ‘‘to eat’’ prohibitive cleanup costs, a legacy left by the previous tenant. Even though firms that have potential pollution problems are supposed to notify the United States Environmental Protection Agency (U.S. EPA) of these problems, site-selection teams need to be alert to certain red flags when considering previously inhabited sites.

- Landfills and other solid waste depositories should be excluded from the selection process. Remediation costs normally exceed land value.
- Any site with underground storage tanks (USTs) should be thoroughly assessed. Older tanks need to be checked for rust and leaks, and all USTs must be evaluated and maintained regularly. Underground gasoline tanks must, by law in many states, be noncorrosive; replacement of old, corrosive tanks is an expensive process.
- Any facility with asbestos will require abatement or encapsulation, which is a costly procedure.
- Sites with potential for soil and groundwater contamination, due to inadequately treated sewage, discharge of manufactured or agricultural wastes, the dumping of toxic wastes, or even runoff from nearby golf courses, should be considered problematic.
- Storm water is considered one of the three leading causes of pollution in U.S. waters. Permits are required for storm water discharge, and contingencies must be made for storm water to be separated from sanitary sewer systems.
- Sites where solid waste disposal is necessary must be located near a licensed outlet. Burning of waste materials is not permitted in most locales.

Site-selection teams must learn all about pertinent regulations that govern the use of a site for specific operations. Teams should never assume the site is ‘‘clean’’; this is an expensive assumption to make.

3.4. Free Trade Zones

One of the considerations in selecting a site is whether or not the site is in a free trade zone (FTZ) and whether or not the FTZ meets the company’s needs. FTZs are secured areas within the United States but they are legally considered outside the company’s territory. The purpose is to attract and promote international trade and commerce. Foreign and domestic goods may be stored, manufactured, and processed duty-free in the FTZ, which is typically a fenced-in area with warehouse facilities and industrial park space and access to all modes of transportation. Subzones, which fall beyond the perimeter of the public zone, are set up to accommodate industries that satisfy specific government criteria (e.g. ability to generate public benefit through employment).

Companies that intend to import products for manufacture or sale should research the feasibility of the FTZ. The advantages are:

- Customs duty and internal revenue tax, if applicable, are paid only when merchandise is transferred from a foreign trade zone to a custom territory for consumption.
- Goods may be exported from a zone free of duty and tax.
- Merchandise may remain in a zone indefinitely, whether or not subject to duty.

Companies should note that certain commodities and industries are excluded from zone consideration: alcoholic beverages, tobacco, firearms, white phosphorous material, sugar, and material operations that may prove to be a detriment to public health and welfare. No retail trade is allowed in FTZs.

3.5. Site Visitation

The team should visit the final round of site candidates. Items either misrepresented or unrepresented in previous information can be determined, and the final assessment may be based on this new information. Extensive research should be done at each site visitation, including in-depth interviews with managers of similar businesses. The project team should keep in mind labor market dynamics, business costs, and support infrastructure.

3.6. Finalizing the Process

Once the site has been selected, productivity should be maintained; in other words, a ‘‘business as usual’’ atmosphere must be generated so as to meet supply chain and customer requirements as seamlessly as possible. The following steps should be taken to ensure that the relocation or expansion plans of the company move forward smoothly:

- Identify and communicate with employees.
- Address training and recruitment issues for the new facility.
- Schedule all official announcements with discretion—premature public disclosure may alienate and demoralize employees.
- Assign either an individual or a team to monitor all aspects of the new facility, not just the construction or the move (taxes, regulations, human resources, and telecommunications all play major roles in successful site accommodation, so these must be considered and handled as diligently as the building or the relocating itself).

Site selection is the next logical step after the strategic distribution network planning, which analyzes the company’s distribution network and develops facilities requirements and customer service requirements based on the warehouse strategic master plan. Site selection is the microanalysis of the company’s needs, and through the use of a structured approach, where criteria are identified, sourced, quantified, weighted, and evaluated, it enables the company to see what the best choice is for expansion or relocation.

The team must be fully cognizant of restrictions and limitations (e.g., pollution controls or energy consumption) placed on parcels of land by governmental bodies before making lasting decisions. Through the careful processes of distribution network analysis, site evaluation and negotiation, companies can select a site that offers flexibility, versatility, and utility for many years.

4. MOVING FROM SITE SELECTION TO CONSTRUCTION

As if site selection is not difficult enough, then comes the actual construction of the facility. On a blueprint, the requirements have little more than conceptual meaning; once brought into reality with bricks and mortar, the structure will either prove immensely effective and efficient or be a capitally intensive reminder of an organization not fully comprehending its strategy.

Building an industrial facility such as a manufacturing plant or distribution center is an arduous undertaking, taking up to two years from start to finish. There should not be a question about why something is being done, who is doing it, or how much more efficient it will make your operations. Unfortunately, many projects are overwrought with conflict, unexpected changes and delays, and ultimately a realization that the building doesn’t reflect, much less satisfy, the needs of the company: today, tomorrow, or next year.

The following will provide an overview of project methodologies and make suggestions on how to choose the right architect/engineering firm and contractors. Construction is not a “let your fingers do the walking through the Yellow Pages” type of endeavor. In order to make a concept a reality, careful planning and implementation of the four major elements of construction, as illustrated in

Figure 6, must be in order. Know what you want out of your manufacturing plant, warehouse, or DC and stick to the plan.

The construction or expansion of a DC or manufacturing facility is a complex undertaking that, if not properly managed, can overwhelm an organization, delay the occupancy of a much-needed facility, and adversely affect a corporation's bottom line. Effective facilities construction requires a balancing act among expectations, cost, scope, schedule, quality, and ongoing operations.

A typical distribution facility can cost over 10 million dollars, take six months to plan, nine months to build, three months for fit-up, and two months to bring on line. It can include the participation of consultants, architects, engineers, realtors, bankers, lawyers, economic developers, state, local, and federal officials, general contractors, subcontractors, and suppliers. During design, construction and start-up, many employees are called on to assume additional responsibilities related to the new facility while continuing to perform their normal jobs. Day-to-day operations must continue, and customer service must not suffer.

No wonder the building process is often viewed by participants as a painful interruption of business as usual rather than as a celebration of growth and financial success for the corporate team.

With proper planning and with use of internal and external resources, the building process can be managed. Projects can be brought in on time, budgets can be maintained, and quality can be achieved. The process will not be painless, but the pain can be reduced if the following steps are taken:

- Establish realistic criteria for site selection.
- Select the contracting or project delivery method that best fits your needs.
- Select the best construction team for your project.
- Vigorously manage the project.

As the site-selection process proceeds, it is easy to lose sight of the ultimate goal: to locate and build a facility in accordance with your strategic master plan. This is the only way that strategic master plan forecast results can be achieved. Resist the temptation to make compromises to the site criteria that will change the building's functionality. Your objective should be to build around the process, from the inside out. Stick to your plan.

4.1. Methods of Project Delivery

With the conceptual design in hand and with a suitable site selected, it is time to decide what contracting method you will use to turn concepts into detail design and detail design into concrete and steel.

There are several methods of project delivery commonly used in today's market. Each method has advantages and disadvantages. Before deciding which method to use for your project, answer the following questions:

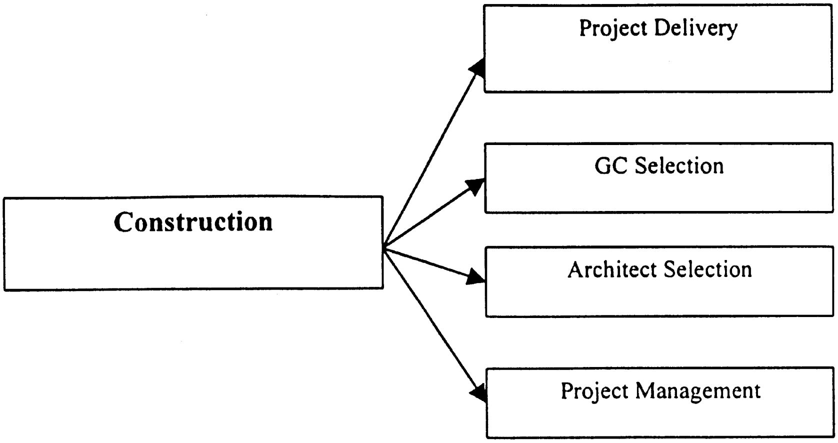


Figure 6 Construction Process.

- *How involved in the details do you want to be?* Different methods of project delivery require different degrees of owner involvement. You must be clear about what your role and responsibilities will be. After the project starts, delays caused by the owner can be costly.
- *What schedule requirements and constraints do you have?* Different contracting methods support different schedule requirements. Make sure that you understand the typical time line for each project delivery method.
- *What budget constraints do you have?* What contracting method best fits your budget?
- *What resources do you have in-house to devote to the project?* Do you have qualified people on staff that can act as your representative, or should you assign this role to a consultant?
- *What are your priorities—scope, schedule, budget, aesthetics, quality?* Each project-delivery method has advantages and disadvantages in these areas. Only through a thorough understanding of the methods can the correct selection be made.

The success of a project is often directly related to the delivery method used. The following section outlines four commonly used methods of project delivery. Once the above questions are answered, the appropriate method for your project can be selected.

4.1.1. Design-Bid-Build (Traditional Method)

The conventional fixed bid method of delivery project approach is characterized by a clear separation between the design phase and the construction phase. The owner hires an architect/engineering firm (A&E) that turns the functional requirements of the building into a detailed set of construction drawings and specifications. These drawings and specifications, as well as schedule requirements and special conditions, become the basis of a request for proposal (RFP) that is issued to general contractors (GC) for bids. Bids are received and evaluated and a fixed price or lump sum contract is awarded to a GC. The GC purchases materials and awards subcontracts as required to satisfy his obligations as detailed in the contract documents.

The GC assumes the entrepreneurial risk of completing the work in accordance with the plans and specifications for the lump sum amount agreed to in the contract. The owner's liability in theory is limited because he has a firm price to do the work. The key thing to remember here is that the contractor has given a lump sum price to perform the scope of work that is detailed in the bid documents; he has not given a lump sum price to do "whatever it takes."

As in any method of project delivery, changes in the scope are grounds for additional compensation. Likewise, changes in the schedule (e.g., delays caused by someone other than the GC or their subcontractor) can be grounds for requesting an extension time and additional compensation. Generally, delays caused by weather are covered under an excusable delay clause in the contract. In this case, the contractor would not be eligible for additional compensation to cover his extended presence on the job but would receive an extension of time equal to the time lost.

This is still the most common method of contracting in today's commercial market. Most public projects, both state and federal, use this traditional method. It is well suited to projects where aesthetics are important and elaborate or very specialized design is required. Examples would be churches, government buildings, schools, hospitals, and multistory office buildings. Figure 8 is a typical organizational chart for a design–bid–build project.

- *Advantages of design–bid–build:*
- Construction drawings and specifications are very complete and depict in detail how the project is to be constructed prior to breaking ground.
- The A&E is retained directly by the owner and represents his interest throughout the project.
- Owner has a good estimate of the final project cost before construction begins.
- A&E and GC are retained separately by the owner with clear divisions of responsibility.
- *Disadvantages of design–bid–build:*
- The entire process takes longer than other methods.
- It generally costs more than other methods.

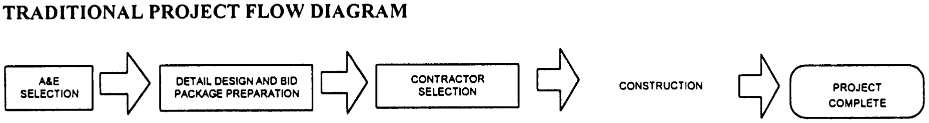

TRADITIONAL PROJECT FLOW DIAGRAM
A&E SELECTION
DETAIL DESIGN AND BID PACKAGE PREPARATION
CONTRACTOR SELECTION
CONSTRUCTION
PROJECT COMPLETE

Figure 7 Design–Bid–Build Method.

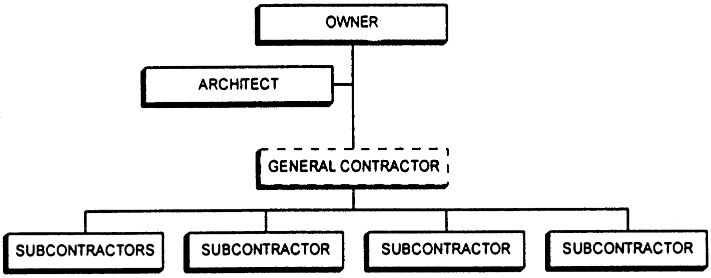


Figure 8 Organization Chart for Design-Bid-Build.

- The A&E and GC may develop an adversarial relationship, with the owner caught in the middle.
- It requires extensive owner involvement on an ongoing basis throughout the project.

4.1.2. Construction Management (CM method)

In the past 20 years, the technique known as construction management (CM) has evolved, flourished, and to some degree lost favor. It was touted as a method of speeding the construction process through fast tracking while giving the owner additional control of the project by eliminating the general contractor. Instead of awarding the project to a GC through a competitive bidding process or through negotiations, the owner hires a construction manager just as he hires an architect. The GC system is eliminated in favor of a system designed to have the architect and CM both work for the owner on a fee basis. A typical organizational chart for a CM project is shown in Figure 9.

The CM provides managerial services and acts as the owner's agent in construction matters. Rather than the owner contracting with a single GC, he contracts directly with multiple prime contractors and specialty subcontractors. The construction manager schedules, coordinates, and directs the day-to-day activities of these contractors. He generally does not perform any work himself, but he may provide essential services normally included in a GC's overhead cost, such as temporary facilities, utilities, cleanup, and security. These services are provided on a cost-reimbursement basis and are not part of the CM's fee.

The construction management system, which is depicted in Figure 10, is best suited to large, complex projects requiring substantial full-time, on-site support staffs. A typical CM staff for a major industrial or public works project might include engineers, accountants, purchasing agents, safety representatives, inspectors, and construction specialist. Extensive owner representation is also required.

In theory, the overall schedule is reduced by maximizing the fast-track approach. As drawings are completed, contracts are awarded and work begins. There is no doubt that the project duration can be reduced by overlapping design and construction. However, in a case where multiple contractors are proceeding simultaneously at breakneck speeds, disputes between contractors and the CM/owner are common. This gives rise to an unnecessarily formal, and often adversarial, project climate. The owner becomes an arbitrator of disputes in a no-win situation. No matter what he does, it will ultimately cost him money. This type of project is typically characterized by disgruntled contractors, continuous change order requests, claims for additional compensation, and sometimes litigation.

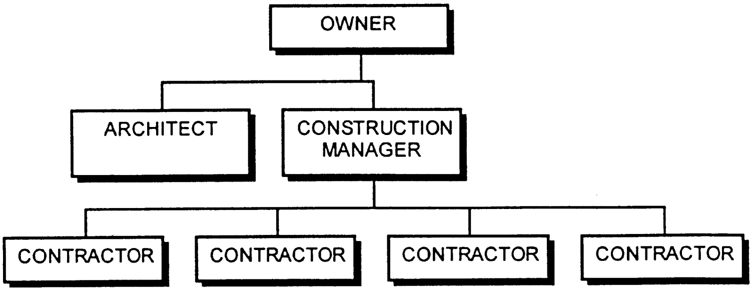


Figure 9 Construction Management Method.

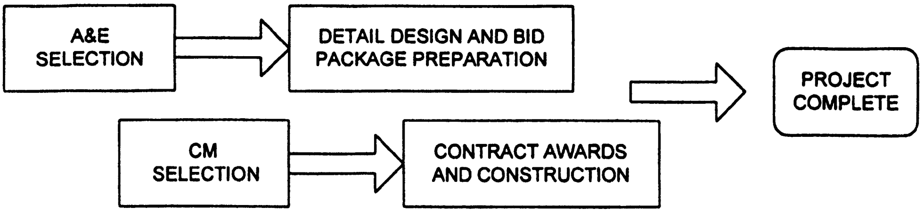

A&E SELECTION
DETAIL DESIGN AND BID PACKAGE PREPARATION
CM SELECTION
CONTRACT AWARDS AND CONSTRUCTION
PROJECT COMPLETE

Figure 10 Organization Chart for Construction Management.

Because the construction manager is retained on a fee basis, he does not warrant or guarantee results, and may have little incentive to keep the cost down. In contrast to a fixed price contract with a GC, here the owner must assume ultimate liability for the final cost of the project no matter what the circumstances.

- *Advantages of a CM system:*
- Total project duration is reduced by fast tracking.
- Reduces owner's staffing requirements for large or complicated projects.
- Architect and construction manager both work directly for the owner.
- *Disadvantages of a CM system:*
- Often leads to disputes between multiple prime contractors.
- Final cost of the project is not known until late in the process.
- Owner's exposure is increased since CM is retained on a fee basis.

4.1.3. Design–Build

The design–build method of project delivery, which is illustrated in Figure 11, has become increasingly popular as an alternative to the traditional (design–bid–build) process. Here the owner contracts with a single source for both the design and construction of the facility. The single source is usually a contractor who specializes in this type of construction. He may have design capabilities in-house or may subcontract this phase to an A&E firm. In either case, the design–build contractor assumes full responsibility for the adequacy of the design and its constructability.

The design–build project begins with the owner (or owner's agent) developing a RFP, including a functional bid specification. This document will include:

- Layout drawings
- Project overview
- Summary of key building characteristics
- Construction completion date requirements

The RFP package is issued to a group of prequalified design–build contractors. Bids are received and evaluated by the owner/agent, and a contract is issued to the successful contractor.

The design–build contractor begins the development of detailed building specifications that are submitted to the owner for approval. The detailed building specifications are completed in sequential packages, phased with the construction sequence. This allows construction to begin prior to completion of the total design process, reducing the total project lead time. A typical design-build project organizational chart is shown in Figure 12.

The most important advantage of design–build is that it can significantly shorten the total project duration by overlapping or fast-tracking design and construction without exposing the owner to change order requests, claims and litigation due to errors, omissions, or ambiguities in the plans or

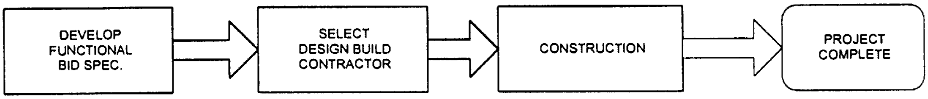


Figure 11 Design-Build Method.

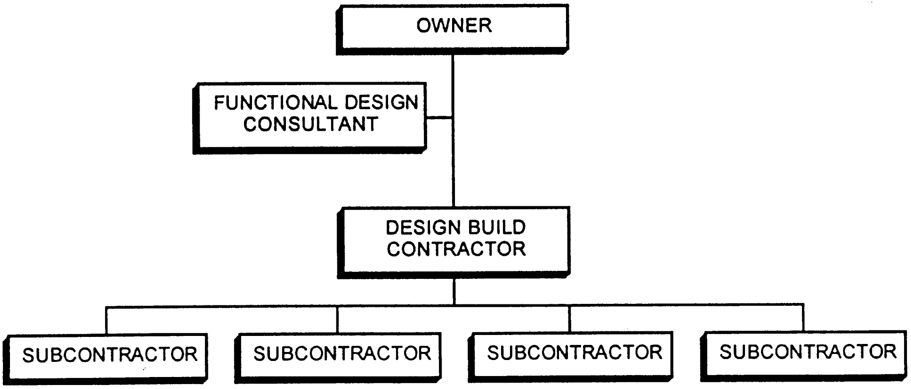


Figure 12 Organization Chart for Design-Build Method.

specifications. Likewise, the owner will not have to referee disputes between his architect and general contractor.

Design–build gives the owner a single source of responsibility and a single point of communications for all project-related issues, including schedule, budget, quality, and warranties.

A disadvantage of the system is that the owner loses the benefits of having an independent architect working directly for him. The designer now works for the contractor. His selection may have been based on price rather than qualifications and experience, and his primary objective may be to turn out a generic design quickly rather than interpret the owners needs and expectations. Drawings prepared by design–build contractors are typically intended for field use rather than owner review and are characterized by a lack of detail and much standardization. It can be difficult for an owner to visualize the details of the finished facility based on these drawings. It can also be difficult to coordinate the purchase, detail design, and installation of owner-furnished equipment and materials, using the information provided by the design–build contractor.

Since the owner does not communicate directly with the designer on a regular basis, his influence is diminished and his control of the details is lost. He also loses the advantages of having the architect act as an unbiased quality control auditor who inspects the work for compliance with plans and specifications.

- *Advantages of design–build:*
  - Gives the owner a single point of responsibility.
  - Minimizes change order request.
  - Promotes continuity between design and construction.
- *Disadvamtages of design–build:*
  - Less owner control—designer is working for contractor.
  - Loss of checks and balances.
  - Drawings may lack details.

4.1.4. Team Design/Construct

The team design/construct project approach, as depicted in Figure 13, is set up with separate contracts between the owner and the A&E and GC, like the design–bid–build process, but the detail design is completed in sequential steps like the design–build process.

The process begins with the owner/agent developing an A&E RFP that is submitted to qualified firms. The RFP can request either a fixed cost bid throughout the project or a fee for time plus expenses. After evaluating the bids and interviewing candidates, an A&E firm is selected and a contract is signed.

Next, the owner/agent and the A&E develop an RFP for the selection of a GC. The request for propsal includes:

- Layout drawings
- Project overview
- Summary of key building characteristics

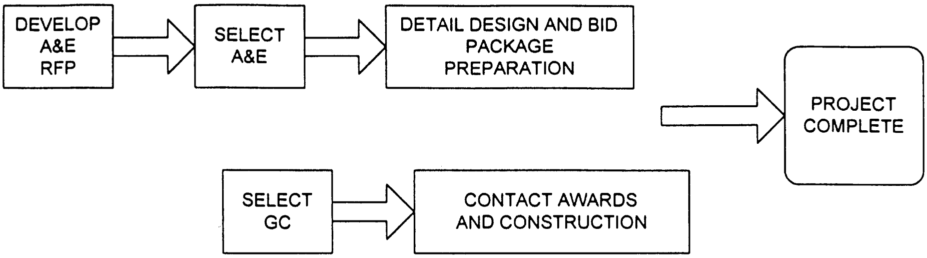

DEVELOP A&E RFP
SELECT A&E
DETAIL DESIGN AND BID PACKAGE PREPARATION
PROJECT COMPLETE
SELECT GC
CONTACT AWARDS AND CONSTRUCTION

Figure 13 Team Design/Construct Method.

- Construction completion date requirements
- The general conditions governing the project
- The general requirement specifications
- The project bid form

The bid form requests the GC to state the fee to cover the project's general conditions and requirements (this is essentially the project overhead) and a fixed profit percentage for the cost of the work. This essentially creates a cost structure of time and materials with an agreed-upon profit margin for the contractor.

Once all bids are received, the owner/agent and A&E evaluate the bids and select a GC to become a member of the project team. While the selection process is taking place for the GC, the A&E is concurrently developing the building construction documents. Six different performance specification bid packages are developed and released at different points in time in order to expedite construction of the project. Each of the packages is reviewed by the entire project team, including the GC. Once all parties agree on a performance specification bid package, the GC solicits fixed bids from appropriate subcontractors.

All subcontractor bids are evaluated and summarized by the GC and reviewed with the owner/agent and A&E. The team mutually selects the subcontractor that will be used. With this process, the owner still receives the cost advantage of a competitive bid and is directly passed all project costs. Once contracts have been awarded for all six packages, a guaranteed maximum price (GMP) is established. Generally, the contract with the GC is structured such that any savings between the GMP and the actual final cost is shared, 75% returned to the owner and 25% to the GC. This keeps the GC aggressively seeking ways to save the owner money. A typical project organizational chart for the team design construct process is presented in Figure 14.

- *Advantages of team design/construct:*
  - All primary team members are on board from the outset of the project.
  - The GC is involved in the design process, allowing valuable cost savings input early on.
  - Reduces project lead time by facilitating a fast-track approach.
  - All project costs are available to owner.
- *Disadvantages of team design/construct:*
  - Total project costs are not established at start of construction.
  - Advantages of the competitive bidding process are lost.

Owners must familiarize themselves with the pros and cons of the various methods of project delivery. Ultimately, he should choose the method that best meets his needs and that he feels the most comfortable with. After contracts are signed and long-term commitments are made, is not the time to discover that you have selected the wrong contracting method for your project. In summary, the four methods are listed in Table 1.

4.2. Selecting an Architect

A&E firms come in all sizes with varying areas of specialties. One may concentrate on retail facilities like shopping centers and department stores; others may specialize in low-rise office complexes, public school facilities, or industrial complexes. Staffs can range from 2 or 3, to more than 100. Large firms typically have in-house engineering specialty capabilities such as civil/site, structural,

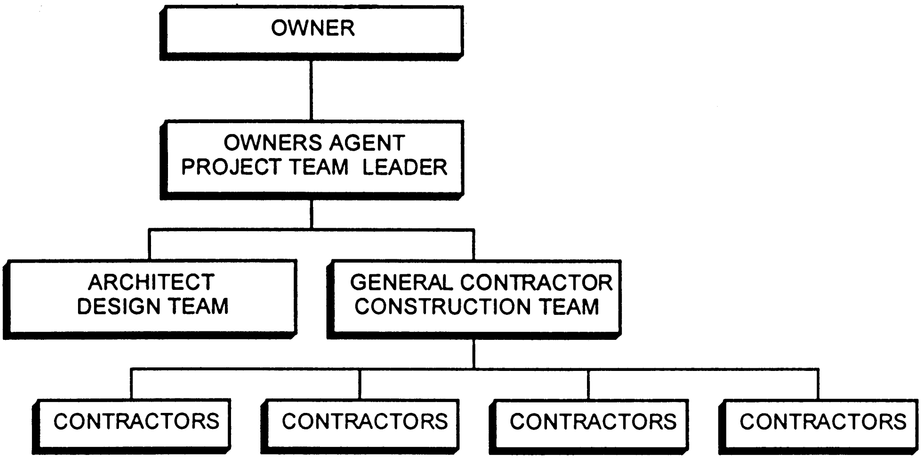


Figure 14 Organization Chart for Team Design/Construct Method.

mechanical, electrical, and so on, while smaller firms usually subcontract these phases to an outside consulting firm.

It is important that you find the right size and type of firm for your project. A very small firm is not capable of delivering the kind of service that a large or complex project requires. While the quality of the design work may be comparable, small firms just can't be as responsive as firms that keep the majority of the work in-house and have a large enough staff to work on multiple tasks at

TABLE 1 Methods for Project Delivery

	Design–Bid–Build	Construction Management	Design–Build	Team Design/Construct
Owner involvement	High throughout process	High throughout process since construction manager is retained on fee basis	Diminished influence since designer works for the contractor	Reduced because owner is represented by third party consultant or agent
Costs	Generally highest among the methods	High with no incentive to keep costs down	Reduced due to fewer change order requests; designs often generic or standard	Though not known up front, costs are managed through GMP process
Schedule	Generally longer than other three methods	Reduced by fast tracking	Reduced by fast tracking or overlapping	Reduced by fast tracking
Scope	Typically used in highly elaborate, aesthetic projects such as schools, churches, and hospitals	Seen in major public works or industrial projects	Not typical in projects requiring very specific, elaborate design	Seen in all types of projects

the same time. When the project gets started, any delay in completing design drawings on time, reviewing and returning contractor and supplier submittals, or responding to questions by any of the project participants can delay the progress of the work and have financial consequences for the owner. Likewise, if a firm has never designed a modern manufacturing or distribution center, it is unlikely that it will have the expertise that your project needs. Large firms may have the advantages of big staffs, but because of their size, workload, and number of clients, they may be unable to give your project the personal attention that a smaller firm can.

Start your selection process by compiling a list of potential candidates. Contact your local branch of the American Institute of Architects (AIA) to obtain a list of firms in your area who specialize in industrial facility design. Ask friends and business associates who have been through the process for their recommendations. Visit facilities in your area that you admire and find out who designed them. If there are firms in your area that specialize in the type of structure that you are planning, add them to your list.

From this initial list, interview only those firms that you think might be able to meet your needs. Consider the following when selecting finalists:

- Experience with similar projects (size, site conditions, functional complexity)
- Size of firm (is it the appropriate size for your project?)
- In-house capabilities (are they sufficient for your project?)
- Length of time in business (financial stability—will they be around next year?)
- Staff experience (experience of those assigned to your project)
- References (check references thoroughly)

Consider Company X: they're hoping to build a new, 1 million ft^2 facility to accommodate their distribution needs. Current management has never contracted out for an A&E firm, so they choose N Enterprises, basically out of the phone book. They're an A&E firm after all! Unfortunately, they have two architects and one structural engineer, each with three years of experience. They're working on two 2000 ft^2 residences currently. Company X fails to call but one reference: the engineer's mother, for whom they built an addition on her home. What's wrong with this picture?

Consider Company Y: they're hoping to expand their existing facility by 50,000 ft^2 to accommodate new manufacturing equipment. Their management wants ''only the best,'' so they hire the leading architect in the country. Although the firm is large and well known throughout the world, its reputation was built on design of large, high-profile office complexes, not industrial facilities. This firm will essentially have to start from scratch. With little previous experience in designing industrial facilities to draw from, the process is likely to take longer and be more costly. The final design is adequate but lacks imagination and shows that the firm really didn't have an understanding of company Y's operation and requirements. The firm is so large and has so many projects going on that company Y's project doesn't get the attention it deserves.

These are extreme examples of poor A&E selection. Keep in mind, though, that it is better to err on the side of caution and do your homework slowly. RFPs and selection matrices will help this process.

From the information gathered, develop a short list of three to five firms to send RFPs. In order to be able to compare responses, the RFP must clearly define the scope of services that the architect is to provide. Basic services include information gathering, preparation of preliminary schematics designs, detailed design development, construction document preparation, assistance in awarding contracts, and contract administrative services during the construction phase.

The RFP should also include:

- A description of the project approach (organization, methods and procedures)
- An overview of the project's objectives (why the facility is being built)
- Design criteria and constraints (what must be included to achieve objectives)
- Schedule requirements and constraints (must start and/or must complete dates)
- Budget parameters (accurate estimate of the intended budget)

Generally, the more relevant and detailed information you can include in your RFP, the more responsive the A&E proposals will be.

The RFP can request either a fixed cost bid, an hourly rate plus reimbursable expenses, or a combination of the two. Whatever method of compensation you choose, make sure you clearly understand the agreement and have an accurate estimate of the final cost of services.

The final selection of an architect should be based not only on technical competence, experience, organization, cost, and schedule, but also on your personal confidence in the firm's ability to meet your expectations. An example of an A&E selection matrix is included in Table 2; each firm is ranked

TABLE 2 A & E Selection Matrix

	Company A	Company B	Company C	Company D	Company E
1. Project understanding	5	5	3	5	5
2. Scope of services	5	4	2	4	4
3. Scope of qualifications	5	5	3	4	4
4. Experience with similar projects	5	3	3	5	3
5. Cost not to exceed	$352,000	$497,000	$430,000	$370,000	$295,000
6. Hourly rate	$65/hr	$70/hr	$89/hr	$53/hr	$61/hr
7. Project staffing	5	4	5	5	3
8. In-house staffing	5	5	5	5	2
9. Schedule	3	2	5	4	3
Total	33	28	26	32	24

on a 1 (poor) to 5 (excellent) scale on a predetermined set of criteria. Results of interviews and written proposals are considered in these assessments. Based on the numbers below, Company A and Company D were in a dead heat for the project. In the end, however, Company A was chosen due to its stronger showing in services, deliverables, and staffing. Price alone should never be the motivating factor in an A&E selection.

Make sure that you have met and feel comfortable with the people who will actually be working on your project. Finally, select a firm whose style, personality, and project approach are compatible with yours.

4.3. Selecting a Contractor

The traditional way of selecting contractors is through the competitive bidding process. This may result in an initial low price, but it does not necessarily result in selection of the best contractor for the project. Price is an important consideration, but other factors, such as experience with similar projects, financial stability and strength, number of years in business, reputation, annual volume, and safety record, must be considered before entering into a long-term binding agreement with a contractor.

For this reason, RFPs should only be issued to a short list of prequalified contractors that meet acceptable standards in all categories, any of which you would be willing to award the project to. Below is an example of a prequalification form that can be used to screen prospective contractors. Here again, we are using a rating system of from 1 to 5, with 5 being the highest possible score. This is only an example; you will need to evaluate what categories to include on your form and what weights to assign to each factor. However, at this stage in the selection process, emphasis should be placed on financial strength, experience, and reputation.

The role of the GC has changed significantly over the years. Where once the GC employed his own tradesmen to perform the work, today most work is done by specialty subcontractors: electrical, plumbing, roofing, and so on. The GC is predominantly a supervisor. He assembles a group of specialists, generally through the competitive bidding process, and then has the responsibility of scheduling and coordinating their activities to ensure that the work is performed correctly and completed on time. Subcontractor administration is the GC's primary concern once construction begins. GCs may subcontract more than 98% of the work to be performed on a project. For this reason, the owner must reserve the right of final approval of all major subcontractor and material suppliers. During final negotiations with GCs, the owner should review the credentials of all proposed subcontractors (e.g., electrical, HVAC, fire protection, etc.) and critical material suppliers such as structural steel fabricators. If there is any question about the ability of any of these participants to perform, the owner should insist that they be replaced.

Obviously, selecting the right contractor is essential to the success of your building project. The right contractor is the one that has the appropriate balance between stability, technical competence, experience, organization, cost and that intangible element: your personal confidence.

4.4. Project Management

Having decided what to build, where to build it and who will do the work, you might be ready to wash your hands of the whole thing and let the process run its natural course. This approach can be disastrous. Construction projects are dynamic creatures requiring constant monitoring and guidance. If left to its own device, the project can take a wrong turn before you realize what has happened.

TABLE 3 GC Selection Matrix

Item	Description	Weight	Rating	Score
1	Number of years in business under present name		3	0.15
2	Annual volume	8%	3	0.24
3	Bondable for this size project (prerequisite)[a]	10%	5	0.50
4	Financial strength	10%	3	0.30
5	Overall experience	10%	3	0.42
6	Experience with similar projects	16%	5	0.80
7	Home office location	5%	1	0.05
8	Qualifications of key employees	8%	3	0.24
9	Safety record	12%	4	0.48
10	Reputation and references	12%	3	0.36
	Total	100%		3.54

[a] *Eliminates contractor:* A performance bond is a guarantee by a surety to the owner that the contractor will complete the work in accordance with the plans and specifications. If the contractor defaults under the terms of the contract, the surety becomes responsible for the contractor's obligations and must complete the work in accordance with the terms of the contract.

Generally, performance bonds can only be obtained by contractors who have a documented history of financial stability and satisfactory performance.

Whether or not a bond is required from the successful bidder, requiring all potential contractors to furnish letters from a surety that they are bondable ensures that only financially sound companies will participate in the process.

Lost time or money spent is rarely fully recoverable. To avoid surprises, cost, schedule, and quality must be constantly monitored.

The nature of the construction process leads to natural differences of opinion among the participants. Everyday issues must be identified and resolved before they develop into problems. Timely decisions must be made so that the project can move forward as scheduled. Placing an activity on hold because approval drawings have not been returned or because the person who needs to make a decision is on vacation can disrupt the flow of a project and prove to be costly to the owner.

The secret to a successful project is effective project management. The individual who has the responsibility to represent the owner has more at stake than meets the eye. He must coordinate, facilitate and take responsibility for the successful completion of the project. No matter which method of project delivery has been chosen, the owner will have ongoing responsibilities, and must assume an active role throughout the building process. The owner must:

- Be represented at all planning and progress meetings
- Develop and maintain a master project milestone schedule
- Provide timely information to the A&E, contractor, or suppliers
- Develop and implement procedures for budget tracking and cost control
- Review payment requests and make recommendation for disposition
- Coordinate and expedite delivery and installation of owner-furnished materials and equipment
- Review drawings, submittals, and product samples as required
- Maintain documents, records, and logs as required to protect the owners interest
- Prepare and submit progress reports and schedule updates to top management
- Review and approve all change order requests

Completing a major project on time and within budget requires an extraordinary amount of time and effort on the part of the owner. Those with limited resources, who do not have experienced construction professionals on staff who can be assigned to the project, should consider securing the services of an outside consultant to act as their agent. An experienced consultant who is familiar with the strategic master plan and understands the role of the facility as a production tool rather than a structural box can be invaluable to the owner. Working as the owner's designated representative, he can move the project along smoothly from one phase to the next without interruption or delay while protecting the integrity of the functional design at each step. This ensures that contractors, suppliers, or other project participants do not make changes to the structure that will affect the process that the building has been designed to house.

For example, the placement of light fixtures, conduits, ductwork, or fire-protection lines below the designated clear height of the building can dramatically affect the designed manufacturing or warehousing process by restricting the travel of material handling equipment. Relocating utilities or other obstructions after the fact can be costly to the owner. Having a qualified representative on site during critical construction phases is a case of spending a penny to save a dollar.

The planning, design, site selection, construction, and start-up of a new or expanded distribution center or manufacturing facility is a linear process that follows logical steps from concept to completion. It is easy to understand the steps; it is not easy to manage the process effectively.

Do not underestimate the difficulties that will arise between concept and completion. Typically, millions of dollars of construction money are at stake, as well as the potential loss of revenues should the facility not be operational as scheduled, not perform as anticipated, or prove to be in the wrong location.

4.5. Summary Points for Construction

Let's review the steps to project success:

- *Establish realistic criteria for site selection:* When we say "realistic," we mean according to the needs identified in a strategic master plan (SMP). The SMP will identify and prioritize facility requirements for a given planning horizon, based on historical data and objective analytical projections. Establish needs and stick to the plan.
- *Select the contracting or project-delivery method that best fits your needs:* The generalized definitions of the four methods described in this monograph are intended only to introduce the reader to the process. Before committing to any method, the owner should have a clear understanding of the rules of the game and the roles and responsibilities of each player. Knowledge is power. Do research. Have unbiased consultants or third parties help you.
- *Select the best construction team for your project:* Your project is only as great as the skills and dedication of its team members. Do not get involved with unnecessary conflict and politics. Be careful; use quantitative and qualitative evaluations. Weight each criterion based on its unique importance to you. Your cautious diligence will pay off.
- *Vigorously manage the project:* You've done a lot of work up to this point; Do not let success slip away by neglecting the bricks-and-mortar phase. Allocate qualified staff to oversee the process, or select third-party agents or consultants to represent your interest during this critical stage. Wiping your hands of the construction process now is guaranteed to create difficulties down the road.

5. CONCLUSION

Developing a comprehensive plan that encompasses site selection and construction ensures that the organizational mission and goals will be addressed in the concept to reality process. Through a combination of awareness, research, reflection, and leadership, organizations can be assured that the property they choose and facility they ultimately build will increase return on assets, enhance competitive advantage, and provide the supply chain with the resources, processes, and methodologies that provide a true competitive advantage.

ADDITIONAL READING

Brockmann, T., "21 Warehousing Trends in the 21st Century," *IIE Solutions,* July 1999.

Brockmann, T., "Warehouse Slotting: Optimizing Item Layout," *Grocery Distribution,* September 1999.

Hudock, B., "Shelve It," *Operations and Fulfillment,* November–December. 1999.

Olsen, R., "Ready, Set, Plan," *Modern Materials Handling,* May 15, 1999.

Purkiss, M., "Planning for Flexibility," *Logistics Europe,* June 1999.

Schaffer, B., "Speed Up: WMS and Conveyors Crossdocking," *Materials Management and Distribution,* May 1999.

Sims, Jr., E. R., "Facilities Planning," in *The Warehouse Management Handbook,* 2nd Ed., Tompkins Press, Raleigh, NC, 1998, pp. 295–318.

Tompkins, J., *Revolution: Take Charge Strategies for Business Success,* Tompkins Press, Raleigh, NC, 1998.

Tompkins, J. A., and White, J. A., Eds., *Facilities Planning,* 2nd Ed., John Wiley & Sons, New York, 1996.

CHAPTER 56

Material-Handling Systems

Yavuz A. Bozer
University of Michigan

1. OVERVIEW 1502
2. CONTAINERIZATION AND UNIT LOADS 1503
3. BASIC MATERIAL HANDLING EQUIPMENT 1504
4. INDUSTRIAL TRUCKS 1505
5. CONVEYORS 1513
6. STORAGE SYSTEMS 1520
7. AUTOMATED SYSTEMS 1524
REFERENCES 1525

1. OVERVIEW

At a basic level, material handling is primarily concerned with the storage and movement of material (in various forms) in/through production and service systems such as factories, warehouses, distribution centers, cross-docks, container terminals, airports, hospitals, and similar mission-oriented facilities. Although the physical movement of material is perhaps the most visible aspect of material handling, as suggested by the following "right definition," material handling goes beyond that. Material handling is "providing the right amount of the right material, in the right condition, at the right place, at the right time, in the right position, in the right sequence, and for the right cost, by using the right method(s)" (Tompkins et al. 1996). Note that using the "right method(s)" includes safety and ergonomic considerations, especially when humans are involved directly or indirectly in the handling system.

It is interesting to note that some publications have used a similar definition (i.e., providing the right amount of the right material at the right time and place) in referring to the just-in-time philosophy of the Toyota Production System, which is now generally known as lean manufacturing. This overlap suggests that providing the right amount of the right material at the right time and place must occur at two levels. One is at a planning level, where decisions such as lot sizes, shipment frequencies, reorder quantities, production or delivery schedules, and so on are made, while the other is at an execution level, where actual product (or parcel) movement and tracking/control takes place. The former would generally fall under production and inventory control for manufacturing facilities (or scheduling and planning for service facilities), while the latter would fall under material handling for either type of facility.

Another interesting aspect of material handling is the fact that, in the context of lean manufacturing, material handling is viewed uniformly as waste. A driving force in lean manufacturing is the elimination of waste. If this is interpreted strictly as the elimination of all material handling, it becomes immediately obvious that no goods would be shipped from factories (in fact, all the incoming raw material and purchased components would accumulate at the receiving dock of a factory), all incoming and outgoing ships would wait indefinitely at container terminals, no trucks would be loaded or unloaded at cross-docks, and so on. Clearly, the intent is to eliminate waste by eliminating all movement and inventories that are not essential for completing the mission of a production or service system. Therefore, planning, engineering, and the successful day-to-day operation of the material-handling system is absolutely necessary to achieve efficiency while meeting the goals of a mission-oriented facility. Significant cost-savings and performance improvements can be realized by eliminating and/or simplifying material handling functions, implementing methods changes, selecting proper handling equipment, and eliminating unnecessary handling operations and inventories.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

With the increasing availability and use of automatic identification technologies (such as bar coding), many people have come to recognize that material and information flow together, or that material flow generates information and vice versa. Clearly, the two flows occur in different forms over different channels. Material flow involves movement of physical objects, while information flow involves movement of bytes or packets; material flow is handled via people, conveyors, lift trucks, hand trucks, and so on while information flow is handled through copper or fiber networks. However, as one type of flow occurs, it changes the state of the other or it triggers certain "events" for the other, and the two flows are in many ways intertwined.

Before automatic identification technologies became widely available, information flow was slow relative to material flow. Material would be moved (typically in large batches and/or over a period of time) before the information system was updated. This resulted in delays and in information that was too old, highly clustered, and often inaccurate to be useful, except for accounting or off-line, limited tracking purposes. Today, with the aid of automatic identification technologies (such as bar coding and voice recognition), used in conjunction with computer networks and communications technologies (such as radio frequency), information systems are often updated on a real-time (or near-real-time) basis and hence have become more accurate and more prominent/useful to a wider range of users, from decision makers to operations managers and end customers. This has also resulted in a higher degree of parallelism between information flow and material flow.

Due to rapid advances in computing technology, the information system has, at the same time, moved from a batch-oriented, centralized mainframe environment to a real-time-oriented, decentralized microprocessor environment, which makes it possible to store/retrieve larger amounts of information at multiple locations while developing more advanced logic to guide, optimize, or interpret material flow. (Despite an occasional misplaced parcel or lost luggage, most readers can readily identify with excellent examples of real-time information and material-flow systems used by overnight delivery providers and major airports.) While information technology and the hardware/software considerations for material-handling systems are well beyond the scope of this chapter, reference will be made to information flow/handling and its implications when appropriate.

2. CONTAINERIZATION AND UNIT LOADS

Perhaps one of the least understood, yet critical, aspects of material handling is the notion of containerization and unit loads. Many design drivers in materials handling, such as frequency of flow, load dimensions and weight, type of equipment that can or cannot be used, are often dictated by the type of container(s) and unit load(s) handled by the system. However, there is no clear-cut definition of a container or a unit load; a container in one application may very well be a unit load in another. Generally speaking, a container serves as a receptacle to keep individual/loose parts or packages together, confined in the same physical space, such as a cardboard box, a wire basket, or a tote box, among others. A unit load, on the other hand, is generally defined as the unit to be moved or stored at one time; it may consist of a single part or a bundle of parts/products, or it may consist of one or more containers. Also, the unit load includes the carrier or support needed to store or move the load. For example, a group of cardboard boxes (i.e., containers) stacked on top of a pallet would constitute a unit load, which includes the pallet itself and, say, stretch wrapping to stabilize the individual boxes.

Another example would be a 6-pack of soft drink cans. The container is the can and the bundle of six cans may be the unit load. Many people would find it difficult or awkward to carry six (loose) cans at one time. When the cans are held together with the white plastic holder, however, it becomes much easier to handle them. This underlines the significance of the unit load concept and the inclusion of support structures in its definition. Of course, a 12-pack carton is another example of a unit load. Multiple 12-pack cartons stacked on a pallet (as one might see in a grocery store) would be yet another example of a unit load. Clearly, multiple types of unit loads may flow through one or more systems as cans are bundled into 6-pack or 12-pack loads, which are then stacked on pallets.

The size and configuration of the unit load often determines how it can be moved and stored. For example, a forklift truck will typically be used if the unit load is palletized. If the unit load is chained, strapped, or stringed, however, a hook or similar lifting device, attached to a crane/hoist, may be used to lift and move the load. The configuration as well as the dimensions and weight of the unit load will dictate or rule out certain types of material handling equipment. Since it is not possible to treat the subject of unit loads and handling equipment in a comprehensive manner due to limited space, here we will show only some of the basic type of equipment typically used with palletized loads or with containers that have smooth surfaces, such as tote boxes, cardboard boxes, and the like. For more information, the interested reader may refer to Chapter 6 in Tompkins et al. (1996), among others.

The size and configuration of the unit load also determines how often it must be moved and stored. For example, if 100 boxes must be moved per hour from point A to point B, and a trip-based material-handling system is used (i.e., devices such as lift trucks must perform trips to move the loads), then 100 trips/hr would be required to move the boxes one at a time. However, if the boxes are palletized, for example, and moved in batches of 20 (i.e., a unit load consists of 20 boxes), then 5 trips/hr would be performed from point A to point B. In many cases, the speed of the handling device would not show a significant difference, whether it is moving 1 box or 20 boxes, as long as the unit load is within the weight/volume capacity of the device and standard safety precautions are followed.

Hence, from a material-handling perspective, it may appear that moving the boxes in batches of 20 is the preferred approach since the device(s) would have to perform only 5 trips/hr instead of 100 trips/hr. However, moving the boxes in batches of 20 also leads to a phenomenon known as transfer lot delay. That is, if the transfer lot size is 20 boxes, the first box at point A must wait for the remaining 19 boxes, the second box at point A must wait for the remaining 18 boxes, and so on, before they can be moved to point B. In many systems (including manufacturing systems), such a delay increases the time required to move the boxes through the system even if the throughput (i.e., the rate at which the boxes are moved through the system—in our case 100 per hour) remains the same. Such an increase in time is also accompanied by a proportional increase in the number of boxes in the system at any one time, which would be work-in-process, WIP, or work-in-progress, depending on the application.

The average time to move a box through the system (say, the cycle time per box) and the average level of WIP would be minimized if the boxes are moved one at a time, *provided there is enough material-handling capacity to move the boxes one at a time.* Whether there is sufficient capacity to achieve one-piece flow or not depends on the volume of flow, as well as the distance, from point A to point B. If the flow volume is high and the distance from A to B nonnegligible, then we may have no choice but to use a conveyor to achieve one-piece flow.

However, conveyors require a large initial investment, and once they are installed, they are difficult to modify or relocate. Hence, as is the case in lean manufacturing, one-piece flow is often achieved by reducing the distance from A to B to such a short distance that the material-handling function is essentially eliminated. That is, no handling device or conveyor is required; instead, either the human operator moves the part with him or her over a very short distance or simple slide/roll mechanisms are used to slide or roll the parts from A to B. This approach is generally known as setting up manufacturing cells to put machines A and B within very close proximity of one another, which works reasonably well for manufacturing applications. In other applications, such as warehouses and airports, however, there are cases where the cell analogy does not apply and the ideal solution may indeed be a conveyor or a trip-based handling system. Furthermore, in facilities such as warehouses and airports, the handling system is also often used to sort the loads, and for high-volume, high-speed sortation, conveyors may be the best solution.

Thus, material-handling decisions are often complex decisions that involve the configuration and size of the unit load(s), the determination of transfer lot size(s), the type of handling systems available (trip-based, conveyors, or robots), the volume of flow, the frequency of flow, and the distances involved. In the next section, some basic handling equipment is presented. We note that the material presented here applies largely to discrete-parts material flow systems, where there is a discrete unit of flow. In some applications, such as oil refineries and sugar-processing plants, the flow is a continuous flow and until the product is unitized (by putting it in, say, a barrel or package), there is no discrete unit of flow. Such systems generally fall under bulk material-handling systems, where flow is often measured as gallons/hr, tons/shift, and so on.

The reader may refer to Fruchtbaum (1988), Shamlou (1988), and Woodcock and Mason (1987), among others, for bulk material-handling. It is interesting to note that some systems may be a combination of bulk and discrete-parts flow. For example, in a soft drink bottling plant, the preparation or mixing of the drink may be in bulk form, but once the drink is filled into cans or bottles via the filling machines, the flow will be a discrete flow. Likewise, in a pasta factory, the preparation of the durum wheat would typically involve bulk handling, but once the dough is mixed and pressed into various forms and shapes through the dies, the remainder of the flow will be in discrete units. (Of course, individual pieces of pasta may still be treated as bulk flow until they are packaged.) Both of the above examples happen to involve food handling and preparation, which is often subject to more strict and specialized handling standards.

3. BASIC MATERIAL-HANDLING EQUIPMENT

In this section, we will present some basic equipment used in various material-handling applications. It is by no means a comprehensive list. Material handling is a dynamic world; new technologies, new equipment, and creative applications of the two continue to change the face of material handling. The reader is, therefore, encouraged to take advantage of the Web to get up-to-date information on material-handling technology, equipment, applications, and vendors. Today, most vendors in the United States have Web pages to describe their capabilities and display their equipment/systems. Many of them include material on information/computer systems, especially in those cases where automated equipment/systems are involved. The reader is encouraged to check, among others, the

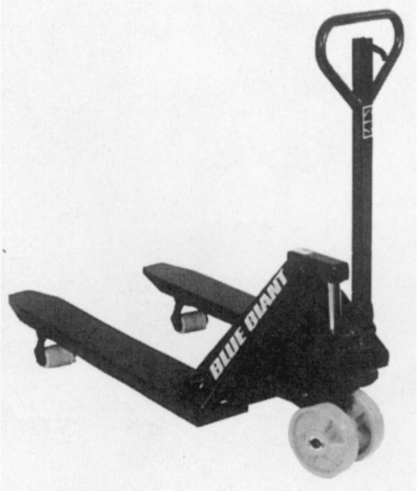


Figure 1 Pallet Jack. (Courtesy of Blue Giant Equipment Corporation)

website of Modern Materials Handling (a trade publication at www.mmh.com/info) and websites such as forkz.com (see www.forkz.com/Business/Industries/Industrial_Supply/Materials_Handling/ for a list of material-handling vendors under various categories).

4. INDUSTRIAL TRUCKS

There is a wide range of industrial trucks available. Going from simple to sophisticated and from wide-aisle to narrow-aisle trucks, a possible list would be as follows: (1) hand truck/cart, (2) pallet jack, (3) walkie stacker, (4) pallet/platform truck, (5) counterbalanced lift truck, (6) straddle truck, (7) reach truck, (8) sideloader, (9) turret truck, (10) storage/retrieval truck, and (11) order picker truck, among others. Examples of 2, 3, 5, 6, 8, 9, and 11 are shown in Figures 1, 2, 3, 4, 5, 6, and

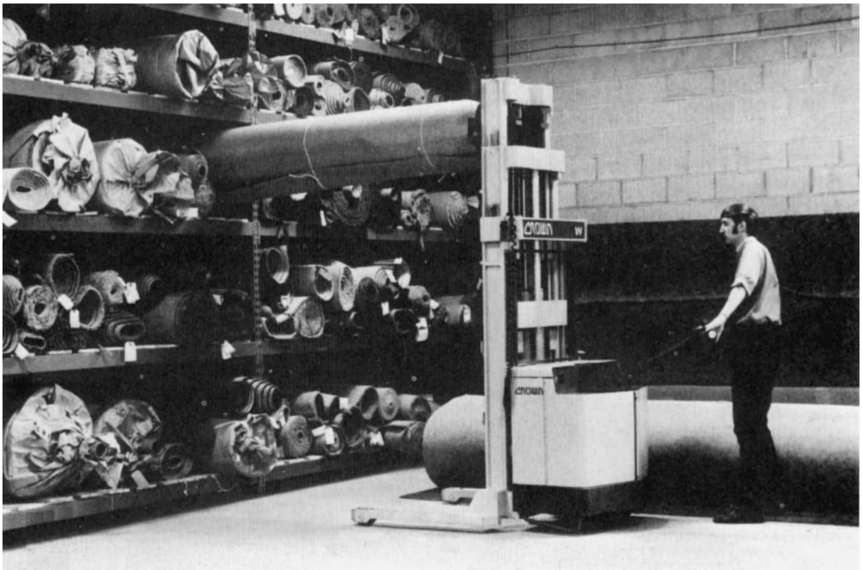


Figure 2 Walkie Stacker. (Courtesy of Crown Lift Trucks)

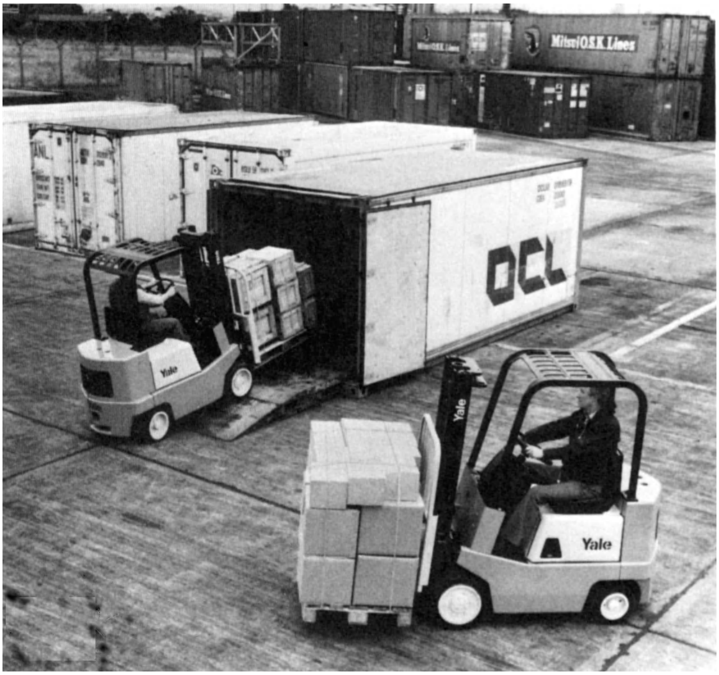
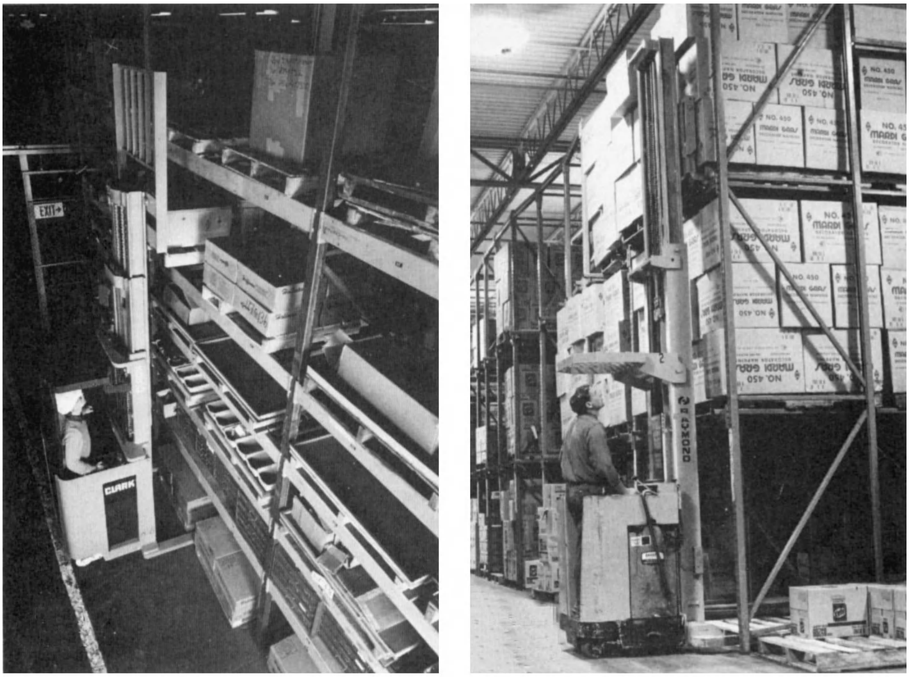

Mitsui O.S.K. Lines
OCL
Yale
Yale
Yale

Figure 3 Counterbalanced Lift Trucks. (Courtesy of Yale Industrial Trucks)


EXIT
CLARK
NO. 450
MARDI GRAS
RAYMOND

(a) (b)

Figure 4 Straddle Truck. (Courtesy of Clark Equipment Company and The Raymond Corporation. Raymond is a registered trademark of The Raymond Corporation.)

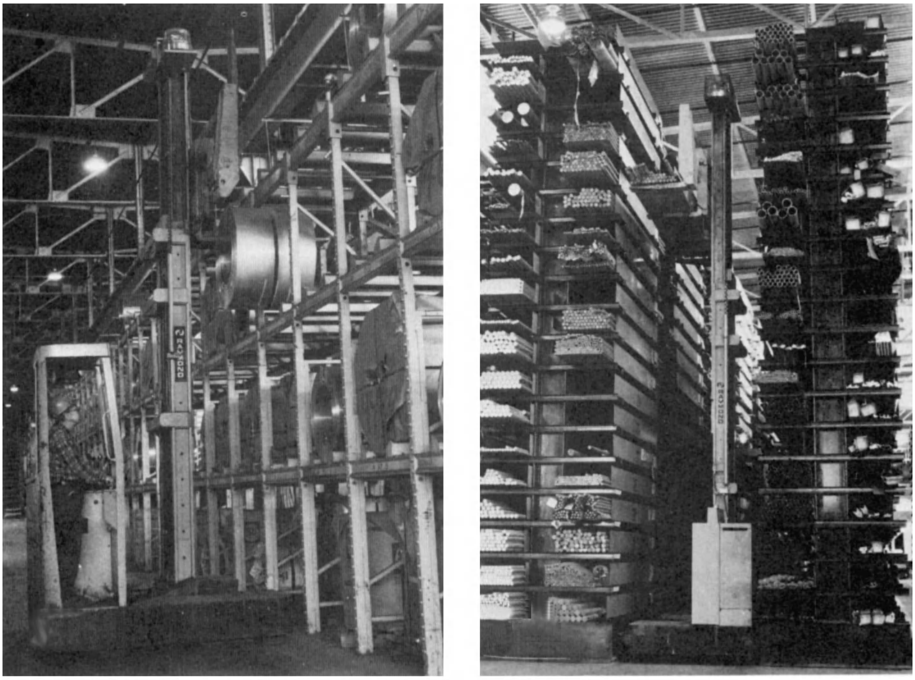

(a) (b)

Figure 5 Sideloader Trucks. (Courtesy of The Raymond Corporation. Raymond is a registered trademark of The Raymond Corporation.)

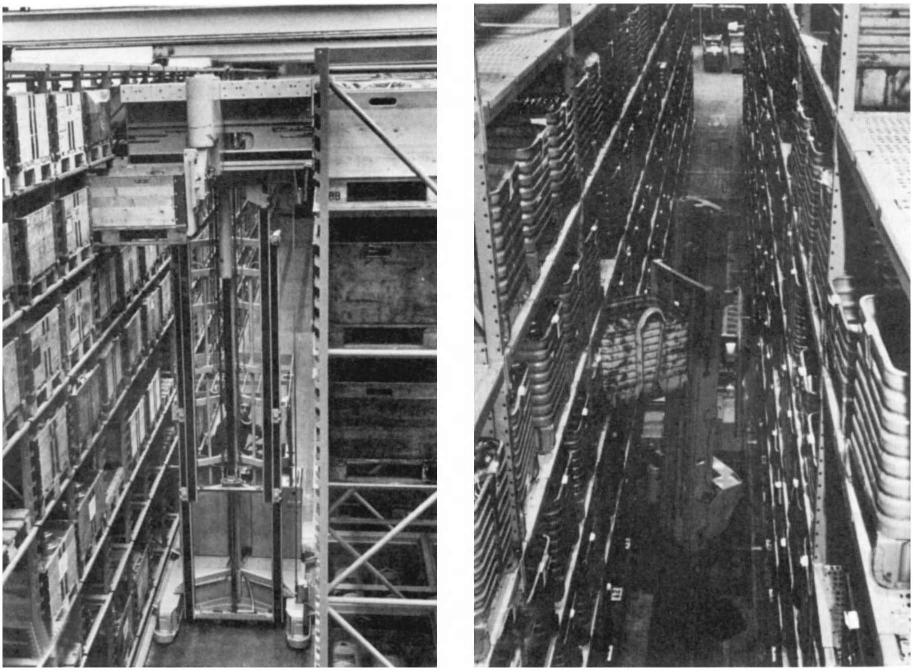

(a) (b)

Figure 6 Turret Trucks. (Courtesy of Lansing Bagnall and Drexel Industries, Inc.)

7, respectively. The simplest truck with a mast (which allows the load to be raised and lowered) is a walkie stacker. The most common truck found in industry, however, is the counterbalanced lift truck, which still requires wide aisles due to its turning radius, and the fact that the truck must face the rack to store or retrieve the load.

Due to its shorter size, the straddle truck has a smaller turning radius than the counterbalanced lift truck; however, outriggers ("long feet") are added to compensate for load weight. Also, the straddle truck still has to face the rack to store/retrieve a load, which results in wide aisles. A reach truck is very similar to a straddle truck except that it is has relatively shorter outriggers and it is equipped with a pantograph mechanism that allows the truck to insert/remove a load into/from the rack without moving the truck back and forth; the truck still has to face the rack, however.

To further reduce aisle width, the sideloader truck was introduced. This truck faces only one side of the aisle; aisle width is reduced by eliminating any turns inside the aisle. It is also especially suited for handling long stock such as lumber or bar stock. Its primary drawback, of course, is the fact that in order to serve the opposite side of the aisle, the truck must leave the aisle, turn around, and reenter facing the opposite side. Serving both sides of the aisle at the same time, while requiring no turns inside the aisle, became possible with the turret truck, which resembles a counterbalanced lift truck but is able to store/retrieve loads without facing the rack. This is accomplished by giving the turret truck a swing mast or swing forks: the body of the truck remains stationary (facing forward in the aisle) while the load is turned to face the rack and the operator remains at ground level. As rack heights increased to better utilize land and cubic space, it became necessary to raise the operator along with the load, which led to the storage/retrieval truck.

Another truck that raises the operator with the load is the order picker truck. However, unlike the above trucks which are designed primarily for unit load in and unit load out, an order picker truck allows the operator to start with an empty pallet (or magazine) and stack various containers or parts on it as he/she picks the appropriate parts/products from unit loads stored in the rack.

A summary of basic features of various industrial trucks, including cost, is presented in Table 1, adapted from (Tompkins et al. 1996). The reader should keep in mind that truck prices and features change over time; nevertheless, the information provided in Table 1 is useful for comparison purposes.

An industrial truck is basically used for moving loads in and out of storage (i.e., they perform a storage/retrieval function) and/or moving loads from one point to other point(s). Counterbalanced lift trucks, and similar low-mast trucks, are also used for loading/unloading trailers at shipping/receiving docks. In the first case, a fundamental question that arises is the throughput capacity of a truck; that is, how many storages and retrievals per hour can a particular truck perform? In the second case, a similar question posed in a slightly different fashion would be: given the throughput requirement (the number of loads that must be moved per time unit), how many trucks do we need? Both questions may appear simple on the surface, but they are not straightforward to answer.

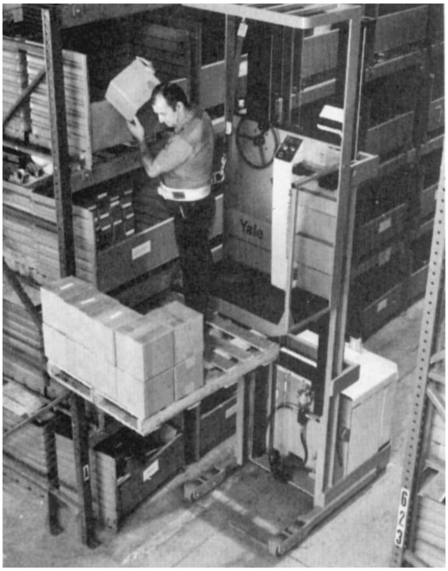


Figure 7 Order Picker Truck. (Courtesy of Yale Industrial Trucks)

TABLE 1 Features of Various Industrial Trucks

	Cost	Lift Height	Aisle Width	Weight Capacity	Lift Speed	Travel Speed
Counter balance	\$30,000	22 ft	10–13 ft	2–10K lb	80 fpm	550 fpm
Straddle	\$35,000	21 ft	7–9 ft	2–6K lb	60 fpm	470 fpm
Straddle reach	\$40,000	30 ft	6–8 ft	2–5K lb	50 fpm	490 fpm
Sideload	\$75,000	30 ft	5–7 ft	2–10K lb	50 fpm	440 fpm
Turret	\$95,000	40 ft	5–7 ft	3–4K lb	75 fpm	490 fpm

Consider first the storage/retrieval case. The throughput capacity of the truck depends on the rack size, the storage policy, the travel speed and type of truck, and the load pick-up/deposit times. Assuming a unit load in, unit load out operation, a truck will handle a load either on a single command (SC) or dual command (DC) basis. With SC operation, the truck starts at the input/output (I/O) point, which is typically located at the lower left-hand corner of the rack, travels to the appropriate rack opening, deposits (or picks up) the load, and returns to the I/O point. With SC storage, the truck picks up the load from the I/O point and returns empty to the I/O point. With SC retrieval, the truck travels empty from the I/O point to the rack, picks up the load, and brings it to the I/O point, where it is deposited.

To avoid empty travel to/from the I/O point, a truck can perform a DC cycle, provided that a storage request and a retrieval request are present at the same time. To perform a DC cycle, the truck starts at the I/O point, picks up the load to be stored, travels to an empty rack opening to deposit the load, then travels directly to the rack opening where the load to be retrieved is located, picks up that load, travels to the I/O point and deposits the load. Hence, two loads are handled on each DC cycle and empty travel occurs only from the storage point to the retrieval point.

The throughput capacity of a truck (operating within an aisle) depends on the expected time it takes to perform a SC and a DC cycle; it also depends on what fraction of time the truck performs a SC or DC cycle. Suppose the storage rack is H ft high and L ft long. Further suppose the truck's lift speed (or vertical speed) is v fpm and its travel speed (or horizontal speed) is h fpm. If the truck is capable of concurrent (i.e., simultaneous) travel in the horizontal and vertical directions, Bozer and White (1984) have shown that the expected SC travel time, say *E(SC),* and the expected DC travel time, say *E(DC),* can be computed from the following equations, assuming that randomized storage is used (i.e., a load is equally likely to be stored in or retrieved from any location in the rack) and the I/O point is located at the lower left-hand corner of the rack:

$$E(SC) = \left(1 + \frac{b^2}{3}\right) T \tag{1}$$

$$E(DC) = (40 + 15b^2 - b^3)\left(\frac{T}{30}\right) \tag{2}$$

where b is the shape factor and T is the scaling factor for the rack. Expressions for alternative I/O point locations can be easily obtained from Eqs. (1) and (2); the reader may refer to Bozer and White (1984).

The shape factor and scaling factor are computed as follows. Let t_h designate the horizontal travel time from the I/O point to the farthest end of the aisle, that is, $t_h = L/h$. Likewise, let t_v designate the vertical travel time from the I/O point to the top of the rack; that is, $t_v = H/v$. Note that $t_h(t_v)$ simply represents the length (height) of the rack in time. We then have:

$$T = \max[t_h, t_v], \tag{3}$$

and

$$b = \min\left(\frac{t_h}{T}, \frac{t_v}{T}\right). \tag{4}$$

In other words, T simply represents the longer (in time) side of the rack, and b is the shorter (in time) side of the rack divided by the longer (in time) side. Note that $0 < b \leq 1$, by definition. If $b = 1$, the rack is said to be square-in-time (SIT).

The above results are based on randomized storage, which is as an approximation of the closest-open-location (COL) rule used in industry. (Under COL, an incoming load is simply stored in the open location closest in time to the I/O point.) If the average rack utilization is high, the above expected travel times are reasonably accurate (not considering acceleration/deceleration of the truck). If the average rack utilization falls below 90% or so, however, the above expected travel times are likely to overestimate the actual expected travel times since the COL rule tends to favor rack openings close to the I/O point, while randomized storage always picks each opening in the rack with equal probability.

Concurrent travel is possible whenever the lift motor (vertical motor) is separate from the travel motor (horizontal motor). While concurrent travel is often limited when the forks are raised (due to safety considerations), most trucks follow a combination of concurrent and sequential travel while operating. A conservative approach would be to assume 100% sequential travel; that is, the truck first moves horizontally down the aisle, once it stops, then the forks are raised (for vertical travel).

With sequential travel, $E(SC)$ and $E(DC)$ can be computed from the following equations assuming randomized storage:

$$E(SC) = (1 + b)T \tag{5}$$

$$E(DC) = (1 + b)(1.333T) \tag{6}$$

where T is the scaling factor [Eq. (3)] and b is the shape factor [Eq. (4)] as before.

The total cycle time for a SC cycle (say T_{SC}) or a DC cycle (say, T_{SC}) is obtained by adding the load pick-up time (P) and the load deposit time (D) to the expected travel time. That is,

$$T_{SC} = E(SC) + P + D \tag{7}$$

and

$$T_{DC} = E(DC) + 2P + 2D \tag{8}$$

Often the load pick-up/deposit time is constant and $P = D$.

As an example, consider a rack 20 ft tall ($H = 20$) and 100 ft long ($L = 100$). The I/O point is assumed to be located at the lower left-hand corner of the rack. Suppose the truck's lift speed is 80 fpm ($v = 80$) and travel speed is 250 fpm ($h = 250$). The load pick-up or deposit time is equal to 0.15 min ($P = D = 0.15$). Using Eqs. (3) and (4), we obtain $T = 0.40$ and $b = 0.625$. That is, the rack's longer side is 0.40 min long and its shape factor is equal to 0.625. Assuming randomized storage and *concurrent* travel, we obtain $E(SC) = 0.4521$ min and $E(DC) = 0.6082$ min from Eqs. (1) and (2), respectively. Hence, for the expected cycle times we obtain $T_{SC} = 0.7521$ min and $T_{DC} = 1.2082$ min. Although the expected DC cycle time is longer than the expected SC cycle time, two loads (one storage and one retrieval) are handled on each DC cycle.

The throughput capacity of the truck (operating in one aisle) depends on the fraction of SC vs. DC cycles it performs. If an operation is defined as a storage or a retrieval, the throughput capacity can be measured in operations performed per hour. Suppose 50% of the operations in the above example are storages and 50% are retrievals. (In the long-term, of course, the two percentages would be equal; however, during certain shifts or segments of the day, the truck may perform more operations of one type.) Further suppose that 30% of the storages are performed on a DC basis (i.e., 30% of the storages are matched with a retrieval and the rest are performed on a SC basis).

If the throughput capacity is denoted by z operations/hr, we have $(2 \times 0.15z)(1.2082 \div 2) + (0.35z + 0.35z)(0.7521) = 60(0.90) = 54$ min/hr, assuming a 90% truck utilization. Note that 30% of 50% yields $0.15z$, which is the fraction of storage operations performed per hour on a DC basis; each storage operation takes $1.2082 \div 2 = 0.6041$ min when performed as part of a DC cycle. Also, each such storage operation is matched with a retrieval operation, which explains the multiplication of $0.15z$ by two. Likewise, 70% of 50% yields $0.35z$, which is the fraction of storage operations performed per hour on a SC basis. The same holds true for SC retrievals. Solving the above equation for z, we obtain $z = 76.30$ operations/hr as the throughput capacity of the truck at 90% utilization. A smaller value for truck utilization may be used to allow for downtime or maintenance, battery recharge or replacement, time for the truck to change aisles, and so on.

For comparison purposes, the reader may verify that, with *sequential* travel, assuming no other changes, the throughput capacity of the truck drops to 61.02 operations/hr. That is, in this particular example, the truck performs approximately 25% more operations per hour with concurrent travel. The number of trucks needed depends on the overall throughput requirement of the system. For example, if the system needs to perform 220 operations/hr, 220/76.30 = 2.88, or 3 trucks, would be required if concurrent travel is assumed. With sequential travel, 3.61, or 4 trucks, would be required. An additional allowance may be required for congestion if the trucks serve a small number of aisles and interfere with each other. Also, we assumed that all the loads are delivered to (and removed from) the I/O point (located at the end of each aisle) via another handling system. Additional truck travel time required can be easily computed if a single, central I/O point is used to serve all the aisles.

Consider next the case where trucks are used for moving loads from one point to other point(s). Given the throughput requirement (i.e., the number of loads that must be moved per time unit), and the layout/configuration of the points served by the trucks, we would like to know how many trucks are needed. It is assumed that one load is moved on each trip.

To determine the number of trucks needed, in addition to the throughput requirement and the location of the points, we need to specify a "dispatching policy," which basically determines which truck moves which load and when. There are a number of centralized and decentralized dispatching rules one can adopt. Centralized rules require a central dispatcher (or computer) to keep track of each load in the system waiting to be moved; such loads are also known as move requests. They also require a means of communication between the dispatcher and each device so that the appropriate device can be dispatched to the appropriate load. Decentralized rules tend to be less efficient but simpler and less expensive. With decentralized rules, each device generally follows a set of fixed instructions (or preprogrammed instructions) to decide which load to move next.

Specifying a dispatching rule is important because the dispatching rule often affects how much empty travel (or "deadheading") each device performs in serving the move requests. Depending on the flow patterns and the location of the points served by the devices, each device may have to perform a nontrivial amount of empty travel. The model we present below assumes first-come-first-served (FCFS) dispatching, which is a simple, centralized dispatching rule used in industry. Under FCFS dispatching, when a device becomes available (i.e., it delivers a load and becomes empty), it is assigned to the oldest move request in the system, regardless of the location of the oldest move request relative to the location of the device. If there are no move requests in the system when the device becomes available, then it idles at its last delivery point and waits for the next move request to arrive. When a move request arrives, if there is an idle device, it is assigned to the move request; if there are two or more idle devices, one of them is randomly picked and assigned to the move request. If there are no idle devices when a move request arrives, then it must wait until it becomes the oldest move request in the system and a device becomes available.

If the device becomes available at point i (or station i) and the oldest move request is located at point j (or station j), the device travels empty from station i to station j. Hence, when a device is assigned to a move request (i.e., when a device is serving a move request), the service time includes the empty travel time needed for the device to reach the move request. Obviously, there are several possible improvements to the FCFS dispatching rule. For example, when a device becomes available at station i, instead of the oldest move request, it could be assigned to the move request closest to station i. Such a rule, known as the shortest-travel-time-first (STTF) rule, is also used in industry, and it tends to reduce empty device travel relative to FCFS dispatching. However, STTF dispatching does not lend itself well to analytical modeling and therefore its use often requires simulation. In contrast, FCFS dispatching can be modeled analytically, and the results obtained with it would at least serve as a benchmark.

Assuming FCFS dispatching, suppose the flow data are given as a from-to chart, where f_{ij} represents the number of loads that must be moved per hour from station i to station j. (Recall that one load is moved on each trip.) Let $t^e_{ij}(t^f_{ij})$ be the empty (loaded) device travel time in minutes from station i to station j. Assuming that it takes P mins to pick-up and D mins to deposit the load, for simplicity we will assume that $t^f_{ij} = t^e_{ij} + (P + D)$. (The model we show below works with any user-defined t^e_{ij} and t^f_{ij} values.)

With FCFS dispatching, using the results presented by Chow (1986), the number of empty trips performed per hour from station i to j, say, e_{ij}, can be shown to be given by the following equation:

$$e_{ij} = \sum_k f_{ki} \left(\sum_k f_{jk}/F \right) \tag{9}$$

where F is the total loaded trips performed per hour, that is, $F = \Sigma_i \Sigma_j f_{ij}$. Note that the empty trips performed per hour from station i to j is proportional to the number loaded trips/hr ending at station i and the number of loaded trips/hr originating at station j. Each time the device completes a loaded trip, it becomes empty by definition. Hence, if more loaded trips end at station i, the device becomes empty more often at station i. Likewise, if more loaded trips originate at station j, more empty devices must be dispatched to station j. Equation (9) is also presented by (Egbelu 1987); however, no dispatching rule is specified. We note that the two terms shown in Eq. (9) can be multiplied to obtain e_{ij} only when the next load to be served by an empty device is independent of the current location of the device, which holds true under FCFS dispatching.

In Eq. (9), we allow the case where j is equal to i; that is, an empty trip may be performed from station i to station i, which implies that when a device delivers a load at station i, it is possible that the oldest move request in the system is located at station i. Depending on the exact locations of the pick-up and deposit points associated with station i, such an empty trip may or may not require a nonnegligible travel time.

Let $\alpha_f(\alpha_e)$ denote the total loaded (empty) device travel required (in min/hr). It is straightforward to obtain α_f and α_e as follows:

$$\alpha_f = \sum_i \sum_j f_{ij} t^f_{ij} \quad \text{and} \quad \alpha_e = \sum_i \sum_j e_{ij} t^e_{ij} \tag{10}$$

where e_{ij} is obtained from Eq. (9) and f_{ij} is user specified. Given the values of α_f and α_e, the number of trucks required, say, N, can be obtained as follows:

$$N = \frac{(\alpha_f + \alpha_e)}{(60 - t)u} \tag{11}$$

where t is the device unavailable time (expressed in min/hr) due to, say, battery replacement or scheduled breaks for lift truck drivers, and so on, and u is the target device utilization during the time that it is available. Of course, N is rounded up to the closest integer value.

The following example will demonstrate the use of the above model. Suppose four stations are to be served by a fleet of lift trucks. Suppose the from-to chart, that is the loads moved per hour, and the empty travel times (in min/trip) are given as follows:

From/to (loaded trips/hr)

	1	2	3	4
1	–	8	4	12
2		–	4	2
3			–	
4				–

Empty travel time (min/trip)

	1	2	3	4
1	0.0	1.0	0.8	1.5
2	1.0	0.0	1.2	0.5
3	0.8	1.2	0.0	1.0
4	1.5	0.5	1.0	0.0

Load pick-up and deposit are each assumed to require 0.50 min. For simplicity, loaded travel times are obtained by adding the pick-up and deposit times to the empty travel times.

The total number of loaded trips performed per hour (F) is equal to 30. Using Eq. (9), we obtain the following matrix for the number of empty trips performed per hour under FCFS dispatching:

From/to (Empty trips/hr)

	1	2	3	4
1	0.0	0.0	0.0	0.0
2	6.4	1.6	0.0	0.0
3	6.4	1.6	0.0	0.0
4	11.2	2.8	0.0	0.0

The reader may verify that the entries in the above matrix sum up to 30 trips/hr as expected. (Recall that each loaded trip is followed by an empty trip, although the duration of an empty trip may be negligible.) Also, we note that no empty trips are performed out of station 1 because no loads are delivered into station 1. (A device becomes empty only after it has delivered a load.) Likewise, no empty devices are dispatched to stations 3 and 4 because these two stations do not send any loads, which means they never request an empty device.

Using the loaded and empty trips performed per hour, from Eq. (10) we obtain $\alpha_f = 65$ min/hr and $\alpha_e = 31.64$ min/hr. That is, the total (loaded plus empty) travel time requirement comes to 96.64 min/hr. Of course, there are only 60 min/hr; hence the preceding result simply implies that we are going to need more than one lift truck. Assuming $t = 5$ min/hr and $u = 0.90$, from Eq. (11) we obtain $N = 1.952$, that is, two lift trucks are needed.

During its available time (i.e., during the 55-minute portion of an hour), the expected state of each lift truck can be broken down as follows: $(65/2)/55 = 59.09\%$ of the time traveling loaded (including load pick-up/deposit); $(31.64/2)/55 = 28.76\%$ of the time traveling empty; and the remaining portion of the time (12.15%) waiting idle for the next move request. The expected device utilization during a 55-minute period comes to $59.09\% + 28.76\% = 87.85\%$, which is less than the target utilization of 90% because two lift trucks are provided when the exact mathematical requirement was 1.952 trucks. Note that each device travels empty almost one-third of the time that it is busy $(28.76/87.85 = 0.3273)$.

One possible approach to reduce empty travel time (or deadheading) is to change the dispatching rule from FCFS to STTF. Generally speaking, STTF reduces empty device travel, and a system that meets throughput under FCFS will continue to meet throughput under STTF. However, with STTF, even if the system meets throughput and all the move requests are eventually served, some move requests placed by certain stations may have to wait longer than usual depending on the layout of the stations. This may occur especially if the stations are clustered. Note that under STTF, a device will continue to serve all the move requests in a cluster of stations, including those that arrive while it is busy, before it is dispatched to a move request placed by a station outside the cluster.

Another possible approach to reduce empty device travel is to eliminate dispatching and run each device according to a predetermined schedule. If there is little or no variance in the system, that is, if the arrival of the move requests can be predicted with reasonable certainty, then running the devices according to a schedule may be the best approach. In fact, in some systems, such as lean manufacturing systems, since each cell or station operates to takt time, the device routes, and each move they make, can be set up ahead of time. With shrinking unit load sizes, this approach has led to what is known as a ‘‘milk run,’’ where each device, on each trip, makes a preplanned number of deliveries to a particular set of stations at specific times. In many instances, ‘‘tuggers’’ with carts attached perform the milk runs. Such handling systems are known as tractor-trailer systems, which were the forerunners to automated guided vehicle (AGV) systems discussed in Section 7.

5. CONVEYORS

Conveyors are perhaps the oldest means of handling material. As is the case with industrial trucks, there is a wide range of conveyors available for different applications. They may take the form of a simple gravity chute (where a load placed at point A simply slides down to point B with gravity) or a sophisticated high-speed sortation conveyor that can sort hundreds of thousands of items per hour. Virtually any type of conveyor, in any type of industry/application, can be designed to handle virtually any type of material, including bulk materials (coal, sugar, etc.), loose items (nuts, bolts, screws, etc.), packaged goods (such as a men’s shirt or lady’s blouse), individual products (such as a light bulb or an engine block in a fixture), and palletized or containerized loads.

Due to limited space and extensive variations in design, only a basic set of conveyors are presented in this chapter. Such a list would include: (1) chute conveyor (Figure 8), (2) belt conveyor (Figure 9), (3) roller conveyor (Figure 10), (4) wheel conveyor (Figure 11), (5) slat conveyor (Figure 12), (6) Chain conveyor (Figure 13), (7) tow-line conveyor (Figure 14), (8) trolley conveyor (Figure 15), and (9) power-and-free conveyor (Figure 16).

Belt conveyors come in a wide variety of forms. The simplest form is a flat belt conveyor, which is often roller supported (i.e., there is a series of rollers placed underneath the belt). Belt conveyors are typically driven by a drive pulley/roller connected to an electric motor. For conveying items that may not be containerized, however, belt conveyors can also be of the slider-bed-supported type, where a flat metal surface is placed underneath the belt. An everyday example of the slider-bed supported belt conveyor can be found at the check-out lanes of most grocery stores.

An important characteristic of belt conveyors is that items or containers placed on such conveyors retain their relative position as they are conveyed, unless, of course, one of the items is intentionally

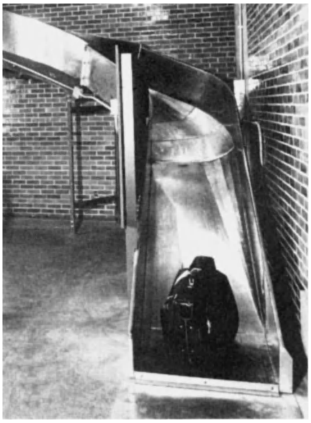


Figure 8 Chute Conveyors. (Courtesy of Standard Conveyor Company)

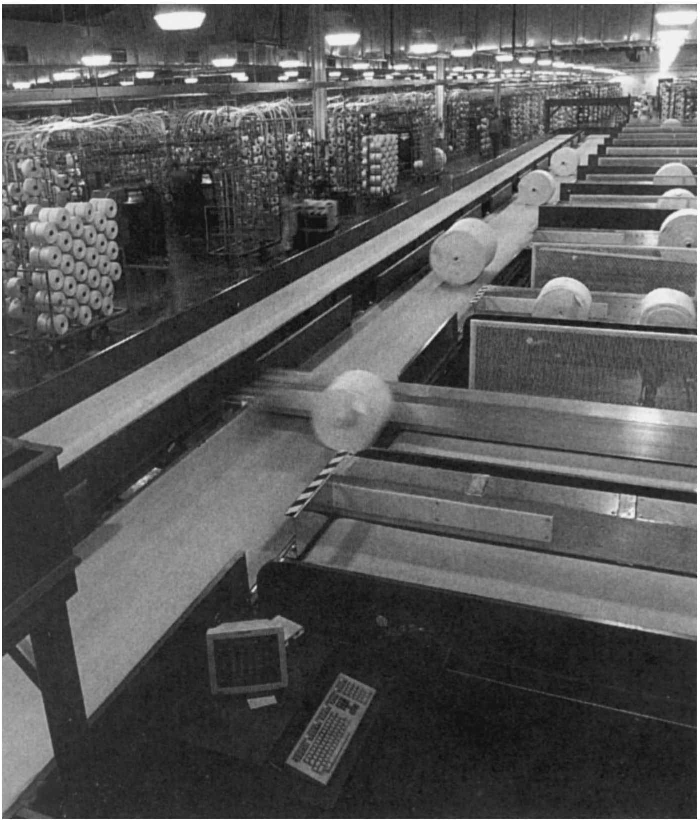


Figure 9 Belt Conveyor. (Courtesy of Litton UHS)

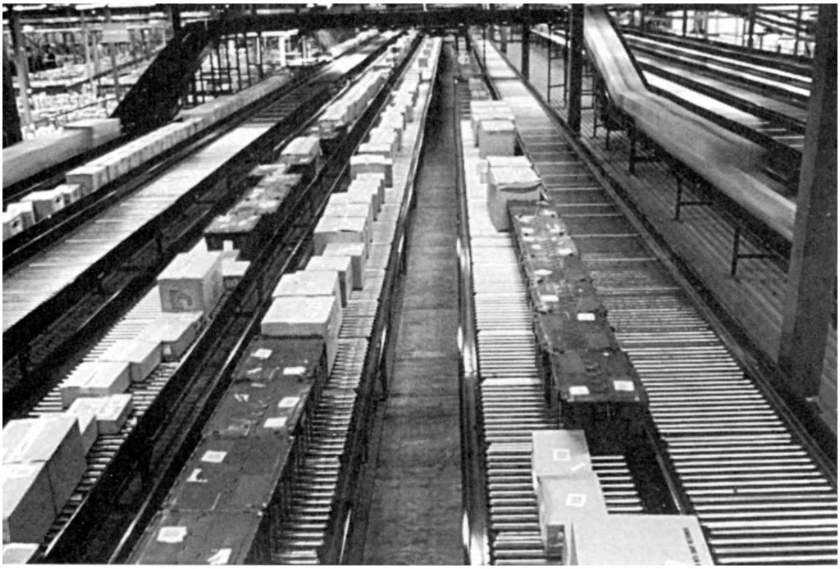


Figure 10 Roller Conveyor.

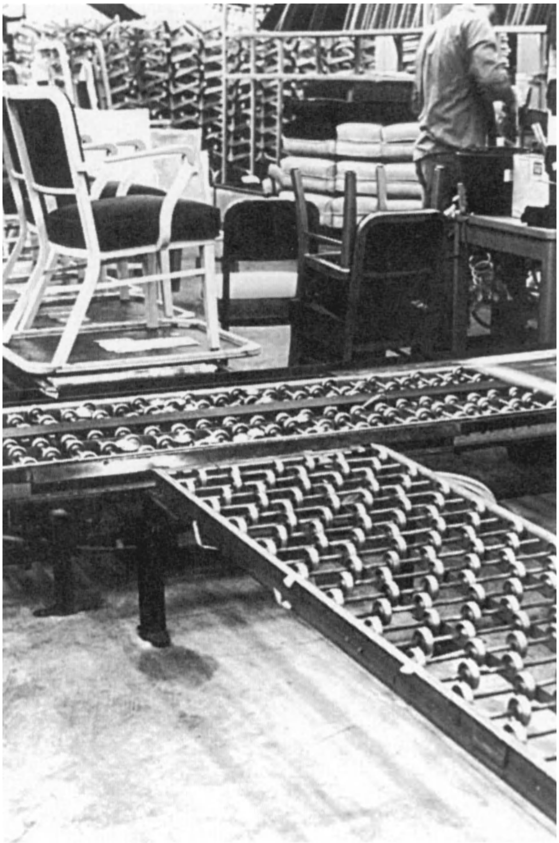


Figure 11 Skate Wheel Conveyor.

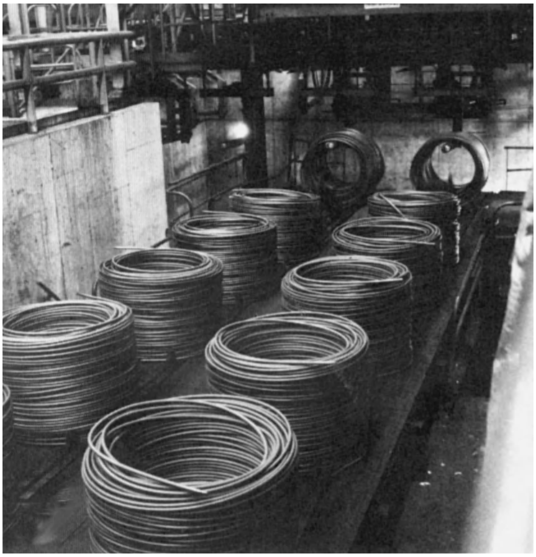


Figure 12 Slat Conveyor. (Courtesy of Acco Babcock)

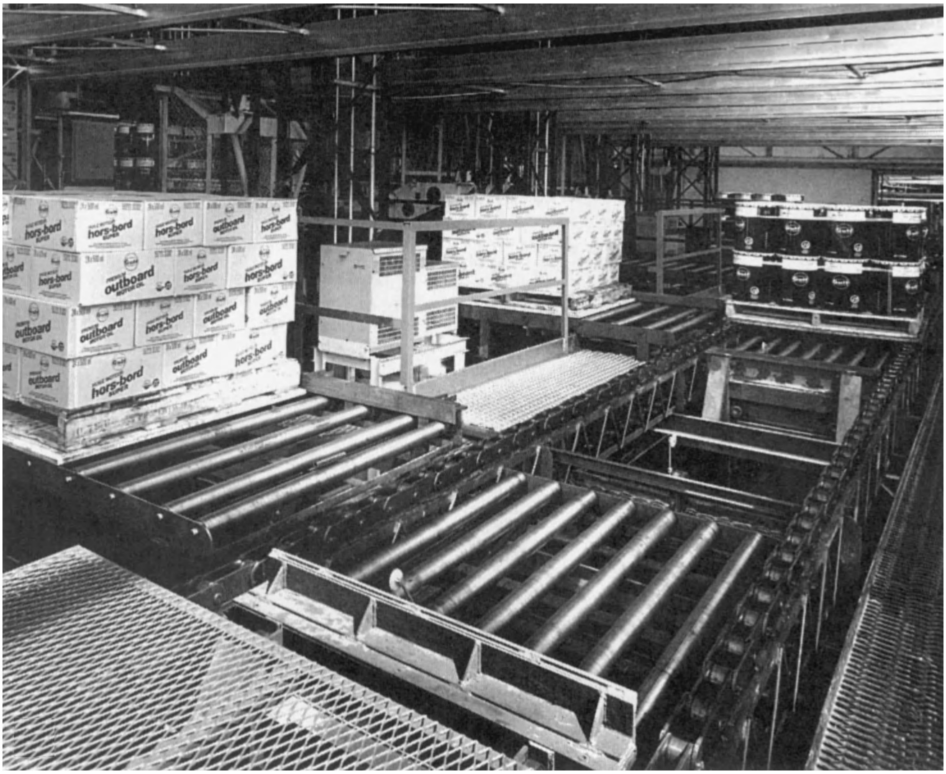


Figure 13 Chain Conveyor. (Courtesy of Jervis B. Webb Co.)

stopped or removed. Belt conveyors can also be used for moving loads between floors; this may be done with a belt incline or a spiral belt conveyor, which consists of sections of powered belt conveyors spiraling up or down. Depending on the application, there is a large variety of plastic belts one may use for horizontal conveying and belt inclines.

Telescoping belt conveyors have been used successfully at shipping/receiving docks to load/unload trailers. (The belt conveyor telescopes into the trailer to minimize walking distances during manual load/unload operations.) In addition, (troughed) belt conveyors are used extensively in handling bulk materials horizontally as well as up/down an incline.

Roller conveyors also come in a wide variety of forms and are used in wide ranging applications. Typically, a roller conveyor is driven either by a belt (placed underneath the rollers), a drive shaft (that runs along the conveyor underneath the rollers, the rollers are connected to the drive shaft with industrial-grade rubber bands), a chain-and-sprocket mechanism (a sprocket is attached to the end of each roller, the chain runs along one end of the conveyor), or an electric motor placed inside the rollers themselves.

Roller conveyors are most useful when the load to be conveyed has a smooth conveying surface. (Loads with irregular conveying surfaces would be better suited for belt or slat conveyors.) As a rule of thumb, for load stability, the roller spacing on the conveyor should be such that the smallest load conveyed is supported by at least three rollers at all times. Otherwise, the load may rock back-and-forth as it travels on the conveyor. In addition to conveying containerized loads, roller conveyors can be used for accumulation, which is a critical function required in some conveyor systems. Zero-pressure-accumulation is a type of technique often used with roller conveyors to accumulate many loads without buckling (i.e., without loads falling off the conveyor as more loads arrive at the accumulation point).

Wheel conveyors (or skate wheel conveyors) are relatively inexpensive because they are not powered (i.e., gravity or a human's push is often used to move the loads). They are quite cost effective in moving light to medium-weight loads over short distances. A collapsing variety, which can collapse and expand and has legs with adjustable height, has been used at shipping/receiving docks to load/

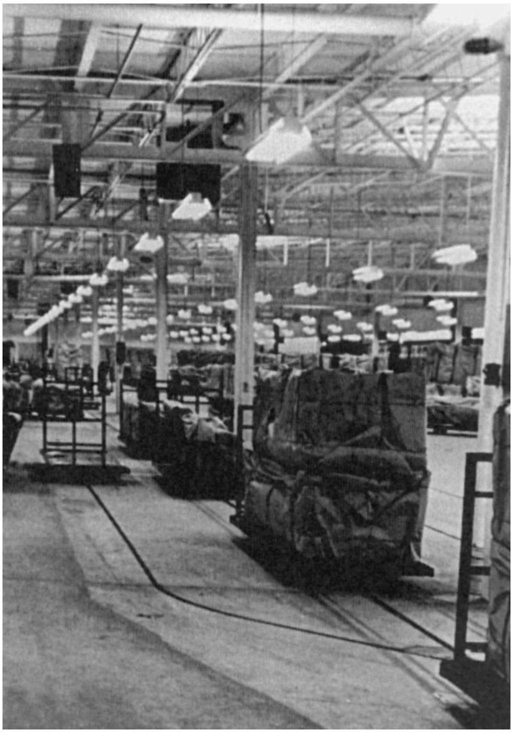


Figure 14 In-floor Tow-Line Conveyor.

unload trailers. The conveyor is stretched into the trailer to minimize walking distances during manual load/unload operations.

A slat conveyor consists of individual slats attached to a chain (or other driving mechanism) that runs underneath the conveyor. Often the slats are made of wood or steel. Slat conveyors, similar to belt conveyors, provide a smooth conveying surface to move heavy and/or irregularly shaped loads that may or may not be containerized. Most conveyors used in the baggage claim areas of airports can be classified as slat conveyors.

Chain conveyors are often used for moving medium-weight to heavy loads over short distances. They are also used for performing 90° transfers, where a set of chain conveyors pops up (or moves up) from underneath the load to lift it an inch or so, moves it laterally off of one conveyor and onto another one, and then moves back down to release the load.

Tow-line conveyors consist of a drive mechanism (often a chain) that may be buried in the floor (in-floor tow line conveyor) or supported by an overhead track (overhead tow-line conveyor). Hooks or dogs attached to the chain move the load forward when the chain is powered. Such motion can also be described as synchronized motion because all the loads attached to the conveyor move at the same time and at the same speed.

Tow-line conveyors generally offer a cost-effective means of moving loads (often placed on carts with nonpowered wheels) over long distances with no human operators. Each cart being pulled by the conveyor can be preprogrammed manually to be diverted off the conveyor automatically at specific points (marked on the floor, often by magnets). Hence, a cart manually hooked onto the conveyor at point A can be diverted off the conveyor at point B or point C with no human intervention. Tow-line conveyors, especially the in-floor variety, are also used in automotive manufacturing, where the automobile bodies are moved through the paint and assembly areas. Such tow-line conveyors are also known as ‘‘drag chain’’ conveyors.

Trolley conveyors, which are often used for moving loads overhead, are very similar to tow-line conveyors. A chain is placed inside an overhead track, which may be I-shaped or U-shaped. The carriers are attached to the chain via trolleys. When the chain is powered, the carriers move forward. Based on the product/part being handled, there is a wide range of designs used for the carrier, or

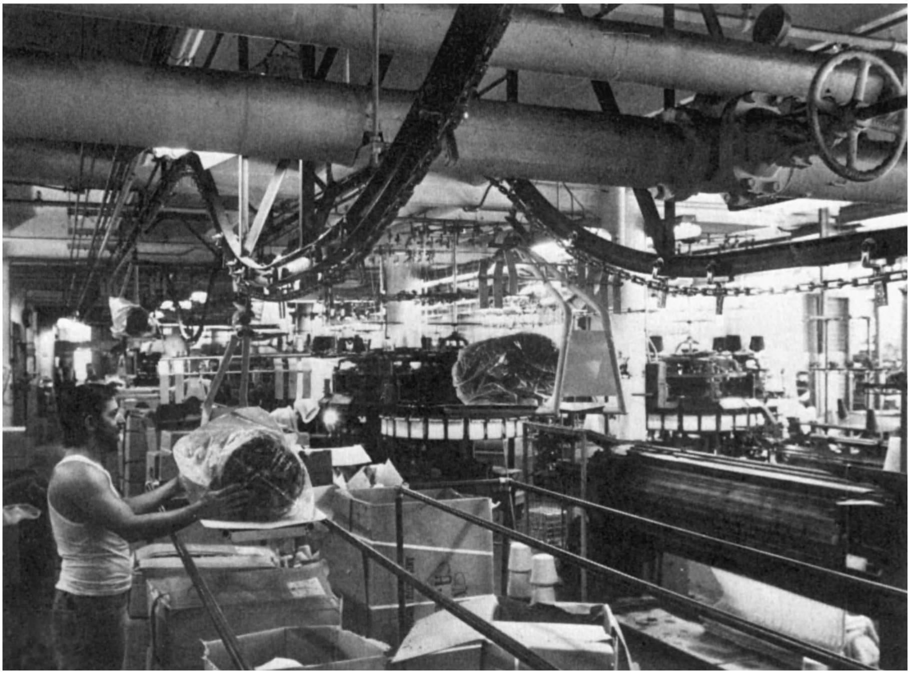


Figure 15 Trolley Conveyor. (Courtesy of Rapistan Systems)

hook attached to the carrier. Again, the type of motion provided is synchronized motion. Trolley conveyors have been used extensively in manufacturing, and similar applications to move parts from, say, the paint department, to workstations and assembly stations where parts are needed. They often travel overhead but can dip to deliver the parts to human operators.

Power-and-free conveyors have also been used in manufacturing and similar applications. Often they are installed overhead, although floor-supported versions (known as inverted power-and-free) are also available. A power-and-free conveyor consists of two tracks: a power track, which is similar to a trolley conveyor in design, and a free track. The carriers or hooks are placed on the free track via trolleys. When the dog on the power track catches (or engages) the trolley on the free track, the load moves forward. To stop motion (or to divert the load), the free track moves away from the power track and the dog disengages. Hence, merge and divert points can be created by using a power-and-free conveyor. Also, loads can be accumulated by disengaging the dog. The power track may run directly above the free track, or the two tracks may run side by side. Note that a power-and-free conveyor provides synchronized motion as long as the dog is engaged with the carrier in the free track.

Conveyors that provide asycnchronized motion are more advanced in design. One example is the cart-on-track conveyor and its variations, which allow the carts to be stopped or moved at different speeds on the conveyor at the same time. (The ‘‘cart’’ is often a fixture to hold the part.) Monorails, which strictly speaking are not classified as conveyors, also provide asynchronized motion since each carrier on a monorail is individually powered through an electric motor attached to the carrier. Power for the electric motor is received from the track, which also serves as a conduit to serve information to/from each carrier on the track. By moving sections of track vertically or horizontally, monorails can move individual carriers between floors, or they can route individual carriers in different directions (i.e., a carrier arriving at an ‘‘intersection’’ may continue straight through, or it may be diverted off to another segment of the track).

There is also a family of special-purpose conveyors known as sortation conveyors. (Interestingly, the word ‘‘sortation’’ is not in the English dictionary, but it has been used in the material-handling literature for a long time.) As the name implies, the primary function of a sortation conveyor is to sort incoming (mixed) items into specific lanes based on specific characteristics of each item. For example, if incoming items represent shipments to various retail stores, then a sortation conveyor can be used to sort all the items so that items destined to store A are diverted into lane A, items

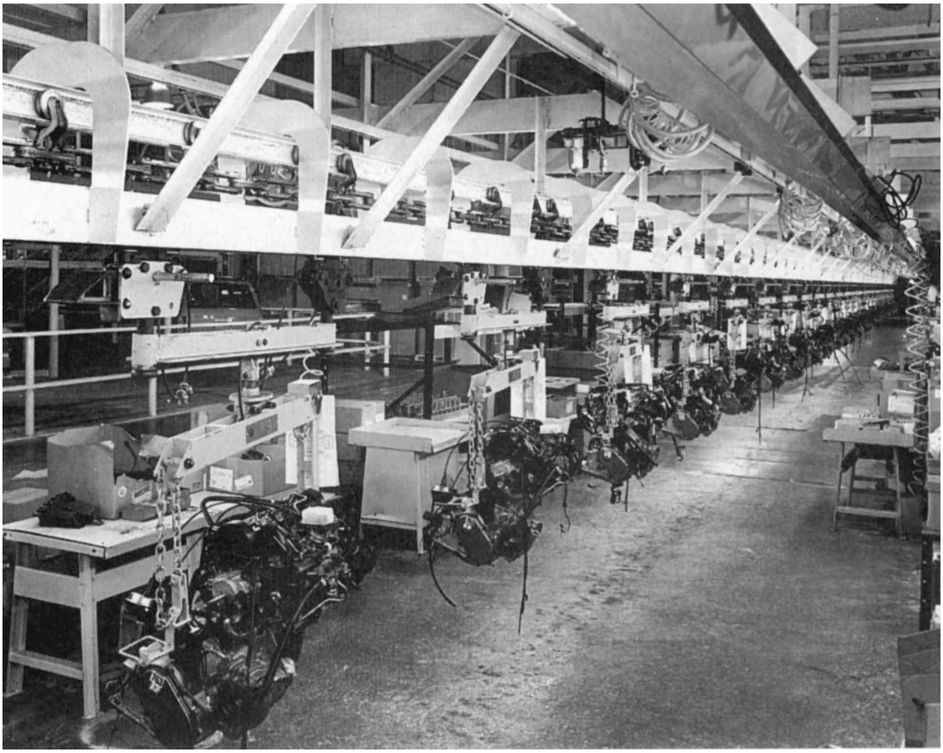


Figure 16 Power-and-Free Conveyor. (Courtesy of Jervis B. Webb Co.)

destined to store B are diverted into lane B, and so on. Naturally, sortation conveyors play a significant role in distribution centers and similar facilities where a large number of items must be sorted by customer, or by destination, or some other user-specified criteria.

With sortation conveyors, the items to be sorted are moved in a fashion similar to those conveyors described above, but their main distinguishing characteristic is the fact that loads can be automatically diverted off at specific points and at high speeds. (As described earlier, loads can also be diverted off at specific points with an in-floor tow-line conveyor, for example, but the tow-line conveyor would be exceedingly slow for most sortation purposes.)

Items to be sorted enter the sortation conveyor at a point known as the ‘‘induction point.’’ (There may be multiple induction points.) In high-volume systems, the induction point is typically fully automated; the item is automatically identified (often via bar coding the item and using laser scanners) and it is automatically ‘‘inserted’’ into the sortation conveyor. The insertion mechanism depends on the type of sortation conveyor used. For example, most sortation conveyors (except for tilt tray and cross-belt sorters, described below) require a minimum amount of clearance (or physical distance) between the items to be sorted; this clearance is automatically inserted at the induction point, which releases the items with the proper clearance between them.

There is a wide variety of sortation conveyors available. Naturally, the sortation rate (number of items sorted per minute) depends on the type of conveyor used. Generally speaking, although it is often not up to the user, it is easier and less costly to achieve high sort rates if all the items are uniform in size and weight and they are packaged/containerized in a consistent fashion. The sortation function slows down or becomes more sophisticated as more variance in item weight/size (or packaging) is encountered.

Basic types of sortation conveyors, their method of operation and basic characteristics (including sort rates), are presented in Table 2 (Kulwiec 1999). As can be seen in Table 2, sort rates may be as small as 20 items/min (or 3 sec per item) and as large as 670 items/min, which is less than 0.1 sec per item! Among the conveyors described in Table 2, the sliding shoe conveyor and the cross-belt sorter are sortation conveyors that have been introduced relatively recently.

TABLE 2 Sortation Conveyor Equipment Guide (Kulwiec 1999)

Sortation Device	Operation Details	Application Notes
Deflector	Stationary or movable arm deflects carton from a conveyor to another line or chute. Usually does not contact conveyor belt or rollers.	Low sorting rate (20–100 sorts per min). Handles up to 150 lb loads. Can be used for bidirectional sorting.
Diverter (pusher)	Moving arm with paddle sweeps across conveyor to push carton or item off opposite side. Pusher face retracts until next divert.	Sort rate is 20–100 sorts per minute (spm). Unit design allows for close grouping of spurs and for dual side sorting.
Pullers (rake)	Rake puller uses rows of chains equipped with tines to lift and pull off carton.	Can be used where space limits installation of a pusher. Sort rates to 80 spm; load capacity to 300 lb.
Pullers (pop-up belt and chain)	Pop-up devices rise up between rollers to change direction of item.	System handles heavy, durable loads, usually built into live roller conveyor. Has bidirectional capability.
Pop-up (wheels)	Powered wheels, usually skewed, rise up to contact bottom of carton and convey it to a spur.	Maximum load to 300 lb, sort rates of 60–150 spm. Gentle impact on load.
Pop-up (rollers)	Powered rollers rise up between conveyor chains or rollers to lift carton above conveyor surface and propel it off to the side.	Provides good load orientation. Sort rates to 150 spm; loads to 500 lb.
Sliding shoe	A series of sliding shoes (or moving slats) slide across conveyor surface to contact product and move it off the main conveyor.	High speed, up to 200 spm. Separates single line of items into multiple discharge lines.
Tilting devices (tilt tray) (tilt slat)	Cartons or hinged products are carried in hinged trays that can tilt in either direction at discharge point. Slat type carries items on flat surface; slats tilt up to effect discharge.	Various products and shapes can be handled. Sort rates range to 365 spm. System operates with elevation changes.
Cross-belt	Product is carried on short belt segments mounted 90° to conveyor line travel. Belt carries off item at discharge point.	Very high sort rates, to over 670 spm. Large number of divert points can be handled. Continuous-loop or train-style operations.

Source: Kulwiec 1999.

In addition to those conveyors listed in Table 2, special-purpose sortation systems have been developed for overnight delivery companies and the United States Postal Service (USPS) for mail/parcel handling. Although a very small fraction of parcels/mail may be lost, delayed, or misdirected, such systems have been used successfully for many years to sort mail and parcels that show considerable variation in size/weight. Given the increasing volume of domestic and international air travel, sophisticated sortation systems have also been developed for some of the world's largest airports, where thousands of pieces of luggage must be moved and sorted among a large number of arriving and departing flights every day.

6. STORAGE SYSTEMS

Various types of racks have been developed to store loads of different sizes and shapes. One of the simplest storage systems is block stacking, where loads (which may or may not be palletized) are simply stacked on top each other with no separate support structure. While block stacking is inexpensive and provides dense storage (i.e., little or no space is lost due to air between the loads), its use is limited in practice because only certain types of loads (such as appliances or furniture packed properly in cartons) can be stacked without crushing/damaging the loads at the bottom and access to individual loads (i.e., load selectivity) is very limited. In fact, loads must be removed from a stack, starting with the load at the top. This leads to last-in-first-out (LIFO) stock rotation, which is not desirable for many storage systems.

A fundamental trade-off that exists in storage systems design is that between storage density (how well the cubic volume is utilized) and load selectivity (how easily and how fast one can store or retrieve each load). Loss in storage density often occurs due to aisles and due to space created between individual loads when they are palletized and/or stored in a rack. Generally speaking, higher-density storage systems offer lower selectivity. A well-designed and well-engineered storage system provides the right level of load selectivity while maximizing the storage density. It must also provide the right level of support for the loads and must be compatible with the device/method used for storing and retrieving the loads.

For example, in many storage systems, stock is rotated on a first-in-first-out (FIFO) basis, which implies that random access to every single load in the system may not be necessary. Hence, the right level of selectivity would ensure that, given a set of loads of the same type, only the oldest load needs to be easily accessible. This concept has, in fact, led to pallet flow racks, where loads are stored in lanes and each lane holds multiple loads of the same type. At any given time, only the loads in the front of each lane (which, by definition, are the oldest loads in each lane) are easily accessible. Pallet flow racks increase storage density by minimizing the space required by aisles. However, some space is wasted within each lane because the number of loads in a lane varies over time and, at any given instant, some lanes may be full while others are nearly empty.

Note that block stacking has another advantage: flexibility. Since there is no permanent rack structure, when stacked loads are retrieved (or if the need for storage space decreases over time), the floor space that opens up may be used for other purposes, provided new loads are not stored in the system shortly thereafter. To protect and support the loads at the bottom while avoiding the cost and loss of flexibility associated with a permanent rack, some systems use portable racks, also known as stacking frames, (Figure 17), which are basically self-contained steel units made up of four posts attached to a deck. (An alternative design is a frame that is attached to the pallet itself.)

With portable racks, the weight of each load in the stack is transferred to the floor. It must be noted, however, that portable racks create air between the loads (storage density decreases) and load selectivity is still very limited. Also, portable racks are typically designed for use with palletized loads, and the time to store or retrieve a load may be affected when portable racks are used.

Hence, to maintain convenient and fast load access while providing support for each load in the system, many storage systems use permanent racks. Basic examples of permanent racks include (1) single-deep selective rack (Figure 18), (2) double-deep rack, (3) drive-in or drive-through rack, (4)

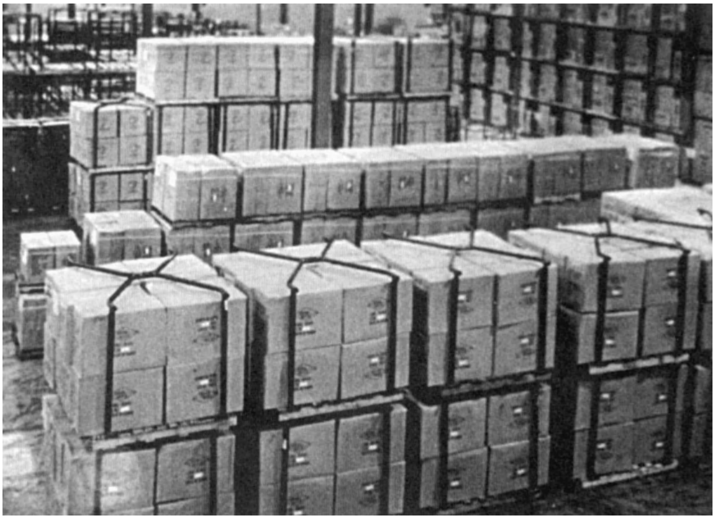


Figure 17 Portable Racks, Stacking Frames.

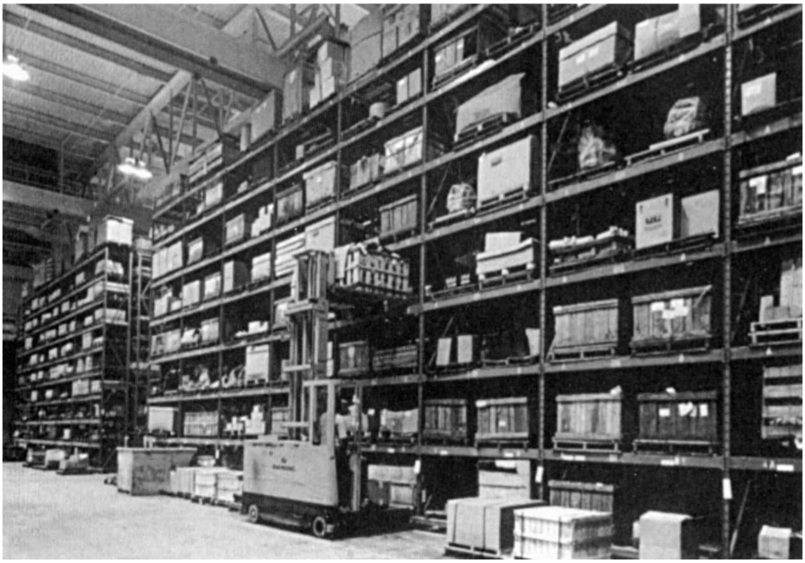


Figure 18 Single-deep Selective Rack.

pallet flow rack (Figure 19), (5) mobile rack (which slides on permanent tracks installed on the floor), and (6) Cantilever rack (Figure 20), which is suitable for storing bar stock. When equipped with wooden decks, cantilevered racks may also be used for storing furniture.

With an increasing desire to utilize the cube well and the availability of lift trucks that have higher reach, there is a tendency to build taller rack structures, especially in regions where land or floor space is very limited. While high cube utilization and going up instead of going out are appropriate

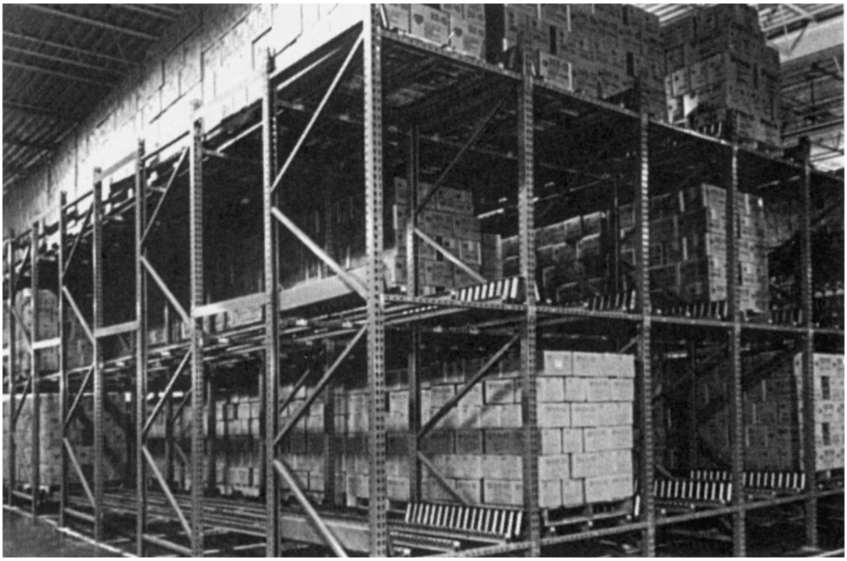


Figure 19 Pallet Flow Rack.

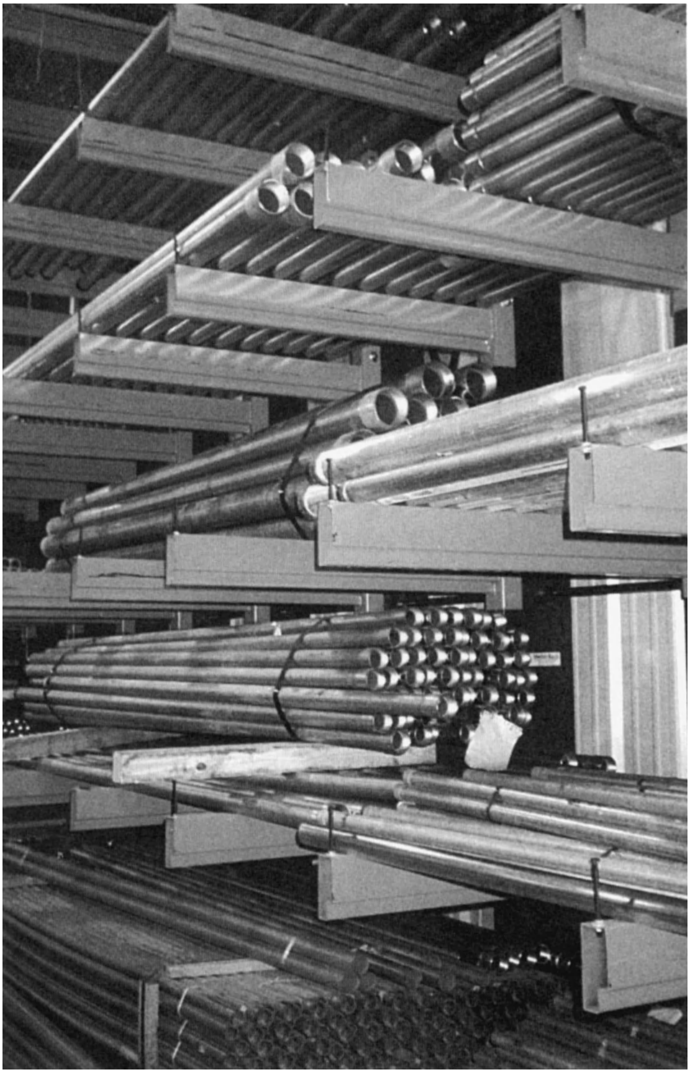


Figure 20 Cantilever Rack. (Courtesy of Jervis B. Webb Co.)

objectives in many instances, it must also be stressed that taller racks can have higher overall cost. As a result, given two storage systems with equal total storage capacity, the one with a taller rack (and smaller footprint) may be more expensive with respect to the total rack cost. (This must of course be traded off against potential savings created by a smaller footprint.)

Two other factors affect the rack cost and rack height. First, racks are more expensive in regions with high seismic activity since they must meet more stringent requirements. Taller racks in such regions may also present a safety hazard. Second, small deviations in the floor surface are amplified when the forks of the truck are raised to the higher tiers of the rack. Consequently, a wavy floor surface may create serious problems for a narrow-aisle truck serving a tall rack. One must also ensure that the floor loading capacity (psi) is sufficient for the rack being considered. A taller rack requires a higher floor loading capacity. Last but not least, one must also consider fire safety and sprinkler systems when selecting the height and type of rack used in a storage system. Tall racks or decked racks may require modifications or enhancements to the sprinkler system.

7. AUTOMATED SYSTEMS

Material handling is often a physically demanding, repetitive task. Although very few handling systems are completely automated, mechanization and automation play a significant role in designing and operating effective and efficient handling systems. Some of the better-known examples of automated material-handling systems include automated storage/retrieval (AS/R) systems, carousels/rotary racks, automated guided vehicle (AGV) systems, and robotic systems.

AS/R systems have been in use since the 1960s, although recent systems are much smaller in size. Each aisle in an AS/R system consists of a storage rack on either side and a S/R machine. (In some systems the S/R machine may not be aisle-captive.) The S/R machine is supported by a guide rail both at the top and bottom; this allows concurrent travel in the horizontal and vertical directions. Load pick-up and deposit is performed via a shuttle mechanism that allows the S/R machine to serve either side of the aisle without turning. Due to the shuttle mechanism, however, the rack used in AS/R systems is a specially-designed rack that is different from a standard single-deep selective (pallet) rack.

AS/R systems are capital-intensive systems. However, they offer a number of advantages, such as low labor and energy costs, high land/space utilization, high reliability/accuracy, and high throughput rates. The expected travel time equations presented earlier for single command (SC) and dual command (DC) cycles under concurrent horizontal and vertical travel [see Eqs. (1) and (2)] apply to AS/R systems as well.

There are many papers in the technical literature for the reader who wishes to obtain more detailed information on AS/R systems. In addition to results such as the expected value and/or distribution of SC and DC travel times (e.g., Chang et al. 1995; Foley and Frazelle 1991; Graves et al. 1977; Hausman et al. 1976; Kouvelis and Papanicolaou 1995; Sarker and Babu 1995), various papers have investigated operational issues such as S/R dwell point strategies, storage-retrieval sequencing, and storage methods (e.g., Chang and Egbelu 1997a, b; Egbelu and Wu 1993; Hwang and Lim 1993; Lee and Schaefer 1996; Peters et al. 1996), twin-shuttle S/Rs (Keserla and Peters 1994; Meller and Mungwattana 1997; Sarker et al. 1994), and storage-retrieval matching or AS/RS control strategies/design models (Han et al. 1987; Elsayed and Lee 1996; Lee 1997; Linn and Wysk 1990; Rosenblatt et al. 1993; Seidmann 1988; Wang and Yih 1997).

Carousels have been used in many storage/retrieval systems involving small to medium-sized parts. A carousel is basically a group of carriers that are suspended via trolleys from an overhead, closed-loop track. (Heavier loads may utilize floor-supported carriers.) Each carrier typically contains a set of shelves to store trays or tote boxes. Using a drive mechanism similar to a trolley conveyor, the carriers can be moved clockwise or counterclockwise around the track. Most readers would recognize a simple version of the carousel used at the dry cleaner, where garments are brought to the operator for retrieval.

Although most carousel applications in industry use human operators to store and retrieve the parts, fully automated systems have also been installed by replacing the human operator with automatic insert/extract devices to remove or insert the trays/tote boxes automatically. Also, depending on the plane of rotation, carousels are classified as horizontal carousels or vertical carousels.

When there are multiple items to be retrieved, the carousel must be turned and stopped several times, once for each item, assuming that the items are located on different carriers. Suppose each carrier in a horizontal carousel contains only one shelf, that is, the carousel is only one-level high. Further suppose several one-level-high carousels are stacked vertically. The result is a ‘‘carousel’’ where each level is powered and operated independently. Such a system is known as a rotary rack, which generally yields higher throughput and can support multiple extract/insert devices at the same time.

Another type of automated material-handling system that has been used successfully in manufacturing and warehousing is the automated guided vehicle (AGV) system. An AGV is basically a fully automated cart that can pick up, route, and deliver unit loads from one point to another (within a network of pick-up/deposit (P/D) points) with no human intervention. The vehicle runs under the control of an on-board computer. (Actually, a vehicle may contain two or more microprocessors on board.) If centralized control is used, that is, if a central computer keeps track of all the vehicle movements and move requests in the system, it is quite common to have each AGV communicate with the central computer via radio frequency (RF) communication. Although alternative guidance technologies exist, the most common is wire guidance, which consists of segments of wire buried in the floor. When energized, the wire generates a magnetic field that is sought by the vehicle.

Guidewire-free AGV systems (also known as self-guided vehicles, or SGVs) are also available from several vendors. SGV systems offer increased flexibility in those systems where the load routings and/or the P/D points change frequently. In a typical SGV system, the aisle structure is maintained by each vehicle as a road map. In moving a load, each vehicle follows the road map according to preprogrammed instructions. New road maps can be prepared off-line and downloaded to each vehicle on an as-needed basis. Initially, the cost differential between AGV and SGV systems was significant. More recently, however, SGV systems have become cost-competitive with AGV systems.

Using the appropriate number of vehicles in an AGV or SGV system is very important. To obtain a quick estimate, one may use the model presented earlier for lift trucks operating under FCFS dispatching. However, due to possible congestion and blocking delays, and to capture more efficient, dynamic dispatching rules, many AGV/SGV systems today are designed via simulation. In fact, many simulation packages in the market have modules or logic built in to facilitate AGV simulation. In addition, the reader will find many papers on AGV systems in the technical literature. These papers address various issues, including system configuration, vehicle dispatching, guidepath design, and gridlock avoidance (see, e.g., Arifin and Egbelu 2000; Borenstein 2000; Bozer and Srinivasan 1991; Bozer and Yen 1996; Ganesharajah et al. 1998; Goetz and Egbelu 1990; Hwang and Kim 1998; Kim et al. 1997, 1999; Srinivasan et al. 1994; Yeh and Yeh 1998).

Robots have also played a significant role in material handling. Perhaps the most common robot used in material handling is the pick-and-place robot, which comes in a variety of configurations. Other robotic or robot-type systems used in material handling include automated item dispensing (similar to, or more advanced versions of, vending machines), palletizers/depalletizers, and gantry robots. Other robots, such as those used in welding and painting, are not considered material-handling robots. For further information on robots, the interested reader may refer to Nof (1999) or Tompkins et al. (1996, Chap. 6), among others.

REFERENCES

Arifin, R., and Egbelu, P. J. (2000), “Determination of Vehicle Requirements In Automated Guided Vehicle Systems: A Statistical Approach,” *Production Planning and Control*, Vol. 11, No. 3, pp. 258–270.

Borenstein, J. (2000), “The Omnimate: A Guidewire- and Beacon-Free AGV for Highly Reconfigurable Applications,” *International Journal of Production Research,* Vol. 38, No. 9, pp. 1993–2010.

Bozer, Y. A., and White, J. A. (1984), “Travel-Time Models for Automated Storage/Retrieval Systems,” *IIE Transactions*, Vol. 16, No. 4, pp. 329–338.

Bozer, Y. A., and Srinivasan, M. M. (1991), “Tandem Configurations for Automated Guided Vehicle Systems and the Analysis of Single-Vehicle Loops,” *IIE Transactions*, Vol. 23, No. 1, pp. 72–82.

Bozer, Y. A., and Yen, C. K. (1996), “Intelligent Dispatching Rules for Trip-Based Material Handling Systems,” *Journal of Manufacturing Systems*, Vol. 15, No. 4, pp. 226–239.

Chang, S. H., and Egbelu, P. J. (1997a), “Relative Pre-Positioning of Storage/Retrieval Machines in Automated Storage/Retrieval Systems to Minimize Maximum System Response Time,” *IIE Transactions,* Vol. 29, No. 4, pp. 303–312.

Chang, S. H., and Egbelu, P. J. (1997b), “Relative Pre-Positioning of Storage/Retrieval Machines in Automated Storage/Retrieval System to Minimize Expected System Response Time,” *IIE Transactions,* Vol. 29, No. 4, pp. 313–322.

Chang, D. T., Wen, U. P., and Lin, J. T. (1995), “The Impact of Acceleration Deceleration on Travel-Time Models for Automated Storage-Retrieval Systems,” *IIE Transactions*, Vol. 27, No. 1, pp. 108–111.

Chow, W. M. (1986), “Design for Line Flexibility,” *IIE Transactions*, Vol. 18, No. 1, pp. 95–108.

Egbelu, P. J. (1987), “The Use of Non-Simulation Approaches in Estimating Vehicle Requirements in an Automated Guided Vehicle Based Transport System,” *Material Flow*, Vol. 4, pp. 17–32.

Egbelu, P. J., and Wu, C. T. (1993), “A Comparison of Dwell Point Rules in an Automated Storage-Retrieval System,” *International Journal of Production Research*, Vol. 31, No. 11, pp. 2515–2530.

Elsayed, E. A., and Lee, M. K. (1996), “Order Processing in Automated Storage/Retrieval Systems with Due Dates,” *IIE Transactions*, Vol. 28, No. 7, pp. 567–577.

Foley, R. D., and Frazelle, E. H. (1991), “Analytical Results for Miniload Throughput and the Distribution of Dual Command Travel Time,” *IIE Transactions*, Vol. 23, No. 3, pp. 273–281.

Fruchtbaum, J. (1988), *Bulk Materials Handling Handbook,* Van Nostrand Reinhold, New York.

Ganesharajah, T., Hall, N. G., and Sriskandarajah, C. (1998), “Design and Operational Issues in AGV-Served Manufacturing Systems,” *Annals of Operations Research*, Vol. 76, pp. 109–154.

Goetz, W. G., and Egbelu, P. J. (1990), “Guide Path Design and Location of Load Pick-up Drop-off Points for an Automated Guided Vehicle System,” *International Journal of Production Research,* Vol. 28, No. 5, pp. 927–941.

Graves, S. C., Hausman, W. H., and Schwarz, L. B. (1977), "Storage Retrieval Interleaving in Automatic Warehousing Systems," *Management Science*, Vol. 23, No. 9, pp. 935–945.

Han, M. H., McGinnis, L. F., Shieh, J. S., and White, J. A. (1987), "On Sequencing Retrievals in an Automated Storage-Retrieval System," *IIE Transactions*, Vol. 19, No. 1, pp. 56–66.

Hausman, W. H., Schwarz, L. B. and Graves, S. C. (1976), "Optimal Storage Assignment in Automatic Warehousing Systems," *Management Science*, Vol. 22, No. 6, pp. 629–638.

Hwang, H., and Kim, S. H. (1998), "Development of Dispatching Rules for Automated Guided Vehicle Systems," *Journal of Manufacturing Systems*, Vol. 17, No. 2, pp. 137–143.

Hwang, H., and Lim, J. M. (1993), "Deriving an Optimal Dwell Point of the Storage-Retrieval Machine in an Automated Storage-Retrieval System," *International Journal of Production Research,* Vol. 31, No. 11, pp. 2591–2602.

Keserla, A., and Peters, B. A. (1994), "Analysis of Dual-Shuttle Automated Storage-Retrieval Systems," *Journal of Manufacturing Systems*, Vol. 13, No. 6, pp. 424–434.

Kim, C. W., Tanchoco, J. M. A., and Koo, P. H. (1997), "Deadlock Prevention in Manufacturing Systems with AGV Systems: Banker's Algorithm Approach," *Journal of Manufacturing Sciences E-T ASME*, Vol. 119, No. 4b, pp. 849–854.

Kim, C. W., Tanchoco, J. M. A., and Koo, P. H. (1999), "AGV Dispatching Based on Workload Balancing," *International Journal of Production Research*, Vol. 37, No. 17, pp. 4053–4066.

Kouvelis, P., and Papanicolaou, V. (1995), "Expected Travel-Time and Optimal Boundary Formulas for a 2-Class-Based Automated Storage-Retrieval System," *International Journal of Production Research,* Vol. 33, No. 10, pp. 2889–2905.

Kulwiec, R. (1999), "Here's a Guide to Basic Types of Sortation Conveyors and Systems, Where They Are Used, and Who Makes Them," in *Modern Materials Handling* (MMH.COM), Equipment Report, Refresher course on sortation conveyors.

Lee, H. F. (1997), "Performance Analysis for Automated Storage and Retrieval Systems," *IIE Transactions*, Vol. 29, No. 1, pp. 15–28.

Lee, H. F. and Schaefer, S. K. (1996), "Retrieval Sequencing For Unit-Load Automated Storage and Retrieval Systems With Multiple Openings," *International Journal Of Production Research,* Vol. 34, No. 10, pp. 2943–2962.

Linn, R. J., and Wysk, R. A. (1990), "An Expert System Based Controller for an Automated Storage-Retrieval System," *International Journal of Production Research*, Vol. 28, No. 4, pp. 735–756.

Meller, R. D., and Mungwattana, A. (1997), "Multi-Shuttle Automated Storage/Retrieval Systems," *IIE Transactions*, Vol. 29, No. 10, pp. 925–938.

Nof, S. Y., Ed. (1999), *Handbook of Industrial Robotics*, 2nd Ed., John Wiley & Sons.

Peters, B. A., Smith, J. S., and Hale, T. S. (1996), "Closed Form Models for Determining the Optimal Dwell Point Location in Automated Storage and Retrieval Systems," *International Journal of Production Research,* Vol. 34, No. 6, pp. 1757–1771.

Rosenblatt, M. J., Roll, Y., and Zyser, V. (1993), "A Combined Optimization and Simulation Approach for Designing Automated Storage-Retrieval Systems," *IIE Transactions*, Vol. 25, No. 1, pp. 40–50.

Sarker, B. R. and Babu, P. S. (1995), "Travel Time Models in Automated Storage Retrieval Systems: A Critical Review," *International Journal of Production Economics*, Vol. 40, Nos. 2–3, pp. 173–184.

Sarker, B. R., Mann, L., and Dossantos, J. R. G. L. (1994), "Evaluation of a Class-Based Storage Scheduling Technique Applied to Dual-Shuttle Automated Storage and Retrieval Systems," *Production Planning and Control*, Vol. 5, No. 5, pp. 442–449.

Seidmann, A. (1988), "Intelligent Control Schemes for Automated Storage and Retrieval Systems," *International Journal of Production Research*, Vol. 26, No. 5, pp. 931–952.

Shamlou, P. A. (1988), *Handling of Bulk Solids: Theory and Practice,* Butterworths, London.

Srinivasan, M. M., Bozer, Y. A., and Cho, M. S. (1994), "Trip-Based Material Handling Systems—Throughput Capacity Analysis," *IIE Transactions*, Vol. 26, No. 1, pp. 70–89.

Tompkins, J. A., White, J. A., Bozer, Y. A., Frazelle, E. H., Tanchoco, J. M. A., and Trevino, J. (1996), *Facilities Planning*, 2nd Ed., John Wiley & Sons, New York.

Wang, J. Y., and Yih, Y. (1997), "Using Neural Networks to Select a Control Strategy for Automated Storage and Retrieval Systems (AS/RS)," *International Journal of Computer Integrated Manufacturing,* Vol. 10, No. 6, pp. 487–495.

Woodcock, C. R. and Mason, J. S., Eds., (1987), *Bulk Solids Handling: An Introduction to the Practice and Technology,* L. Hill, Glasgow and Chapman & Hall, New York.

Yeh, M. S., and Yeh, W. C. (1998), "Deadlock Prediction and Avoidance for Zone-Control AGVS," *International Journal of Production Research*, Vol. 36, No. 10, pp. 2879–2889.

CHAPTER 57

Storage and Warehousing

Jerry D. Smith
Tompkins Associates

1. INTRODUCTION 1527
- 1.1. Warehousing Defined 1527
- 1.2. The Value of Warehousing in Today's Economy 1528

2. STORAGE AND WAREHOUSING 1528

3. SCIENTIFIC APPROACH TO WAREHOUSE PLANNING 1528
- 3.1. Requirements for Successful Warehousing 1528
- 3.2. Warehouse Objectives 1529
- 3.3. Contingency Planning 1530
- 3.4. Strategic Master Planning 1530
  - 3.4.1. Qualities of a Strategic Master Plan 1531
  - 3.4.2. Strategic Master Planning Methodology 1531

4. STORAGE SPACE PLANNING 1532
- 4.1. Define the Materials to Be Stored 1532
- 4.2. Determine Storage Philosophy 1534
- 4.3. Determine Alternative Storage Method Space Requirements 1535

5. WAREHOUSE LAYOUT PLANNING 1538
- 5.1. Objectives of a Warehouse Layout 1538
- 5.2. Layout Planning Methodology 1538
- 5.3. Generate Alternative Layouts 1538
- 5.4. Evaluate the Alternative Layouts 1539

6. WAREHOUSE EQUIPMENT PLANNING 1541

7. WAREHOUSE OPERATIONS AUDIT 1544
- 7.1. Operations Audit Performance Categories 1544
- 7.2. Operations Audit Methodology 1546

ADDITIONAL READING 1547

1. INTRODUCTION

Storage and warehousing operations are a critical part of maintaining a profitable business. With over 300,000 large warehouses and 2.5 million employees in the United States, the cost of American warehousing is over 5% of the gross national product. In the past few years, the field of warehousing has begun to receive the attention it deserves. However, the warehouse management has been asked to increase customer service, reduce inventories, increase productivity, handle a large number of stock-keeping units, and improve space utilization. Warehouse management has realized that these conflicting objectives require a much more professional approach than previously adopted. It is critical that today's warehouse management follow this approach to achieve the results expected by today's upper management.

1.1. Warehousing Defined

The functions performed by a warehouse are:

1. Receiving the goods from a source
2. Storing the goods

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

3. Picking the goods when they are required
4. Shipping the goods to the appropriate customer

Oftentimes, a distinction is made between a finished-goods warehouse and a raw-materials storeroom. The fact is, however, that the functions performed in a finished-goods warehouse—receive, store, pick, and ship—are identical to the functions performed in a raw-materials storeroom. Consequently, both are warehouses. The only true distinctions between the two are the source from which the goods are received and the user to whom the goods are shipped. A raw-materials storeroom receives goods from an outside source, stores the goods, picks the goods, and ships the goods to an inside user. A finished-goods warehouse receives goods from an inside source, stores the goods, picks the goods, and ships the goods to an outside user. Likewise, an in-process inventory warehouse receives goods from an inside source, stores the goods, picks the goods, and ships the goods to an inside user, while a distribution warehouse receives goods from an outside source, stores the goods, picks the goods, and ships to an outside user. The differences among these various warehouses are restricted to the perspectives of the sources, management, and users of the warehouses. If the primary functions of an activity are receive, store, pick, and ship, the activity is a warehouse, regardless of its position in a company's logistics. The tools and techniques presented in this chapter can be successfully used to plan and manage that activity.

1.2. The Value of Warehousing in Today's Economy

It is important to ponder the question ''Does warehousing add value to a product?'' The traditional school of thought has concluded that no, warehousing does not add value to a product; warehousing is strictly a cost-adding activity that is a necessary evil. In firms that follow this school of thought, warehousing costs are typically classified as indirect costs. Often these cost categories are spread over the direct costs of the firm in such a way that the cost of warehousing is not distinguishable.

2. STORAGE AND WAREHOUSING

To convince yourself of the value of warehousing, consider the value of the refrigerator in the home. The refrigerator is essentially a warehouse. You purchase food at the supermarket, deliver the food to the refrigerator, store it in the refrigerator, pick the food from the refrigerator as needed, and ship the food to some location where it will be processed or consumed. What is the value of the refrigerator? What is the value of having milk where needed, when needed? If the answer is not yet clear, consider the costs that would be incurred if you did not own a refrigerator. What are the costs of not having milk for cereal at breakfast in the morning? Some of the costs are hunger from not eating, indigestion from eating dry cereal, the inconvenience of having to go to the supermarket before breakfast, and the actual expense of going to the supermarket before breakfast.

The true value of warehousing lies in having the right product in the right place at the right time. Thus, warehousing provides the time-and-place utility necessary for a company to prosper.

Without a complete and accurate understanding of the value of warehousing, companies have failed to give warehousing the same scientific scrutiny as the other aspects of their business. For a profession as important as warehousing, this is not acceptable. A more scientific approach must be taken to the warehouse of today.

3. SCIENTIFIC APPROACH TO WAREHOUSE PLANNING

Warehouse planning is more than pouring a concrete slab and installing some rack and tilting up some walls. Warehouse planning is not a static, one-time activity. The changing, dynamic environment within which warehouses are planned quickly renders existing plans obsolete. Therefore, warehouse planning must be a continuous activity in which the existing plan is constantly being scrutinized and molded to meet anticipated future requirements. For a warehouse to accomplish its objectives, warehouse managers must consider the variable warehouse resources and mold them into an effective plan. A successful warehouse maximizes the effective use of the warehouse resources while satisfying customer requirements.

3.1. Requirements for Successful Warehousing

To be successful into the 21st century, warehouse planning must be accomplished within the framework of a clear, long-term, consistent vision of where the warehousing operations are headed. The following strategies should form the basis of this vision.

1. *Professionalism:* Warehousing will be viewed as a critical supply chain enabler and a competitive strength, not as an inert facility.
2. *Customer awareness:* Serving the customer is only the foundation; satisfying them is vital to their continuing to patronize your organization.

3. *Measurement:* Warehouse standards will be established, performance will be measured against these standards, and timely actions will be taken to overcome any deviations.
4. *Operations planning:* Systems and procedures will be put into effect that allow the warehouse manager to plan the operations proactively rather than reactively respond to external circumstances.
5. *Supply chain network:* Warehouses will not be viewed as independent operations but as an element of the overall, well-planned supply chain.
6. *Third party/outsourcing:* More intelligent use of third-party logistics (3PL) is the norm so that organizations can focus on core competencies.
7. *Pace:* The reduction of lead times, shorter product lives, and increased inventory turnover result in an increase in the pace of the warehouse.
8. *Variety:* More different SKUs and more special customer requirements result in an increase in the variety of tasks performed in the warehouse.
9. *Adaptability:* Due to the increase in warehouse pace and variety, all warehouse systems, equipment, and people will be able to handle products that vary in size and weight.
10. *Uncertainty:* All uncertainty will be minimized; discipline will be increased.
11. *Integration:* Integration needs to be understood as not only a method of improvement within, or even between, processes but rather as a method of improvement of the whole process.
12. *Inventory accuracy:* Inventory above 99% is the norm with real-time warehouse management systems, bar coding, and electronic order processing. Annual physical inventories are eliminated and cycle counting is fully embraced.
13. *Space utilization:* Space utilization will be enhanced through dynamic slotting or the placement of product in a facility for the purpose of optimizing material handling and space efficiencies.
14. *Housekeeping:* There is efficiency in order in the warehouse. The certainty that the work areas are safe, free of congestion, and properly organized enables personnel to move through a day's work with just the work on which to concentrate.
15. *Orderpicking:* The criticality of orderpicking will be understood, and procedures and layouts will be designated to maximize orderpicking efficiency.
16. *Business process continuous improvement* (*BPCI*)*:* The power of the people will be unleashed via a methodical process of continuous improvement.
17. *Continuous flow:* There will be a clear focus on pulling product through the logistics system and not building huge inventories.
18. *Warehouse management systems:* Real-time, bar code-based, RF communication Warehouse management systems (WMS) will be required to meet today's requirements.
19. *Change:* Organizations that will usher in the new century successfully will be the organizations that proactively embrace change.
20. *Leadership:* There must be a balance between the control aspects of management and harnessing the energy of change to create peak-to-peak performance of leadership.

3.2. Warehouse Objectives

The resources of a warehouse are space, equipment, and personnel. The cost of space includes not only the cost of building or leasing space but also the cost of maintaining the space. Typically, the cost of space in a warehouse is $0.20 to $0.30 per cubic foot per year for taxes, insurance, maintenance, and energy. A company that is ineffectively using its available cubic space is incurring excessive operating costs.

The equipment resources of a warehouse include data-processing equipment, dock equipment, unit load equipment, material-handling equipment, and storage equipment, all of which combine to represent a sizable capital investment in the warehouse. In order to obtain an acceptable rate of return on this investment, the proper equipment must be selected and it must be properly used.

Oftentimes, the personnel resource of the warehouse is the most neglected resource, even though the cost of this resource is usually the greatest. Approximately 50% of the costs of a typical warehouse are labor related. Reducing the amount of labor, pursuing higher labor productivity, good labor relations, and worker satisfaction, will significantly reduce warehouse operating cost.

Customer requirements are simply the demand to have the right product in good condition at the right place at the right time. Therefore, the product must be accessible and protected. If a warehouse cannot meet these requirements adequately, then the warehouse does not add value to the product and, in fact, very likely subtracts value from the product.

Therefore, the following objectives must be met for a warehouse to be successful:

1. Maximize effective use of space.
2. Maximize effective use of equipment.
3. Maximize effective use of labor.
4. Maximize accessibility of all items.
5. Maximize protection of all items.

The two distinct types of continuous warehouse planning needed to result in an efficient and effective warehouse operation are contingency planning and strategic master planning.

3.3. Contingency Planning

Contingency planning is a *defensive* tool used to guard against a predictable future change in warehouse requirements whose timing is extremely difficult, if not impossible, to anticipate. In other words, a contingency plan answers the question ''What do I do *if* some unexpected event or condition arises?'' Contingency plans are needed to guard against the following short-term situations:

1. Equipment downtime
2. Labor problems
3. Surges of activity
4. Material supply disruptions
5. Other emergencies

Contingency planning is not crisis management or putting out fires, which entail developing solutions to problems *after* the problems occur. Proper contingency planning develops the action plan to the fullest extent possible *before* the problem occurs. Consequently, proper contingency planning can significantly reduce the lead time required to correct or accommodate the unexpected event. One does not wait until after a fire starts in the warehouse to install a sprinkler system; instead, one installs the sprinkler system long before as a contingency against a fire whose timing is unpredictable. Likewise, formal contingency plans can protect the warehouse for other conceivable circumstances with unpredictable timing.

To develop contingency plans for a warehouse, use the following procedures:

1. Make a list of the conceivable ''bad things'' that can happen to or within the operation.
2. Rank the bad things with the events having the greatest probability of occurrence, and/or the most adverse consequences if they do occur, at the top of the list.
3. Starting with the highest-ranked bad thing, carefully determine, in as much detail as possible, the proper steps and actions that should take place to resolve, eliminate, and deal with the consequences to the warehouse operations of the bad thing if and when it occurs.
4. Review these steps and actions with the key warehouse people and refine them based on the inputs received.
5. Publish the resulting contingency plans in print and drill those persons responsible for executing the plans at the time of need in the details of the plans so that everyone knows exactly when, how, and by whom the plan is to be executed.
6. Periodically review and update the contingency plans to keep them current with existing conditions in the operation.

3.4. Strategic Master Planning

Strategic master planning is an *offensive* tool designed to guard against a predictable future change in warehouse requirements whose timing can be anticipated. Strategic master planning is directed at forecasting future warehousing needs sufficiently in advance of the actual requirement to allow enough lead time to efficiently and effectively meet those needs.

Warehousing strategic master plans are needed to accommodate:

1. Forecasted growth or decline in the throughput
2. Space, labor, and equipment deficiencies
3. Product mix changes
4. Inventory increases or reductions
5. Warehouse control problems

Most of these ''problems'' do not develop overnight. Future inventory levels and product mixes typically can be predicted, based on historical and future business plans, years in advance. Granted, forecasting with long planning horizons is risky. Forecasts are often inaccurate. Nevertheless, the forecast is the best available information concerning the future we have, and it is folly not to use that information to advantage. With today's costs of warehouse space, labor, and equipment, more and more decision makers are demanding that future warehouse requirements be expressed in quantitative terms rather than in subjective, qualitative assessments of needs. That is what strategic master planning is all about.

Contingency planning and strategic master planning are complementary. Strategic master planning without effective contingency plans will subject the warehouse to unanticipated problems that do not show up in a forecast of future requirements. Likewise, the absence of good strategic master planning will subject the warehouse to a continuous barrage of ''fires'' to be dealt with by contingency plans, many of which could have been avoided through proper insight and strategic planning. In either situation, the absence of one planning approach severely limits the effectiveness of the other.

3.4.1. Qualities of a Strategic Master Plan

A warehousing strategic master plan is a set of documents describing what actions must be accomplished and when they must be accomplished to satisfy the warehousing requirements of an enterprise over a given planning horizon. A closer examination of this definition reveals the important attributes of a good warehouse strategic master plan.

First of all, a good warehouse strategic master plan is a formal set of documents. It should not consist simply of ideas, thoughts, possibilities, desires, and so forth that are casually recorded ''somewhere,'' if at all. A good plan is a formal set of documents that have been created, collected, edited, and so forth specifically as a strategic master plan of action. Common components of this set of documents include an implementation plan, a descriptive narrative, scaled facility drawings, and supporting economic cost and justification data.

Second, a good warehouse strategic master plan is action oriented. Where possible, the plan should set forth very specific actions to be taken to meet requirements rather than simply stating the alternative actions available to meet those requirements. The strategic master plan is established based on a set of premises concerning future production volumes, inventory levels, manpower levels, available technology, and so forth. As long as these premises are clearly stated as a part of the strategic master plan, and understood, problems should not arise with regard to implementing actions that prove to be based on false premises.

The strategic master plan should be time phased to indicate when each recommended action should be implemented to meet changing warehousing requirements. Typically, scaled facility drawings should accompany each recommended action to illustrate what the facility will look like after a given action has been implemented.

Finally, a good warehouse strategic master plan should encompass a specified planning horizon. It should have a definite beginning point and a definite ending point. Typically, the planning horizon is stated in terms of years. A five-year master plan might serve the years 1991 through 1996.

3.4.2. Strategic Master Planning Methodology

The general methodology for developing a warehouse strategic master plan consists of the following seven-step procedure:

1. Document the existing warehouse operation.
2. Determine and document the warehouse storage and throughput requirements over the specified planning horizon.
3. Identify and document deficiencies in the existing warehouse operation.
4. Identify and document alternative warehouse plans.
5. Evaluate the alternative warehouse plans.
6. Select and specify the recommended plan.
7. Update the warehouse master plan.

Step 1 involves obtaining or developing scaled drawings of the existing warehouse facilities and verifying their accuracy. The accuracy of existing drawings should never be assumed. It should always be physically verified on the warehouse floor.

Existing warehouse equipment should be identified and documented. The labor complement of each area of the warehouse should be determined and the general responsibilities of each person documented. Existing standard operating procedures should be scrutinized and compared against what actually takes place on the shop floor. The first step of the master planning process establishes a baseline against which recommendations for improvement can be compared.

Step 2 involves defining what materials will be stored in the warehouse and the volume anticipated during the planning horizon. Items to be stored in the warehouse should be classified into categories according to their material-handling and storage characteristics.

Forecasts or production schedules should then be used to predict the storage volumes and turnover rates of each category of material over the specified planning horizon. Ideally, these volumes would be stated in terms of the unit loads in which the materials would be stored and handled.

Step 3 involves identifying potential areas of improvement in the existing warehouse operation. The potential for improvement may exist because the operation lacks sufficient capacity to handle future requirements or because existing facilities, methods, equipment, and/or labor forces are not the most efficient or effective available.

Step 4 deals with identifying alternative facility, equipment, procedural, and/or personnel plans that will eliminate or minimize the deficiencies identified in the existing warehouse operation. From these alternative plans of action will come the specific time-phased plan of action to be recommended for meeting the warehouse requirements over the given planning horizon.

Step 5 of the master planning process involves performing both an economic and a qualitative assessment of the alternative plans of action. The economic evaluation should consist of a time-value-of-money assessment of the total life-cycle costs of competing alternative plans of action. The qualitative assessment of alternatives requires that the alternatives be subjectively compared on such attributes as personnel safety, flexibility, ease of implementation, maintainability, potential product damage, and so forth.

Step 6 involves selecting the best of the alternative plans of action implicated by the economic and qualitative evaluations and specifying the recommended warehouse strategic master plan. The master plan will document the space, equipment, personnel, and standard operating procedure requirements of the warehouse over the planning horizon. In addition, scaled facility drawings should be included showing the recommended warehouse layout for all revisions recommended by the plan of action.

The first six steps of this procedure will result in a warehouse strategic master plan. The strategic master planning process, however, will not be complete. In fact, it will never be completed, since strategic master planning is a continuous activity. Step 7 is the process, therefore, of updating the master plan. By its very nature, a strategic master plan is inaccurate. Since it is based, to a large extent, on predictions of the future, the warehouse strategic master plan will require updating as better information concerning the future is obtained. Consequently, it should never be used as a precise tool but only as a valuable guideline for planning future warehouse operations.

4. STORAGE SPACE PLANNING

Space planning is the part of the science of warehousing concerned with making a quantitative assessment of warehouse space requirements. As is true of any science, space planning possesses a very specific methodology. The space planning methodology consists of the following general steps:

1. Determine what is to be accomplished.
2. Determine how to accomplish it.
3. Determine space allowances for each element required to accomplish the activity.
4. Calculate the total space requirements.

The first two steps of the space planning process define the activity and techniques, equipment, information, and so on to be used in performing that activity. Step 3 involves determining the space requirements of each element that goes into performing the activity. In warehousing, these elements might include personnel and personnel services, material handling and material storage equipment, maintenance services, and utilities. Finally, step 4 combines the space requirements of the individual elements to obtain total space requirements.

Storage-space planning is particularly critical because the storage activity accounts for the bulk of the space requirements of a warehouse. Inadequate storage-space planning can easily result in a warehouse that is significantly larger or smaller than required. Too little storage space will result in a world of operational problems, including lost stock, inaccessible material, poor housekeeping, damaged material, safety problems, and low productivity. Too much storage space will breed poor use of space so that it appears that all the available space is really needed. The result will be high space costs in the form of land, construction, equipment, and energy.

To avoid these problems, storage-space planning must be approached from a quantitative viewpoint, as opposed to a qualitative assessment of requirements. The following sections present the scientific methodology of storage-space planning, which when followed, will generate a quantitative and defensible assessment of storage-space requirements.

4.1. Define the Materials to Be Stored

The first step in storage-space planning is to define what is to be accomplished; that is, to define the materials to be stored. A useful tool in defining the materials to be stored is the storage analysis chart (SAC) given in Table 1. Columns 1–5 of the SAC define what materials are to be stored,

TABLE 1 Storage Analysis Chart

Company	A.R.C., Inc.	Date	March 18, 1991	Raw Materials		In-Process Goods	
Prepared by	J. Smith	Sheet	1 of 1	Plant Supplies		Finished Goods	

	Unit Loads				Quantity of Unit Loads Stored			Storage Space			
Description (1)	Type (2)	Cap (3)	Size (4)	Weight (5)	Max. (6)	Avg. (7)	Planned (8)	Method (9)	Specs (10)	Area (ft^2) 11	Ceiling Height Required (12)
Steel pipe plug 1.00 in. diameter × 0.50 in.	Wooden crate	3200 pieces	2 ft × 2 ft × 4 ft	825 lbs	14	8	12	Pallet rack	25 ft × 10 ft × 3 ft	66	9 ft
Aluminum bar 2.75 in. × 2.50 in. × 16 ft	Bundles	25 bars	12.5 × 14 in. × 16 ft	1625 lbs	30	17	30	Cantilever racks	Four-Arm dual rack 5 ft × 16 in. ×x 8 ft	160	8 ft
Stainless steel bar 0.875 in. × 12 ft	Bundles	36 bars	6 in. × 6 in. × 12 ft	900 lbs	7	4	7	Cantilever rack	Four-Arm dual rack 4 ft × 12 in. × 10 ft	48	10 ft
Rubber O rings 0.75 in. diameter	Cartons	40,000 O rings	12 in. × 18 in. × 3 ft	125 lbs	2	1	2	Storage shelf	Metal frame 12 ft × 2 ft × 8 ft	24	8 ft
Brass bar 0.75 in. diameter × 12 ft	Bundles	36 bars	6 in × 6 in × 12 ft	720 lbs	15	8	14	Cantilever rack	Four-Arm dual rack 4 ft × 12 ft × 6 ft	48	6 ft

columns 6–8 specify how much is to be stored, and columns 9–12 define how the materials are to be stored. The information requirements for columns 1–5 of the SAC can be obtained by physically surveying the existing storage areas. The survey would proceed by identifying, generically classifying, measuring, and weighing the unit loads presently in the storage areas.

Columns 6 and 7 of the SAC list the maximum and average number of unit loads of each category of material that should be on hand. Column 8 cites the planned inventory level of each type of material for which storage area will be planned. Determining the proper inventory level is directly related to the storage philosophy that will be used for each category of materials. The different storage philosophies and the decision process one should use to determine the proper planned inventory level will be discussed in the next section of this chapter.

The last four columns of the SAC define the physical characteristics of the storage area being planned. These physical characteristics include the method of storage and the space requirements of that method.

4.2. Determine Storage Philosophy

Once the maximum and average inventory levels have been recorded, the inventory level that will be used as a basis for planning required storage space must be determined. The planned inventory level depends on the philosophy followed in assigning material to storage space. There are two major material-storage philosophies: fixed (or assigned) location storage and random (or floating) location storage. In fixed-location storage each individual stock-keeping unit will always be stored in a specific storage location. No other stock-keeping unit may be stored in that location, even though that location may be empty.

With random-location storage, any stock-keeping unit may be assigned to any available storage location. A stock-keeping unit in location A one month might be stored in location B the following month and a different stock-keeping unit stored in location A.

The amount of space planned for a stock-keeping unit is directly related to the method of assigning space. If fixed-location storage is used, a given stock-keeping unit must be assigned sufficient space to store the maximum amount of the stock-keeping unit that will ever be on hand at any one time. For random-location storage, the quantity of items on hand at any time will be the average amount of each stock-keeping unit. In other words, when the inventory level of one item is above average, another item will likely have an inventory level that is below average; the sum of the two will be close to the average.

Oftentimes, the storage philosophy chosen for a specific stock-keeping unit will not be strictly fixed-location storage or random-location storage. Instead, it will be a combination of the two. A grocery store is an excellent example of combination, or hybrid, location storage. Fixed-location storage is used in the front room of a grocery store where the consumers shop. Pickles are assigned a fixed location, and only pickles will be stored in that location. Pickles will not be found in any other location in the front room of the grocery store. In the back room, or storeroom, of the grocery store, however, the excess, or overstock, merchandise is usually stored randomly. Pickles may be found in one location one week and in a different location the next week. Because combination-location storage is based on a mixture of fixed-location storage and random-location storage, its planned inventory level falls between the fixed-location quantity and random-location quantity. At what point between the fixed-location and random-location quantity the planned inventory level falls is dependent on the percentage of inventory to be assigned fixed locations.

To summarize, the planned inventory level recorded in column 8 of the storage analysis chart in Table 1 should be equal to the maximum inventory level (column 6) for fixed-location storage, the average inventory level (column 7) for random-location storage, or a value between the maximum and average quantities for combination-location storage.

Little has been said at this point about the advantages of one storage philosophy over another. Should the storage philosophy be fixed-, random-, or combination-location storage? Unfortunately, an unequivocal answer to this question does not exist. Choosing one storage philosophy over another means making a number of trade-offs, which must be evaluated. Table 2 presents a qualitative comparison of fixed-, random-, and combination-location storage for three extremely important criteria: use of space, accessibility to material, and material handling.

Use of space in a fixed-location system is poor because space for the maximum amount of inventory that will ever be on hand has been allocated although actual on-hand inventory will normally approach the average inventory level. Therefore, a great deal of empty space is common in fixed-location storage. Random-location storage is extremely space efficient because the space requirements are only about 15% above the average amount of inventory expected on hand. Use of space for combination-location storage is better than it is for fixed location storage and worse than it is for random-location storage because the space requirements are based on a planned inventory level somewhere between the fixed-location and random-location quantities.

TABLE 2 Comparison of Storage Philosophies

Philosophy Criteria	Fixed Location Storage	Random Location Storage	Combination Location Storage
Use of space	Poor	Excellent	Good
Accessibility to material	Excellent	Good, if there is a good material locator system; poor otherwise	Good
Material handling	Good	Good	Poor

Material in fixed-location storage has excellent accessibility because a given storage location contains only one stock-keeping unit. The location of every item is fixed: it is known. Blocked stock is avoided and every stock-keeping unit is readily accessible. Accessibility to material in random-location storage is good as a long as a good material-locator system exists. The material-locator system keeps track of the present location of every item in storage. Once a specific stock-keeping unit is committed to a storage location, no other stock-keeping unit can be placed in that location until the original stock-keeping unit is completely removed. However, if a material-locator system does not exist, or is poorly designed and maintained, then accessibility to material in a random-location storage system will be extremely poor. Blocked stock, lost material, and obsolete material will inevitably result. Even with a good material-locator system, accessibility to material will never be as good in random-location storage as in fixed-location storage. Accessibility to material in combination storage is good if a good material-locator system exists for the randomly-stored portion of storage, or if the percentage of inventory stored randomly is small.

Fixed-location storage and random-location storage score equally well on the material-handling criterion. In each philosophy, material is received, placed into storage, retrieved from storage, and shipped to a user. The flow of material is straightforward and economical. Material is received, placed in the random-location storage area, retrieved, placed into the fixed-location storage area, retrieved, and shipped to a user. Consequently, combination-location storage involves several extra handling steps not required by either fixed-location storage or random-location storage.

In summary, fixed-location storage trades efficiency in use of space for easy accessibility to material; random-location storage trades accessibility to material for efficiency in use of space; and combination-location storage trades material-handling efficiency for middle-of-the-road efficiency in use of space and accessibility to material. However, a clear-cut decision still cannot be made on the best storage philosophy. Perhaps the only general conclusion that can be drawn is that poor use of space by fixed-location storage is a big factor. Compared to the use of space by random-location storage for the same materials, fixed-location storage will generally require 65–85% more space. With the escalating costs of money, land, and construction, few firms can afford to build a fixed-location storage warehouse, which would be 75% larger than that required for random-location storage. The expense of developing and maintaining an effective material-locator system for random-location storage—when compared with these costs—is easily justified. Consequently, one should always carefully evaluate random-location storage before deciding to use fixed-location storage. Rarely will the gains in accessibility to material made by fixed-location storage be enough to offset its high space costs. Occasionally, however, efficient use of space is not a critical factor, so fixed-location storage is preferred. For example, when the items to be stored are extremely small and/or extremely valuable, accessibility to them and accountability for them may be all-important. Few jewelers care about the use of space when they are storing diamond rings.

4.3. Determine Alternative Storage Method Space Requirements

The space requirements of a storage alternative are directly related to the volume of material to be stored and the use-of-space characteristics of the alternative. The two most important use-of-space characteristics are aisle allowances and honeycombing allowances. Aisle allowance is the percentage of space occupied by aisles within a storage area. Aisles are necessary within a storage area to allow accessibility to the material being stored. The amount of aisle allowance depends on the storage method, which dictates the number of aisles required, and on the material-handling method, which dictates the size of the aisles. Expected aisle allowance must be calculated for each storage alternative under consideration.

Honeycombing allowances are the percentage of storage space lost because of ineffective use of the capacity of a storage area. Honeycombing is illustrated in Figure 1. Honeycombing occurs whenever a storage location is only partially filled with material. The unoccupied area within the storage

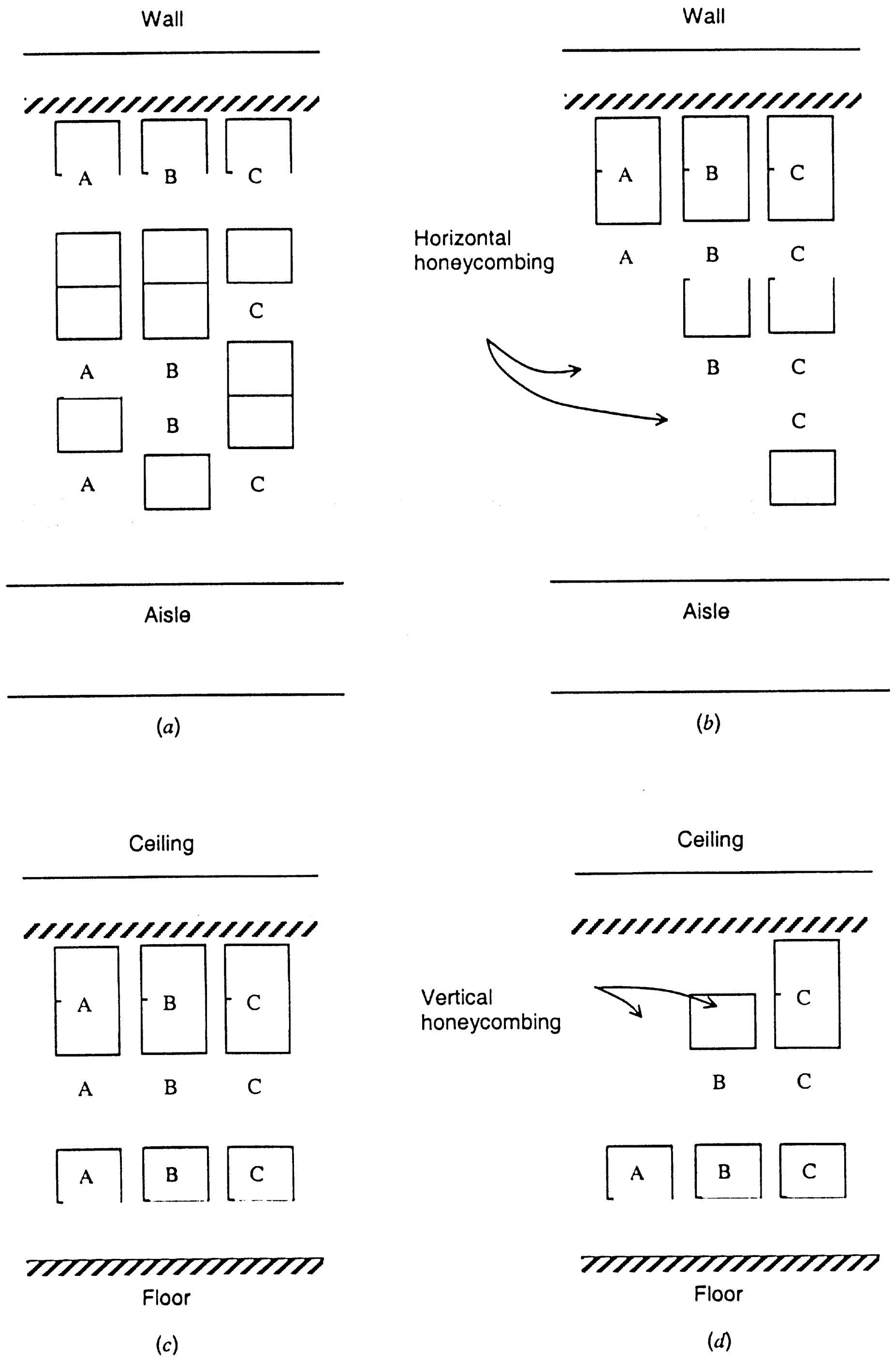

Wall
A
B
C
A
B
C
B
A
C
Aisle
(a)
Wall
A
B
C
Horizontal honeycombing
A
B
C
B
C
C
Aisle
(b)
Ceiling
A
B
C
A
B
C
A
B
C
Floor
(c)
Ceiling
Vertical honeycombing
C
B
C
A
B
C
Floor
(d)

Figure 1 Horizontal and Vertical Honeycombing. (*a*) Plan view of bulk storage area—no honeycombing. (*b*) Plan view of bulk storage area showing horizontal honeycombing. (*c*) Elevation view of bulk storage area—no honeycombing. (*d*) Elevation view of bulk storage area showing vertical honeycombing.

location is honeycombed space. Honeycombing may occur horizontally and vertically. For example, Figure 1(*a*) presents a plan view of a bulk storage area in which material can be placed four units deep. Because the bulk storage area is full, no honeycombing occurs. In Figure 1(*b*), however, two units of product A and one unit of product B have been removed, leaving three empty slots. No other items can be placed in these slots until the remaining units of A and B have been removed (otherwise, blocked stock will result); so these slots are horizontal honeycombing losses. Figure 1(*c*) is an elevation view of a bulk storage area in which material can be stacked three units high. Here again, the storage area is full and no honeycombing occurs. In Figure 1(*d*), however, two units of product A and one unit of product B have been removed, leaving three empty slots. To avoid blocked stock or poor stock rotation, no other units can be placed in these slots until the remaining units of A and B have been removed. Consequently, the empty slots are vertical honeycombing losses. Horizontal and vertical honeycombing losses will occur. Efforts to totally eliminate honeycombing may improve space utilization but will assuredly result in increased material-handling costs related to double handling loads, material damage, and lost productivity. Honeycombing, while it should be minimized, must be considered a natural and allowed-for phenomenon of the storage process. For each storage alternative under consideration, the expected honeycombing allowance must be estimated.

Once the aisle and honeycombing allowances for a storage method alternative have been determined, a space standard can be calculated for that storage method. A space standard is a benchmark

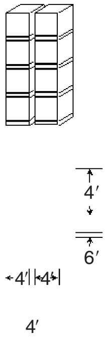
Item A requires special environmental control. A special storage area must be established to house the maximum quantity of item A, which is to be stored on pallets, four pallets high. In a bulk storage analysis chart, how much space should be allocated for the storage of item A?

Case size = 2 ft × 1 ft × 1 ft (height) = 2 ft^3

Palletized = 48 in. × 48 in. × 48 in. pallet × 4 tiers high = 64 ft^3 (32 cases/pallet)

Step 1: The aisle allowance (AA) has been estimated from a proposed layout to be 10%. The honeycombing allowance (HA) has been estimated to be 25%.

Step 2: The pallet height is 6 in. and the clearance between stacks is 4 in. The total space required for one four-pallet-high stack of item A is therefore:



Stack width × stack depth × stack height = (4 ft + 0.33 ft) × 4 ft × [(4 ft + 0.5 ft) × 4] = 312 ft^3 for 128 cases of item A

Step 3: The inclusion of allowances for aisles and honeycombing results in the following space standard:

$$\frac{312\text{ ft}^3}{(1-\text{AA})(1-\text{HA})(128\text{ cases})} = \frac{312\text{ ft}^3}{(1-0.10)(1-0.25)(128\text{ cases})} = 3.61\text{ ft}^3/\text{case}$$

Step 4: A total storage space required for the maximum anticipated volume of item A, using the proposed storage method, is:

Total storage space required = 300,000 cases × 3.61 ft^3/case
= 1,083,000 ft^3
or 60,167 ft^2 having a clear stacking height of 18 ft

Figure 2 Example Space Standard Calculation for Storage Area.

that defines the amount of space required per unit of product stored. Given the space standard and the total inventory of a class of items to be stored, the total space required for that class of items may then be calculated. Figure 2 presents an example illustrating the calculation and use of space standards.

5. WAREHOUSE LAYOUT PLANNING

5.1. Objectives of a Warehouse Layout

Before layout planning can begin, the specific objectives of a warehouse layout must be determined. In general, the objectives of a warehouse layout are:

1. To use space efficiently
2. To allow the most efficient material handling
3. To provide the most economical storage in relation to costs of equipment, use of space, damage to material, and handling labor
4. To provide maximum flexibility in order to meet changing storage and handling requirements
5. To make the warehouse a model of good housekeeping

The astute observer will notice that the first three objectives are essentially identical to the overall objectives of a warehouse. Recall that the objectives of a warehouse are:

1. To maximize effective use of space
2. To maximize effective use of equipment
3. To maximize effective use of labor
4. To maximize accessibility of all items
5. To maximize protection of all items

It is true that the objectives of a warehouse layout are redundant. This shows the importance of layout planning to warehouse planning. Without a good warehouse layout, it is impossible to have a good warehouse. The objective of layout planning is to arrange and coordinate the space, equipment, and labor resources of the warehouse. Poor layout planning can undermine superior space, equipment, and personnel planning. Put another way, accomplishing the objectives of warehousing depends on having a good layout. If the warehouse layout is bad, the warehouse as a whole will be bad. Conversely, if the warehouse as a whole is bad, chances are the warehouse layout is bad.

The fourth objective of a warehouse layout recognizes the fact that warehousing exists not within a static, unchanging environment but within a dynamic, ever-changing environment. If the mission of a warehouse changes, the warehouse layout should very likely change, too, to adapt to the new mission. However, a good warehouse layout possesses the flexibility to absorb minor variances in expected storage volumes and product mixes with few or no alterations required. This flexibility allows the warehouse to function even if the forecasts on which it was planned prove to be wrong, as they inevitably do.

The last objective of warehousing follows the principle that there is efficiency in order. Good housekeeping is essential to good warehousing; a good warehouse cannot exist without good housekeeping. Yet good housekeeping by itself will not ensure a good warehouse. If the space, equipment, personnel, and layout are not properly planned, all the housekeepers in the world could not get a warehouse to function. But poor housekeeping will surely undermine good space, equipment, personnel, and layout planning.

5.2. Layout Planning Methodology

Warehouse layout planning methodology consists of two steps:

1. Generate a series of warehouse layout alternatives
2. Evaluate each alternative against specific criteria to identify the best warehouse layout

These two steps are discussed in the following sections.

5.3. Generate Alternative Layouts

Generating alternative warehouse layouts is as much art as science. The quality of the layout alternatives will largely depend on the skill and ingenuity of the layout planner. This fact is crucial to the most common approach to generating layout alternatives: template juggling. The word ‘‘juggle’’ means ‘‘to skillfully manipulate a group of objects to obtain a desired effect.’’ Consequently, template juggling is the skillful manipulation of a group of templates, models, or other representations of warehouse space, equipment, and personnel in order to obtain a warehouse layout that meets objectives. In other words, template juggling is a trial-and-error approach to finding the proper arrangement and coordination of the physical resources of the warehouse.

The quality of the alternatives created from template juggling depends on the creativity of the layout planner. Unfortunately, layout planners often either lack creativity or do not attempt to express their creativity. Many layout planners approach the problem with a preconceived idea about what the solution should be. They tend to base the layout planning process on that preconceived solution. As a result, creativity is stifled. Oftentimes, the layout chosen for a new warehouse looks exactly like the layout used for the old warehouse. The generation of layout alternatives thrives on the creativity of the layout planner, yet many layout planners withhold this basic and essential ingredient.

The generation of warehouse layout alternatives should be accomplished by the following procedure:

1. *Define the location of fixed obstacles.* Some objects in a warehouse can be located only in certain places, and they can have only certain configurations. These objects should be identified and placed in the layout alternative first, before objects with more flexibility are located. Some fixed obstacles are building support columns, stairwells, elevator shafts, lavatories, sprinkler system controls, heating and air conditioning equipment, and, in some cases, offices. Failure to consider the location of these types of items first will prove disastrous. The warehousing corollary to Murphy's law states: "If a column can be in the wrong position, it will be." Don't be the layout planner who designs a warehouse and buys the storage and material-handling equipment only to find that when the equipment is installed, the location of the building columns makes an aisle too narrow for the handling equipment.
2. *Define the location of the receiving and shipping function.* Oftentimes, the configuration of the warehouse site will dictate the location of the receiving and shipping functions. When this is not true, however, the receiving and shipping location decision becomes an important one. Receiving and shipping are high-activity areas and should be located so as to maximize productivity, improve material flow, and properly utilize the warehouse site. The location of access roads and railroad tracks, if rail service is required, are important considerations in locating receiving and shipping. The question of whether receiving and shipping should be located together or in different areas of the warehouse must be addressed. Common receiving/shipping docks can often result in economies of scale related to sharing space, equipment, and personnel. Separate receiving and shipping areas may, on the other hand, be best to ensure better material control and reduce congestion. Energy considerations are important. Where a choice exists, receiving and shipping docks should not be located on the side of the building that faces north. Avoiding this location reduces the amount of heat loss in the winter from northerly winds entering the warehouse through the open dock doors. The preferred location of the receiving and shipping docks is the south side of the warehouse, with east and west as second and third choices. The particular weather patterns around each warehouse site should be examined, however, to identify the prevailing wind direction at that particular site. Then the docks should be located away from the prevailing wind.
3. *Locate the storage areas and equipment, including required aisles.* The types of storage areas and equipment to be used will dictate to some extent the configuration of the storage layout and the aisle requirements. Be sure to make allowances for the fixed obstacles in the facility. Main warehouse aisles should connect the various parts of the warehouse. The cross-aisle at the end of the storage area may need to be wider than the aisles within the storage area, depending on the type of material-handling equipment used. For example, a side-loading fork truck that can operate with a 7-foot-wide storage aisle may require 12-foot-wide cross-aisles at the ends of the storage aisles to allow maneuvering into and out of the storage aisle.
4. *Assign the material to be stored to the storage locations.* This step in the generation of layout alternatives ensures that storage allowances have been made for all the items to be stored. In addition, it allows the performance of a mental simulation of the activities expected within the warehouse.
5. *Repeat the process to generate other alternatives.* Once a warehouse layout alternative has been established, following the four steps just outlined, the process must be repeated many times to generate additional layout alternatives. Different layout configurations, building shapes, and equipment alternatives should be used. The creativity of the layout planner should be taxed to ensure that each succeeding layout alternative is not essentially identical to the first layout alternative generated.

5.4. Evaluate the Alternative Layouts

A number of warehouse layout philosophies exist to serve as guidelines for the development of an effective warehouse layout. Each warehouse layout alternative should be evaluated against the specific criteria established for each of these warehouse layout philosophies.

1. *Popularity philosophy:* In a typical warehouse, it is not unusual to find that 85% of the product throughout is attributable to 15% of the items, another 10% of the product throughput is attributable to 30% of the items, and the remaining 5% of the product throughput is attributable to 55% of the items. Consequently, the warehouse contains a very small number of highly active items (often called A items), a slightly larger number of moderately active items (often called B items), and a very large number of infrequently active items (often called C items). The warehouse layout philosophy on popularity suggests that the warehouse should be planned around the small number of highly active items that constitute the great majority of the activity in the warehouse. The popularity philosophy maintains that the materials having the greatest throughput should be located in an area that allows the most efficient material handling. Consequently, high-turnover items should be located as close as possible to the point of use. The popularity philosophy also suggests that the popularity of the items helps determine the storage method used. Items with the greatest throughput should be stored by methods that maximize the use of space. For example, if bulk storage is used, high-turnover items should be stored in as deep a space block as possible. Because the items are moving into and out of storage at a relatively high rate, the danger of excessive honeycombing losses is reduced and excellent use of space will result from the high-density storage. Low-throughput items in deep bulk storage blocks will cause severe honeycombing losses because no other items can be stored in that location until the low-throughput item is removed.
2. *Similarity philosophy:* Items that are commonly received and/or shipped together should be stored together. For example, consider a retail auto parts distributor. Chances are that a customer who requires a spark plug wrench will not buy, at the same time, an exhaust system tail pipe. Chances are good, however, that a customer who buys the spark plug wrench might also require a condenser, points, and spark plugs. Because these items are typically sold (shipped) together, they should be stored in the same area. The exhaust system tail pipe should be stored in the same area in which the mufflers, brackets, and gaskets are stored. Sometimes, certain items are commonly received together, possibly from the same vendor; they should be stored together. Similar types of items should be stored together. They will usually require similar storage and handling methods, so their consolidation in the same area results in more efficient use of space and more efficient material handling. An exception to the similarity philosophy arises whenever items are so similar that storing them close together might result in order-picking and shipping errors. Examples of items that are too similar are two-way, three-way, and four-way electrical switches, which look identical but function quite differently. Storing these items close together will inevitably result in order-picking and shipping errors.
3. *Size philosophy:* The size philosophy suggests that heavy, bulky, hard-to-handle goods should be stored close to their point of use. The cost of handling these items is usually much greater than that of handling other items. This is an incentive to minimize the distance over which they are handled. In addition, if the ceiling height in the warehouse varies from one area to another, the heavy items should be stored in the areas with a low ceiling and the lightweight, easy-to-handle items should be stored in the areas with a high ceiling. Available cubic space in the warehouse should be used in the most effective way while meeting restrictions on floor loading capacities. Lightweight material can be stored at greater heights within typical floor loading capacities than heavy materials can. The size philosophy also asserts that the size of the storage location should fit the size of the material to be stored. Do not store a unit load of 10 ft^3 in a storage location capable of accommodating a unit load of 30 ft^3. A variety of storage location sizes must be provided so that different items can be stored differently. In addition to looking at the physical size of an individual item, one must consider the total quantity of the item to be stored. Different storage methods and layouts will be used for storing 2 pallet loads of an item than will be used for storing 200 pallet loads of the same material.
3. *Product-characteristics philosophy:* Some materials have certain attributes or traits that restrict or dictate the storage methods and layout used. Perishable material is quite different from nonperishable material, from a warehousing point of view. The warehouse layout must encourage good stock rotation so that limitations on shelf life are met. Oddly shaped and crushable items, subject to stacking limitation, will dictate special storage methods and layout configurations to use available cubic space effectively. Hazardous material such as explosives, corrosives, and highly flammable chemicals must be stored in accordance with government regulations. Items of high value or items commonly subject to pilferage may require increased security measures such as isolated storage with restricted access. The warehouse layout must be adapted to provide the needed protection. The compatibility of items stored close together must also be examined. Contact between certain individually harmless materials can result in extremely hazardous reactions and/or significant product damage. Specific steps must be taken to separate incompatible materials. Oftentimes, the easiest way to accomplish this objective is through the warehouse layout.

5. *Space-utilization philosophy:* This philosophy can be separated into four areas: conservation of space, limitations on use of space, accessibility of material, and orderliness.
    - **(a)** The conservation-of-space principle asserts that the maximum amount of material should be concentrated within a storage area, the total cubic space available should be effectively used, and the potential honeycombing within the storage area should be minimized. Unfortunately, these objectives often conflict. Increased concentration of material will usually cause increased honeycombing allowances. Therefore, determining the proper level of space conservation is a matter of making trade-offs among the objectives that maximize use of space.
    - **(b)** Limitations on use of space must be identified early in the layout planning process. Space requirements for building support columns, trusses, sprinkler system components, heating system components, fire extinguishers and hoses, and emergency exits will affect the suitability of certain storage and handling methods and layout configurations. Floor loading capacities will restrict storage heights and densities.
    - **(c)** The warehouse layout should meet specified objectives for material accessibility. Main travel aisles should be straight and should lead to doors in order to improve maneuverability and reduce travel times. Aisles should be wide enough to permit efficient operations, but they should not waste space. Aisle widths should be tailored to the type of handling equipment, using the aisle and the amount of traffic expected.
    - **(d)** The orderliness principle emphasizes the fact that good warehouse housekeeping begins with housekeeping in mind. Aisles should be well marked with aisle tape or paint; otherwise, materials will begin to infringe on the aisle space, and accessibility to material will be reduced. Void spaces within a storage area must be avoided, and they must be corrected when they do occur. If a storage area is designed to accommodate five pallets, and, in the process of placing material into that area, one pallet infringes on the space allocated for the adjacent pallet, a void space will result. Because of this, only four pallets can actually be stored in the area designed for five pallets. The lost pallet space will not be regained until the entire storage area is emptied.

The alternative warehouse layouts should be evaluated by comparing each against specific expectations relative to the layout philosophies as discussed here. The layout planner must determine which layout philosophies are most important under the specific circumstances and attempt to maximize the extent to which the recommended layout adheres to those philosophies. Remember, however, that warehousing exists within a dynamic environment; therefore, the layout chosen as best today may not be so as conditions change. The extent and timing of changing requirements in the future should be forecast and a warehouse master plan established to compensate effectively for the changing mission of the warehouse.

6. WAREHOUSE EQUIPMENT PLANNING

Like space planning, effective equipment planning must follow a very specific methodology. The general steps to this methodology are:

1. Specify what functions the equipment must perform.
2. Identify equipment alternatives.
3. Evaluate the equipment alternatives.
4. Select the equipment.

This methodology is appropriate for equipment planning for all warehouse activities: receiving, shipping, storage, order picking, and data processing.

The first step in the equipment planning methodology is to define the function the equipment must perform: what must the chosen equipment be able to do in order to accomplish the desired objective? This question is crucial, and it must be thoroughly answered before one begins to identify alternatives. Failure to adequately specify the objective the equipment must accomplish, and the minimum capabilities the proper equipment should have to achieve that objective, will often result in selection of equipment that fails to solve the real problem. It is amazing how often poor specification of requirements provides a brilliant solution to the wrong problem.

Unfortunately, no standard guidelines exist that guarantee a thorough specification of the capabilities desired of the equipment. What is desired of the equipment will vary, not only from warehouse to warehouse but also from activity to activity performed in a given warehouse. Although each circumstance will require different answers to different questions, the *types* of questions that will allow adequate specification of the capabilities the equipment must have are virtually the same for all activities within a warehouse. Table 3 lists the types of questions one might have to answer before specifying begins. The specifications of the capabilities desired of equipment should then be detailed; they should be in writing; then they should be reviewed by each member of warehouse management who will be affected by the equipment in order to identify any overlooked specifications.

TABLE 3 Typical Information Required to Specify Capabilities Required of Equipment

Objective of Equipment	Typical Questions That Must Be Answered
Unload incoming truck shipments	1. What types of trucks will be serviced?
	2. What types of unit loads will be handled?
	3. How heavy are the unit loads that will be handled?
	4. What combination of unit loads might be found on a given shipment?
	5. Where will the unit loads be deposited after unloading?
	6. What constraints in maneuvering space must be met?
	7. Is lifting capability required? To what heights?
	8. What productivity rates must be achieved?
	9. What other activities will this equipment be required to perform?
Place materials into storage racks	1. What type of storage rack will be used?
	2. What type of unit load will be handled?
	3. How heavy are the unit loads that will be handled?
	4. How high must the unit loads be lifted?
	5. From where will the unit loads be obtained?
	6. What constraints in storage-aisle width must be met?
	7. What constraints in maneuvering space outside the storage area must be met?
	8. What other activity occurs simultaneously in the operating area of this equipment?
	9. What other activities will this equipment be required to perform?
Retrieve materials from storage rack	1. What types of loads will be retrieved? Full unit loads? Full cases? Individual pieces?
	2. How much do the loads that will be retrieved weigh?
	3. What type of storage rack will be used?
	4. How high off the floor is material stored?
	5. What constraints in storage aisle width must be met?
	6. What constraints in maneuvering space outside the storage area must be met?
	7. What order-picking philosophy will be used: zone picking, full-order picking, simultaneous picking of multiple orders?
	8. Where will the materials be deposited after retrieval?
	9. What productivity rates must be achieved?
	10. What other activities occur simultaneously within the operating area of this equipment?
	11. What other activities will this equipment be required to perform?
Load materials into carriers for shipment	1. What types of loads will be handled? Unit loads? Loose cartons? What combination of loads?
	2. How heavy are loads that will be handled?
	3. From where will loads to be handled be obtained?
	4. Is lifting capacity required? How much?
	5. What types of carriers will be loaded?
	6. What maneuvering space constraints must be met?
	7. What productivity rates must be achieved?
	8. What other activities must this equipment perform?

The next step in equipment planning is to identify specific equipment alternatives that meet the needed specifications. This step is critical because if the ideal equipment for the job is never identified, then obviously the ideal equipment will not be selected.

At this point in the equipment planning process, the intent should not be to identify the specific make or model of each alternative but rather to identify generic categories of alternatives. First, one must compare the various generic equipment alternatives in order to identify the best alternative; then, in step 4 of the equipment planning process, the specific makes and models in that generic category are compared.

Unfortunately, choosing the best equipment alternatives is easier said than done because of the enormous variety of warehouse equipment on the market today. The number of combinations of equipment that can be made to achieve a certain goal is virtually limitless. A great deal of ingenuity and foresight is often required to predict the impact of integrating several types of equipment into a warehouse system. Consequently, the identification of warehouse equipment alternatives is an art as well as a science. The art does not necessarily have to be inborn; it can be acquired by keeping abreast of the capabilities of existing warehouse equipment and new innovations in the state of the art. Excellent sources of continuing education on warehouse equipment are the many trade publications on warehousing and material handling; trade shows, where equipment manufacturers show and discuss their wares; and seminars and conferences on warehousing and material handling.

A proper evaluation of warehouse equipment alternatives must be a *quantitative* comparison of the alternatives. This is not to say, however, that the choice should be based solely on dollars and cents or that the many *qualitative* attributes of the alternatives—flexibility, reliability, or maintainability, for example—have no bearing on the merits of one alternative or another. On the contrary, superior qualitative attributes of an alternative will often overshadow its apparent economic inferiority to the point where this alternative is chosen as best. In fact, where such a decision is made, the ‘‘economically inferior’’ alternative is judged not to be really economically inferior because its qualitative attributes do indeed have economic value. The problem, then, is an inability to express this qualitative value in quantifiable, economic units (dollars and cents). One solution to this problem is to discontinue the attempt to quantify economically the qualitative attributes of an alternative and instead quantify the qualitative attributes in judgmental units and convert the economic units—dollars and cents—into these judgmental units to form a common base for comparison of alternatives.

The first step in the economic evaluation of equipment alternatives is to identify and estimate the relevant costs of each alternative over its useful life. Relevant costs are usually divided into two categories: investment costs and annual operating costs. Investment costs are incurred to obtain the equipment; they occur on a one-time or periodic basis. The most common investment cost is the purchase price of the equipment. Typically, investment costs are depreciable, and they are often subject to capital investment tax credits.

Annual operating costs are the recurring expenses that keep the equipment in operation. Typical annual operating costs are the wages of operating personnel, the costs of equipment maintenance, and the taxes and insurance costs incurred by owning the equipment. Annual operating costs are generally not depreciable.

Once the relevant life-cycle costs of the alternatives have been identified and estimated, a detailed time-value-of-money analysis must be performed for each alternative. Economic analysis techniques are presented in Chapters 52 and 54 of this Handbook.

Consideration of the qualitative attributes of the equipment alternatives may very well result in selection of equipment that might not have been chosen if the economic analysis had been the sole basis of comparison. However, a casual discussion of the intangible, qualitative attributes of the equipment alternatives that causes a reversal of a decision originally based on the tangible dollar costs of the alternatives will often not withstand the scrutiny of those who must review the decision. Consequently, the discussion of the qualitative aspects of equipment alternatives must be explicit and well documented.

Some qualitative factors that are often looked at are:

Ability of the equipment to fit into and serve warehouse operations

Versatility and ability to adapt to day-to-day changes in products and fluctuations in productivity requirements

Flexibility (ease of changing or rearranging the installed methods)

Limitations imposed by the equipment on the flexibility and ease of expansion of the layout, building, or both

Use of space

Safety and housekeeping

Working conditions and employee satisfaction

Ease of supervision and control
Availability of trained operators
Frequency and seriousness of potential breakdowns
Ease of maintenance and rapidity of repair
Volume of spare parts that must be stocked
Availability of repair parts
Quality of product and risk of damage to materials
Ability to pace, or keep pace with, productivity requirements
Personnel problems: training capability, disposition of unnecessary workers, job description changes, and union contracts or work practices
Availability of needed equipment
Tie-in scheduling, inventory control, and paperwork
Effect of natural conditions: land, weather, etc.
Potential delays from required synchronization and peak loads
Supporting services required
Time required to get into operation, i.e., to complete installation, training, and debugging
Availability of capital or investment money
Promotional or public-relations value

The final step of the equipment planning process is to select the specific equipment. The selection process is as follows:

1. Sell management on the proposed equipment and obtain approval for any capital appropriations required.
2. Compose detailed specifications of the equipment required.
3. Identify vendors who can potentially provide the equipment.
4. Prepare and distribute a vendor bid package.
5. Receive and evaluate the vendors' bids.
6. Select and order the equipment.

Evaluating vendors and their specific equipment requires the same decision process that should be used when selecting the desired generic equipment. The decision should be based on a combination of economic and qualitative factors. The obvious economic factor that must be considered is the invoice price, which will include purchase price, sales tax, freight costs, installation costs, and so on. The qualitative factors to be considered will include many of the same factors as the evaluation of generic-equipment alternatives, such as the volume of spare parts that must be kept on hand, the ease of maintenance, the rapidity of repair, and the availability of repair parts and service. In addition, other factors specifically related to vendor selection are the availability of the equipment, installation and debugging services provided, warranties, and the reputation of the manufacturer and its local representative. Consequently, equipment and vendor selection should never be based solely on a low invoice bid. The judgmental units used to select generic categories should be used here.

7. WAREHOUSE OPERATIONS AUDIT

Critical to the ongoing success of storage and warehousing operations is the continuous analysis and evaluation of the day-to-day operations to identify opportunities for improvement. A formal, periodic audit of the existing operations, conducted in a systematic manner, can be an effective tool for achieving continuous improvements. While a number of specific methodologies can be utilized to conduct such an audit, the following discussion will define one approach that has been successfully used.

7.1. Operations Audit Performance Categories

The operations audit is a process that evaluates 10 categories of performance in the warehouse. Utilizing information on the performance of the warehouse in these categories, a quantitative overall performance measure is determined. The 10 categories assessed by the operations audit are:

1. Customer service
2. Control systems
3. Inventory accuracy

4. Space utilization
5. Labor productivity
6. Layout
7. Equipment methods
8. Equipment utilization
9. Building facilities
10. Housekeeping and safety

The first category, customer service, is a primary concern to the warehouse management and upper management. Rating of customer service is based on how well the warehouse performs against its corporate service goals. These goals may include order-to-delivery cycle, order-to-ship cycle, and out-of-stock occurrences.

Control systems is the second category to be evaluated. This is by definition not just computer controls. Evaluation of controls looks at what paperwork is used, how data integrity is used, what duplication of efforts and paperwork exists, how special requests are serviced, and how effective is the use of computer controls, if available. The assessed need for increased computer control is based on the ability of existing manual or computer-controlled operations to adequately control the warehouse. Some indicators that enhanced computer control is needed are the inability to find material, excessive time required to find material, increased obsolescence, and inefficient labor utilization. In most cases a top-rated warehouse has a real-time, online, order entry system that develops truck loads, batches items for picking, preroutes and preposts picking, and manages labor with real-time instructions via data terminals in the warehouse.

Inventory accuracy is critical because many other categories can be affected by poor inventory control. The rating assigned is based on performance against corporate goals. The accuracy of inventory count for all items in total should be considered, as well as the percentage of different items (SKUs, stock-keeping units) found to be accurate in counting. Lack of accuracy on small, inexpensive items can have as big an impact on customer service as on the larger, more expensive items. Item count, dollars on hand, and total part count are all 99% or better, and cycle counting is performed in a top-rated warehouse. Initially, though, the rating assigned should be against the corporate goal, with the goal being improved as consistency is achieved.

Space utilization is calculated for the entire warehouse, based on the storage method being used. The quantity of positions occupied vs. the total available positions is used to calculate the utilization for each type of storage in the warehouse. The utilization of each area, the square footage of the area, and the total square footage of the warehouse are then used to calculate the overall utilization. This overall utilization is compared to the maximum efficient utilization, usually 80–90%, to determine the operating utilization. A review of the projected growth can then be utilized to determine the life expectancy of the warehouse. This life expectancy, the time until operationally full, is used to determine the rating. Typically, a new warehouse takes between 12 and 24 months to design, construct, and occupy. If the life expectancy is less than this time, a low rating is assigned. Also, if the life expectancy is beyond five to six years and growing, a low rating is assigned. Also, if the life expectancy exists, the warehouse space should be reviewed for use by other functions, or for sale.

Labor productivity, category 5, means different things to many warehouse managers. The rating is based on a review of the operating procedures for the warehouse. Each of the major functions in the warehouse—receive, store, pick, and ship—are evaluated. The procedures are reviewed to determine how effectively they support high labor productivity. The existence and use of labor standards is also considered. Effective procedures and proper use of standards is needed for a top rating.

Category 6, layout, is integral to the successful performance of other categories. The objectives of a proper warehouse layout are:

1. To use space effectively
2. To allow the most efficient material handling
3. To provide the most economical storage in relation to costs of equipment, use of space, damage to material, and handling labor

Rating the layout of the warehouse is based on how well these objectives are met. Effective use of space for storage, operational, and support functions is considered here. The transport and storage of material is analyzed to determine how well the layout supports reduced handling costs and increased labor productivity.

Equipment methods refer to the appropriateness of the types of equipment and the use of this equipment in the warehouse. At least two major types of equipment exist in every warehouse: storage equipment and handling equipment. Rating is based not on how much each is utilized, but how well. For example, storage equipment should contain the proper items based on physical characteristics (e.g., size, weight, fragility) and activity level. Handling equipment is also evaluated based on these same characteristics, as well as how the equipment interfaces with storage and delivery points.

Equipment utilization is calculated for each group of equipment in the warehouse. This may include forklift trucks, storage/retrieval systems, conveyors, carousels, and automatic guided vehicles. The utilization, together with the variation in demand, is considered to assign a rating. Too high a utilization can be as detrimental as too low. The operational utilization must be considered with the actual run time for a vehicle. A forklift should travel with a load the majority of the time. High use of vehicles traveling empty is not acceptable. In a top-rated warehouse, the utilization is at or just below the operational maximum with the growth to meet the life expectancy considered.

Building facilities are often overlooked areas of the warehouse. Building facilities include:

Dock capacity
Lighting
Personnel services (offices, restrooms, break areas, etc.)
Fire protection
Outside space (truck aprons, service areas, etc.)

The utilization of docks depends on two factors: the turnaround time to load or unload and the arrival pattern of trucks at the docks. Typically, the dock utilization should be 70–80%. Also, the use of proper dock equipment is evaluated. Dock locks or chocks, levelers, and light should be available. Lighting should not only be sufficient to support operations but also be located properly to avoid equipment interferences. Personnel services should be properly located and of sufficient size and quantity to support the staff. Fire protection is rated based on the type of facility, the equipment methods used, and the type of material stored. Outside space is most often overlooked in building facilities. Proper allowances for truck access, personnel access, and other services can affect safety and efficiency.

The last category is housekeeping and safety. There is a strong relationship between housekeeping and safety. Poor safety conditions do not necessarily mean housekeeping is poor, but poor housekeeping always impacts negatively on safety. Housekeeping is reviewed in several specific areas. Material should be put away, not lying on the floor and in aisles. Empty pallets, cartons, or tools should be stored neatly. The warehouse should be clean; the rack should be aligned properly on working aisles. The issue of safety in the warehouse is a direct function of professionalism. The equipment operators should be trained, certified, and periodically recertified. The equipment should be in proper working order. Lighting and other environmental conditions should be proper for the work. Personnel access to high-traffic areas for equipment should be limited. All material must be stored properly. Bulk materials should be stacked properly, not exceeding allowable load heights. Pallets in racks should have the proper amount of overhang and be loaded within the capacity limits. The rating is based on these rules being followed throughout the warehouse.

7.2. Operations Audit Methodology

Each performance category is rated on a scale of 1 to 5, with 5 being the highest. The rating is based on both quantitative and qualitative assessments. The auditor should record specific factors or indi-

TABLE 4 Calculation of Performance Index

Category	Rating	Target Rating	Weight	Category Score	Target Score
Customer service	1 2 3 4 5	5	40	120	200
Control systems	1 2 3 4 5	4	30	120	120
Inventory accuracy	1 2 3 4 5	5	30	150	150
Space utilization	1 2 3 4 5	4	20	60	80
Labor productivity	1 2 3 4 5	5	20	80	100
Layout	1 2 3 4 5	5	20	100	100
Equipment methods	1 2 3 4 5	5	10	30	50
Equipment utilization	1 2 3 4 5	5	10	50	50
Building facilities	1 2 3 4 5	5	10	50	50
Housekeeping/safety	1 2 3 4 5	5	10	40	50
Total				800	950
Performance index					84%

TABLE 5 Warehouse Class

Warehouse Class	Performance Index	Rating ≤3
(A+) Excellent	95–100%	0
(A) Very Good	90–94%	2
(B) Good	85–89%	3
(C) Average	80–84%	3
(C−) Below average	70–79%	4
(D) Poor	<70%	>4

cators used to obtain the assigned rating. After rating of each category, determination of overall performance is straightforward. Before the results of the audit are calculated, two additional actions must be performed. The first is to assign weights to the performance categories. Each category varies in importance to a particular warehouse operation. To account for this, a weight is assigned to each category, reflecting its relative importance. The total of all weights is 200, which together with the maximum rating of each category (5) yields a maximum possible score of 1000 points. Second, for a particular warehouse the realistic maximum rating for a category may be less than 5. Due to the warehouse size or activity level, for instance, a computer control system may not be practical. Therefore, a target rating is assigned to each category. In future audits of the same warehouse, this target rating may change to reflect consistent high levels of performance or changes in activity and size.

The category rating and target ratings are multiplied by the category weight to obtain a category score and target score. The total of the category scores is then divided by the total target scores to obtain a performance index (PI). Table 4 shows the results of a typical audit and calculation of the PI. In this particular example, a performance index of 84% is achieved. The performance index is one part of the analysis to determine the warehouse class. The other part of the analysis needed to determine the class is based on the consistency of ratings. The number of categories with a rating of 3 or less is obtained. In the example, this is equal to 3. Table 5 shows how the warehouse class is determined for the example in Table 4.

The first step is to identify a class based on the performance index. In each class, there is a limit to the number of ratings that are less than or equal to 3. In the example, the performance index is 84%, which is a class C, which has a limit of three ratings less than or equal to 3. This warehouse has three ratings less than or equal to three, so the class does not change. Exceeding the number of ratings less than or equal to three causes a drop in class, but the inverse is not true. Class cannot improve over the performance index level achieved.

The results of the audit show warehouse management how well the warehouse is doing and provide a tool to communicate this performance. The breakdown of ratings for each category can be used as a road map to plan improvements. The operations audit should be performed on an annual or semiannual basis to track the results of improvement efforts. Following each audit, new goals for improvement in each category must be set, and plans made to implement required operations changes.

ADDITIONAL READING

Ackerman, K. B., *Practical Handbook of Warehousing,* 4th Ed., Traffic Service Corporation, Washington, DC, 1999.

Frey, S. L., *Warehouse Operations: A Handbook,* M/A Press, Beaverton, OR, 1983. A. T. Kearney, Inc., *Measuring Productivity in Physical Distribution,* National Council of Physical Distribution Management, Chicago, 1978.

Nelson, R. A., *Computerizing Warehouse Operations*. Prentice Hall, Englewood Cliffs, NJ, 1985.

Tompkins, J. A., and Smith, J. D., Eds., *The Warehouse Management Handbook,* 2nd Ed., Tompkins Press, Raleigh, NC, 1998.

Tompkins, J. A., and White, J. A., *Facilities Planning,* 2nd Ed., John Wiley & Sons, New York.

CHAPTER 58

Plant and Facilities Engineering with Waste and Energy Management

JAMES R. ROSS
Resource Management Systems (RMS Inc.)

1. **OBJECTIVES AND CONTENT OF THE CHAPTER** **1550**
2. **SCOPE OF PLANT AND FACILITIES ENGINEERING** **1550**
   - 2.1. Definition of Plant Engineering 1550
   - 2.2. Emerging Concept of Enterprise Asset Management 1550
   - 2.3. Relationships between Plant Engineering and Other Departments 1550
     - 2.3.1. Relationship between Plant Engineering and Maintenance 1550
     - 2.3.2. Relationship between Plant Engineering and Production or Operations 1550
     - 2.3.3. Relationship between Plant Engineering and Upper Management 1551
     - 2.3.4. Relationship between Plant Engineering and Product and Process Design and Introduction 1551
   - 2.4. Roles of Plant and Facilities Engineers 1551
   - 2.5. Industry Characteristics That Affect Plant Engineering 1552
3. **INTEGRATING INDUSTRIAL ENGINEERS INTO PLANT AND FACILITIES ENGINEERING** **1553**
   - 3.1. Why Industrial Engineers Make Good Plant Engineers 1553
   - 3.2. Problems That May Face an Industrial Engineer as Plant Engineer 1553
   - 3.3. Facilities Management as a Resource-Utilization Issue 1553
   - 3.4. Training Needed by Industrial Engineers to Become Good Plant Engineers 1553
     - 3.4.1. Knowledge Needed in Other Engineering Disciplines 1553
     - 3.4.2. New Skill Requirements for Plant and Facility Engineers 1554
     - 3.4.3. Training as a Motivator 1555
     - 3.4.4. Assessing Training Needs for Plant Engineering Employees 1555
     - 3.4.5. Types of Training 1556
4. **MANAGING PLANT AND FACILITIES ENGINEERING** **1557**
   - 4.1. Organization and Management of the Plant or Facilities Engineering Function 1557
     - 4.1.1. Strategy for Plant Engineering 1557
     - 4.1.2. Size of Operation 1557
     - 4.1.3. Type of Operation 1557
     - 4.1.4. Managerial Style and Structure 1557
     - 4.1.5. Area of Responsibility 1557
     - 4.1.6. Availability of Qualified Personnel 1558
     - 4.1.7. Budgetary Constraints 1558
   - 4.2. Applying Industrial Engineering Techniques to Plant Engineering Problems 1560
     - 4.2.1. Determining Personnel Requirements 1560
     - 4.2.2. Decision Making and Problem Solving/Data Analysis 1561

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

4.2.3. Benchmarking Plant Engineering 1561
4.2.4. Productivity and Quality in Plant Engineering 1561
4.2.5. Work Measurement Techniques in Plant Engineering 1562
4.3. Financial Aspects of Plant and Facility Management 1562
4.3.1. Plant Engineering as a Profit Center 1562
4.3.2. Budgeting 1562
4.3.3. Costing 1562
4.3.4. Cost Control and Reduction 1563
4.4. Conducting a Facility Survey 1564
4.4.1. How to Conduct a Facility Survey 1564
5. OPERATIONAL ISSUES FOR PLANT AND FACILITY ENGINEERS 1565
5.1. Plant and Facility Design and Construction 1565
5.1.1. Design Using CAD/Computerized Layout Techniques 1565
5.1.2. Building Codes Compliance and Use of Standards 1565
5.2. Plant and Facility Maintenance 1566
5.3. Facility Management/Building Automation Systems 1566
5.4. Buildings and Grounds 1566
5.5. Safety and Loss Control 1567
5.5.1. Safety Management 1567
5.5.2. Loss-Control Programs 1568
5.6. Plant and Facilities Security 1568
6. WASTE-MANAGEMENT CONCERNS FOR PLANT ENGINEERS 1569
6.1. Management and Legal Issues on Waste Management 1569
6.1.1. Legal Issues on Waste Management 1569
6.1.2. Waste Management as a Productivity or Resource-Utilization Issue 1569
6.1.3. Environmental and Waste-Management Productivity and Benchmarking Measures 1570
6.1.4. Economic Considerations on Waste Management 1570
6.2. Waste Streams and Waste Handling 1570
6.2.1 Solid and Hazardous Waste Streams 1570
6.2.2. Solid Waste Handling 1571
6.3. The Industrial Engineering/Environmental Methodology 1571
7. TECHNOLOGICAL CONCEPTS FOR PLANT ENGINEERING 1572
8. MANAGING ENERGY 1572
8.1. Why Energy Is an Important Resource 1572
8.1.1. Why Engineers Should Be Concerned about Energy 1573
8.1.2. What Are Enterprise Resources? 1573
8.1.3. Energy Productivity 1573
8.1.4. Energy Myths 1573
8.2. Energy and Utility Concerns for Plant Engineers 1574
8.2.1. The Energy Process 1574
8.2.2. The Energy System 1574
8.2.3. Managing Utility and Service Systems 1574
8.2.4. Demand and Power Factor 1575
8.3. Financial Considerations about Energy Management 1576
8.3.1. Assessing the Cost of Energy 1576
8.3.2. Justifying Energy-Conservation Opportunities Using Activity-Based Costing 1576
8.3.3. Using Life-Cycle Costing to Assess Return on Energy Projects 1577
8.3.4. Impact of Utility Deregulation 1577
8.4. Relationship Between Energy and Environment 1577
8.4.1. Pollution from Energy Production and Waste Heat Recovery 1577
8.4.2. Cogeneration 1577
8.5. Establishing Strategies for an Effective Energy-Management Program 1577

8.5.1. Strategies and Tactics for Major Energy Improvements 1577
8.5.2. Starting an Energy-Management Program 1578
8.6. Steps in an Energy Assessment 1578
8.7. Energy-Improvement Possibilities 1579
9. SUMMARY: CREATING EXCELLENCE IN PLANT AND FACILITIES ENGINEERING 1582
REFERENCES 1582
ADDITIONAL READING 1583

1. OBJECTIVES AND CONTENT OF THE CHAPTER

This chapter is intended to identify linkages between industrial engineering and plant and facilities engineering, to create an understanding of the scope and breadth of plant and facilities engineering, and to explain selected activities and issues of which industrial engineers assigned to plant and/or facilities engineering duties need to be aware. Where possible, industrial engineering methodologies and techniques will be applied to the design, installation, and management of plant facilities, utilities, and service systems to improve cost effectiveness, productivity, quality, operations, and environment.

While many fundamental principles from the previous edition of this Handbook remain the same, technology, analytical tools, and managerial style that have changed the world in which plant and facilities engineers live and work have been added to make the chapter as current as possible.

2. SCOPE OF PLANT AND FACILITIES ENGINEERING

2.1. Definition of Plant Engineering

Plant engineering may be defined as the entity responsible for providing and maintaining a safe, productive work environment in a constant state of readiness in support of the organization's mission in a cost-effective manner. Facilities engineering deals more specifically with the building, its equipment, utilities, grounds, and closely associated issues rather than those functions that directly support production or general operation (Rosaler and Rice 1983; Rosaler 1994, Higgins 1988).

2.2. Emerging Concept of Enterprise Asset Management

Although the definition of plant and facilities engineering remains the same from the previous edition, a key difference is that plant engineering in many organizations has moved from a discrete functional area within a corporation to providing an enterprise asset management function. The plant or facilities engineer is the steward of assets and resources used by a company and consequently must assume greater responsibilities than in the past. The facility is now viewed as a value-added key to business productivity and competitiveness. As a workplace, the facility is integrated with all other business functions. Specific techniques to facilitate these changes and improve the plant or facility engineer's qualifications are presented throughout this chapter (Davis et al. 1999).

2.3. Relationships between Plant Engineering and Other Departments

2.3.1. Relationships between Plant Engineering and Maintenance

Maintenance is often the operational arm of the plant engineering function, and it may consume over 50% of a plant engineer's total time. Plant engineering and facilities engineering have commonalities with maintenance in problem-solving methods, scheduling, assignment of tasks, and preventive maintenance on the facility and equipment that are used plant-wide. Because many plant engineers are in direct charge of maintenance, procedures for managing both maintenance and plant engineering functions become intermingled. If maintenance is a separate function from plant and facilities engineering, the plant engineer must form strong relationships with maintenance managers to get quick response when critical maintenance problems arise. Chapter 59 of this Handbook describes maintenance in detail (Tatum 1997).

2.3.2. Relationship between Plant Engineering and Production or Operations

In the past, plant and facilities engineering was a separate entity that provided service to production and operations management. With the advent of team management, plant and facilities engineering personnel may be assigned to work directly with production and operations people. The linkage between plant engineering people and production or operations has become indistinct, especially in manufacturing, where line mechanics may report directly to production managers or team leaders. With greater emphasis on just-in-time production, plant engineering must have greater flexibility to meet fast-track schedules and reduce process cycle time.

Allocation of effort between plant engineering and production depends on the type of operation. Generally speaking, operation and maintenance of plant equipment, such as air compressors, chillers, boilers, electric systems, and HVAC systems remain the sole responsibility of plant engineering, but where production equipment is involved, the trend is to have maintenance readily available to assist in repairing production equipment quickly. Production workers are often trained to do first-line or routine maintenance tasks including preventive maintenance. If specialists such as electricians, instrument mechanics, and electronics technicians are needed, these are normally supplied by the central maintenance group, but on occasion these people are also integrated with production operations.

2.3.3. Relationship between Plant Engineering and Upper Management

Despite the blurring of assignments relative to plant and facilities engineering, the plant and facilities engineer will be held accountable for the proper maintenance and operation of facilities-related systems. It is incumbent upon the plant and facilities engineers to communicate with the managers of all departments to obtain their input on situations that need attention and to report progress to upper management using a variety of measures, some of which are described under the section on productivity improvement in this chapter. Ideally, the plant engineer should report directly to the plant manager or general manager, depending on the senior executive in charge of the facility. Because plant engineering is an extremely important and vital function, lines of communication with decision-makers must be as short as possible to minimize reaction time for decisions, optimize allocation of resources, solve problems, and maintain an effective operation.

2.3.4. Relationship between Plant Engineering and Product and Process Design and Introduction

The concept of concurrent engineering in product design and planning is intended to avoid the ‘‘throw it over the wall’’ approach by including everyone concerned with design, marketing, engineering, manufacturing, human resources, warehousing, packaging, and distribution of the product on the product-planning committee. The plant engineer should be involved from the inception of each new product as a member of a team or committee and can contribute to successful product development or introduction in the following ways:

Participate in value analysis studies and help to verify new product manufacturability.
Assist in developing the manufacturing process, flow, and plant layout.
Direct a team to determine facilities needed to produce the new product.
Identify maintenance, utilities, or service requirements for the new product.
Specify, obtain quotes, procure, or fabricate new equipment.
Advise on tooling design and cost with manufacturing engineering.
Determine level of maintenance staffing needed to support the new product.
Manage equipment relocation or new equipment installation.
Concentrate on problem resolution during the startup phase and beyond.
Contribute creative ideas throughout the process.

2.4. Roles of Plant and Facilities Engineers

In some organizations, the plant engineer manages all engineering, maintenance, shops, security, utilities, buildings, and grounds. In other organizations, the plant engineer directs design and construction of new facilities. Often, peripheral functions such as plant safety and security, fire protection, recycling, waste disposal, property records administration, risk management, and pressure vessel inspection and maintenance are included in the plant engineering function. Where no one else is available, the plant or facilities engineer may be called upon to perform other diverse functions that are not engineering but must be done by somebody. This is a compliment to the flexibility of the typical plant engineer. The facilities engineer’s scope deals more specifically with the building, its equipment, utilities, and grounds rather than those functions that directly support production. The proliferation of team management in the last few years has changed the character and style of plant and facilities engineering. These changes are addressed throughout this chapter.

Although called an engineer, the plant or facilities engineer in a large corporation may manage a multimillion-dollar operation, a greater responsibility than that of the president of many small companies. By contrast, a small plant may have one engineer whose responsibilities include plant and facilities engineering. Both large or small plants need to have the functions of plant engineering performed. If these functions are not performed, deterioration in equipment, buildings, grounds, and other facilities can cause cost penalties when repair of damage due to neglect is required later. Every organization pays for plant engineering, whether or not it is actually performed. Plant engineering is a ‘‘pay me now or pay me later’’ profession.

The classical functions of management for plant engineers who have managerial responsibility include planning, organizing, motivating, analyzing, controlling, instructing, delegating, disciplining, and communicating, the same as for managing any other type of enterprise. Not all plant engineers have managerial responsibility, but most have technical responsibilities. Some typical responsibilities of plant engineering managers and technical plant engineers include:

Set departmental strategic and tactical objectives.
Develop policies for the department.
Recruit employees to build an effective organization to perform essential functions.
Devise motivational programs for plant engineering employees.
Participate in development and functioning of team management activities.
Plan, schedule, and assign work directly or though supervisors or team leaders.
Prepare and control capital and expense budgets.
Survey condition of the facility, issue work orders for correction of deficiencies.
Acquire a cutting-edge knowledge of all facility equipment and practices.
Conceptualize, prioritize, and manage improvement or major repair projects.
Procure and install new equipment.
Keep an updated layout of the plant.
Develop and execute productivity and quality measurement systems.
Assess training needs and train employees to perform tasks correctly and quickly.
Establish a program of preventive and predictive maintenance.
Manage maintenance work order, cost, and information systems.
Conduct value analysis studies on equipment, parts, and supplies.
See that material inventory is carefully controlled.
Build the status of plant engineering within the organization with peers.
Communicate with top management to get support for the plant engineering function.
Develop and follow procedures as required under ISO 9000 or ISO 14000.
Manage an integrated information technology/data-management system including CMMS.
Assure that computer support systems are adequate for present and anticipated needs.
Coordinate telecommunications systems installation and maintenance.
Oversee contracting and outsourcing of services.
Assure dependable and cost-effective utility services.
Maintain utility systems in optimum condition.
Direct the safety program and proactively find and correct hazards.
Maintain fire-protection systems in a constant state of readiness.
Perform environmental audits and correct deficiencies quickly.
Negotiate with insurance carriers for maximum protection at the best rate.
Manage the security program for the facility.
Establish and maintain the buildings and ground program for the facility.
Conduct an effective energy-management and conservation program.

This list of responsibilities is intended not to be all inclusive, but to describe the diversity of duties that plant and facilities engineers encounter. Additional duties not included above may be suggested throughout this chapter (Tomlingson 1988a,b).

2.5. Industry Characteristics That Affect Plant Engineering

Although plant and facilities engineering may have originated in manufacturing, there are many other industries and businesses that need plant and facilities engineers. Each of these organizations has a set of unique problems that may not be common to manufacturing. Some examples of nonmanufacturing facilities that require plant or facilities engineers are:

- Public buildings have plant engineers to ensure that heating ventilation and air conditioning, utilities, building maintenance, telecommunications, security, and computerized building management are available and properly maintained.
- Military installations have extensive and unique needs for technologies that are available through plant engineers.
- Research laboratories have, in addition to the usual plant engineering activities, reduction and disposal of hazardous biological ands chemical agents that require highly specialized handling.
- Hospitals and other health-care facilities have long had plant engineers to ensure the facility is in top condition, as the lives of patients depend on a perfectly functioning facility.
- Electric utility generating stations rely heavily on plant engineers to maintain equipment and facilities under adverse conditions.
- Coal mining, oil drilling, and other mineral extraction all require extensive maintenance and plant engineering assistance, also under adverse conditions.

Industrial engineers can bring a wealth of skills to plant and facilities engineering positions in any of the above-named facilities and others not previously mentioned. In preparing this chapter, efforts have been made to make materials as generic as possible so that they apply to any industry or business to which an industrial engineer functioning as a plant or facilities engineer may be assigned.

3. INTEGRATING INDUSTRIAL ENGINEERS INTO PLANT AND FACILITIES ENGINEERING

3.1. Why Industrial Engineers Make Good Plant Engineers

An industrial engineer is an excellent choice to be a plant or facilities engineer. By definition, an industrial engineer is a systems designer who integrates materials, machines, people, technology, and energy to produce goods or services in a productive manner. IE's possess a wealth of analytical tools that can be creatively focused on plant engineering problems. The academic course known as facilities planning and design is fundamental for an industrial engineer becoming a plant engineer. The IE plant engineer can overcome a lack of specific knowledge about technical aspects of plant engineering by keeping many details in clear focus simultaneously and by applying good reasoning, system-integration abilities, creativity, judgment, communication skills, flexibility, adaptability, and quality consciousness, excellent human relations skills, and the principles of industrial engineering.

3.2. Problems That May Face an Industrial Engineer as Plant Engineer

Although an industrial engineer brings an impressive array of skills to the job as a plant engineer, long-time employees, for example, may not readily accept a person who has not previously held a plant engineering position. Selling oneself to the existing workforce should take high priority. The new plant engineer must also be a fast learner, and take steps to become familiar with details of the facility itself, in addition to the engineering, business, and human resource aspects of the new job. Seldom is time available to step back and reflect on these issues, as plant engineering, by design, is a pressure job in which the incumbent must take charge immediately.

3.3. Facilities Management as a Resource-Utilization Issue

As the manager responsible for a multimillion-dollar facility, the plant engineer is accountable for utilization of all resources, under the command of the enterprise manager. When the facility is available for use, people, machines, equipment, and the facility itself are well utilized. If the facility fails, all resources are underutilized and materials can be wasted as well. The plant engineer must also be cognizant of energy use at all times and optimize its use.

3.4. Training Needed by Industrial Engineers to Become Good Plant Engineers

3.4.1. Knowledge Needed in Other Engineering Disciplines

The many technological, analytical, and managerial changes that have occurred concurrently both complicate and simplify the work of a plant engineer. An industrial engineer functioning as plant or facilities engineer must acquire knowledge outside of industrial engineering. Some examples of technologies and equipment related to mechanical, electrical, and civil engineering with which the industrial engineer may be unfamiliar include, but are not limited to, the following:

Mechanical	Electrical	Civil
Piping and valving	Primary power system	Site design
Boiler operation/maintenance	Transformers	Surveying/contours
Machine design	Electrical distribution system	Soil test results
Heating, ventilation	Machine hookup	Structural design
Air conditioning	Building automation	Building envelope
Hydraulics, pneumatics	Electrical maintenance	Water, sanitation
Instrumentation and control	Cogeneration	Waste management

In a large organization, specialists in each of these areas may be available, but in a small organization, the plant engineer may be required, without prior training, to address the above issues plus many more. To achieve a level of competence and expand his or her knowledge base, the engineer can acquire additional skills and knowledge through intensive independent study, academic courses, focused short courses, retraining with outside experts, courses sponsored by equipment vendors, or in some cases, conversation with vendors.

3.4.2. New Skill Requirements for Plant and Facility Engineers

In addition to the above, successful plant engineers will be required to be skilled in the following technologies and techniques:

3.4.2.1. Management Skills Classical management skills of planning, organizing, motivating, controlling, and communicating will continue to be essential, but knowledge of emerging managerial techniques will be the key to future success. Team management in many different forms is growing rapidly, and hence it behooves the plant engineer to become skilled in that managerial style. Authoritarian leaders may survive in some organizations if that is the prevailing managerial style, but with changing workforce attitudes, a more democratic/participative managerial style using team approaches is likely to emerge in the future. A description of team management appears in Chapter 37.

3.4.2.2. Project Management Plant engineers are often called upon to manage major construction or installation projects plus innumerable repair projects. While the level of formality differs from large to small projects, the methodologies of concurrent activity, prioritization, and cost control remain the same. Project-management software is available to optimize utilization of resources as well as remove some routine charting and computational work associated with managing projects. See Chapter 46.

3.4.2.3. Training Skills As the plant engineering field becomes more complex and the availability of trained people diminishes, the need for training increases. Training of engineers and hourly employees in everything from basic maintenance techniques to computer building-control systems will help maintain productivity. The topic of training is addressed in detail later in this chapter due to its current and future importance.

3.4.2.4. Computers Applications for computers in plant engineering organizations are almost endless. As industry becomes more computerized, plant engineers must learn computer skills to compete. A few computer systems applicable to plant engineering are:

Computerized maintenance management systems
Management information systems
Computer-automated facilities management systems
Computer-aided design
Computer-assisted manufacturing
Computer numerical control (CNC)
Programmable logic controllers (PLCs)
Machine-specific computer controls

See Chapters 46, 59, and 72 for more details on the above.

3.4.2.5. ISO 9000 In recent years, the use of ISO 9000 has proliferated. ISO 9000 is often mistaken for a quality system, but it is in fact an organized way to document procedures used in managing an enterprise or department, including quality. Plant engineers are often involved in documenting maintenance procedures and systems, such as work order processing, equipment inspections, predictive or preventive maintenance, and data collection using the documentation discipline of ISO 9000. If a computerized maintenance management system (CMMS) is available, converting data to ISO 9000 requirements should be relatively easy. Although the benefits of documenting procedures can be significant, conversion to ISO 9000 is very time consuming and expensive, and records must be open to auditors who verify that documented procedures are actually followed. See Chapter 74.

3.4.2.6. Quality The need for continuous improvement through application of statistical techniques and the permeation of quality thinking into all aspects of plant and facilities engineering are now a reality. It is no longer possible to accept ''pretty good''quality of workmanship or to solve problems by guesswork. The only acceptable approach is to identify the problem, find the root cause, and eliminate the cause. An attitude of quality must permeate every aspect of organization, including plant and facilities engineering. See Chapters 66, 67, and 73.

3.4.2.7. Energy Plant engineers are often responsible for energy within a facility. With deregulation of energy, great opportunities for cost reduction exist. Energy conservation must also have a high priority to improve profits. A section of this chapter is devoted to energy management and conservation.

3.4.2.8. Telecommunications Many plants and office buildings have telecommunications systems to support electronic commerce, voice and data transmission, computer networking, and many emerging technologies. Transmission may be by wire, cellular, radio, or satellite. Because technology in this field requires frequent upgrading of equipment and supporting software, a successful plant engineer must work closely with information systems engineers and equipment vendors to maintain the state of the art in telecommunications.

3.4.2.9. Environment More stringent environmental laws and regulations continue to force plant engineers to spend massive amounts of time and money to avoid legal penalties for failure to comply with the laws. Despite this, reduction of hazardous waste, solid waste, air pollution, and water pollution offers significant cost-saving opportunities for industrial, plant, and facilities engineers whose key function is the effective utilization of all enterprise resources. More details on waste reduction appear later in this chapter and in Chapters 16 and 19 of this Handbook.

A new tool to ensure environmental compliance and waste reduction is ISO 14000, in which environmental procedures are documented. The documentation and auditing requirements to maintain ISO 14000 certification, although somewhat cumbersome and time consuming, tend to pressure organizations to comply with their own procedures. Environment must be a consideration in every management and engineering decision.

3.4.2.10. New Maintenance Techniques Advances continue in maintenance technology, including more sophisticated techniques of vibration-signature analysis, tribology, motor meggering, oil and wear particle analysis, laser alignment, infrared thermography, acoustic leak detection, and nondestructive testing. These techniques give plant engineers tools to diagnose and predict and/or prevent equipment failures. The concepts of equipment durability, availability, maintainability, and reliability-centered (or based) maintenance are converting maintenance from a reactive to proactive mode. Concurrently, the emphasis has shifted from a power organization to an empowered organization, a shift largely brought about by total productive maintenance. Successful plant engineers must have the vision to utilize all available techniques and technologies to improve maintenance in a new, more demanding environment.

3.4.2.11. New Problem-Solving and Analysis Techniques For many years, the typical five-step problem-solving technique seemed adequate, but new problem-solving techniques seem to appear almost weekly. These techniques center around finding the real problem and the root causes of the problem, leading to a more effective solution. There is seldom a shortage of problems to be solved by a plant engineer, but how the engineering handles those problems can make the difference between success or failure. New problem-solving techniques such as kaizen, 5 ws, 8d, 5s, and others are described in other chapters in this Handbook and in supplemental books and training programs.

3.4.3. Training as a Motivator

Training is intended to develop and enhance employee understanding of current or new responsibilities and skills, help employees keep abreast of new technologies, and improve employee morale, performance, and productivity. It can also be a motivator and reward to employees. Training should also be regarded as an investment in the future of the company by helping management retain valuable employees. It is usually more cost effective to retrain an employee than to hire someone new.

3.4.4. Assessing Training Needs for Plant Engineering Employees

To determine training needs, the plant engineer may ask questions including the following to make an initial assessment:

Are some plant engineering services lacking?
Is quality of work and service substandard?
Are primitive work methods instead of modern techniques being used?
Is there dependence on one individual for certain essential services?
Are some people failing to reach their full potential?
Are some skills lacking or in short supply?
Are new skills needed for plant expansion or new processes?
Are excessive accidents occurring?
Is labor turnover abnormally high?
Is downtime on equipment excessive?
Do some people lack training in teamwork?
Do people have problems in using computers and technology?

A ''yes'' answer to the above questions can verify the need for training.

A skills inventory that lists each piece of equipment or operation and the skills required for operating or maintaining that equipment can indicate the need for training to fill voids in available skills (Peele and Chapman 1989).

3.4.5. Types of Training

The plant engineer must determine the type, level, method, material, and media to be used in the program. A blend of the following types of training will normally be required for a comprehensive training effort:

3.4.5.1. Orientation An orientation program gives basic information about the organization, its history, mission, organization, benefits, policies, products or services, and mode of operation. Time spent in orientation is well worth the cost because it makes the employee feel part of the organization. Omission of this step may lead to labor turnover and low productivity.

3.4.5.2. On-the-Job Training While employees may learn the job themselves, monitored on-the-job training administered by supervisors can save months of unlearning bad habits acquired from self-instruction. The job instruction training (J.I.T. or show and tell) technique, with adequate explanation, is effective for training in manual skills and repetitive or semirepetitive operations. The J.I.T. method is as follows:

1. *Prepare for training:* Learn the job, write the method if possible, see what the employee already knows, explain value of training, establish rapport with the trainee.
2. *Present training:* Demonstrate and explain each step in the job, point out quality and safety expectations, trouble spots, shortcuts, specifications, and other key points, repeat process as required to ensure understanding.
3. *Practice training:* Let employee do the job while explaining to the supervisor the steps just learned, make corrections as necessary.
4. *Pursue:* Follow up to make frequent checks to ensure training has been absorbed, retrain as required, compliment trainee on progress.

Typical plant engineering tasks suitable for this training approach include scheduled lubrication, boiler water testing and blowdown, lift truck battery filling, belt tightness testing, sprinkler system testing, recurring parts replacement, setup and changeover, and repetitive preventive maintenance activities. If a video camcorder is available, a training tape can be recorded to explain the task, for use when new employees are being trained.

3.4.5.3. Internal or External Training The decision on where to obtain training depends on in-house training skills, organizational needs, educational level of trainees, and expected quality of training. Commercial self-study training programs can be integrated into an in-house program to save development time. External training can be cost effective if new skills must be learned quickly. Equipment suppliers may offer training on equipment-specific topics (boilers, air conditioning, computers). Packaged courses on maintenance and plant engineering, in programmed instruction and/or interactive video or computer, can reduce course training development and presentation time (Phelps 1988).

Apprentice training for machinists , electricians, boiler operators, mechanics, or other jobs requiring exceptional skill levels is one solution for skill deficiencies. An apprentice program may be established by assessing the need for such training, determining interest of employees in becoming certified in their job disciplines, designing the work processes and training sessions, obtaining approval for the Department of Labor or other accrediting agency, presenting the program, and rewarding employees who complete the program.

Training of the new plant engineer and all plant engineering personnel is an excellent investment that deserves high priority for the benefit of the organization.

4. MANAGING PLANT AND FACILITIES ENGINEERING

4.1. Organization and Management of the Plant or Facilities Engineering Function

The plant engineering organization is shaped by many factors that vary widely among enterprises, including the following:

4.1.1. Strategy for Plant Engineering

First and foremost, be in harmony with the strategic plan of the enterprise. Not all enterprises have the same strategic plan for plant engineering, so the organization must be shaped to support the unique objectives for the specific enterprise. Functions assigned to plant engineering can lead companies into businesses the mission statement does not recognize. A plant engineer may, by default, become the manager of such businesses as real estate, vehicle repair, electric generating, telephone, building construction, machine building, and scrap metal. Management should make a conscious decision to get into or stay out of a business depending on corporate objectives, skills availability, funding sources, and volume and type of business. Without the strategic decision, the company may be in unwanted businesses it can neither manage nor run profitably. Many companies assign side businesses to contractors. Despite higher unit prices, this may be more cost effective than running an unfamiliar business.

4.1.2. Size of Operation

In a larger organization where delegation is possible, the plant or facilities engineer performs chiefly managerial functions, such as assigning and reviewing the work of others and communicating with superiors, peers, and subordinates. In a smaller organization, the plant engineer may perform many of the engineering, managing, and implementation steps individually.

4.1.3. Type of Operation

The plant engineering organization must be patterned to achieve the mission and objectives of the operation being served. A large, continuous chemical complex will have different organizational needs than a five-day-per-week automotive parts manufacturer or an office building. In a small, light manufacturing plant, the plant engineer often has full responsibility for maintenance, support equipment, and the facility itself, with no subordinate employees. Plant engineers in large plants usually have executive status, but inroads by teams and the level of technology and automation are changing organizational relationships and structures. The degree of automation also has an impact on organization of plant engineering.

4.1.4. Managerial Style and Structure

Until recently, plant engineering organizations were essentially hierarchical, but now team management techniques are being used increasingly in industries and businesses. This is particularly true when total productive maintenance is used. In this concept, maintenance people who were formerly in a separate department are integrated into manufacturing or operating departments, where they are part of a close-knit team rather than outside. While maintenance and plant engineering people may lose some of their individuality and visibility, the benefits of teams in improving productivity are well documented. One advantage of TPM is that intradepartmental conflicts are avoided and everyone is motivated toward a single goal of maintaining production in the most efficient manner.

Another new paradigm is that plant and facilities engineering operates as a professional-services firm composed of multifunctional teams of experts who provide services to other departments within the organization. Charges against departmental or activity budgets are made for services rendered.

4.1.5. Area of Responsibility

The organization of the plant engineering activity depends on top management's perception of its responsibilities in relation to other departments. Plant engineering often acquires unwanted or inappropriate, but necessary, functions that do not fit within other departments. In a TPM situation, traditional reporting relationships of plant engineering personnel may be blurred because individuals may report to operating or administrative departments rather than directly to the plant engineer. The plant engineer is obliged to adopt a new organizational structure and concentrate on communicating with manufacturing or other departments that manage plant engineering people. The plant engineering function should report to the plant manager, general manager, chief engineer, or assistant plant manager. This direct line to key executives can facilitate obtaining the resources necessary to keep the facility operating optimally.

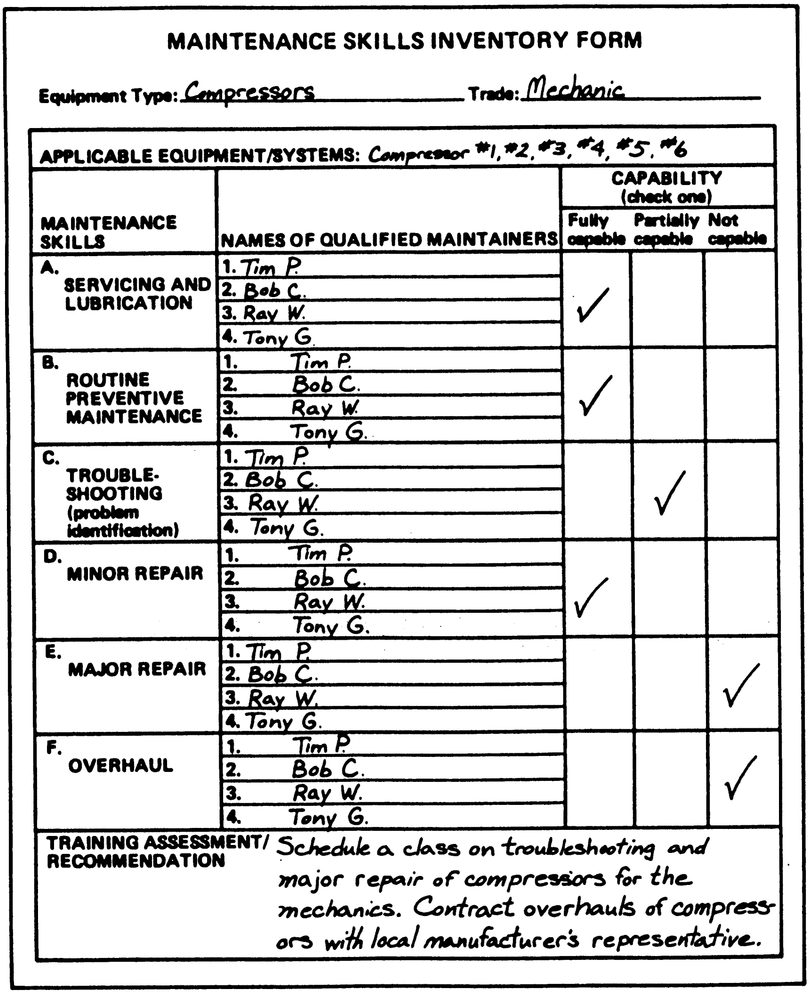
4.1.6. Availability of Qualified Personnel

The availability of qualified personnel relative to operational requirements may force a more pragmatic approach to organization. In highly technical operations, unavailability of qualified people to perform specialized tasks may cause the plant engineer to seek such expertise outside of the organization.

4.1.7. Budgetary Constraints

If upper management views plant and facilities engineering functions as just more overhead, it may be excessively frugal when authorizing budgets. Top managers must realize that plant and facilities engineering are interdependent with production and that adequate funding to perform an optimum job is essential because "You pay for plant engineering whether you have it or not."

Graphical representations of the typical hierarchical organizational structure followed by one of many team configurations are shown in Figure 2 (Lewis and Marron 1973).



MAINTENANCE SKILLS INVENTORY FORM

Equipment Type: Compressors **Trade:** Mechanic

APPLICABLE EQUIPMENT/SYSTEMS: Compressor #1, #2, #3, #4, #5, #6

MAINTENANCE SKILLS	NAMES OF QUALIFIED MAINTAINERS	CAPABILITY (check one)		
		Fully capable	Partially capable	Not capable
A. SERVICING AND LUBRICATION	1. Tim P. 2. Bob C. 3. Ray W. 4. Tony G.	✓		
B. ROUTINE PREVENTIVE MAINTENANCE	1. Tim P. 2. Bob C. 3. Ray W. 4. Tony G.	✓		
C. TROUBLE-SHOOTING (problem identification)	1. Tim P. 2. Bob C. 3. Ray W. 4. Tony G.		✓	
D. MINOR REPAIR	1. Tim P. 2. Bob C. 3. Ray W. 4. Tony G.	✓		
E. MAJOR REPAIR	1. Tim P. 2. Bob C. 3. Ray W. 4. Tony G.			✓
F. OVERHAUL	1. Tim P. 2. Bob C. 3. Ray W. 4. Tony G.			✓

TRAINING ASSESSMENT/RECOMMENDATION Schedule a class on troubleshooting and major repair of compressors for the mechanics. Contract overhauls of compressors with local manufacturer's representative.

Figure 1 Example Skills Inventory for Plant Engineering Personnel. (Courtesy of Plant Engineering Magazine)

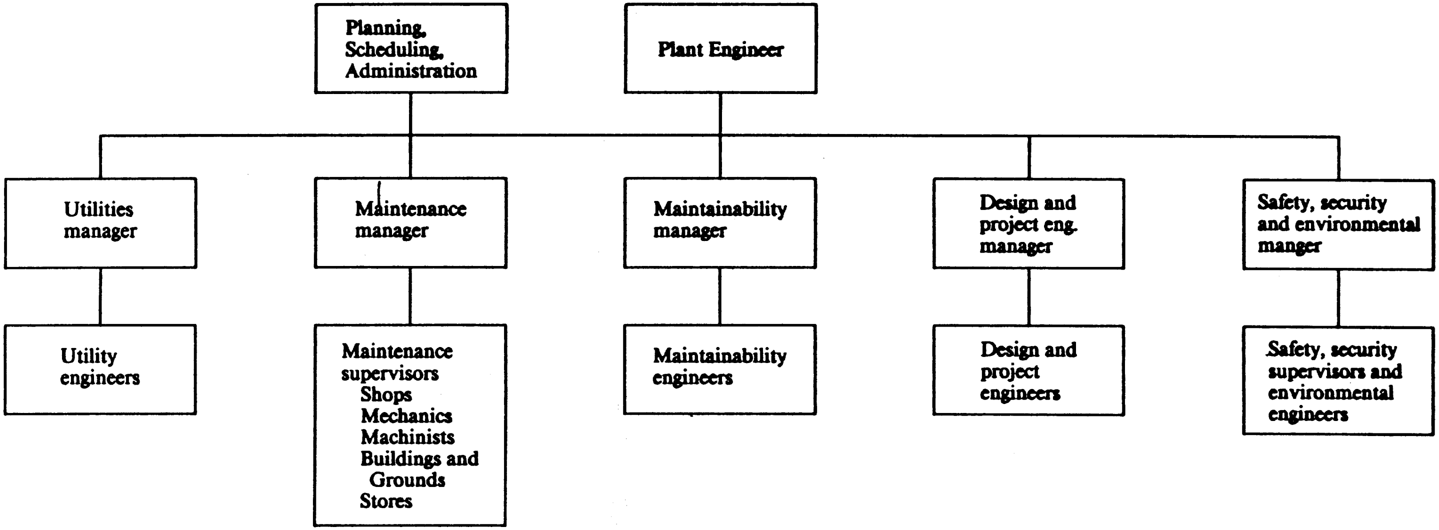

Planning, Scheduling, Administration
Plant Engineer
Utilities manager
Maintenance manager
Maintainability manager
Design and project eng. manager
Safety, security and environmental manger
Utility engineers
Maintenance supervisors
Shops
Mechanics
Machinists
Buildings and Grounds
Stores
Maintainability engineers
Design and project engineers
Safety, security supervisors and environmental engineers

Figure 2 Classical Plant Engineering Organization Chart.

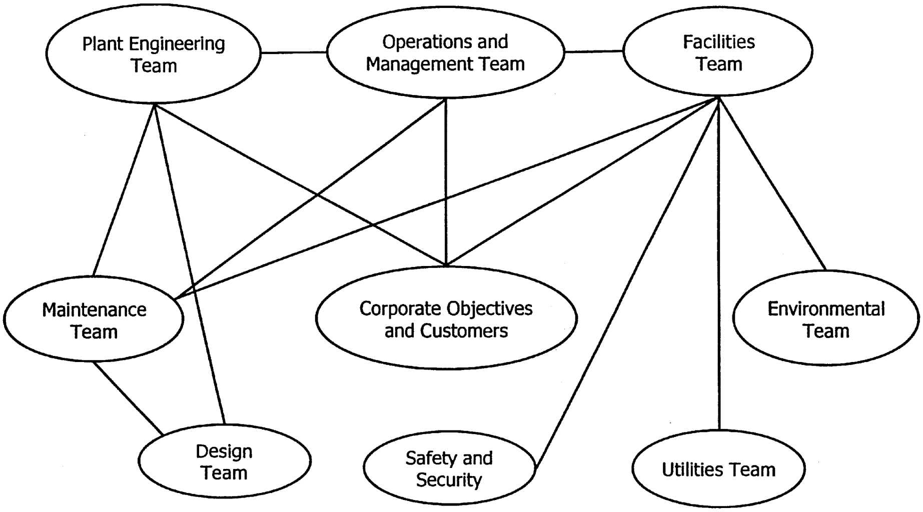


Figure 3 Team-Oriented Organization Chart.

4.2. Applying Industrial Engineering Techniques to Plant Engineering Problems

4.2.1. Determining Personnel Requirements

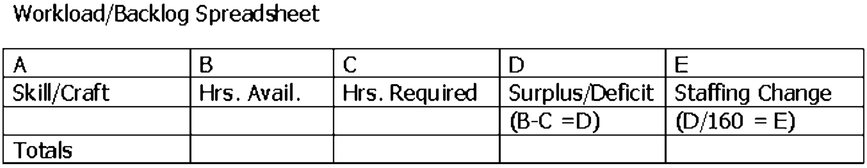
Determining personnel levels is best done by determining the workload for each activity in the plant or facilities engineering department through time study, work sampling, work order estimation, predetermined times, standard data, video camcorder, or other accepted work measurement technique. Although plant engineering work is often regarded as impossible to measure, there are many repetitive or semirepetitive operations that can be measured with reasonable accuracy. The fact that *not all* operations can be measured accurately should not deter the plant engineer from obtaining times for those that can be measured.

When the forecast workload becomes available, the plant engineer should compare total hours needed with those available and compute the difference. A workload and backlog report (Figure 4) prepared on any spreadsheet software or by a CMMS and issued monthly to management can help to justify requests for additional employees.

Station jobs (such as boiler operator), which involve continuous attendance vs. jobs where the workload varies according to assignment or demand for service, should be measured according to the actual work content in the operation. If other work can be done on a station job without sacrificing the primary mission, productivity of the operation can be improved.

Factors affecting personnel requirements include type of industry, size and age of the plant, degree of automation, equipment complexity, amount of capital work, level of preventive maintenance, quality requirements, degree of excellence expected and the qualification, availability of labor, and training level of the existing workforce (Palko 1989). Difficulty in finding qualified people to fill plant engineering and maintenance jobs motivates some companies to invest in retraining to improve employee competence.

Because plant engineering and maintenance labor cost are visible, these functions are frequent targets of cost cutters. Invisible to top management and buried in overhead or other categories are



Workload/Backlog Spreadsheet

A	B	C	D	E
Skill/Craft	Hrs. Avail.	Hrs. Required	Surplus/Deficit	Staffing Change
			(B-C =D)	(D/160 = E)
Totals				

Figure 4 Example Workload and Backlog Report.

costs of excessive downtime, obsolete equipment, utility outages, poorly maintained machines, or other causes of low productivity. Identifying such problems using activity-based costing or value stream analysis allows adequate staffing for plant engineering activities to be justified.

4.2.2. Decision Making and Problem Solving/Data Analysis

As a manager of a significant part of the operation, the plant engineer must be an effective decision maker and problem solver. Unlike the manager of a production operation, the plant engineer is typically faced with new challenges, unusual problems, and nonrecurring decisions almost every minute of every day. Solutions and decisions must be made quickly and accurately for the benefit of the organization. The decision-making and problem-solving process is as follows:

1. Identify the problem, not symptoms—use statistics to find the real problem.
2. Collect all available causes and other data—make fishbone diagrams.
3. Analyze data—develop creative ideas.
4. Select and test alternatives.
5. Pick the best solution—test and implement the solution.

Seldom does a hands-on plant engineer have the luxury of time to implement the formalized approach. Using the above thought process coupled with other industrial engineering knowledge, the plant engineer can intuitively and informally develop solutions to many diverse problems instantaneously. While the above basic method for problem solving is valid, familiarity with kaizen, Ishikawa, 8D, 5 why, brainstorming, decision tree, and other solution-generation methods is advantageous.

4.2.3. Benchmarking Plant Engineering

A plant engineer who wants an organization on the cutting edge of technology should consider benchmarking the organization against others in comparable fields. The benchmarking process involves sharing information with other companies and emulating best practices in critical areas (Gulati and Lach 1997; Raymond 1993).

4.2.4. Productivity and Quality in Plant Engineering

Plant engineering in both manufacturing and service facilities and operations use the same resource inputs of capital, equipment, humans, materials, and energy, but outputs can be quite different. The outputs of manufacturing are chiefly discrete products measured in pieces, tons, gallons, dollars, or other similar measures. Measures of output in a service facility may be in industry-specific terms such as meals served, patients treated, students enrolled, cars parked, nights occupied, or the dollar value thereof.

Productivity measurement is expressed as:

$$\text{Productivity} = \frac{\text{Output}}{\text{Input}}$$

Using the basic formula, meaningful measures of productivity for a variety of inputs and outputs can be developed to track progress toward a more productive plant engineering activity. Examples of productivity measures for plant engineering concerns are:

$$\text{Productivity (Facility)} = \frac{\text{\$ Value of Production}}{\text{\$ Facility Cost}} \text{ or } \frac{\text{Pieces Produced}}{\text{Ft}^2 \text{ Floorspace}}$$

$$\text{Productivity (Equipment)} = \frac{\text{productive hours}}{\text{hours scheduled}} \text{ or } \frac{\text{pieces produced}}{\text{equipment hours}}$$

The use of industrial engineering techniques in plant engineering can be the key to improvement of productivity. Work measurement, work sampling, computerized maintenance management, equipment down time analysis, predictive maintenance, proper layout, effective floor utilization, and effective strategy planning for facility utilization all tend to improve productivity of resources. Due to the close relationship between maintenance and plant engineering, the following productivity measures for maintenance can be applied or modified to measure plant engineering effectiveness:

MTBF—mean time between failures
MTTR—mean time to repair
FMEA—failure mode effect analysis
Uptime of equipment or downtime avoided

Accidents avoided/accident rate
Energy losses eliminated
Material costs reduced

The use of productivity formulas is limited only by the imagination of the user. Formulas must be relevant, meaningful to the user, and motivate positive action and continuous improvement in productivity (APC 1981; Steele 1997).

The success of the plant engineering function must be measured not only in quantity but also in quality of work produced and quality of the organization. There are many assessment tools for quality that address plant engineering. If the organization is competing for the Malcolm Baldrige Award, Shingo Prize, North American Maintenance Excellence, or other awards of excellence, going through the process will motivate maintenance and plant engineering to improve performance. Assessment tools such as Tompkins Associates' Maintenance Scoreboard can highlight gaps that when remedied can lead to significant improvements.

4.2.5. Work Measurement Techniques in Plant Engineering

Although there is a widely held perception that maintenance and plant engineering work is not measurable, there have been many successful applications of work measurement to these functions. A number of consulting firms offer maintenance standards, based on MTM or other predetermined time systems, that can be applied to maintenance and plant engineering work. Measurement of plant engineering activities uses a combination of work sampling, time study, and time recording. These times can be used to determine the level of effort and the number of the people required to perform plant engineering tasks.

In some instances, maintenance standards have been developed from extensive time study and converted to standard data, which are then applied to specific jobs being performed by mechanics. The actual time taken by the mechanic is then compared to the computed standard based on work content and the percentage is used to calculate a day work or incentive efficiency. Station jobs such as boiler operators may use a simple reporting system which indicates that assigned work is being done. See Chapter 54 for more information on work measurement.

4.3. Financial Aspects of Plant and Facility Management

4.3.1. Plant Engineering as a Profit Center

Most plant engineering, maintenance, and facilities engineering activities have been viewed by accountants and management as expense items or cost centers. Only recently has the concept of viewing these entities as profit centers been promulgated. While this may seem like a minor change, it can add to the profitability of the enterprise through reducing or controlling cost of the facility. Plant engineering and maintenance activities are viewed in some circles as value-added components of the value stream. By applying preventive and predictive maintenance and effective plant engineering practices, plant engineering departments can increase availability of the facility and equipment, thus adding real value to the organization.

4.3.2. Budgeting

Plant engineers are required to submit budgets for materials, suppliers, labor, and capital investment. The budget is a list of proposed sources of income and expected expenditures. When approved, the budget may be viewed as a bank account against which charges can be made. In some cases, departments are charged as a professional service, and it is possible for plant engineering to earn a surplus.

Many organizations still use the analysis-of-variance method to determine whether cost objectives are being met within budgeted amounts. This approach gives some indication of actual vs. planned expenditures, but because data are released months after actual events occur, it is difficult to trace how activities could have been performed better.

Activity-based budgeting is ideal for tracking plant engineering functions. Identifying activities actually performed along with the cost drivers allows real-time information to be generated and problem areas identified quickly. Cost distributions between plant engineering and other departments can be done more equitably, and non-value-added activity can be spotted more easily.

4.3.3. Costing

As noted in the section on budgeting, it is incumbent on plant engineers, like all managers, to maintain close control of costs, and make decisions based on accurate cost data. While other chapters in this Handbook detail costing systems, and other cost issues have been noted elsewhere in this chapter, a plant engineer should utilize activity-based costing data to identify non-value-added activity, both within the plant engineering function and throughout the organization. Plant engineers are often blamed for breakdowns, power outages, and other situations that produce non-value-added activity. Elsewhere in this chapter it is noted that a well-run plant engineering function adds value and can enhance profitability for the organization. Activity-based costing is one of the best methods available for collecting cost data and assigning responsibility to the appropriate entity. To implement activity-based costing for plant engineering, the principles are the same as for any other part of the organization, but data from the computerized maintenance management system may be more readily available than in other departments.

4.3.4. Cost Control and Reduction

Plant engineering can add dramatically to the profitability of the company by controlling and reducing costs. A well-run plant engineering organization should establish an annual cost-reduction objective and measure itself against actual achievement. Although careful planning and execution of plant engineering activities can lead to cost reduction and control, the following are specific initiatives plant engineers can take to reduce costs:

1. Ensure cost-effective building designs that have:
   - **(a)** Minimum enclosing ratio (wall area/floor area)
   - **(b)** Minimum partitions
   - **(c)** Efficient heating and air conditioning
   - **(d)** Best utilization of the site (possibly through multiple floors)
   - **(e)** Low maintenance costs
   - **(f)** Standard components
   - **(g)** Lowest-cost material to fulfill function and aesthetics
2. Prepare accurate cost estimates of all construction and repair work.
3. Obtain competitive bids from qualified/reliable contractors.
4. Control construction and repair costs to avoid overruns.
5. Avoid engineering change orders with contractors to control costs.
6. Conduct work sampling studies on plant engineering personnel to solve problems and improve efficiency.
7. Apply methods improvement/work simplification techniques to plant engineering operations.
8. Assist other departments in building equipment jigs, fixtures for cost-improvement projects.
9. Practice effective maintenance management using computerized maintenance management systems if appropriate.
10. Provide optimum machine maintenance to improve productivity and avoid downtime.
11. Train and retrain plant engineering personnel to perform efficiently.
12. Apply incentives to plant engineering activities to motivate higher levels of performance.
13. Motivate plant engineering personnel to improve attitudes and reduce absenteeism and labor turnover.
14. Provide a safe workplace for all employees.
15. Involve plant engineering personnel in teams, problem-solving groups, or other participative groups to solve plant engineering problems. Implement total productive maintenance principles throughout the department
16. Reduce parts and stores inventory.
17. Conduct value engineering studies on repair parts or supply items.
    Use activity-based costing to identify and eliminate non-value-added activity.
18. Use good environmental practices to reduce or recycle solid waste.
19. Apply productivity and quality measures to plant engineering work.
20. Reduce machine setup time through careful planning.
21. Justify capital expenditures on strategic and competitive bases, not solely on discounted cash flow or hurdle rates.
22. Ensure employee comfort through environmental controls and good ergonomic design.
23. Take energy audits and control demand and power factor charges.
24. Automate energy, heating and cooling, security, and related items using computerized building automation systems.

25. Contract out expensive, undesirable, or hazardous operations.
26. Use the most modern maintenance techniques available.

4.4. Conducting a Facility Survey

When a plant engineer assumes responsibility for an existing facility, it is advisable to conduct a facility assessment to determine the condition and appropriateness of the major facility systems and components. The objective is to develop a plan for correcting, upgrading, or retrofitting these systems. The survey requires preparation, observation, analysis, planning, and action.

4.4.1. How to Conduct a Facility Survey (Piper 1988a)

1. Obtain drawings of facility and find location of critical components.
2. Peruse records of all systems, components, and equipment—repair history, material, specifications, purchase date, age of component.

Facility Component	Condition	Action	Location
Facility Site			
Roads, parking lots, walk ways			
Curbs, gutters, storm drains			
Grass and trees			
Flower beds and trees			
Building Structure and Envelope			
Trusses			
Columns			
Decking			
Exterior walls			
Footers			
Foundation			
Floors			
Roofs, flashing			
Roof drains			
Windows, frames			
Interior, Finishes			
Interior doors			
Panic hardware			
Wall coverings			
Ceilings and windows coverings			
Restrooms			
Interior lighting			
Mechanical Equipment			
Boilers, pumps, economizer			
Boiler chemical system			
Blowdown system			
Gas lines, meters, regulators			
Oil tanks, lines , pumps			
Chillers			
Cooling towers			
Water heaters			
Gas distribution systems			
Air compressors, dryer, tanks controls			
Air conditioning, handling units			
Units heaters, Unit ventilators			
Exhaust fans			
Dust collectors, fans, ducts			
Air makeup units			
Control systems			
Electrical			
Primary service			
Transformers			
Switchgear			
Distribution system, duct			
Breaker panels			
Motors			
Lighting			
Building automation system			
Battery chargers			
Uninterruptible power supply			
Meters			
Energy controller			
Electronic Infrastructure			
Telecommunications center			
Telecommunications system			
Satellite communication system			
In=Plant radio, pager, voice system			
Computer data busway, LAN			
CNC download network			
Facsimile systems			
Service Systems			
Elevators, escalators			
Dock lifters			
Bridge cranes, jib cranes			
Waste handling systems			
Trash compactor			
Waste incinerator			
Recycling baler			
Wastewater, sewer system			
Lunch room, cafeteria			
Nurse station, first aid			
Security, Fire Protection			
Perimeter fences			
Security gates			
Guard house			
Security lights			
Security/key card readers			
Motion detectors			
Video cameras and monitors			
Doors, locks			
Manual/automatic alarms			
Sprinkler pipe, valves pump, tanks			
Hoses, nozzles, hose enclosures			
Extinguishing systems			

Condition/Action/Location Codes

Condition Codes
A. Deteriorated
B. Bad structure
C. Broken
D. Rotting
E. Inoperative
F. Obsolete
G. Malfunction
H. Leaking
I. Inefficient
J. Unsafe
K. Corroded
L. Dirty
M. Clogged
N. Energy loss
O. Environmental risk

Action Codes:
1. Repair now
2. Repair routine
3. Test more often
4. Upgrade/replace
5. Cleanup/cleanout
6. Repaint/redecorate
7. Improved efficiency
8. Improve safety
9. Improve environment
10. Retrain operators
11. Adjust controls
12. Reassign operator

Location Options:
Building #
Grid or Column #
Department/Section #
Machine #

Figure 5 Facility Survey and Action Form. (Adapted from *Building Operation Management*)

3. Cite potential trouble spots.
4. Assemble a team of specialists to survey the facility.
5. Prepare an itinerary and schedule for touring the facility.
6. Develop or use a prepared checklist to avoid missing key items.
7. Take a tour or a series of tours, focusing on one or more components during each tour, noting observed conditions.
8. Review corporate strategy to ascertain future use of the facility.
9. Discuss condition of the facility with other engineering and operations personnel, building occupants, or anyone who can provide insight into the condition of the facility or equipment.
10. Generate a report with lists of problems and deficiencies plus recommended corrective actions.
11. Prioritize the lists by classifying problems as critical, serious, routine, or minor.
12. Make cost estimates of required repairs as a part of the report.
13. Issue work orders to have critical and serious problems corrected immediately or within a short time.
14. Schedule other repairs or problem correction as funds and manpower become available.
15. Repeat the survey every three to six months to determine further deterioration or level of correction from previous survey; take further corrective action as required.

An assessment survey would include a thorough review of the facility systems and components shown in Figure 5. Using the above approach, the new plant or facilities engineer can gain control of assigned responsibilities and develop a plan for resolving identified problems as soon as possible after assuming the new duties.

5. OPERATIONAL ISSUES FOR PLANT AND FACILITIES ENGINEERS

5.1. Plant and Facility Design and Construction

A plant or facilities engineer responsible for design, construction, and startup of new facilities or operations manages the project by maintaining liaison among architects, consulting engineers, contractors, and suppliers to ensure the project is on schedule, within budget, according to specifications, and completed according to the terms of the contract.

5.1.1. Design Using CAD/Computerized Layout Techniques

The plant engineer may be called upon to design a new facility or the flow and layout of an existing facility. With the use of CAD and computerized layout techniques such as CORELAP, multiple options can be evaluated qualitatively and quantitatively to select the optimum solution. Performance of the selected solution can be evaluated further using techniques of simulation and optimization discussed in Chapters 93–102.

5.1.2. Building Codes Compliance and Use of Standards

Building codes are detailed listings of design and performance criteria that must be observed before building occupancy is approved. Codes generally describe types of construction, building limitations, environmental requirements, safety systems, repair and alteration procedures, permits, and fee structures and penalties. Codes can be both an asset and liability. While they may add cost to a building, codes compliance generally ensures a safe building. If an inspector makes unreasonable demands, the code can be used to refute these demands. Enforcement and approval procedures vary depending on the jurisdiction in which the facility is located.

Plant engineers must know which code applies to facilities in each locality. The best-known codes publishing organizations are Building Officials and Codes Administrators International (BOCA), International Conference of Building Officials (ICBO), and Southern Building Code Congress International (SBCC). Further information can be obtained from the Internet.

There are also many published standards for various parts of the building or equipment available from such agencies as the American National Standards Institute (ANSI), the American Society for Testing Materials (ASTM), the American Society of Mechanical Engineers (ASME), the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE), the National Fire Protection Association (NFPA), Underwriters Laboratories (UL), Factory Mutual Engineering Corporation (FMEC), and the Occupational Safety and Health Administration (OSHA). The use of standards in planning and constructing a new or revised facility ensures that safety considerations as well as other building conventions are observed (Steiner 1988).

Standards cover almost every conceivable component of buildings and equipment and are normally followed by architects when the original structure is built. The plant engineer making revisions should be especially careful to follow standards to maintain the integrity of the building, its processes, or components.

The building permitting and approval process contains the following procedural steps:

1. Determine the type of structure needed.
2. Select the site.
3. Determine zoning of site/apply for variances to zoning board.
4. Submit architectural and engineering plans to building codes administrator (BCA) for approval.
5. Obtain a temporary building permit.
6. Apply for permanent permit, obtain full permit.
7. Select contractor(s).
8. Obtain environmental permits.
9. Begin construction.
10. Cooperate during periodic inspections by BCA.
11. Complete building.
12. Pass final inspection or make revision to comply with codes.

5.2. Plant and Facility Maintenance

Whether the plant or facilities engineer is directly in charge of maintenance, this activity is of paramount importance. Throughout this chapter are noted a number of maintenance issues, and the relationship between maintenance and plant engineering is explained in Section 2.3.1. The key issue relating to maintenance for plant engineer is to ensure that all facilities and services are available for production or operations depending of the type of facility involved. In some cases maintenance reports to the plant engineer, while in other cases maintenance is a separate department or is integrated into operations. Details of maintenance are described in Chapter 59 of this Handbook. See also sections throughout this chapter relating to maintenance.

5.3. FACILITY MANAGEMENT/BUILDING AUTOMATION SYSTEMS

Both manufacturing and service facilities can benefit from the installation of building automation systems that monitor energy consumption and automatically adjust heating and air conditioning systems. These computer-based systems may also permit real-time monitoring and management of security, video surveillance, fire detection, utilities, transportation equipment, elevators, room occupancy, equipment operation, lighting or other facility management-related functions. Newer systems have user-friendly software that can be learned by inexperienced personnel in a relatively short time. Some systems use voice synthesizers for communicating with security, fire, management, or maintenance people. An emerging technology known as computer-assisted facility management not only includes the above-mentioned features but also has the ability to do maintenance planning, facility design, space planning, and property management. CAFM, coupled with computerized maintenance management systems, gives the plant or facilities engineer the resources to manage the facility effectively. To make best use of building automation systems, the plant engineers should audit facility performance by collecting and analyzing data collected from the system. Building automation systems are also capable of managing multiple facilities and controlling processes. Data management has become fully integrated with direct digital control of functions and interconnected systems directed by building specific protocols (Tatum 1990; Katzel 1998).

In addition to saving labor, these systems reduce energy consumption, avoid security breaches, reduce fire damage, and provide other intangible benefits that make justification for automated building systems feasible. In selecting a building automation system, it is essential to design the system properly, select systems to meet organizational needs, purchase quality equipment with user-friendly software and system flexibility, and obtain long-term customer service from a reliable vendor (Piper 1999).

5.4. Buildings and Grounds

As part of their duties, plant engineers are typically in charge of the buildings and grounds department that employs the janitorial staff and groundskeepers. While this may seem far afield from engineering, many opportunities for making the functions efficient are available. The key goal is to maintain an orderly workplace both inside and outside of the facility that positively impacts productivity in the operation. While everyone needs to be involved in keeping order and cleanliness throughout the facility, buildings and grounds employees facilitate and support other workers in maintaining a clean and orderly workplace.

The janitorial workforce is typically involved in cleaning the the plant or facility, handling refuse, and performing sanitation duties. In some instances, power sweepers, scrubbers, and powered handling devices are operated by the janitorial staff.

Some of the more common functions in groundskeeping are caring for flowers, trees, or shrubs, mowing grass and weeds, cutting out fence rows, patching holes in roads or parking areas, opening drainage ditches or culverts, and in winter removing snow from parking lots, sidewalks, or shrubbery, to name a few. In performing these functions, equipment used ranges from rakes and shovels to pruning shears, power mowers, chain saws, tractors, backhoes, front-end loaders, snow plows, and trucks. The equipment used depends on the size of the facility and the assignment of duties by management.

The plant engineer can make significant cost savings by ensuring that buildings and grounds services are provided efficiently by taking the following actions:

1. Recognize the value of employees who perform buildings and grounds work.
2. Set a budget to cover the grounds maintenance needs of the organization.
3. Train people in use of equipment and certify operators of larger equipment, especially in safe use of lawnmowers, chain saws, tractors, and other dangerous machinery.
4. Issue safety shoes, goggles, hard hats, ear muffs, gloves, back supports, and other safety equipment and ensure that equipment is used to protect employees.
5. Simplify tasks by providing adequate equipment for all assigned tasks. (See also Section 6.2 on waste handling.)
6. Apply work measurement to find time for janitor and groundskeeper work, measure efficiency, and determine the correct number of people assigned to the size and type of facility being served (Use work sampling, time study, and published standard data.)
7. Provide adequate supervision to plan and schedule the job to ensure that all necessary work is done on the correct frequency and within time available.
8. Periodically inspect quality of work being performed, reward work well done, and make corrections if work is unsatisfactory.
9. Conduct work sampling studies to identify potential problems and solve problems as they are recognized.
10. Institute teams to reduce reliance on supervision.
11. Keep equipment maintained properly and practice preventive maintenance with reporting/documentation that PM has been performed.
12. Consider outsourcing all or portions of the buildings and grounds functions to contractors whose qualifications, credit ratings, and performance records have been thoroughly screened.

Proactively creating a well-organized buildings and grounds staff can greatly improve productivity and quality of life in the facility and minimize problems associated with this function (Ross 1996).

5.5. Safety and Loss Control

5.5.1. Safety Management

Plant engineers often manage the safety program in plants or other facilities. Although industrial engineers receive a course in industrial safety, they may need to fill knowledge gaps by further study and experience. A comprehensive and effective safety program should include the following process steps, none of which may be omitted:

1. Well-defined management strategy, policy, and commitment to safety
2. Goal for safety improvement set and resources allocated
3. Designated responsibility/authority/accountability for safety
4. Formal rules and procedures for safety
5. Engineering of safety into every process and product
6. Safety training and special awareness programs
7. Supervision committed to and competent in safety
8. Effective safety communication and promotion
9. Positive safety attitudes through motivation and rewards
10. Safety inspections using committees
11. Immediate correction of unsafe acts and conditions

12. Disciplinary action administered for willful unsafe acts
13. Well-equipped, trained, and staffed first aid function
14. Thorough accident investigation with effective remedial action
15. Meaningful safety performance measures with critical analysis
16. Effective preventive and corrective maintenance programs
17. Continual assessment of the safety program
18. An iterative program of continuous safety improvement

The results of an effective safety program not only can be measured in injuries prevented and lives saved but also may have a distinct bottom-line impact on the corporation or entity. Companies that do not have effective safety programs often pay the equivalent cost in terms of workers' compensation and lost productivity. The inference can be made that a safe workplace is a productive workplace (McElroy 1964).

Because an effective safety program in the plant or facility must include training, the ultimate objective of training is to influence the attitudes of employees around the clock. Off-the-job accidents cost industry many times more than on-the-job injuries. Although workers' compensation is not involved, absenteeism from off-the-job injuries results in productivity losses and insurance cost increases. Employees injured off the job may, with the assistance of unscrupulous lawyers, sometimes claim the injury occurred during work time. Such fraud can be combatted by effective record keeping and supervisory vigilance.

A plant engineer charged with the responsibility for safety is obligated to comply with OSHA regulations, but must as a professional obligation administer an effective safety program beyond OSHA regulations that concentrate on conditions. Accidents are caused by people, and a well-trained, safe worker can avoid unsafe conditions and attendant accidents. (See also Chapter 43.)

5.5.2. Loss-Control Programs

Administration of the plant loss-control program may fall to the plant engineer or may be handled jointly with the personnel manager. Loss control includes the insurance program plus fire, security, and safety. The objective of loss control is to provide uninterrupted operations and minimize losses of life and property from fire, theft, accidents, and other such occurrences. Safety and security are addressed elsewhere in this chapter, and a brief treatise on fire protection follows.

Many organizations have regular visits by insurance inspectors who identify potential fire hazards. Although some inspectors may go to extremes (such as sprinklers under desks), it is advisable to follow recommendations by insurance inspectors where feasible. If inspections do not occur, the plant engineer should conduct a fire risk survey to identify fuels, ignition sources, fire propagation routes, fire detection and extinguishing systems, and life-saving measures. The adequacy of fire-detection and extinguishment systems should also be assessed and corrective measures taken immediately.

Sprinklers should be tested regularly to ensure that all systems are ready should a fire strike. If a booster pump is in place, it should also be tested. Because as booster pumps are high horsepower, the testing alone can add to electrical demand charges. Having the pump on a safe circuit and testing during off-peak loads can save some of the demand charges.

A well-trained fire brigade composed of plant employees can respond to a fire within 1–4 minutes in most facilities, compared with 5–15 minutes for a municipal fire department. Most fires can be controlled in the first 5 minutes if a rapid response occurs. Fire brigade members should be recruited from maintenance or operations ranks dispersed throughout the facility. Training should consist of first aid, evacuation procedures, and the location and use of fire extinguishers, hoses, sprinkler valves, and other equipment. Municipal fire departments should also be acquainted with the plant layout, fire-fighting equipment, and special hazards.

5.6. Plant and Facilities Security

The security function usually includes fire prevention and reporting, crime prevention and detection, risk management, and administration of security personnel. A security survey should be taken to determine the scope of the security function, the condition of this service, and the steps necessary to bring security up to standard. Outside guards may be necessary to prevent unauthorized entry to critical manufacturing areas, laboratories, computer rooms, and other sensitive areas. The level of security depends on the sensitivity and confidentiality of the work, the labor situation, the local crime situation, and the proximity of fire and police protection (Piper 1988b).

Special protection should be given to computer records. Not only should access to computer areas be denied to unauthorized persons, but codes and passwords should be carefully restricted. Computer rooms should be fire resistant, with special halon extinguishing systems for protection of electronic circuitry and magnetic media, the loss of which could be catastrophic for the business.

By using the checklist that follows, the plant or facilities engineer can cite deficiencies in the security system and take action to correct deficiencies (Piper 1988b; Pearlman and Cana 1999):

Is a security plan currently in effect?
Is a key control system rigidly enforced?
Are penalties for unauthorized entry or use of keys enforced?
Are safe combinations, computer access codes, and other sensitive information closely controlled?
Have security personnel been thoroughly screened and trained?
Are security patrol schedules revised regularly to avoid established patterns?
Are employees carefully screened for past criminal behavior?
Are security personnel trained in sprinkler system operation and cutoff? In fire extinguisher use?
Are computer operating and records rooms, laboratories, and other sensitive areas secured by modern personnel identification systems (card access, hand print, eye retina)?
Is the perimeter of the plant or facility protected to limit access?
Are gates kept closed when employee access is minimal?
Are fences maintained and inspected regularly?
Are electronic fire, movement, and detection monitoring devices (ultrasonic, seismic, infrared, contact) installed and operational?
Does the electronic building control system automatically notify fire and police as well as plant security personnel of incidents?
Are video monitors/CCTV and intrusion-detection monitors used for constant surveillance of critical areas?
Are positive identification systems for personnel installed at plant entrances?
If contract guards are used, does the guard service have a blemishless record for guard selection and training, service and reliability?
Does a disaster plan and organization exist with delegated responsibilities? Do security personnel know names of all managers, employees, agencies, and emergency services to be contacted in a disaster?
Does a public relations plan exist and is a spokesperson designated to provide information about the disaster to the press and public?

6. WASTE-MANAGEMENT CONCERNS FOR PLANT ENGINEERS

6.1. Management and Legal Issues on Waste Management

Waste management, environmental law, and other environmental issues must be addressed by plant and facility engineers. The information that follows should guide a new plant engineer with an industrial engineering background in gaining control of the waste management and environmental program as quickly as possible.

Industrial engineers by definition eliminate waste, and they can apply industrial engineering methodology to waste reduction. Waste was paid for as part of a purchased item.

Although many managers take a nonchalant attitude toward waste management, management must realize that effective waste management can improve profitability. To prevent waste, management at all levels must learn to think of the *total system or process* and include waste management in all decisions. They should understand that the beginning of the waste stream is not at the trash dock, but in product design, purchasing, engineering, or even top management. It should never be assumed that any waste item cannot be eliminated, utilized more effectively, or recovered.

6.1.1. Legal Issues on Waste Management

Too often, management makes decisions about waste management based solely on avoidance of fines resulting from many environmental laws in and court decisions instead of making rational decisions based on good management practice. It behooves the plant and facilities engineer to become intimately familiar with all applicable environmental laws and regulations. Space in this chapter does not permit a review of laws, but some assistance can be found in Chapter 19 of this Handbook.

6.1.2. Waste Management as a Productivity or Resource-Utilization Issue

Industrial engineering by definition maximizes the utilization of resources including capital, machines, material, people, data, energy, and technology by devising innovative systems for production or service. The objective of productivity improvement is to maximize the utilization of these resources.

Waste avoidance or reduction maximizes the material resource and reduces labor and equipment requirements as a result of less handling and disposal effort.

6.1.3. Environmental and Waste-Management Productivity and Benchmarking Measures

The effectiveness of the waste-management and environmental program can be measured by applying productivity measures internally to the organization or externally as benchmarking measures to compare performance with similar organizations. Measures include:

Labor productivity (output/man hour)
Handling equipment productivity
Operations productivity
Energy productivity (output/M BTU)
Labor content of waste management activities
Cost of waste handling and disposal
Value of waste handled and discarded
Volume of waste handled
Weight of waste handled
Production downtime incurred (or avoided)
Complaints from EPA, OSHA (or avoided)
EPA fines assessed (or avoided)
Operating costs reduced
Material losses (or losses avoided)
Energy saved or converted
Defects avoided
Quantity or value of waste sold or exchanged

6.1.4. Economic Considerations on Waste Management

Industrial engineers learn to weigh alternative proposals using engineering economy principles. Business school graduates are often taught that there must be a short-term, bottom-line impact for an expenditure to be made. Solutions to waste-management problems may be evaluated using life-cycle costing principles with data supplied by activity-based costing. Applying waste-management and disposal costs to the justification process improves the justification dramatically. While waste disposal is a non-value-added activity, finding ways to eliminate or reduce waste at the source can produce savings through one or more of the following:

Lower labor cost
Less material waste
Less expensive raw materials
Reduced material-handling cost
Lower energy costs
Improved product quality
Reduced maintenance costs
Profit through waste exchange
Reduced long-term liability for improper disposal, spills, and accidents
Avoidance of fines for environmental noncompliance
Reduced transportation, tipping fees, and disposal costs

6.2. Waste Streams and Waste Handling

6.2.1. Solid and Hazardous Waste Streams

Industrial waste streams contain nonhazardous materials such as paper, wood, metals, plastics, fibers, and food waste. Paints, some plastics, and metals may be hazardous or nonhazardous depending on their composition. Many chemicals are considered hazardous and require special disposal methods under stringently controlled conditions.

6.2.2. Solid Waste Handling

The typical solid waste stream consists of nonhazardous packaging material, cafeteria and restroom trash, office waste paper, floor sweepings, and waste materials from processing operations. When possible, material should be sent to a recycling area, with nonrecyclables going to a disposal area.

Engineers often give detailed consideration to material handling of product and raw material but often ignore handling of waste material within a plant. Consequently, default material-handling methods for waste material remain primitive and labor intensive. While every effort should be made continually to eliminate waste material, an IE-oriented plant or facilities engineers should devise cost-effective or innovative methods for handling waste material that has not yet been eliminated.

Waste handling should be viewed as a non-value-added but necessary process. All waste-handling activities should be documented or mapped to determine who is spending time to handle waste and how much this activity is costing at each step by applying work measurement. From these cost data, more cost-effective methods of handling and disposal can be developed.

Simply designing a route for waste handlers to follow and monitoring their methods can improve efficiency in waste handling. Savings opportunities exist if the study finds that production people are spending time handling or disposing of waste materials while the production operation remains at a standstill. Lift trucks from production areas are sometimes used for transfering trash to disposal area, but this practice interferes with production when they are needed to deliver or remove palettes from production operations. A lift truck should be assigned to waste handling to avoid this situation, especially if distances to the disposal area is long.

Depending on the type of operation, such handling devices as dumping hoppers, tilt carts, four-wheeled carts, trailer trains, rolling waste cans, scrap conveyors, chutes, and pallets are used to move waste material for recycling or disposal. At the disposal area, waste material is sometimes placed on an open trailer or dumped into a trash compactor. A skid steer loader may also be used to load a trainer or compactor. Mechanization is desirable to reduce costs to the extent possible if waste-reduction efforts have not fully eliminated waste.

If a contractor is engaged to handle waste in the plant, the handling equipment may be owned by the contractor, who may work on a fixed-fee basis either by tonnage or hours expended. While this approach may reduce equipment investment, the contractor should indemnify against liabilities for contractor personnel and ensure that the company is compensated for damage to equipment or interruptions to production caused by the contractor.

6.3. The Industrial Engineering/Environmental Methodology

The industrial engineering-based environmental methodology outlined below is the systems approach to waste management through which industrial and plant engineers can solve environmental problems in a cost-effective and productive manner. The methodology is based on the premise that waste begins at the top management decision-making process. When top management realizes where the value stream containing waste begins, commitment to waste reduction and elimination of non-value-added activity may occur. The revised IE environmental methodology is as follows (Ross 1989, 1999):

1. Help top management understand where the waste stream begins and get support for waste-reduction and environmental-improvement programs.
2. Outline clear objectives and scope for the environmental program.
3. Get everyone involved at all levels of the organization.
4. Handle the legal issues of environment as a top priority as follows:
   Become familiar with all applicable laws.
   Take an environmental audit to find problems before regulators arrive.
   Be sure all paperwork is submitted on time to avoid fines.
   Implement effective environmental controls to keep the organization in the lowest possible environmental risk category.
   If an inspector arrives at your facility, be cooperative and don't try to hide anything (if you're caught, it will cost you dearly).
   Implement changes punctually.
   Protest unfair citations or excessive fines.
5. Organize and train teams to address waste-reduction issues.
6. Make a process map of all activities performed in all processes.
7. Identify, quantify, and prioritize waste streams at any point in each process.
8. Implement effective waste-management tactics, including the following, which appear in descending order of value:
   Eliminate or reduce waste streams at the source.

Redesign the product to reduce waste—use value analysis.
Change processes, conditions, and procedures to reduce waste.
Reevaluate reality of quality requirements that may produce waste.
Purchase good-quality, nonpolluting materials.
Insist on returnable containers—set up a container or pallet pool.
Exchange waste with other companies.
Find another product to use waste productively.
Segregate waste and reduce each type.
Recover resources from waste.
Reclaim, reuse, recycle wastes even if they only break even.
Find secondary outlets for waste.
Improve material handling of waste; avoid makeshift handling methods.
Improve maintenance procedures on equipment to reduce waste
Mix wastes into compost for landscaping and ground cover.
Cogenerate waste material to make electricity and reduce waste volume.
Incinerate waste material for process, water, or comfort heating.
Dispose of waste that cannot be reclaimed to a landfill as a lost resort.

9. Take a long-range perspective; don't look for a quick fix.
10. Do justifications based on activity-based and life-cycle costing, including environmental and social costs.
11. Train people in the organization to reduce waste and reward people for exceptional effort in the environmental area.
12. Benchmark with similar organizations and emulate best waste practices.
13. Seek outside assistance such as universities and trade associations.
14. Develop or recognize economic incentives for waste reduction.
15. Make the process of waste reduction iterative and repeatedly review processes to make further waste reductions.
16. Evaluate results and make changes in the program as necessary.

By using the methodology outlined above and keeping cognizant of the many industrial engineering-related resource conservation issues, plant engineers can improve the environment, resource productivity, organizational profitability, and quality for all stakeholders.

7. TECHNOLOGICAL CONCEPTS FOR PLANT ENGINEERING

Throughout this chapter, reference has been made to many electronic and computer applications with which plant and facilities engineers need to be familiar. Notable among these is the computerized maintenance management system (CMMS), which is often used by plant engineers to schedule and control maintenance and operations in plant engineering areas. This system is described in some detail in Section 3.4.

The second electronic system under the control of plant engineers is the facility management or building automation system, which usually includes an energy controller. These systems are described in Section 5.3.

Depending on the type of operation, process-control computer systems interconnected with manufacturing, engineering, and administration are often needed. Key to interconnectivity is the installation of a fieldbus through which data are transmitted among users. Manufacturing execution systems operate on a distributed control network ring (Kamal Zafar 1998).

Another group of electronic systems used by plant engineers pertains to preventive and predictive maintenance. A description of these systems appears briefly in Section 3.4 and is described more fully in Chapter 59, which pertains specifically to maintenance.

The importance of understanding new technologies has been emphasized in this chapter by outlining the knowledge and skills that a successful plant and facilities engineer needs upon assuming that position. Because the above-mentioned systems are highly varied and extremely complex, it is not within the scope of this chapter to describe these technologies in detail. By being aware of the many possibilities for improvement through technological concepts, the plant engineer can take steps to become competent in the use of electronic equipment and/or to be able to manage persons who are familiar with the details.

8. MANAGING ENERGY

8.1. Why Energy Is an Important Resource

Until 1973, when the oil embargo occurred, energy was regarded as an uncontrollable overhead item, but rapid escalation in energy costs changed this perception and elevated energy to the status of a key resource to be productively utilized. Since energy has become plentiful, many enterprises have simply adjusted their budgets to higher prices and have forgotten the easily implemented opportunities for energy conservation and cost reduction.

This section offers an integrated approach for finding, evaluating, prioritizing, and implementing energy conservation opportunities and energy and utility system improvements. Plant or facilities engineers or other managers should devote a significant amount of time to energy management. If techniques described here are applied, significant energy-saving opportunities should be found.

8.1.1. Why Engineers Should Be Concerned about Energy

Engineers should be aware of and take action to overcome these conditions:

Failure to recognize energy as one of the five key resources to be managed
Missed conservation opportunities sapping profits
Emphasis by engineers on other resources (labor, machines)
Lack of concern about energy in capital decisions
Ignoring that energy price increases continue to occur
A lack of understanding of the total cost of energy
Ignoring that energy savings drop directly to the bottom line
Poor administrative controls of energy costs
Poor maintenance of energy equipment and systems

8.1.2. What Are Enterprise Resources?

Resources are needed to activate any enterprise. Most enterprises have the following resources at their disposal:

Labor/manpower
Material
Machines
Technology/data
Energy

How and in what proportion resources are consumed depends on the type of activity or operation in which the enterprise is engaged. Although energy may be a small portion of running a garment factory or manual assembly plant, a metal fabricating shop with large punch presses or a heavy chemical plant may have a large proportion of its budget in energy. Energy is sometimes an essential part of the process (melting, heat treating, welding, chemical processing), while at other times it may facilitate the operation (heating, comfort, driving machines).

8.1.3. Energy Productivity

As a key enterprise resource, energy productivity can be computed. By definition, productivity measures the output vs. the input, and the index is found by applying formulas similar to the following:

$$\text{Energy productivity index} = \frac{\text{\$ value of plant outputs}}{\text{kWh}} \text{ or } \frac{\text{\$ value added by plant pperations}}{\text{\$ energy cost}}$$

Other combinations may reveal trends in particular situations. Productivity measures plotted over time can show trends in utilization of the energy resource relative to a baseline that can signal problems or show progress in energy reduction.

8.1.4. Energy Myths

To begin an energy conservation program, engineers must overcome myths that inhibit consideration of many opportunities that if implemented would produce sizeable savings for this valuable resource. Some myths are:

Energy costs are a small part of the budget.
We have no big energy consumers.
We’ve already minimized energy consumption.
Its not in the budget.
We don’t have time to reduce energy.

Its cheaper to let machines run.
Energy is plentiful.

8.2. Energy and Utility Concerns for Plant Engineers

8.2.1. The Energy Process

The entire effort to obtain and maintain an adequate, dependable, and cost-effective supply of energy is a process that is more extensive than most people realize. The process should be conceptualized in terms of a business process in which all steps from conception to termination are considered. Although the energy process may vary with the type of enterprise, some typical steps in the energy process include:

1. Planning the process
2. Determining energy needs for process, building heat, and other uses
3. Selecting of energy form
4. Negotiating rates with utilities
5. Designing the optimum system installation
6. Anticipating future needs
7. Detailing maintenance requirements of the system (periodic, preventive, predictive)
8. Conducting daily operation of the system
9. Accommodating environmental concerns
10. Using waste heat or material for energy
11. Conducting energy-related waste disposal
12. Administering the energy effort (invoice processing, etc.)
13. Taking regular energy assessments for continuous improvement in energy conservation
14. Justifying and replacing equipment
15. Replacing energy sources with more efficient or lower-cost sources
16. Iteratively reassessing and reengineering the process

8.2.2. The Energy System

The energy system includes not only the equipment in the plant or facility, but utilities that supply energy (gas, electric, water) to the plant. The system also includes secondary energy supplies such as steam and compressed air. The distribution system, the protective devices, and all equipment needed to supply energy to the process or operation of the facility, equipment, or operations are parts of the energy system. Aside from maintaining good relationships with utility companies, the main concern of a plant engineer or plant manager is to keep the in-plant system functioning effectively at all times.

While this chapter focuses predominately on energy conservation and cost reduction, the importance of system operation and maintenance cannot be overemphasized. We tend to take energy for granted, but continual vigilance is needed to ensure that energy is available for the process to run, the building to be heated, and paychecks to be printed on time. Some action items on energy system operation and maintenance are included in the energy-assessment procedure in Sections 8.6–8.8.

8.2.3. Managing Utility and Service Systems

The effective management of utilities and services is a primary responsibility of most plant and facilities engineers. The scope of utility and service systems includes:

Electric, gas, and water supplied by outside utility systems
Piping systems
Steam generation and distribution systems
Chiller systems, cooling towers
Heating, ventilation, refrigeration, and air conditioning systems
Building instrumentation and control systems
Pollution-control systems, dust-collection systems
Telephone and communication systems
Compressed air systems

To be able to manage these complex systems, the plant engineer should (Rospond 1999):

1. Become familiar with all utility and service equipment
2. Establish contact with all utilities; gain their cooperation
3. Conduct an audit of each system to determine current condition and needed corrective measures
4. Determine availability of installed spares, spare components and replacement parts, and backup systems
5. Ascertain ownership of utility equipment such as substations, transformers, and lift pumps
6. Develop effective preventive and corrective maintenance programs
7. Implement continuous improvement of utility and service operations
8. Upgrade training of employees responsible for these systems
9. Justify and install state-of-the-art equipment where possible
10. Monitor power quality, upgrade system for digital computer systems
11. Improve transformer efficiency

See the Appendix for specific improvement possibilities.

8.2.4. Demand and Power Factor

Demand and power factor are often major invisible electric costs. Many electric utilities do not show data on these items on electric bills, and it is incumbent upon the energy manager to obtain these data. Electric utilities base demand charges on an integrated peak demand in kilowatts over the highest 30 minutes during any 1-, 6- or 12-month period, depending on the utility's policy. Power factor charges occur when inductive or capacitive loads get out of phase with current supplied, as shown in Figure 6.

Power factor is the difference between current used in kilowatts (kW) and current supplied in kilovolt amperes (kVA) as shown in this example.

A small factory has a monthly demand of 237 kW and 324 kVA as measured by utility meters.

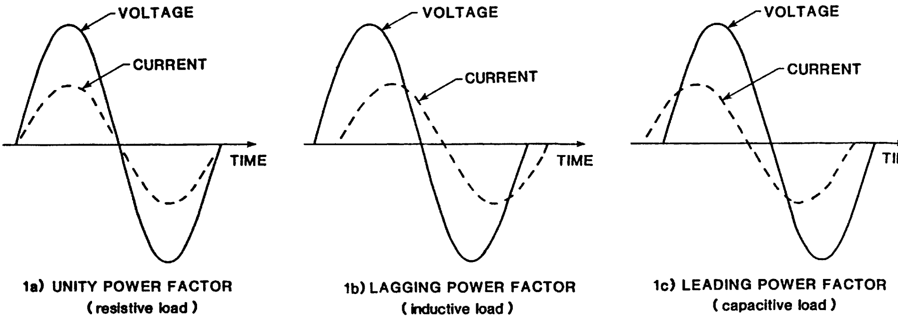
$$\text{Power factor} = \frac{\text{kW demand}}{\text{kVA metered}} = \frac{237}{324} \times 100 = 73\%$$

If the utility requires 85% power factor to avoid penalty,

$$0.85 \times 324 \text{ kVA} = 275 \text{ kW billing demand}$$

$$\begin{aligned} \text{Billing demand} &= 275 \text{ KW} \\ \text{Actual demand} &= \underline{-237 \text{ kW}} \\ \text{Excess demand} &= 38 \text{ kW} \end{aligned}$$

Excess demand is charged at utility demand charges of $4.00–10.00 per kW depending on rate structures. Even at $4.00 the cost of 38 kW/month would be $152.00, or $1824.00/year.



1a) UNITY POWER FACTOR (resistive load)
1b) LAGGING POWER FACTOR (inductive load)
1c) LEADING POWER FACTOR (capacitive load)

Figure 6 Graph of Reactive/Out of Phase Current.

When capacitors are installed, power factor increases to 93% (well above the 85% level), thus avoiding the penalty. The payback is as follows:

$$\frac{\$1200 \text{ capacitor cost}}{\$1824 \text{ power cost saving}} \times 12 \text{ months/year} = 8 \text{ months}$$

Adding capacitors means the circuit capacity is thus increased to allow adding more load. Note: Demand and power factor calculations in each utility area differ, hence the above example would have to be revised to meet local conditions.

Demand and power factor charges are often a major percentage of the total power cost and to a large extent are controllable. Demand can be controlled by shedding loads when peak demands are approached, either manually or through the use of a programmable energy controller. Power factor can be eliminated by replacing inefficient motors or installing capacitors that limit reactive current to levels acceptable to the utility. Controlling demand and minimizing reactive current makes significant reductions in power costs obtainable. Solutions to reducing costs of these items are included in Sections 8.6–8.8

8.3. Financial Considerations about Energy Management

8.3.1. Assessing the Cost of Energy

The real cost of energy is not just the price per gallon, Mcf or kWh but in executing the process that provides energy to run the business. Because energy is an essential resource for almost any manufacturing or processing operation, massive effort must be expended to keep the supply of energy available in the proper quantity and quality at the minimum cost consistent with keeping the enterprise in effective operation.

The most costly aspect of energy is unavailability. When energy becomes unavailable, costs soar and adverse consequences result:

Production interruptions
Missed schedules
Lost customers
Penalties for shutting down customers production lines
Spoiled work, rework, non-value-added activity
Costly cleanout of ruined product from processing equipment

Likewise, the benefits of effective energy management can be measured in terms of lost production avoided, productivity improvement, continued customer satisfaction, and profitability increases. The real cost of energy is in executing the process that provides the energy supply to run the business.

By having the cost of energy failures highlighted through the use of activity-based costing, the engineering manager can determine the cost–benefit ratio of operating the system to avoid future failures.

Another caveat in determining the real costs and benefits of effective energy management is where to focus effort. Prevalent philosophies today include ‘‘pick the low-hanging fruit’’ and ‘‘don’t sweat the small stuff.’’ Often these philosophies lead to missing the really big winners. While there may be some merit to each of these philosophies, the real savings in energy systems may be from finding less obvious, high-return savings that will provide long-term solutions to complex problems.

8.3.2. Justifying Energy-Conservation Opportunities Using Activity-Based Costing

While energy costs are typically considered overheads, costs can be traceable to specific activities within the operation. Applying activity based costing, life-cycle costing, or other systems that provide accurate data on energy use allows opportunities for conservation and system upgrade to be identified and justified.

Too often, worthwhile energy projects are killed by the one-year payback syndrome. In addition to being too strict, the one-year payback concept ignores the time value of money, a fundamental of engineering economy. In many energy and environmental projects, the payback is nowhere close to a one-year undiscounted payback.

To get top management’s attention, the project author must get outside of the narrow confines of labor or utility savings and include the benefits of a dependable, cost-effective energy process to justify energy related projects. Management reacts best to $$$$ and ####, so it is essential that you speak their language. The basis, for these savings is in keeping records using activity-based costing. If energy problems cause lost production, customer dissatisfaction, or excessive costs, ABC cost data that quantify these problems should get top management project approval.

Especially useful in justifying energy improvements is rework, reinspection, repackaging, or any other re-word that connotes non-value activity caused by energy failures or inadequacies.

8.3.3. Using Life-Cycle Costing to Assess Return on Energy Projects

Life-cycle costing is also useful in justifying energy projects because all benefits and costs throughout the expected life of the project or equipment are predicted, and disposal costs at the end of the project are included. Placing a value on environmental conservation, energy availability, and system dependability in a LCC calculation can further assist in justifying energy projects.

Data from ABC can also be used to substantiate life-cycle costing. The use of activities as the basis for cost justification gives a true representation of what actually happens in the organization, and can demonstrate interdependencies of energy with other activities in the organization. See Chapters 88–92 for more detail on costing.

8.3.4. Impact of Utility Deregulation

Until recently, utility companies had monopolies in certain areas of the country, but as a result of deregulation, it is now possible to purchase gas or electric from other utility companies and have it delivered through systems of the former utility company. The impact on energy users is that opportunities for significant savings now exist as a result of deregulation. Energy managers should investigate the possibilities for accessing savings from deregulated utilities to improve profitability of the company.

8.4. Relationship between Energy and Environment

8.4.1. Pollution from Energy Production and Waste Heat Recovery

Energy and environment are inextricably intertwined issues. Energy production is one of the chief causes of air pollution, and conversely, energy production can be a major solution to environmental problems. Acid rain caused by burning fossil fuels and discharging pollutants such as nitrous oxide and sulfur dioxide results in harm to aquatic life, animals, vegetation, and humans.

Some solutions for both energy and environmental problems lie in burning trash or other waste material after all recyclables have been removed. While this approach reduces waste streams by over 90%, so-called environmentalists raise a loud cry about burning anything, even though benefits far outweigh costs. The ultimate solution to energy-related pollution is conservation.

8.4.2. Cogeneration

The concept of cogeneration is both environmentally friendly and energy cost effective. Definitions of cogeneration include topping and bottoming systems. A topping system primarily generates electricity, and an alternative use is found for the exhaust steam. The bottoming system produces heat to facilitate a process, and the excess is captured for electric power generation or other uses as a byproduct.

Typical cogeneration methods include use of waste process gases to drive gas turbines for electric generation, recapture of low-pressure steam for electric generation or driving machinery, and use of waste steam or heat for electric generation and peak shaving. Incineration of waste materials as fuel for steam generation, process heat, district heating or air conditioning, or electric generation is a growing cogeneration option.

8.5. Establishing Strategies for an Effective Energy-Management Program

8.5.1. Strategies and Tactics for Major Energy Improvements

1. Convince your management of the value of energy conservation and systems improvement.
2. Get energy improvement into the budget.
3. Develop a strategic objective to have the best installed energy systems and make the most efficient use of energy possible.
4. Develop and implement a specific plan for taking energy assessments.
5. Identify every energy-consuming device in the facility.
6. Conduct a comprehensive analysis of all energy equipment.
7. Develop a plan to upgrade technology of all energy equipment.
8. Review energy bills and find demand power factor and contract rates.
9. Negotiate with utilities to get lower rates and better breaks on charges.
10. Audit past bills, and get adjustments where possible.
11. Find real costs of energy deficiencies and unavailability.

12. Justify new technologies and permanent solutions to energy problems.
13. Reengineer the energy process, and stop doing business as usual.
14. Make energy an iterative process that is under frequent review.
15. Use steps in the detailed procedure that follows.

8.5.2. Starting an Energy-Management Program

Central to an energy-improvement effort is an energy assessment. Whether techniques of energy improvement are called surveys, audits, or assessments depends on objectives and individual preferences. An audit focuses on the quantitative aspects of energy consumption and finds savings opportunities by analysis of current energy use patterns and consumption data. A survey usually focuses on qualitative opportunities that produce easily implemented improvements. An assessment usually follows a structured approach to identify, evaluate, and implement energy saving projects or system improvements. Often the terms are interchangeable, and all approaches yield good results. In the remainder of this chapter the term *assessment* will be used most frequently Managers or engineers finding ways to reduce energy requirements should not be concerned with terminology as much as results.

The methodology shown in the Appendix provides a step-by-step approach to energy conservation and cost reduction. It includes practical solutions, many of which can be implemented at little or no cost. Examples will be discussed along with application of the methodology. By following the steps in the methodology, the energy-analyst should find numerous energy-saving opportunities.

Generally, an energy assessment will include finding energy-saving and system-improvement opportunities. The procedure in the Appendix shows the detailed steps to assess energy opportunities. Before the assessment starts, one person and/or a multidisciplinary team should be assigned to conduct the assessment, and each member should be trained in their specific roles. The composition of the team depends on skills available in the organization. Normally a combination of engineers, supervisors, operators, and human resource people have sufficient diversity of background to generate creative ideas for energy improvements.

8.6. Steps in an Energy Assessment

Prepare for the assessment:
  Make a commitment to energy improvement.
  Select an individual or a team to improve energy.
  Collect data for energy bill analysis.
  Identify operations and components of the energy system to include in the assessment.
  Obtain instrumentation to do a credible technical assessment.
  Determine who controls those operations or components.
  Obtain agreement from operations chiefs to proceed with the assessment.
  Get a computerized energy analysis program if available.

Collect data on energy consumption, practices:
  Review energy bills for past two years.
  Record demand and power factor data.
    Get utility to show power factor and demand on monthly bills.
    Find criteria for power factor and demand charges.
  Input to computer spreadsheet and make charts and graphs, convert to BTU/caloric equivalent.

Take plant tours to identify energy opportunities:
  Observe present/potential energy waste and system problems.
  Record nameplate data, load rating, etc.
  Review the energy system.
  Find all energy system equipment, including invisible.
  Fill out energy data sheets.
  Compute theoretical load, compare with billed load.
  Make inferences about discrepancies.

Analyze data from steps 1 and 2.
  Correlate data with production or activity level.
  Find the big energy consumers.
  Apply activity-based costing to find energy-related delays.

- Develop and evaluate energy conservation solutions.
  - Revise team assignment to develop solutions.
  - Use teams to get ownership of improvements.
  - Use checklists to find energy savings.
  - Follow energy-improvement opportunities (see Section 8.7).
  - Apply analytical techniques:
    - What, where, when, who and why?
    - Eliminate, combine, resequence, and simplify operations.
  - Apply systematic creative thinking, brainstorming.

- Evaluate solutions, select best alternatives:
  - Conduct economic evaluation.
  - Use ABC and LCC.
  - Consider people aspects.
  - Assess feasibility.

- Quantify and prioritize energy-improvement opportunities.
  - Avoid the one-year payback syndrome.
  - Show costs of energy unavailability.

- Present findings to management:
  - Develop each recommendation fully.
  - Prepare a coherent report.
  - Show why management should accept your recommendations.
  - Express recommendations in management's language ($$$$ and ####).
  - Rehearse the presentation, and present the report succinctly.
  - Gracefully accept management's decision.

- Implement improvements and monitor results:
  - Assign responsibility for implementation.
  - Establish measures of specific improvements.
  - Verify savings.
  - Chart energy productivity.
  - Modify improvements if required.
  - Avoid false savings due to external changes.

8.7. Energy-Improvement Possibilities

Facility

- Install enough insulation, weather stripping.
- Seal off leakage through windows, cracks.
- Evaluate refenestration (window replacement).
- Install vestibules at doors.
- Seal around dock doors/use flap doors to seal out cold air.
- Reduce solar gain (insolation) to reduce cooling load.
- Close all doors and other openings in winter.

Energy infrastructure in the plant

- Give the energy infrastructure and other energy delivery systems constant attention.
- Make someone responsible for operating and maintaining energy systems in the facility.
- Check ownership of interface devices with the utility supplier.
- Set up a preventive maintenance program for every part of the system, including:
  - Contacts, insulation, dust and dirt on equipment, heat buildup in ducts and switches.
- Note failure points and correct the root problem immediately.
- Reengineer the system to bring it to state of the art.

Electrical

Upgrade electrical distribution systems.
Control peak load demand, shed loads using prioritized controller.
Run equipment off peak where possible.
Retrofit for higher voltages.
Correct power factor by using energy-efficient motors or capacitors.
Turn off equipment when not needed.
Monitor and upgrade power quality for digital equipment.

Natural Gas

Check for leaks and explosion hazards.
Check piping for corrosion.
Check for improper installations.

Utility deregulation

Become familiar with new regulations and procedures.
Contact potential energy suppliers for quotes, interview alternate suppliers.
Beware of bogus suppliers offering unrealistic deliverables.
Carefully evaluate economics of all offers; determine the *real cost*.
Negotiate rate reductions.
Take full advantage of deregulation.

Lighting

Reduce number of fixtures.
Avoid electrician's dream (excessive lighting).
Install more efficient lighting fixtures, electronic ballasts, and long-life bulbs.
Use task lighting, reduce lighting in nonproduction areas.
Turn off lights when not in use.
Balance lighting heat load with air conditioning.
Install occupancy sensors or more switches to turn off unneeded lights.
Utilize photocell and/or timers on lighting, especially outdoors.
Tie lighting into building control systems.

Process

Run process equipment only when needed.
Avoid short runs on process equipment.
Consider energy in scheduling production.
Make energy a prime consideration in replacing process equipment.
Schedule operations/production around energy considerations.
Reduce run times of equipment to bare minimums.
Use waste heat to run the process or supplement primary energy sources.

Heating, ventilation

Select heating equipment carefully for maximum efficiency.
Install programmable thermostats, energy controllers.
Control heating using computer building controllers.
Avoid pulling out heat, make up air loss with waste heat.
Burn waste material for building heat.
Don't heat seldom used space.
Use radiant heat in isolated spots.
Install ceiling fans to bring heated air to the floor level.
Change to cheaper energy sources.
Keep filters, coils, and ductwork clean.

Expand comfort zone in summer and winter.
Use passive/low-energy cooling where possible.

Air conditioning systems (see also heating and air ventilation ideas)
Use outside air to cool building before using ACU.
Clean ACU filters.
Use timers to control HVAC.
Install energy-management systems to control HVAC.
Replace refrigerant compressors or chillers with more efficient units.
Redesign system for best efficiency and maximum output/kwh.
Evaluate gas cooling as an alternative to electric cooling.
Conduct vigorous preventive/predictive maintenance on system.
Insulate ductwork and piping in AC system.

Indoor air quality
Eliminate pollutants (smoke, dust, vapors)
Introduce properly filtered fresh air
Do more frequent changes of the air in the building
Change controls, improve dampers, use variable speed blowers
Rebalance system to accommodate varying demands

Boilers, steam, hot water
Check fuel air ratio, NOX
Check flame pattern
Check stack temperature
Check CO content in stack
Insulate pipes
Preheat makeup water with waste heat
Heat air using heat exchangers.
Return steam condensate and keep it warm.
Repair hot spots in fire box.
Correct steam leaks and repair traps.
Install automatic controlled blowdown.
Implement effective boiler water treatment.
Consider cogeneration or alternate fuels.

Compressed air system
Repair air leaks.
Run at lowest possible pressure.
Reclaim waste heat for winter heating or heating restroom water.
Install refrigerant air dryers and maintain effectively.
Install cooling towers for water-cooled compressors.
Discontinue running cooling water down the sewer.
Replace energy-inefficient motors and compressors.
Keep maintenance records to justify new equipment.
Automate compressor system controls to minimize energy use.
Evaluate load/unload against intermittent stop/start cycle.
Avoid improper use of compressed air (blowing chips).
Obtain high-quality intake air using filters and coolest possible air.
Practice effective preventive maintenance.
Consider alternative energy to power compressors.
Install remote monitoring of air compressors with diagnostics.

Motors/drives

- Replace old motors with energy-efficient motors.
- Install capacitors on poor power factor motors.
- Turn off motors when not needed (unless startup demand negates savings).
- Install variable-speed drives where possible.

9. SUMMARY: CREATING EXCELLENCE IN PLANT AND FACILITIES ENGINEERING

A key goal of this chapter has been to motivate industrial engineers assigned to plant and facilities engineering duties to strive for excellence. By keeping mindful of organizational strategy and applying effective industrial engineering and management techniques, including those described in this chapter, the industrial engineer can function as a plant or facilities engineer to create continuous improvement in productivity and quality for the benefit of the organization.

REFERENCES*

American Productivity Center (APC) (1991), *How to Measure Productivity in Your Organization,* APC, Houston.

Davis, G., Szigeti, F., Atherton, A. (1999), ‘‘Why Cutting Costs Isn’t Enough,’’ *Building Operating Management,* Vol. 46, No. 4, pp. 44–52.

Gulati, R., and Lach, A., (1997), ‘‘Maintenance Benchmarking and Survey Results,’’ in *Proceedings of the 14th International Maintenance Conference,* Institute of Industrial Engineers, Atlanta.

Higgins, L. R. (1988), *Maintenance Engineering Handbook,* McGraw-Hill, New York.

Kamal Zafar, S. (1998), ‘‘Integrated Enterprise Proves Key to Flexible Manufacturing,’’ *InTech,* Vol. 45 No. 7, pp. 42–46.

Katzel, J. (1998), ‘‘Optimizing Building Automation System Performance,’’ *Plant Engineering,* Vol. 52, No. 9, pp. 44–50.

Lewis, B. T., and Marron, J. P., (1973), *Facilities and Plant Engineering Handbook,* McGraw-Hill, New York, pp. 123–148.

McElroy, F. E., Ed., (1964), *Accident Prevention Manual for Industrial Operations,* 6th Ed., National Safety Council, Chicago.

Palko, E. (1989), ‘‘Maintenance Manpower Staffing,’’ *Plant Engineering,* Vol. 43, No. 13, pp. 55–57.

Pearlman, A., and Cana, O., (1999), ‘‘A Smoother Road,’’ *Building Operating Management,* Vol. 45, No. 8, pp. 70–74.

Peele, T. T., and Chapman, R. L. (1989), ‘‘Designing a Maintenance Training Program,’’ *Plant Engineering,* Vol. 41, No. 23, pp. 46–49.

Phelps, C. (1988) ‘‘Contemplating the Importance of Maintenance Training,’’ *Building Operating Management,* Vol. 35, No. 4, pp. 48–52.

Piper, J. (1988a), ‘‘Facility Assessment Survey,’’ *Building Operating Management,* Vol. 35, No. 9, pp. 98–100.

Piper, J. (1988b), ‘‘Formulating a Security Program,’’ *Building Operating Management,* Vol. 35, No. 6, pp. 74–78.

Piper, J. (1999), ‘‘Beyond Building Automation,’’ *Building Operating Management,* Vol. 46, No. 8, pp. 76–78.

Raymond, L. (1993), ‘‘Benchmarking,’’ in *Proceedings of the Utilities and Energy Industries Conference,* Institute of Industrial Engineers, Atlanta.

Rosaler, R. C. (1994), *Standard Handbook of Plant Engineering,* 2nd Ed., McGraw Hill, New York, pp. 1-4–1-10.

Rosaler, R. C., and Rice, J. O. (1983), *Standard Handbook of Plant Engineering,* McGraw-Hill, New York.

Rospond, K. M. (1999a), ‘‘Defensive Power Planning,’’ *Building Operating Management*, Vol. 46, No. 12, pp. 23–28.

---

* Space limitations in this chapter preclude a complete list of references, but additional references or clarification of chapter content may be obtained by contacting the author at (931) 528-1175.

Rospond, K. M. (1999b), "Energy Efficiency and IAQ," *Building Operating Management,* Vol. 46 No. 12, pp. 27–30.

Ross, J. R. (1989), "An Integrated Industrial Engineering Methodology for Environmental Problems and Issues," in *Proceedings—Spring Conference of the Institute of Industrial Engineers,* Institute of Industrial Engineers, Atlanta.

Ross, J. R. (1996), *EEPE Division Energy Workshop—Minneapolis,* Institute of Industrial Engineers, Atlanta.

Ross, J. R. (1999), "Energy—The Forgotten Resource," in *Proceedings of National Manufacturing Week Conference,* Reed Exhibitions, Norwalk, CT.

Steele, W., (1997), "Introduction to RCM and Condition Monitoring," in *Proceedings of the 14th International Maintenance Conference,* Institute of Industrial Engineers, Atlanta.

Steiner, V. M. (1988), "Building Codes—Bane or Blessing?" *Plant Engineering,* Vol. 42, No. 11, pp. 88–91.

Tatum, R. (1990), "Protocols for the 'Smart' Building," *Building Operating Management,* Vol. 37, No. 1, pp. 52–58.

Tatum, R. (1997), "Improving Workplace Performance," *Building Operating Management,* Vol. 44, No. 5, pp. 57–76.

Tomlingson, P. D. (1988a), "Methods to Evaluate Maintenance with Positive Results," *Plant Engineering,* Vol. 42, No. 3, pp. 168–171.

Tomlingson, P. D. (1988b), "Evaluating Maintenance Performance," *Plant Engineering,* Vol. 42, No. 15, pp. 106–110.

ADDITIONAL READING

Ashe, J. T., "Air Cleaners: Rx for 'Sick Building' Syndrome," *Plant Engineering,* Vol. 43, No. 2, 1989, pp. 76–78.

Caplan, J. S., and Burrows, W. "Building an Analyst Based RCM Organization within Indianapolis Power & Light," *Reliability,* Vol. 6, Issue 2, 1999, pp. 11–19.

Dhillon, B. S., and Reiche, H., *Reliability and Maintainability Management,* Van Nostrand Reinhold, New York, 1983.

Dunn, R. L., "Advanced Maintenance Technologies," *Plant Engineering,* Vol. 41, No. 12, 1987, pp. 80–87.

Foster, J. W., Phillips, D. T., and Rogers, T. R., *Reliability, Availability, and Maintainability,* M/A Press, Beaverton, OR, 1981.

Gordon, J. R., Mondy, R. W., Sharplin, A., and Premeaux, S. R., *Management and Organizational Behavior,* Allyn & Bacon, Boston, 1990.

Hall, R. E., and Bengard, M. L., "Automation Systems for the Medium-Sized Building," *Building Operating Management,* Vol. 34, No. 10, 1987, pp. 20–26.

Harmon, R. L., *Reinventing the Factory,* Free Press, New York, pp. 335–338.

Hicks, D., *Activity Based Costing: Making It Work for Small and Mid-Sized Companies,* 2nd Ed., John Wiley & Sons, New York, 1999.

Katzel, J., "Comprehensive PM System Includes Predictive Maintenance Aids," *Plant Engineering,* Vol. 41, No. 18, 1987, pp. 92–94.

Keithly, D., "Cutting Costs through Substation Ownership," *Plant Engineering,* Vol. 43, No. 11, 1989, pp. 64–66.

Kelley, A., "Energy Benchmark for Buildings," *Building Operating Management,* Vol. 44, No. 5, 1997, pp. 57–76.

Kelley, A., "Energy Benchmark for Buildings," *Building Operating Mangement,* Vol. 46, No. 9, 1999, pp. 77–92.

Lang, R. G., "A 10 Step Program to Improved Power Quality," *Plant Engineering,* Vol. 53, No. 4, 1999, pp. 84–88.

Mangano, J. M., and Dumdie, D. P., "Foundation Fieldbus Has Arrived and Is Here to Stay," *InTech,* Vol. 45, No. 2.

Moore, D. M. (1989) "Understanding Direct Digital Control," *Building Operating Management,* Vol. 36, No. 12, pp. 48–51.

Nagle, B., "Improving Building IAQ," *Building Operating Management,* Vol. 35, No. 12, 1989, pp. 34–38.

Netherton, D., "Standard to Define RCM," *Maintenance Technology,* Vol. 12, No. 6, 1999, pp. 17–24.

Newbrough, E. T., *Effective Maintenance Management,* McGraw-Hill, New York, 1967, chap. 1.

Ross, J. R., *EEPE Division Environmental Workshop—Los Angeles,* Institute of Industrial Engineers, Atlanta, 1992.

Ross, J. R., “Plant and Facilities Engineering,” in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, 1992.

Ross, J. R., “Indoor Air Quality,” *Energy, Environment and Plant Engineering Newsletter,* Institute of Industrial Engineers, Atlanta, 1994.

Ross, J. R., “Productivity of Grounds Keepers,” *Energy, Environment and Plant Engineering Newsletter,* Institute of Industrial Engineers, Atlanta, 1997.

Ross, J. R., “Selecting and Implementing a CMMS,” *Energy, Environment and Plant Engineering Newsletter,* Institute of Industrial Engineers, Atlanta, 1997.

Salvendy, G., Ed. (1982), *Handbook of Industrial Engineering,* 1st Ed., John Wiley & Sons, New York, pp. 1.4.1–1.4.12.

Solid Waste Technology Staff, “Begin by Reducing,” *Solid Waste Technologies,* December 1999, 22 (compiled from U.S. EPA sources).

Wisner, P., “Accounting for Maintenance Costs,” in *Proceedings of the 13th International Maintenance Conference,* Institute of Industrial Engineers, Atlanta.

CHAPTER 59

Maintenance Management and Control

Ralph W. “Pete” Peters
Tompkins Associates, Inc.

1. **OBJECTIVES AND CONTENT OF THE CHAPTER** **1586**
2. **INTRODUCTION** **1586**
    - 2.1. The Maintenance Process: A Key Element of Plant and Facilities Engineering 1586
    - 2.2. The Scope of Maintenance and Physical Asset Management 1587
    - 2.3. Effective Maintenance Management Requires Industrial Engineering Principles 1588
    - 2.4. Maintenance as an Internal Business Opportunity and Profit Center 1588
3. **THE 25 REQUIREMENTS FOR EFFECTIVE MAINTENANCE MANAGEMENT** **1588**
4. **BENCHMARKING THE TOTAL MAINTENANCE OPERATION** **1593**
    - 4.1. Internal vs. External Benchmarking 1593
    - 4.2. The Scoreboard for Maintenance Excellence 1593
        - 4.2.1. Global Best Practices as External Benchmarks 1593
        - 4.2.2. Conducting an Assessment of the Total Maintenance Process 1594
        - 4.2.3. Internal Benchmarking: The Key to Validation of Results and ROI 1597
5. **PERFORMANCE MEASUREMENT: KEY TO MAINTENANCE CONTROL** **1597**
    - 5.1. Techniques for Measuring Maintenance Activities 1597
    - 5.2. The ACE Team Benchmarking System 1598
    - 5.3. Establishing Key Performance Indicators 1601
    - 5.4. Developing a Maintenance Excellence Index 1604
6. **INFORMATION TECHNOLOGY TO SUPPORT MAINTENANCE MANAGEMENT AND CONTROL** **1605**
    - 6.1. Introduction to Computerized Maintenance Management Systems/Enterprise Asset Management 1605
    - 6.2. Overview of CMMS/EAM System Functionality 1606
    - 6.3. Benchmarking the CMMS/EAM Installation 1609
        - 6.3.1. The CMMS Benchmarking System 1609
        - 6.3.2. The CMMS Benchmarking System Rating Process 1610
        - 6.3.3. Conducting the CMMS Benchmark Evaluation 1610
7. **A REVIEW OF OTHER SELECTED MAINTENANCE BEST PRACTICES** **1610**
    - 7.1. Continuous Reliability Improvement 1610
        - 7.1.1. Asset Facilitation 1611
    - 7.2. Preventive Maintenance 1611
    - 7.3. Predictive Maintenance 1612
        - 7.3.1. Vibration Analysis 1613
        - 7.3.2. Shock Pulse 1613
        - 7.3.3. Spectrometric Oil Analysis 1613
        - 7.3.4. Standard Oil Analysis 1614

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

- 7.3.5. Ferrographic Oil Analysis 1614
- 7.3.6. Infrared Thermography 1614
- 7.3.7. Ultrasonic Detection 1614
- 7.4. Maintenance Storeroom Operations and MRO Materials Management 1615
  - 7.4.1. Storeroom Inventory Management 1616
- 7.5. Planning and Scheduling 1616
  - 7.5.1. The Overall Craft Effectiveness (OCE) Factor 1616
  - 7.5.2. Getting Started with Planning and Scheduling 1618
  - 7.5.3. Focus on Customer Service 1618
  - 7.5.4. Measure Effectiveness of the Planning Function 1618
- 7.6. Reliability-Centered Maintenance 1618
  - 7.6.1. The Evolution of Reliability-Centered Maintenance 1618
  - 7.6.2. Overview of the RCM Process 1618
- 7.7. Total Productive Maintenance 1619
- 7.8. Operator-Based Maintenance 1620

8. MAINTENANCE MANAGEMENT FOR THE NEW MILLENNIUM 1620
- 8.1. The Emergence of the Chief Maintenance Officer (CMO) 1621
- 8.2. Growth of Reliability-Improvement Technologies 1621
- 8.3. The Role of The Internet 1621
- 8.4. MRO Materials Management 1622
- 8.5. The Growth of Contract Maintenance 1622

9. CONCLUSION 1622

REFERENCES 1622

1. OBJECTIVES AND CONTENT OF THE CHAPTER

The key objectives of this chapter are to provide a firm understanding of the importance of the physical asset management and maintenance process, to review the key requirement for maintenance success, and to review some of today's best maintenance practices. This chapter includes how the results of continuous maintenance improvement can be measured, the methods to use for measurement, and how the results and ROI can be validated. The IE principles and practices that can support maintenance process improvement will be highlighted. Current strategies and trends in computerized maintenance management systems (CMMS) and enterprise asset management (EAM) systems technology to support physical asset management will also be covered.

An important look at maintenance management for the new millennium and the emerging role of a chief maintenance officer (CMO) and new technologies, will be presented. The chapter will help the new and experienced IE, the engineering manager, the operations manager, as well as the CEO, to understand better their own maintenance operation. It will outline specific methodologies to help improve mission-essential maintenance operations. It will also help all to view the maintenance process as a profit center and key contributor to total operations success

2. INTRODUCTION

2.1. The Maintenance Process: A Key Element of Plant and Facilities Engineering

The maintenance process is one of the most important elements within the overall scope of plant and facilities engineering function. It is a broad-based technology area for the industrial, commercial, institutional, and business communities. As covered within a previous chapter, the broad scope of plant and facilities engineering is continuously changing and growing in importance as a key contributor to the success of a total operation. Elements within it, such as maintenance, energy management and regulatory compliance, have all evolved to new levels of importance for the total operation.

Plant and facilities engineering is a multidisciplinary field of engineering concerned with the physical infrastructure of industrial, commercial, institutional, healthcare, and business facilities. It embraces the design, installation, operation, maintenance, modification, construction, modernization, and protection of physical facilities and equipment used to produce a product or provide a service (Dunn 1997). It includes, but is not necessarily limited to, the following areas of physical asset management:

- Design of facilities and systems
- Construction of facilities and systems
- Installation of facilities and systems
- Start-up of systems
- Operation of systems
- Maintenance of facilities and systems
- Retrofit of facilities and systems
- Environmental controls
- Safety and health
- Security and fire protection
- Production processes and equipment (in industry)
- Regulatory compliance (local, state, and federal)
- Energy management and building control
- Administration, supervision, organization, planning
- Other support functions as required by the enterprise owner or manager

2.2. The Scope of Maintenance and Physical Asset Management

Maintenance is about the care of physical assets: production equipment, plant facilities, office buildings, hospitals, trucks, cars, forklifts, and computers. It is very important to understand fully the scope of the maintenance management process because maintenance is much more than just the care of physical assets used in a production operation. The maintenance management and control functions, operations, and activities that may be required within a maintenance operation are very broad. Likewise, the application of the actual maintenance process—the hands-on, wrench time of doing the work—applies to many different types of operations. The maintenance processes in discrete manufacturing, continuous processing equipment, facility and building systems, property management, and fleet operations maintenance all have uniquely different challenges and types of assets. There are other unique applications of the maintenance process in areas such as health care, research facilities, and pharmaceutical operations. Physical assets, whether production equipment, a health care facility, or the assets that make up a trucking fleet or an airline, all have maintenance requirements to perform their primary function at a reasonable cost over their expected economical life. A maintenance operation may include, but also is not limited to, the following (Dunn 1997):

- Installation and maintenance of all utilities systems and components for electricity, water, steam, gas, oil, compressed air, communications, data networks, etc.
- Provision of services for construction, maintenance, and repair of buildings and structures
- Provision of services for alterations and modifications of buildings, structures and manufacturing type assets
- Installation, operation, and maintenance of all heating, ventilating, air conditioning, and refrigeration systems and components
- Installation, relocation, modification, maintenance, and repair of other equipment and systems, as determined by coordination with tenants
- Testing of electrical systems and backup utilities
- Maintenance of appropriate equipment records and histories
- Management of maintenance planning, scheduling, and work execution
- Implementation of preventive and predictive maintenance strategies and practices
- Management of spare parts and materials storerooms and inventories
- Coordination of federal, state, local, and insurance licensing inspections and compliance
- Management of special projects as required
- Provision of housekeeping, janitorial, and custodial services
- Preparation and control of budgets
- Control of maintenance contracts
- Maintenance of reports from inspectors and insurance carriers
- Work management: processing of work requests; preparation of project cost estimates; planning and scheduling of work; provision of required parts, materials, and equipment; and maintenance of all related labor and equipment records

- MRO materials management: management of maintenance repair operations (MRO) parts, materials, and special equipment for which the maintenance operation is held accountable
- Preparation of appropriate reports, statistics, and recommendations on activities
- Planning, coordination, and scheduling of predictive, preventive maintenance and other asset reliability programs.
- Maintenance of appropriate records pertaining to labor, MRO items, and asset history
- Ensuring compliance with all applicable life safety, building codes, and regulatory items
- Ensuring maintenance of all fire protection and security systems
- Operation and maintenance of all utilities systems and equipment
- Operation and maintenance of backup utility systems.
- Maintenance of environmental monitoring systems
- Organization, administration, and supervision of safety activities in accordance with all federal, state, and local requirements
- Initiation of safety-related work requests
- Compliance with all safety procedures.

2.3. Effective Maintenance Management Requires Industrial Engineering Principles

The maintenance process should be viewed as an internal business opportunity. Since Tompkins Associates introduced the concept of ‘‘maintenance as a profit center’’ (Tompkins 1999), this modern view of maintenance has emerged in literature and in practice. Business process continuous improvement (BPCI) and most of the traditional IE principles have application to maintenance operations. When we consider that there are countless contract service providers for all types of maintenance services competing to replace or supplement internal maintenance departments, it becomes apparent that internal maintenance with the help of IE techniques can be improved.

2.4. Maintenance as an Internal Business Opportunity and Profit Center

Traditional thinking about maintenance has changed dramatically. Maintenance was once considered a necessary evil, but it is now being viewed as a key contributor to profit in a manufacturing or service providing operation. For example, what if the net profit ratio of an operation is 4%? What does a 4% net profit ratio mean in terms of the amount of equivalent sales needed to generate profits? A net profit ratio of 4% requires $25 of equivalent sales for each $1 of net profit generated. Therefore, when we view maintenance in these terms, we can readily see that a small savings in maintenance can mean a great deal to the bottom line and equivalent sales. From Table 1, maintenance as a profit center is illustrated, showing that only a $40,000 savings is required to translate into the equivalent of $1,000,000 in sales. There are many more areas, such as the value of increased asset uptime, increased net capacity and just-in-time throughput, increased product quality and increased customer service, that all contribute to the bottom line and subsequently to profit (Peters 1994a).

3. THE 25 REQUIREMENTS FOR EFFECTIVE MAINTENANCE MANAGEMENT

When an organization prepares for business process continuous improvement and the application of IE technology and techniques, it should include the evaluation and improvement of the current maintenance processes. A number of key principles and practices that are fundamental to the BPCI journey.

TABLE 1 Maintenance as a Profit Center

Maintenance Savings to Impact Net Profit	Equivalent Sales Required for Generating Net Profit
$1	$25
$1,000	$25,000
$10,000	$250,000
$20,000	$500,000
$30,000	$750,000
$40,000	$1,000,000
$80,000	$2,000,000
$120,000	$3,000,000
$200,000	$5,000,000

They provide the foundation upon which to develop improvements. It is important to understand the requirements for maintenance success because they in turn provide the foundation for today's best maintenance practices.

These fundamental principles become the cornerstone for achieving, maintaining, and continually improving the maintenance process. Organizations that are establishing best practices for world-class maintenance will be actively pursuing the following 25 requirements for effective maintenance management (Peters 1994b).

1. *Priority:* The process of performing maintenance and managing physical assets will be recognized as a top priority within successful organizations.
   - Maintenance will be viewed as a top-priority operation, not as a necessary evil. It will be viewed as another area that contributes directly to the bottom line when a strategy for continuous maintenance improvement is adopted. The future capable leader will have identified top priority areas for improvement, based upon a total benchmark evaluation of the maintenance operation, and investments will be made to implement best practices.
2. *Leadership and understanding:* Maintenance leaders must understand the challenges of maintenance and provide effective maintenance leadership with a vision of continuous maintenance improvement.
   - Maintenance leadership must continually develop the skills, abilities and attitudes to lead maintenance into the future. Maintenance leaders must completely understand the 25 requirements for effective maintenance management and develop priorities for action. Maintenance leaders must create understanding within the organization about maintenance and develop a vision of continuous maintenance improvement shared throughout the organization.
3. *PRIDE in Maintenance*. Maintenance operations in the future capable company will experience fundamental improvements in work ethic, attitude, values, job performance, and customer service to achieve real pride in maintenance excellence.
   - Tangible savings and improvements will occur as a result of continuous maintenance improvement. The successful maintenance operations will experience other fundamental improvements that develop more PRIDE: People Really Interested in Developing Excellence in Maintenance. Improvements in work ethic, performance, attitudes, teamwork, and concerns for customer service will occur. Successful maintenance operations will have leadership that instills PRIDE in maintenance with a vision of maintenance excellence that creates inspiration, cooperation, and commitment throughout the organization.
4. *Maintenance profession:* The profession of maintenance will gain greater importance as a key profession for success within all types of organizations as the role of chief maintenance officer becomes well established.
   - Maintenance leaders will be recognized as critical resources that are absolutely necessary for the success of the total operation. The chief maintenance officer (CMO) within large multisite operations will create and promote standard best practices. The complexity and importance of the maintenance and physical asset management will continue to grow. New technologies and added responsibilities will require a higher level of technical knowledge and skills.
5. *Maintenance personnel:* A significant upgrade in the level of personnel involved with maintenance will take place to keep pace with new technologies and responsibilities.
   - Maintenance operations within future capable companies will achieve a significant upgrade in the skill level of maintenance crafts people in order to keep pace with new technology and responsibilities. Successful maintenance operations will continually upgrade the skill level of crafts people through more effective recruiting with higher standards and through more effective craft training programs. Pay increases will be more directly linked to performance and demonstrated competency levels in required craft skills.
6. *Craft skills development:* Successful maintenance operations will continually assess craft training needs and provide effective skills development through modern technical learning systems.
   - A complete assessment of craft training needs will be accomplished to identify priority areas for skill development that is competency based, to provide demonstrated technical capabilities for each craft skill. The successful maintenance operations will develop an ongoing program for craft skill development. Continuous maintenance education, based on modern technical learning systems, will be viewed as a sound investment and an important part of continuous maintenance improvement.

7. *Adaptability and versatility:* The maintenance crafts person will become more versatile and adaptable by gaining value with new technical capabilities and multi-craft skills.
   - The development of more crafts people with multiskills will occur to provide greater versatility, adaptability, and capability from the existing workforce. Multiskilled personnel will have added value and will be compensated according to well-defined policies. Craftspeople will become more adaptable, versatile, and valuable as a result of ongoing programs for craft skill development.
8. *Teamwork:* Maintenance will be team players and maintain a leadership-driven, team-based approach to continuous maintenance improvement.
   - Maintenance leadership will accept its role as a top priority operation and will set the example as team players within the organization. The strategy for continuous maintenance improvement will be a leadership-driven, team-based approach that captures the knowledge, skills, and ideas of the entire maintenance workforce. Cross-functional teams with representatives from maintenance, operations, engineering, etc. will be formally chartered to address improvements in equipment effectiveness, reliability, and maintainability.
9. *Maintenance and operations:* Maintenance and operations will become integrated, and function as a supportive team through improved planning, scheduling, and cooperative team-based improvement efforts. Operations will be viewed as an important internal customer.
   - Improved planning and scheduling of maintenance work will provide greater coordination, support, and service to manufacturing type operations. Maintenance and operations of all types will recognize the benefits of working together as a supportive team to reduce unplanned breakdowns, increase equipment effectiveness, and reduce overall maintenance costs. Operations will be viewed as an important internal customer. Operations will gain greater understanding of the 25 requirements of effective maintenance management and accept its important partnership role in supporting maintenance excellence.
10. *Pride in ownership:* Equipment operators and maintenance will develop a partnership for maintenance service and prevention and take greater pride in ownership through operator-based maintenance.
    - Equipment operators will assume greater responsibilities for cleaning, lubricating, inspecting, monitoring, and making minor repairs to equipment. Maintenance will provide training support to operators to achieve this transfer of responsibility and to help operators with early detection and prevention of maintenance problems. Operators will develop greater pride in ownership of their equipment with their expanded responsibilities.
11. *Equipment effectiveness:* A leadership-driven, team-based approach will be used by maintenance and operations to evaluate totally, and subsequently improve all factors related to equipment effectiveness. The goal is maximum availability of the asset for performing its primary function.
    - Continuous improvement of equipment effectiveness will address major losses due to equipment breakdowns, set-up/adjustments, idling/minor stoppages, reduced speeds, process defects, and reduced yields. Equipment improvement teams will be established to meet on a regular basis to identify and resolve equipment-related problems. They will work constructively as cross-functional teams to exchange and implement ideas for improving equipment effectiveness. They will use techniques such as continuous reliability improvement (CRI) and reliability-centered maintenance (RCM). Chronic problems will be analyzed using tools such as statistical process control, graphs, process charts, and cause-and-effect analysis. Maintenance operations within successful future capable companies will use a total team effort by operators, engineering, operations staff, and maintenance to identify and resolve root causes of equipment problems.
12. *Maintenance and engineering:* Maintenance and engineering will work closely during systems specification, installation, startup and operation to provide maintenance with the technical depth required to maintain all assets and systems.
    - Engineering will provide technical resources and support to ensure maintenance has the total technical capability to maintain all equipment and systems. Engineering will play a key support role with maintenance in improving the effectiveness of existing equipment. Maintenance and engineering will work closely in developing specifications for new equipment. During installation and start-up, maintenance and engineering will also work closely to ensure operating specifications are achieved.
13. *Reliability and maintainability:* Machines and systems will be specified, designed, retrofitted, and installed with greater reliability and ease of maintainability.
    - Equipment design will focus on maintainability and reliability and not primarily on performance. Design for maintainability will become an accepted philosophy that fully rec-

ognizes the high cost of maintenance in the life-cycle cost of equipment. The causes for high life-cycle costs will be reduced through the application of good maintainability and reliability principles during design. Design will be focused on life-cycle reliability by identifying potential problems before they are designed into the equipment. Equipment design will include a higher level of internal diagnostic capabilities and provide for greater use of expert systems for troubleshooting. Maintenance will work closely with equipment designers to share information about problems with existing equipment and to provide possible maintenance-prevention solutions for new equipment.

14. *Modularity:* Physical assets and systems will be modularly designed so that failures can be easily identified and repaired quickly.

- Overall maintainability will be further improved through modular design of physical assets and systems. Highest-failure parts and components should be the most accessible, easily identified and designed for easy repair. Components should be designed for easy disassembly and assembly using the lowest skill level possible. Modularity of design will be an important part of the design for maintainability philosophy.

15. *Obsolescence:* The life-cycle costs of physical assets and systems will be closely monitored, evaluated, and managed to reduce total costs.

- Successful maintenance operations will achieve significant reductions in total life-cycle costs through an effective design process prior to purchase and installation. During the equipment's operating life, systems will be developed to monitor equipment costs continually. Information to identify trends will be available to highlight equipment with high maintenance costs. Action can be taken to address critical high-cost areas in order to reduce future costs. A complete equipment history of repair costs will assist maintenance in making decisions on equipment replacement, equipment overhaul/retrofit, and overall equipment condition.

16. *Redundancy:* Critical assets and systems will have backups provided so that if something fails, a secondary asset or system will take over.

- Critical operations will be identified where backup equipment or systems are economically justified. Redundancy of critical equipment and systems will ensure that continuous operation is achieved when something fails. Maintenance will focus attention on critical operations in order to increase equipment effectiveness, reduce unplanned breakdown and increase the effectiveness of preventive/predictive maintenance.

17. *Uncertainty:* Uncertainty will be minimized through effective preventive/predictive maintenance programs and through continuous application of modern predictive maintenance technology and expert systems.

- Effective preventive/predictive maintenance programs will be used to anticipate and predict maintenance problems in order to eliminate the uncertainty of expected breakdowns and high repair costs. Predictive maintenance will not be limited solely to the detection of failure but will proactively identify and eliminate the root causes of chronic problems. Preventive/predictive maintenance programs will be adequately staffed to cover all major assets within the operation. Maintenance will maintain current technical knowledge and experience for applying a combination of predictive technologies that is best suited for the specific application or system.

18. *Computerized maintenance management and enterprise asset management:* Systems that support the total maintenance operation will improve the quality of maintenance and physical asset management and be integrated with the overall business system of the organization.

- Computerized maintenance management systems (CMMS) will provide greater levels of manageability to maintenance operations. CMMS will cover the total scope of the maintenance operation, providing the means to improve the overall quality of maintenance management. Enterprise asset management (EAM) will provide a broader scope of integrated software to manage physical assets, human resources, and parts inventory in an integrated system for maintenance management, maintenance, procurement, inventory management, human resources, work management, asset performance, and process monitoring. Vast amounts of data associated with maintenance tasks will come under computer control and be available as key information for planning, scheduling, backlog control, equipment history, parts availability, inventory control, performance measurement, downtime analysis, etc.

19. *Maintenance information system:* The maintenance information system and database will encompass the total maintenance function and provide real-time information to improve maintenance management.

- The implementation of CMMS and EAM provides the opportunity for improved maintenance information systems. With CMMS and EAM, the maintenance information system can be developed and tailored to support maintenance as a true business operation. Information to support planning, scheduling, equipment history, preventive/predictive maintenance, storeroom management, etc. can be established to improve decision making and overall maintenance management. Improved maintenance information will allow for an open information flow to exist between maintenance, operations and all departments within the organization. Maintenance will become an important part of the overall information flow and be kept well informed about current and future operational plans.

20. *Maintenance storeroom:* The maintenance storeroom will be orderly, space efficient, labor efficient, and responsive and provide the effective cornerstone for maintenance excellence.

- The maintenance storeroom for maintenance repair operations (MRO) items will be recognized as an integral part of a successful maintenance operation. Initial storeroom design or modernization will include effective planning for space, equipment, and personnel needs while providing a layout that ensures efficient inventory control and includes maximum loss control measures. It will be professionally managed and maintained in a clean, orderly, and efficient manner. The trend will be towards larger centralized storerooms with responsive delivery systems to eliminate crafts people waiting or traveling to get parts. An effective maintenance storeroom catalog will be maintained to provide a permanent cross-reference of all storeroom items and serve as a tool for identifying and locating items.

21. *Maintenance inventory:* The proper quantity of the proper spare parts will be on hand due to progressive MRO procurement and internal storeroom controls, all to support maintenance excellence.

- The implementation of CMMS and EAM will include an inventory system that totally supports the requirements of maintenance and the storeroom. Maintenance inventory will be managed to ensure that the right part is available at the right time without excessive inventory levels. Information from all available sources will be used to determine optimum stock levels. A continuous review of stock levels will be made to eliminate excess inventory and obsolete parts. Inventory reductions will be achieved through more partnerships with suppliers and vendors that establish joint commitments to purchase based on responsive service and fast delivery. Positions within MRO material management and procurement will increase in their importance and level of technical knowledge to perform effectively

22. *Working environment:* Successful maintenance operations will be safe, clean, and orderly because good housekeeping is an indicator of maintenance excellence.

- Maintenance leaders will provide a working environment where safety is a top priority, which in turn allows maintenance to set the example throughout the organization. Maintenance shop and work areas will be clean and orderly. Good housekeeping practices in maintenance will provide the basic foundation for safety awareness. Maintenance will provide support throughout the organization to ensure that all work areas are safe, clean, and orderly.

23. *Environment, health, and safety:* Maintenance must provide proactive leadership and support to the organization's regulatory compliance actions.

- Maintenance leaders must maintain the technical knowledge and experience to support compliance with all state and federal regulations under OSHA, USEPA (Clean Air Act), the U.S. Department of Transportation, and the Americans with Disabilities Act. The issue of indoor air quality must receive constant attention to eliminate potential problems. Maintenance must work closely with other staff groups in the organization, such as quality and safety, to provide a totally integrated and mutually supportive approach to regulatory compliance.

24. *Maintenance performance and service:* Broad based measures of maintenance performance and customer service will provide a continuous evaluation of the value of maintenance.

- CMMS and EAM will allow for a broad range of measurement for maintenance performance and service. Investment in maintenance best practices will require valid return on investment. Projected savings will be established and results will be validated. Measures will be developed in areas such as labor performance/utilization, compliance to planned repair and preventive/predictive maintenance schedules, current backlog levels, emergency repair hours, storeroom performance, asset uptime and availability, etc. Leaders of successful maintenance operations will continuously evaluate performance and service in order to manage maintenance as a business. They will adopt the philosophy of continuous maintenance improvement and have a method to measure progress.

25. *Maintenance planning and scheduling:* Maintenance customer service and the utilization of available craft time will improve through more effective planning and scheduling systems.

- The development of more effective planning and scheduling systems will be a top priority for the future capable maintenance operation. As reductions in breakdown repairs occur through more effective preventive/predictive maintenance, the opportunity to increase planned maintenance work will result. Maintenance and operations work closely to schedule repairs at the most convenient time. Maintenance will become more customer oriented and focus on achieving greater customer service by completing scheduled repairs on time. The utilization of craft time will increase as levels of planned work increases and as the uncertainties and inefficiencies associated with breakdown repairs are reduced.

World-class maintenance starts with a total commitment to a strategy of continuous maintenance improvement with maintenance as a top priority within the organization. It is the realization that maintenance is a key contributor to profit. It is the realization that maintenance best practice, plus people assets, plus MRO assets and information technology asset, all combine for the success and improvement of the total maintenance operation. It is the application of business process improvement techniques such as industrial engineering that generates the results and enhances the improvement process. Section 7, A Review of Other Selected Maintenance Best Practices, will give the reader greater understanding of maintenance management technology, principles and practices.

4. BENCHMARKING THE TOTAL MAINTENANCE OPERATION

Benchmarking can be defined as the continuous process of measuring our products, services, and practices against our toughest competition. Often our toughest competitor is our own organization, which does not understand the true value of maintenance, the high cost of deferred maintenance, or the dangers of gambling with a run-to-failure strategy of continuous reactive maintenance. It is important to realize that there are two basic forms of benchmarking, internal and external, and to understand why both are essential during the journey to maintenance excellence.

4.1. Internal vs. External Benchmarking

External benchmarking within maintenance allows for taking the global view of identifying best practices and determining how they can be transferred and applied successfully within your own unique maintenance operation. External benchmarking provides for developing broad-based comparisons with other maintenance operations in terms of best practices, standard operating procedures, and industry-wide statistical data. *The Scoreboard for Maintenance Excellence*, reviewed within this chapter, is today's most comprehensive guide for external benchmarking. It covers 18 best practice categories and 200 evaluation criteria (Peters 1994).

Think global—start local: Internal benchmarking starts locally within the maintenance operation at the shop floor. It focuses on measuring the successful execution of best practices such as CMMS, preventive and predictive maintenance, maintenance planning and scheduling, and effective maintenance storeroom operations (Peters 1997).

Internal benchmarking is about developing specific internal metrics or performance indicators. It is about determining progress from an internal baseline or starting point and measuring the progress toward a performance goal specific to your type of maintenance operation. For example, an internal benchmark could be the current level of maintenance-related downtime hours for a critical asset or the maintenance cost per unit of output, such as cost per ton, cost per carton, or cost per equivalent standard hour if a standard cost system is in place (Peters 1998).

4.2. The Scoreboard for Maintenance Excellence

4.2.1. Global Best Practices as External Benchmarks

There are many maintenance best practices that can serve as the global external benchmarks. External benchmarking is about gaining the knowledge and understanding of best practices and then applying them within the maintenance operation to help pursue and gain a manufacturing or service edge. Today's best maintenance practices are in areas such as:

- Preventive/predictive maintenance
- Continuous reliability improvement
- Reliability-centered maintenance
- Maintenance parts/materials control
- Maintenance storeroom operations
- Work order and work control
- Maintenance planning/scheduling
- Maintenance budget and cost control

- Operator-based maintenance
- Team-based continuous improvement
- Improving and measuring equipment effectiveness and reliability
- Craft skills development
- Maintenance performance measurement
- Computerized maintenance management systems
- Continuous maintenance improvement

Effective benchmarking should start locally, with a total evaluation of current maintenance practices and procedures, and then lead to the development of a strategic maintenance plan. The strategic maintenance plan provides the road map for applying maintenance best practices through a long-term process of continuous maintenance improvement. Benchmarking at its best is when maintenance effectively measures its level of services and practices and develops its own unique benchmarking criteria with high standards for maintenance excellence.

4.2.2. Conducting an Assessment of the Total Maintenance Process

Today's most comprehensive benchmarking guide, *The Scoreboard for Maintenance Excellence*, has been discussed in this chapter to support external benchmarking of your maintenance operation. *The Scoreboard for Maintenance Excellence* includes 18 maintenance best practice categories and 200 benchmark evaluation items. It has been used by over 4000 international organizations of all types to include translations into a number of languages. A complete copy of this very comprehensive document, with a guide for doing a self-assessment, can be downloaded from *www.tompkinsinc.com*. A summary of the evaluation categories and the number of evaluation items per category is included in Table 2.

Table 3 provides a general assessment of the overall benchmark rating, which has a maximum point value of 2000 points; 10 points maximum for each of the 200 benchmark evaluation items. Overall ratings of excellent (90–100%), very good (80–89%), good (70–79%), average (60–69%), and below average (less than 60%) are defined with general assessment comments.

Using *The Scoreboard for Maintenance Excellence*: It is recommended that the benchmark evaluation be conducted through a team effort within an organization or through an outside resource that can provide an objective and unbiased evaluation. When using an outside consulting resource, it is important to select a firm with broad-based experience in maintenance management as well as knowledge and experience with CMMS implementation.

When using an internal team, it is important to select knowledgeable members from maintenance as well as representatives from the storeroom, purchasing, accounting, production, and engineering

TABLE 2 Maintenance Evaluation Summary

Section	Evaluation Category	Evaluation Items
A.	Maintenance and organization culture	10
B.	Organization and administration	12
C.	Work authorization and work control	10
D.	Budget and cost control	11
E.	Maintenance planning and scheduling	12
F.	Maintenance storeroom	16
G.	Preventive and predictive maintenance	22
H.	Lubrication program	11
I.	Overall equipment effectiveness	9
J.	Operator-based maintenance	8
K.	Engineering support	9
L.	Safety, housekeeping, and regularity compliance	12
M.	Craft skills development	9
N.	Maintenance performance measurement	9
O.	Maintenance supervision/leadership	6
P.	Computerized maintenance management systems (CMMS)	13
Q.	Maintenance facilities, equipment, and tools	7
R.	Continuous maintenance improvement	14
	Total evaluation criteria	200

TABLE 3 General Assessment of Overall Benchmark Rating

General Assessment of Overall Benchmark Rating	
Total Point Range	Overall Rating Summary
1800 to 2000 (90–100%)	*Excellent:* Practices and principles in place for achieving effective maintenance and world-class performance based on actual results. Reconfirm overall maintenance performance measures. Maintain strategy of continuous maintenance improvement. Set higher standards for maintenance excellence and measure results.
1600 to 1799 (80–89%)	*Very good:* Fine-tune existing operation and current practices. Reassess progress on planned or ongoing improvement activities. Redefine priorities and renew commitment to continuous maintenance improvement.
1400 to 1599 (70–79%)	*Good:* Reassess priorities and reconfirm commitments at all levels to maintenance improvement. Evaluate maintenance practices and develop and implement plans for priority improvements. Ensure that measure to evaluate maintenance performance and results are in place. Initiate strategy of continuous maintenance improvement.
1200 to 1399 (60–69%)	*Average:* Conduct a complete assessment of the maintenance operation and current practices. Determine total costs/benefits of potential improvements. Develop and initiate strategy of continuous maintenance improvement.
Less than 1200 (<60%)	*Below average:* Same as Average, plus depending on the level of the rating and major area that is Below average, immediate attention may be needed to correct conditions adversely affecting on life, health, safety, and regulatory compliance. Priority to key issues, major equipment, or increasing costs that are having a direct impact on the immediate survival of the business.

and computer systems. An outside consultant can also be used as part of the team approach to performing the benchmark evaluation. It is recommended that maintenance provide the team leader to support overall coordination of the benchmarking evaluation.

A self-assessment will provide benefits; however, an assessment conducted by an outside resource provides a greater sense of the big picture in terms of objectivity and completeness. Should you want to begin with an internal study of maintenance, here are some guidelines to consider when using *The Scoreboard for Maintenance Excellence.*

- *Obtain Leadership Buy-in:*
  - Establish a firm commitment from company leadership to take action based upon results of the assessment.
  - Gain commitment from company leadership for the necessary resources.
- *Charter maintenance excellence team:*
  - Establish a maintenance excellence strategy team to guide and promote improved maintenance practices within the operation.
  - Utilize a team-based approach with a cross-functional assessment team specifically chartered for conducting and preparing the results of your assessment.
  - Have at least one team member with solid background in each of 18 evaluation categories.
- *Define and weight benchmarking criteria:*
  - Gain complete understanding of each evaluation criterion (200 total).
  - Modify existing evaluation criteria as required.
  - Add evaluation criteria as required.
  - Ensure that all team members understand the scoring process and develop consistency in scoring each area.
- *Develop action plan:*
  - Determine information required, persons to interview, and observations needed prior to start of assessment.
  - Develop schedule and implementation plan for the assessment.

- *Conduct assessment:*
  - Assign team members to specific evaluation categories (ideally, in two-person teams for each category).
  - Conduct kickoff meeting, firm up schedules, etc.
  - Conduct the assessment, record observations, and assign scores to each evaluation criteria.
- *Analyze and review results:*
  - Review all scoring for consistency.
  - Develop final results of the assessment and document in a written report.
  - Determine strengths/weaknesses, priorities for action, and benefits.
  - Present results to company leadership, define benefits and get commitment for investments.
- *Develop path forward for maintenance excellence:*
  - Develop a strategic maintenance plan (SMP) and implement best practices.
  - Measure benefits and validate ROI.
  - Maintain a continuous maintenance improvement process (i.e., repeat assessment process).

Each of the 200 benchmark items will include evaluation criteria to provide a means to rate each item with up to 10 points maximum. Examples from *The Scoreboard for Maintenance Excellence* are shown in Tables 4 and 5 to illustrate the two methods for establishing benchmark rating values. Ratings are established as follows:

- *Based on point value:* The benchmark item is rated as excellent (10 or 9 points), very good (8 points), good (7 points), average (6 points), below average (5 points) and poor (4 points or less), as shown in the example in Table 3. This method of rating requires that a consistent judgment be made on assigning point values, particularly in the areas rated average and below average.
- *Based on specific conditions:* The benchmark criteria will describe a number of specific conditions that are assigned point values. Table 4 provides two examples of this type of rating from the maintenance storeroom category. For example, benchmark item 6 rates whether storeroom inventory accuracy is being measured and is over 95% for a score of 10. In turn, inventory accuracy levels within 90–95%, 9 points, 80–89%, 8 points, 70–79%, 7 points, and below 70%, 5 points.

The benchmark evaluation score: As the benchmark evaluation is completed, it is important to realize that the total point value is not an absolute value but represents an important baseline value unique to your maintenance operation. It provides a baseline value as to where you are with best

TABLE 4 Example for Benchmark Rating Based on Point Values

Item #	Benchmark Item	Benchmarking Criteria	Rating
A.	Maintenance and Organization Culture		
1.	The organization's vision, mission, and requirements for success include physical asset management and maintenance as a top priority.	The organization has written mission statement/goals that include maintenance and/or preventive maintenance as a top priority and key goal. Rated as: Excellent—10 or 9, Very good—8, Good—7, Average—6, Below average—5, Poor—4 or less.	
2.	Senior management is visible and actively involved in continuous maintenance improvement and is obviously committed to achieving maintenance excellence.	Management commitment is rated as: Excellent—10 or 9, Very good—8, good—7, Average—6, Below average—5, Poor—4 or less.	

TABLE 5 Example for Benchmark Rating Based on Specific Conditions

Item #	Benchmark Item	Benchmarking Criteria	Rating
F. Maintenance Storeroom			
5.	Inventory accuracy is determined by an effective cycle-counting program.	• Cycle counting used—10 • Count once per year—7 • Count occasionally—5 • Do no inventory counts—0	
6.	Inventory accuracy is regulary measured and is 95% or above.	Inventory accuracy 95% or above—10 90%–95%—9 80%–89%—8 70%–79%—7 <70%—5	

practices from *The Scoreboard for Maintenance Excellence.* In Section 5, very specific metrics for internal benchmarking and measurement of progress will be reviewed.

At this point, the benchmarking process has helped define current strengths and weaknesses within a maintenance process. Benchmarking helps to identify where we are with best practices that are essential to an effective asset management. By using a benchmarking guide such as *The Scoreboard for Maintenance Excellence*, you can easily see the status of current practices:

- Work order/work management
- Equipment/asset management
- Preventive/predictive maintenance
- Planning and scheduling
- Parts inventory management/purchasing
- Budgeting and cost control

The benchmark evaluation may reveal weaknesses related to the primary CMMS modules. A wide range of scores have come from the benchmark assessments conducted by Tompkins Associates and others using *The Scoreboard for Maintenance Excellence*. The important point was not the numeric score but that each organization had taken the first step with a total benchmark evaluation of their maintenance operation and used the results to develop and implement a strategic plan for improvement.

4.2.3. Internal Benchmarking: The Key to Validation of Results and ROI

Measurement of improvement internal to a maintenance operation is the real value of internal benchmarking. It's knowing exactly where we are with applying best practices, gaining a commitment from the organization to implement those needed, and then measuring the results. Internal benchmarking provides the means to validate results of maintenance improvement, define the ROI, and manage and control maintenance as a business.

It makes little sense to develop and worry about industry-wide benchmark statistics when, for example, an operation does not measure its own maintenance-related downtime. Industry-wide benchmark statistics can be very misleading when we try to compare our maintenance operation to them. On the other hand, when we understand today's best practices for effective maintenance, apply them, and measure the results, we are providing real value (Peters 1998).

5. PERFORMANCE MEASUREMENT: KEY TO MAINTENANCE CONTROL

5.1. Techniques for Measuring Maintenance Activities

Maintenance work, by its very nature, seldom follows an exact pattern for each occurrence of the same job. Traditional work-measurement techniques are not readily adaptable to maintenance-type work. Therefore, exact methods and exact times for doing most maintenance jobs cannot be established as they can for discrete manufacturing-type work. However, the need for having reliable performance measures for maintenance planning becomes increasingly important as the cost of maintenance labor rises and the complexity of production equipment increases.

Various methods for establishing maintenance performance standards have been used, including:

- *Reasonable estimates:* A knowledgeable person, either a supervisor or planner, uses his or her experience to provide their best estimate of the time required. This approach does not scope out the job in much detail to determine method or special equipment needed.
- *Historical data:* The results of past experience are captured via the CMMS or other means to get average times to do a specific task. Over time, a database of estimated time is developed that can be updated with a running average time computed for the tasks.
- *Predetermined standard data:* Standard data tables for a wide range of small maintenance tasks have been developed. A standard data example is shown for one task for the pipefitting craft of cutting with a pipe machine. Standard data represent the building blocks that can then be used to estimate larger, more complex jobs. Each standard data table provides what the operation is, what is included in the time value, and the table of standard data time for the variables that are included. The variables for the example standard data sheet is the size of the pipe from ½ in. diameter up to 8 in. diameter.

Example: Maintenance Standard Data—Pipefitting

Code PF-2-1

SKILL: Pipefitting

OPERATION: Pipe—Cut with Pipe Machine

Includes:

1. Pick up pipe, place in chuck, and tighten chuck jaws.
2. Measure pipe for cut and position carriage.
3. Hand feed parting tool to cut piece off pipe.
4. Hand feed parting tool to remove burrs.
5. Loosen chuck jaws, remove pipe from chuck, and set aside.
6. Dispose of scrap or set aside unused piece.

The three techniques described previously require that an outside party establish the standards that are then imposed upon the maintenance force as an estimated time on the work order. These approaches to craft performance measurement often bring about undue concern and conflict between management and the maintenance workforce over the reliability of the standards. Rather than progressing forward together in a spirit of continuous improvement, the maintenance workforce in this type environment often works against management's program for maintenance improvement. Some other very progressive options to the challenging task of work measurement within nonrepetitive maintenance operations are available.

5.2. The ACE Team Benchmarking System

As a means to overcome many of the inherent difficulties associated with developing maintenance performance standards, the ACE (A Consensus of Experts) team benchmarking system relies primarily on the combined experience and estimating ability of a group of skilled crafts personnel. The objective is to determine reliable planning times for a number of selected benchmark jobs. This system places a high emphasis on continuous maintenance improvement and the changing of planning times to reflect improvements in performance and methods as they occur.

Generally, the ACE team benchmarking system parallels the UMS (universal maintenance standards) approach in that the range of time concept and slotting are used once the work content times for a representative number of benchmark jobs have been established. The ACE team benchmarking system focuses primarily on the development of work content times for representative benchmark jobs that are typical of the craft work performed by the group (Peters 1996a).

Once a number of benchmark job times have been established, these jobs are then categorized onto spreadsheets by craft and task area and according to work groups which represent various ranges of times. A spreadsheet is then set up for four work groups, such as for jobs in the E, F, G, and H time slots. In this case, work group E would be for benchmark jobs ranging from 0.9 hr up to 1.5 hr and assigned a standard time (slot time) of 1.2 hr. Likewise, work group F would be for benchmark jobs ranging from 1.5 hr up to 2.5 hr and assigned a standard time of 2.0 hr. This spreadsheet would

TABLE 6 Standard Data: Pipe Size ½ in. to 8 in.

Pipe Size	Normal Minutes per Cut
1/2 in.	0.9
3/4 in.	1.0
1 in.	1.1
1-1/4 in.	1.2
1-1/2 in.	1.3
2 in.	1.4
2-1/2 in.	1.6
3 in.	1.8
3-1/2 in.	2.2
4 in.	2.5
5 in.	3.4
6 in.	4.4
8 in.	6.2

also include brief descriptions of the benchmark jobs that have work group times of 1.2 (E), 2.0 (F), 3.0 (G), and 4.0 (H) hr respectively. All of the work groups, and their respective time ranges, are shown in Table 7.

Spreadsheets provide the tool: After spreadsheets have been prepared based on the representative benchmark jobs from various craft/task areas, a planner/analyst now has the means to establish planning times for many different maintenance jobs using a relatively small number of benchmark jobs as a guide for work content comparison. By using work content comparison combined with a good background in craft work and a knowledge of the benchmark jobs, a planner now has the tools to establish reliable performance standards consistently, quickly, and with confidence. The ACE team benchmarking system allows a wide range of tasks to be estimated with a 95% confidence level by using work content comparison to a small but representative sample of well-defined benchmark jobs.

TABLE 7 ACE System Time Ranges and Work Groups

	Time Range		
Work Group	From	Standard Time (Slot time)	To
A	0.0	0.1	0.15
B	0.15	0.2	0.25
C	0.25	0.4	0.5
D	0.5	0.7	0.9
E	0.9	1.2	1.5
F	1.5	2.0	2.5
G	2.5	3.0	3.5
H	3.5	4.0	4.5
I	4.5	5.0	5.5
J	5.5	6.0	6.5
K	6.5	7.3	8.0
L	8.0	9.0	10.0
M	10.0	11.0	12.0
N	12.0	13.0	14.0
O	14.0	15.0	16.0
P	16.0	17.0	18.0
Q	18.0	19.0	20.0
R	20.0	22.0	24.0
S	24.0	26.0	28.0
T	28.0	30.0	32.0

Since the actual times assigned to the benchmark jobs are so critical, it is very important to use a technique that is readily acceptable. The ACE team benchmarking system provides such a technique, since it is based on the combined experience of a group of skilled crafts personnel, and their consensus agreement on the range of time for the benchmark jobs. The following is a recommended approach for using the ACE team benchmarking system (Peters 1996a).

1. *Select benchmark jobs:* Review past historical data from work orders and select representative jobs that are normally performed by the craft groups. Special attention should be paid to determine the 20% of total jobs (or types of work) that represent 80% of the available craft manpower. Focus on determining repetitive jobs where possible in all craft areas.
2. *Select and train experts* (*ACEs*): It is important to select experienced crafts people and/or supervisors who, as a group, have had experience in the wide range of jobs selected as benchmark jobs. All craft areas should be represented in the group. In order to ensure that this group understands the overall objectives of the maintenance planning effort, special training sessions should be conducted to cover the procedures to be used, reasons for establishing performance measures, etc. A total of 6 to 10 craft people is the recommended group size.
3. *Develop major elemental breakdown for benchmark job:* For each benchmark job that is selected, a brief element analysis should be made to determine the major elements of the total job. This listing of the major steps of the job should provide a clear, concise description of the work content for the job *under normal conditions*. It is important that the work content for a benchmark job be described and viewed in terms of what is a *normal* repair and not what may occur as a rare exception.

   An excellent resource to consider for doing the basic element analysis for each benchmark job is the crafts (ACEs) that are selected for doing the estimating. Brief training on methods/operations analysis can be included in the initial training for the ACEs. This process leads to the question "Are we using the best method, equipment, or tools for the job?" Often significant methods improvements are discovered and implemented as a result of this step.

   Major exceptions to a routine job should be noted if they are significant; generally an exception will be analyzed as a separate benchmark job, along with an estimate of time required for such repair. This portion of the procedure ensures that the work content of each benchmark job is clearly defined so that each person doing the estimating has the same understanding about the nature and scope of the job. When the benchmark jobs are finally categorized into spreadsheets, the work content description developed in this step is used as key information about the bench job on the spreadsheet.
4. *Conduct first independent evaluation of benchmark jobs:* Each member of the group is now asked to review the work content of the benchmark jobs and to assign each job to one of the UMS time ranges or slots. Each member of the group provides an independent estimate that represents an unbiased personal estimate of the pure work content time for the benchmark job.

   It is important here for each member of the group to remember that only the work content of the benchmark job description is to be estimated, and not the make ready and put away activities associated with the job. This part of the procedure is concerned only with estimating the pure work content, *excluding* things such as travel time, securing tools and parts, delays, and personal allowances.

   The estimate should be made for each job under these conditions:

   (a) An average skilled craftsman is doing the job giving 100% effort, i.e., a fair day's work for a fair day's pay.

   (b) The correct tools are available at the job site or with the craftsman.

   (c) The correct parts are available at the job site or with the craftsman.

   (d) The machine is available and ready to be repaired.

   (e) The craftsman is at the job site with all of the above and proceeds to complete the job from start to finish without major interruption.

   The work accomplished under these conditions therefore represents the pure work content of the job to be performed. Establishing the range of time estimate for this pure work content is the prime objective of the first evaluation. It is important for each estimator to remember that to develop a planning time requires pure work content time plus additional time allowances to cover make ready and put away-type activities associated with each job.

5. *Summarize first evaluation:* Results of the first evaluation are then summarized to check the agreement among the group as to the time range for each benchmark job. A coefficient of concordance can be computed from the results. A value of 0.0 denotes no agreement, while a value of 1.0 denotes complete agreement, or consensus among the ACEs.

   Group members who are significantly higher or lower than the rest of the group for a particular benchmark job are asked to explain their reasons for their high or low estimates. This information will then be used during the second evaluation to refine the next estimates from the entire group.
6. *Conduct second independent evaluation of benchmark jobs:* A second evaluation is conducted using the overall results from the first evaluation as a guide for the group. Various reasons for high or low estimates from the first evaluation are provided to the group prior to the second evaluation. This allows for adjustment to individual estimates if the other experts' reasons are considered to be valid.
7. *Summarize second evaluation:* Results of the second evaluation are then summarized to evaluate changes or improvements in the agreement or consensus among the ACEs as to the time range for each benchmark job. This evaluation should produce improved agreement among the group. If an extreme variance in time range estimates still exists, further information regarding the work content of the job may be needed.
8. *Conduct third independent evaluation if required:* This evaluation is required only if there remains a wide variance in the estimates among the group.
9. *Conduct group session to review final results:* This session serves to finalize the results achieved and to discuss any of the high or low estimates that have not been resolved completely. A final group consensus on all time ranges is the objective of this session.
10. *Develop spreadsheets:* The benchmark jobs with good work content descriptions and agreed-upon time ranges can now be categorized onto spreadsheets. From these spreadsheets, which give work content examples for a wide range of typical maintenance jobs, a multitude of individual maintenance performance standards can be established by the planner through the use of work content comparison. The basic foundation for the maintenance planning system is now available for generating consistent planning times that will be readily acceptable by the maintenance work force that developed them.

The ACE Team Benchmarking System approach combines the DELPHI technique for estimating along with the inherent and inevitable ability of most people to establish a high level of performance measures for themselves. As used in this application, the objective is to obtain the most reliable, reasonable estimate of maintenance-related work content time from a group of experienced crafts personnel.

This approach allows for independent estimates by each member of the group, which in turn builds into a consensus of expert opinion for a final estimate. The final results are therefore more readily acceptable because they were developed by skilled and well-respected crafts people from within the work unit. Application of the ACE team benchmarking system promotes a commitment to continuous maintenance improvement and provides reliable planning time for a wide range of maintenance activities.

5.3. Establishing Key Performance Indicators

This section will outline and describe key performance indicators that can be used to measure the overall effectiveness of a maintenance operation.

- *Percent craft utilization* (*CU*): Evaluates actual wrench time (hands-on time) for craft labor. Provides one of the two key elements for measuring the overall craft effectiveness (OCE). Measures the overall increase in craft labor wrench time due to a proactive, planned maintenance strategy with effective planning and scheduling, positive impact from the PM/PdM program, effective MRO materials management service, and improved CMMS.
- *Percent craft performance* (*CP*): Evaluates actual craft performance against a reasonable/reliable planned time for a planned repair job or task such as PM inspections. Where craft labor utilization measures effectiveness, this measure addresses the efficiency factor for overall craft effectiveness (OCE). This measure is improved by having effective craft skills to do the job, along with the motivation to work efficiently. It is directly impacted by shop working areas, having the right personal and special tools available, and safe working conditions.
- *Overall equipment effectiveness* (*OEE*): A world–class metric that originated from the total productive maintenance (TPM) movement, which evaluates critical equipment in terms of equip-

ment availability, equipment performance, and the quality of output. Improving OEE focuses on eliminating the six major losses:

Availability issues:	1. Breakdowns
	2. Set-up and adjustments
Performance issues:	3. Idling and minor stoppages
	4. Reduced speed
Quality Issues:	5. Process defects
	6. Reduced yield

The average OEE factor is in the 40–50% range before an improvement process starts. A world-class OEE factor is around 85%, which means that all three elements must be around 95%, i.e. 0.95 (availability) × 0.95 (Performance) × 0.95 (Quality) = 0.857 × 100 = 85.7% OEE factor.

$$\text{The OEE factor} = \text{availability } \% \times \text{performance } \% \times \text{quality } \%$$

- *The overall craft effectiveness* (*OCE*) *factor:* The OCE factor relates to craft labor assets, as compared to the metric for OEE, which measures the combination of equipment asset availability, performance, and quality output. The OCE factor focuses upon measuring and improving the value-added contribution that people assets make to total asset management. Section 7.5 includes more detailed information about the OCE factor.
- *Number of call-backs* (*or percentage of craft rework*): Measures the quality of maintenance repair work and provides one key indicator for maintenance customer service. It helps focus on doing the repair right the first time.
- *Schedule compliance* (*percent jobs completed as scheduled*): Evaluates the overall effectiveness of executing the planned work on the agreed-upon schedule by the craft workforce.
- *Percent planned work:* Measured by either the percent work orders planned or percent actual craft hours on planned work. This metric evaluates the overall effectiveness of the planning process as well as the impact of all maintenance best practices to promote a proactive maintenance repair strategy, i.e., PM, PdM, reliability improvement actions, effective MRO support, etc.
- *Percent work orders with planned time:* Provides one very key measure for the maintenance planner position. Evaluates the ability and effectiveness of the maintenance planner/coordinator position to establish reliable planning times, using maintenance standard data or other available data for determining planning times. The objective for the planning process is to have as many jobs scoped and planned as possible, to the level that reliable estimates are established.
- *Percent planned work orders generated from PM/PdM program:* Provides valuable feedback that the PM/PdM effort is helping avoid catastrophic failure and is truly providing benefits. Measures how well PM/PdM program is detecting deficiencies before catastrophic failure or downtime occurs. The reliability improvement goal, however, is not to continue to fix before failure but to eliminate root causes of failure and in turn the failure rate. However, this is an important metric as a client's PM/PdM program is reinforced or renewed and in turn begins to provide measurable results.
- *Percent PM/PdM compliance:* Measures the execution and compliance to completing scheduled PMs, PdM data collection, lube services, scheduled structural/process inspections, calibrations, etc. Typical completion of scheduled PM/PdM actions required within a week's window of time is used as criterion for compliance. This metric can also apply to instrumentation calibration and to any regulatory inspections such as crane, fire protection testing/PM, etc.
- *Number of stock outs:* Evaluates the MRO inventory/materials management process capability to have the on-hand quantities needed for stock items that are normally stocked in the maintenance storeroom. Provides an excellent countermeasure to ensure planned inventory reduction goal is not detrimental to storeroom customer service. Does not measure nonstock items that require requisitioning/purchasing as direct purchases or availability of project-related items.
- *Percent inventory accuracy:* Maintains accurate fiscal accountability of stocked items to ensure total confidence in current inventory levels and dollar value of MRO inventories. Measures the effectiveness of the cycle-counting process and storeroom control.

- *Dollar value of MRO inventory reductions:* Helps to ensure that proactive MRO inventory management practices and those well-planned MRO inventory reductions are given proper credit and recognition within the organization. This reduction may also serve to offset additions to inventory of more useful items and critical spares.
- *Percent asset utilization:* A good metric that evaluates how well the overall capacity of an asset is being used in the operation. It should be supplemented by measurement of overall equipment effectiveness (OEE).
- *Maintenance cost per unit of output:* Measures bottom-line maintenance cost per unit of output and evaluates net improvements related to maintenance improvements on total operation costs. Excellent metric for process-type industries and/or for discrete manufacturers with major single product output and a strong standard cost and production-reporting system.
- *Maintenance cost as percent of total operations cost:* Provides an overall comparison of how maintenance cost impacts total cost.
- *Other metrics to consider:*
  - Overtime
  - Craft time charged to work orders
  - Overdue work orders
  - Emergency craft hours

TABLE 8 Sample Maintenance Excellence Index

MAINTENANCE EXCELLENCE INDEX

Performance Measurement Categories

	Craft Productivity			Cost	Planning & Scheduling		MRO Mat'l Management		PM Compliance			Asset Uptime		
A. Performance Metric	Craft Utilization	Craft Performance	% Job Completed as Scheduled	Maintenance Cost Per Carton	% Planned Hours	% Work Orders with Standard Time	$ of Stockouts	% Inventory Accuracy	Mobile PM Compliance	Conveyor PM Compliance	Bldg Sys PM Compliance	Sorter % Utilization	Forklift % Utilization	
B. Current Month Performance	55	75	85	0.106	70	80	6	85	90	80	95	98	96	Performance Level
C. Performance Goal	80	95	95	0.082	85	90	2	95	100	100	100	100	98	10
	75	90	90	0.086	80	85	3	90	95	95	95	98	95	9
	70	85	85	0.090	75	80	4	85	90	90	90	96	92	8
	65	80	80	0.094	70	75	5	80	85	85	85	94	89	7
	60	75	75	0.098	65	70	6	75	80	80	80	92	86	6
	55	**70**	70	0.102	60	65	7	70	75	75	75	90	**83**	5
	50	65	65	0.106	55	60	**8**	**65**	**70**	70	**70**	**88**	80	4
D. Baseline Performance in **BOLD**	45	60	**60**	**0.110**	**50**	55	9	60	65	65	65	86	77	3
	40	55	55	0.114	45	50	10	55	60	**60**	60	84	74	2
	35	50	50	0.118	40	45	11	50	55	55	55	82	71	1
	30	45	45	0.122	35	40	12	45	50	50	50	80	68	0
E. Current Performance Scores	5	6	8	4	7	8	6	8	8	6	9	9	9	
G. Weighted Value of Performance Metric	12	8	6	14	8	6	8	10	5	7	2	9	5	
H. Performance Value, *Score (F) x Weight (G)*	60	48	48	56	56	48	48	80	40	42	18	81	45	

F. PERFORMANCE LEVEL SCORES

SCORES X WEIGHT

= **670**

I. TOTAL MEI PERFORMANCE. VALUE: January 2000

J. Total MEI Values Over Time	Dec 98	Jan 99	Feb 99	Mar 99	Apr 99	May 99	Jun 99	Jul 99	Aug 99	Sep 99	Oct 99	Nov 99	Dec 99
	490	510	525	539	550	570	585	590	630	649	650	690	646

MAXIMUM POSSIBLE MEI PERFORMANCE VALUE 1000 TOTAL POINTS

- PM/PdM coverage, i.e., the equivalent craft hours invested in PM/PdM tasks, which can also be shown as equivalent crafts people
- Response time to priority jobs
- Closed work orders without time
- Craft hours charged back to customers
- Jobs by one-person crew
- Percent craft hours by work order type
- Percent craft hours by work/cost center

5.4. Developing a Maintenance Excellence Index

Often performance measurement is something new to the maintenance operation, but it is highly recommend that a performance-measurement system be put in place. Justification for investments in maintenance best practices must be validated. One approach used by Tompkins Associates has been to help clients create a maintenance excellence index (MEI) that includes 10–15 key performance indicators. These metrics, with agreed-upon weighted values, are developed into a composite total performance value (Peters 1998). The metrics that are selected should cover areas such as:

√ Craft Productiivty
√ Budget/cost
√ Planning and scheduling
√ Parts inventory management
√ PM compliance
√ Asset uptime/OEE
√ Asset utilization
√ Customer service

- *The Maintenance Excellence Index* (*MEI*): The sample MEI shown in Table 8 provides a composite index that integrates a number of key metrics. Each metric can also be monitored and trended individually. The MEI helps keep in focus the fact that the success of maintenance operations depends on many factors. Therefore, one or two metrics can't provide the total performance picture. The MEI concept provides a complete approach to maintenance performance measurement. If a maintenance process improvement has been justified on craft productivity increases, parts inventory reduction, maintenance cost per unit of output reduction, or value of increased asset uptime, the original projection of benefits and savings can be validated with the MEI.
- *The basic format of the MEI calculations:* The total MEI performance value is the composite score of all metrics, considering current performance of each metric as compared to the goal and the weighted value of each individual metric. The total possible score for the total MEI performance value is 1000. The key steps in developing and using the MEI are as follows:

Step	Description	Comments
A	Performance metrics	From 10 to 15 metrics are selected and agreed upon by the organization.
B	Current month performance	This is the acutal performance level for the metric for the reporting month. This value will also be noted in one of the incremental values blocks below the performance goal. This value will correspond to a value for F, the performance level scores which go from 10 down to 1.
C	Performance goal	This is the preestablished performance goal for each of the MEI metrics. For example, if the current month's performance is at the performance goal level, the performance level score for that goal will be a 10, the maximum score.
D	Baseline performance	The baseline performance level prior to start of MEI performance measurement

Step	Description	Comments
E	Current performance score	Depending on the current month's performance, a performance level score (F) will be obtained. This value then goes to the current performance score row and serves as the multiplier for (G) the weighted value of the performance metric.
F	Performance level score	Values from 10 down to 1, denote the level of current performance compared to the goal. If current performance achieves the predetermined goal, a performance value of 10 is given. Each metric is broken down into incremental value, from the baseline to the goal. Each incremental value in the column corresponds to a performance level value. This value becomes the current performance score.
G	Weighted value of the performance metric	The values along this row are the weighted value or relative importance of each of the metrics. These values are obtained via a team process and a consensus on the relative importance of each metric that is selected for the MEI. All of the weighted values sum to 100.
H	Performance value score	The weighted values (G) are multiplied by (E) the current performance scores to get the performance value score (H).
I	Total MEI performance value	The sum of the performance value scores for each of the 15 metrics and the composite value of monthly maintenance performance on all MEI metrics.
J	Total MEI performance values over time	Location for tracking total MEI performance values over a number of months

Investments in maintenance and total operations improvements should be measured and bottom-line return on investment validated. The proven methodology of the MEI utilizes existing data to provide a very effective multifactored system to measure mission-critical maintenance and its contribution to total operation performance and success. Organizations will make a wise decision to start this process, for example, as a new CMMS system is purchased and begins implementation. Positive trends on the Maintenance Excellence Index will validate the projected results: tangible benefits and savings that were used to justify the capital investment for this and many other types of maintenance-improvement projects.

6. INFORMATION TECHNOLOGY TO SUPPORT MAINTENANCE MANAGEMENT AND CONTROL

6.1. Introduction to Computerized Maintenance Management Systems/Enterprise Asset Management

The purchase of information technology in the form of a CMMS/EAM system is not a quick fix or a panacea. Many excellent systems are available, but surveys have consistently shown that only 80% of CMMSs are actually being used. In turn, typically only 30% of CMMSs functional capabilities are being used. Expenditures for CMMSs were $553 million in 1995 and have grown to over $1 billion in 2000. Therefore, gaining full value from your CMMS investment should be a top priority.

Basically, CMMS/EAM provides computer software programs designed to assist in the planning, management, and administrative procedures required for an effective maintenance process. The need for a maintenance management system, whether manual or computerized, is determined by the need to perform effective maintenance on assets, equipment and facilities. We can think of CMMS/EAM as providing the business system for maintenance operations.

Choosing a CMMS/EAM and determining how it will operate and what it will do should occur only after recognizing that a system is needed or that an existing system needs improvement. The need for, and use of a CMMS/EAM is not specific to any one industry or type of application. Anyplace where maintenance is done is a candidate user, and all users have the same basic system needs as dictated by the maintenance process itself. The scope of maintenance is broad and technically challenging. Likewise, the scope of today's CMMS/EAM is broad based to include a number of integrated modules to manage the maintenance process.

- *CMMS/EAM facilitates best practices:* Implementation of a CMMS is most successful in an organization that has committed to a total reevaluation and internal benchmarking of current practices and procedures, as reviewed previously in this chapter. Organizations should realize that lack of a CMMS is never the main problem. The main problem is the lack of best practices. A CMMS provides the framework to fully integrate best practices into the maintenance process; to provide the systems, procedures, and information to manage maintenance as a business. The effective use of a CMMS is important, but it is only part of the process for improving maintenance management. The maintenance organization plays a big part; it needs to be refined before a CMMS is put in to support it. Successful implementation of a CMMS will facilitate applying best practices discussed previously, and will provide tangible benefits and savings. Improved performance of the following maintenance activities can be expected to support justification of a CMMS/EAM (Peters 1999):
- *Improved work control:* The work order module is the heart of any CMMS. It provides the basis for work management, cost tracking, equipment history, and performance reporting. Effective CMMS supports improved control of work requests by craft, monitoring backlogs, determining priorities, and scheduling of overtime.
- *Improved planning and scheduling:* Unplanned maintenance work typically costs two to three times as much as planned work. A CMMS provides the systems and procedures to establish a more effective maintenance planning and scheduling function, which is a key contributor to improved craft labor utilization.
- *Enhanced preventive and predictive maintenance* (*PM/PdM*): Automatic scheduling of repetitive PM activities is possible through CMMS. PM tasks and inspection frequencies can be documented on the PM module and printed as part of the PM work order. A CMMS enhances PM by providing a method to monitor failure trends and highlight major causes of equipment breakdowns and unscheduled repairs.
- *Improved parts availability:* A well-organized stockroom with accurate inventory records, stock locator system, stocking levels, and a storeroom catalog can significantly improve the overall maintenance operation. Having the right part at the right time is the key to effective maintenance planning and increased maintenance customer service.
- *Reduced storeroom inventory:* A CMMS provides the means for more effective management and control of maintenance parts and material inventories. Information for decisions on inventory reduction is readily available to identify parts usage, excess inventory levels, and obsolete parts.
- *Improved failure and repair analysis:* A CMMS provides the means to track work order and equipment history data related to types of repairs, frequencies, and causes for failure. It allows maintenance to have key information on failure trends that leads to eliminating root causes of failures and improving overall equipment reliability.
- *Increased budget accountability:* A successful CMMS requires greater accountability for craft labor and parts/materials through the work order and storeroom inventory modules. This increased level of control provides greater accountability of the overall maintenance budget by individual pieces of equipment and by using department or work center.
- *Increased capability to measure performance:* The CMMS database provides a vast source of maintenance information to allow more effective measurement of maintenance performance and service. Successful CMMS applications will establish internal benchmarks to provide a measurement for improvements in the areas of craft labor utilization, PM/PdM compliance, downtime, store inventory control, backlog, level of planned maintenance, etc.
- *Increased level of maintenance information:* A major benefit of a CMMS comes from developing the historical database that becomes readily available as critical maintenance information. Many types of maintenance data are normally available, but it is often not in the form that helps manage and control maintenance operations. Now such data can be turned into valuable maintenance information with an effective CMMS.

6.2. Overview of CMMS/EAM System Functionality

This section provides an introduction to the overall scope, functionality summaries, and the primary CMMS modules that make up a typical integrated system. Many other system functions are available.

The modules described here are provided in various levels of functionality by over 100 CMMS/EAM vendors. These modules and the respective functionality they provide to the user comprise the basic elements needed for effective management of the maintenance process (Peters 1999).

- *Work order/work management module:*
  - Manages and controls approvals of work requests for correcting faults or improving an asset's condition. May be configured to require that certain types of work requests go through a review or approval cycle before they are submitted to maintenance.
  - Opens, edits, and records work orders as history. Basic work order information on the work request might include the equipment/asset number to be worked on, a description of the problem or work to be done, who requested the work, when they wanted it completed, the priority of the work, and the date the request is made.
  - This module records the basic work order information and supports editing the details on the work needed, parts and materials required, sequence of work, job steps, and other pertinent data.
  - Planning and scheduling of work orders may be available to associate each step with specific equipment, craft foreman, estimated and standard duration, lists of tasks, and special tools and equipment required.
  - Provides linking of craft resources, parts requisitions, bill of materials, routes and checklists, safety documentation/permits, material, and labor costing to a job step.
  - Work order completion is recorded with final information and closed to work order history file.
  - Reporting may be available to work orders in current backlog as well as in the work order history file.
- *Equipment/asset management module:*
  - Provides listing and basic data for equipment/assets owned and maintained. May include breakdown of asset information related to hierarchies of system/subsystems.
  - Information recorded may include equipment/asset number, location, maintenance priority, cost, manufacturer, model and serial numbers, and the dates of purchase and installation.
  - Defines assemblies and spare parts lists and supporting information, such as calibration data, which may also be kept in this module.
  - May include detailed support documentation associated with the asset: drawings, specifications, technical reference manuals, performance standards, extended descriptions, etc. Can be referenced for presentation on screen or printed on demand.
  - Support documentation data may be stored as free text, formatted, or a combination of both.
  - Parts lists usually are formatted and within an integrated system are linked with the inventory management module for data validation.
  - Operational data such as hour meter or temperature readings may be captured in this module to allow usage information to trigger the need for maintenance work or inspection.
  - Tracks failure information at the asset system/subsystem level through information entered when closing work orders and provides capability to compute various key performance indicators.
  - Tracks detailed cost history data by type of cost associated with asset system/subsystems.
- *Planning and scheduling:*
  - Manages the detailed planning of the labor, materials, and all other resources needed to complete a work order.
  - Permits defining and sequencing the specific job steps, identifying the required craft labor types, determining estimated times, as well as all other resources, such as special tools and equipment, contract services, etc.
  - Planned work orders for which all planned resources are available may then be scheduled. This may be a separate module or part of the work order/work-management module.
  - Provides ability to allocate and level craft resources for individual jobs as well as larger projects.
  - Supports building a schedule that includes corrective, preventive, and predictive work in a future period. Allows performing availability check on one or more work orders before issuing a final schedule.
- *Preventive maintenance/predictive maintenance (PM/PdM) module:*
  - Manages work orders for predefined repetitive and rigidly scheduled PM tasks to preserve good equipment condition.

  - Opens and edits work orders and automatically schedules and records when jobs are completed. Jobs done under PM usually include inspections, lubrications, and changes of finite-lifetime items such as filters or seals.
  - An important assignment of PM is discovering needed corrective work, which results in writing the required work order and creating a planned job to make the correction.
  - PM may exist as a separate module or be a subfunction within the work order module.
  - Allows for links to process control or PdM systems to obtain critical readings and determine current conditions that initiate actions such as an inspection work order or a troubleshooting call.
  - Predictive readings are interpreted by software/hardware devices to determine whether a problem exists and what is its nature. If a problem is found, a corrective work order is written for the needed work. Any or all parts of this process may be automated.
- *Inventory management module:*
  - Manages MRO parts and materials inventory with the necessary record keeping of items received, stocked, and dispersed as well as their locations in the stockroom.
  - This module provides the information needed for managing the stock of parts and materials used in performing maintenance work in multiple stockrooms.
  - Provides data showing what items are to be ordered, what are available, and what are on order. Data in this module are directly supportive of the planning and scheduling effort.
  - This module is typically integrated with the purchasing module to fully support MRO procurement.
  - Warranty information and records may be included as a separate module or within this module, coded by vendor with description of warranties and warranty coverage. May include printing of exception reports, warranty claims and follow-up information.
- *Procurement/purchasing module:*
  - This module provides the mechanism to ensure that proper resources of stock material, non-stock items, raw materials, and outside services are available in an adequate manner to ensure adequate planning of maintenance work.
  - Provides for the creation and processing of purchase orders. This module manages the purchasing function, beginning with the automatic creation of a purchase order when the reorder point of a stock item is reached.
  - Data include a list of primary and alternative vendors and prior purchase history information along with vendor performance.
  - Reports available would include current items on order, delivery status, and vendor status and history.
  - May include processes to manage quotations, provide invoice matching, and manage blanket orders and service contracts.
- *Budgeting and cost:*
  - Development and management of cost estimates and budgets for maintenance projects and their distribution to various other operating departments.
  - Data managed in this module would include the details of the maintenance budget and provide current cost experience at cost center, project number, equipment location, system/subsystem level, etc.
  - Reports would be designed and used for control of maintenance expenses, charge-backs to customers, and/or activity codes, variance reporting, etc.
  - This module may be structured for easy linkage and distribution of applicable costs to the general ledger.
- *Project management:*
  - Typically works in conjunction with the work management module for grouping of work orders into specific projects. Where more than one work order is needed to describe a job, grouping as a project is convenient and helps to ensure that all jobs will be done.
  - May offer the ability to time-phase a job and to do sophisticated interactive planning for needed resources such as labor, materials, and production facilities.
  - Provides ability to manage projects from a capital project or a larger-scale construction management point of view.

- *Standard maintenance procedure/job library:*
  - A compilation of standard maintenance repair procedures to be used on work orders for specific jobs. The sequence of job step activities, labor requirements, parts, and materials required are listed, along with special permitting and safety requirements.
  - Serves as a valuable guide to work order planning for larger jobs and is updated whenever a standard repair procedure is changed or created.
  - May include a library of standard times needed to perform specific maintenance tasks. This can be used in planning the estimated time to perform the repair job steps.
- *Personnel:*
  - Provides complete management of in-house crafts personnel as well as contract individuals.
  - Maintains a roster of available employees, their skills, and their training for job-assignment purposes.
  - Maintains a history of where employees have worked for compliance with quality and safety programs.
  - Union agreement restrictions on assignments and past assignment histories, would also be used for scheduling of the work force.
- Data within the personnel module could be shared with, and in some cases updated by, other enterprise systems, such as payroll, labor reporting, safety, and contract management.

These basic modules provide the foundation for an effective maintenance management system. Many other system functions are offered, depending on the CMMS sophistication and intended application. CMMS software packages are now being sold by more than 300 vendors for applications on personal computers, minicomputers, and mainframes.

It is very important that a CMMS/EAM be viewed as an information technology tool for more effective maintenance management and not as a panacea that provides a quick fix. Effective installation of a CMMS/EAM is a maintenance best practice that must be tailored to the specific needs of each unique maintenance operation. Implementing a CMMS/EAM also forces the issue for applying other maintenance best practices.

An effective CMMS/EAM requires that a work order system be in place for control of work requests, monitoring of backlogs, control of maintenance labor, maintaining equipment history, and so on. Full utilization of the basic CMMS/EAM modules requires a sound commitment to establishing best practices for PM/PdM, maintenance planning and scheduling, and maintenance storeroom inventory management.

6.3. Benchmarking the CMMS/EAM Installation

Today's information technology for computerized maintenance management offers the maintenance leader an exceptional tool for managing and controlling the maintenance process as an internal business and profit center. However, maintenance surveys and benchmark evaluations conducted by Tompkins Associates and others validate that poor utilization of existing CMMS/EAM systems is a major improvement opportunity.

6.3.1. The CMMS Benchmarking System

To help improve utilization of this information technology for maintenance, many things were needed, especially a tool for benchmarking the actual CMMS/EAM installation. There was a real need to see how well the systems were working, and being utilized in the real world of maintenance management and control. There was also the requirement to validate benefits expected to reach the projected ROI. These needs, in turn, led to the development of the CMMS Benchmarking System by Tompkins Associates in 1997. It is designed as a methodology for developing a benchmark rating of an existing CMMS (Class A, B, C or D) to determine how well this tool is supporting best practices and the total maintenance process. The system is not designed to evaluate or rate levels of CMMS functionality or to compare vendors of various systems (Peters 1999).

The CMMS Benchmarking System can also be used as the process to benchmark the future success of a CMMS system that is now being installed. It can be very easily tailored to include more advanced CMMS functionalities included with a large, multisite installation. By benchmarking an existing CMMS/EAM installation, it can be determined if the operation is receiving the maximum benefits possible from their investment.

The CMMS Benchmarking System includes a total of nine evaluation categories, along with 50 specific evaluation items. The major categories and evaluation items per category are as follows:

Benchmark Evaluation Category	Evaluation Items
CMMM data integrity	6
CMMS education and training	4
Work control	5
Budget and cost control	5
Planning and scheduling	7
MRO materials management	7
Preventive and predictive maintenance	6
Maintenance performance measurement	4
Other uses of CMMS	6
Total evaluation items	50

6.3.2. The CMMS Benchmarking System Rating Process

Each evaluation item that is rated as being accomplished satisfactorily receives a maximum score of 4 points. If an area is currently being "worked on," a score of 1, 2, or 3 points can be assigned, based on the current level of progress being achieved. For example, if spare parts inventory accuracy is 92%, compared to the target of 95%, a score of 3 points is given. A maximum of 200 points is possible. A CMMS rating of "Class A" is within the 180–200-point range, or 90% + of total possible points. The complete CMMS Benchmarking rating scale is shown below:

CMMS Benchmarking Rating Scale	
Class A	180–200 points (90% +)
Class B	140–179 points (70%–89%)
Class C	100–139 points (50%–69%)
Class D	0–99 points (up to 49%)

6.3.3. Conducting the CMMS Benchmark Evaluation

The CMMS Benchmark Evaluation can be conducted internally by the maintenance leader or via an internal team effort of knowledgeable maintenance people. Other options include using support from consultants as an independent and objective maintenance benchmarking resource. Used by over 4000 operations, the Tompkins Associates *Scoreboard for Maintenance Excellence has* proven to be today's most comprehensive benchmarking tool for evaluating the total maintenance process. In combination with a complete CMMS Benchmark Evaluation and implementing results from *The Scoreboard for Maintenance Excellence,* maintenance leaders in all types of operations can move toward greater value-added service to the customers of maintenance.

Complete copies of the CMMS Benchmarking System, along with *The Scoreboard for Maintenance Excellence,* are available for download from www.tompkinsinc.com or by calling 1-800-789-1257.

7. A REVIEW OF OTHER SELECTED MAINTENANCE BEST PRACTICES

Previous sections reviewed a number of best practices such as performance-measurement techniques, internal benchmarking, selection of metrics, assessment of the total maintenance operation and CMMS/EAM. This section will review eight more best-practice areas. Readers must remember that the scope of plant and facilities engineering is very broad and technical. The material presented in this chapter is the maintenance management piece; how to lead, manage and control the execution of maintenance. The sciences and technologies that underlie the practice of maintenance are found in publications such as the Maintenance Engineering Handbook (Higgins 1995) and in many other handbooks specifically related to mechanical, electrical and reliability engineering. The following best-practice areas can all contribute significantly to bottom-line results.

7.1. Continuous Reliability Improvement (CRI)

Continuous reliability improvement is the application of the traditional reliability centered maintenance (RCM) processes, using the best from total productive maintenance (TPM) and going beyond these traditional approaches to include all maintenance resources: equipment/facility, craftspeople/operators, MRO material management, and maintenance information resources. CRI includes effective team processes to create synergy and serve as people asset multipliers. CRI also includes maximizing the capability of the asset through comprehensive asset facilitation.

CRI, developed by Tompkins Associates, is a total maintenance operations improvement process to support the total operation. The CRI approach to leadership driven teams is defined in Tompkins (1998). CRI focuses team processes on continuous reliability improvement opportunities, considering:

- *Physical asset:* Use of reliability improvement technologies; reliability-centered maintenance, preventive/predictive maintenance, and knowledge-based/expert systems for maintenance of the physical asset. Asset facilitation to gain maximum capacity at the lowest possible life-cycle cost.
- *MRO material resources:* Effective maintenance repair operations (MRO) parts, supplies and materials for quality repair with effective storeroom and procurement processes.
- *Information resources:* Quality information resources for maintenance management and control from CMMS, EAM, ERP, vendor and customer.
- *Craft resources:* Quality skill improvement by people assets for the maintenance process to support customer service to the total operation.
- *Operator resources:* The added value of equipment operators instilled with pride in ownership at the most important level—the shop floor.
- *Synergistic team processes* as multipliers of people assets.

7.1.1. Asset Facilitation

This very comprehensive process is a component of CRI. It is conducted for critical assets and, typically, discrete manufacturing equipment. The objective is to maximize the capability of the asset to perform its primary functions. It takes a complete look at the requirements for maintenance, set-up, operations, changeover and shutdown, and documents these into standard operating and maintenance procedures. The results of a typical asset-facilitation process would include the following:

- Development of a quality operations and maintenance guide for the asset that provides complete operating and maintenance guidelines, procedures, and documentation for:
  - Equipment safety and housekeeping requirements
  - Defining operator and extrusion departmental requirements for the 5S Process
  - Defining requirements for subsequent 5S evaluations and ratings
  - Asset documentation references and drawings, either manual or electronic medium, using electronic document imaging capabilities.
  - Preventive and predictive maintenance requirements
  - Setup/validation/changeover/operator inspection procedures
  - Start-up, operating, and shutdown procedures for the asset
  - Quality control requirements
  - Operator-based maintenance (OBM) tasks with details on:
    - Recommended operator-level PM tasks and lubrication services
    - Other well-defined maintenance tasks to be performed by the operators
    - Operator training and certification requirements; production; related tasks
    - Operator training and certification requirements for maintenance related (OBM) tasks
    - Develop visual management to support OBM and production operation tasks
    - Quality requirements for ISO/QS compliance

7.2. Preventive Maintenance

Preventive maintenance is an interval-based surveillance method in which periodic inspections are performed on equipment to determine the progress of wear in its components and subsystems. When wear has advanced to a degree that warrants correction, maintenance is performed on the equipment to rectify the worn condition. The corrective maintenance work can be performed at the time of the inspection or following the inspection as part of planned maintenance. The decision of when to perform the corrective repair related to PM inspections depends on the length of shutdown required for the repair (Tompkins 1998).

Consideration is given to the impact of the shutdown on the operation caused by the repair vs. how immediate is the need for repair. If it is judged that the worn component will probably continue to operate until a future repair can be scheduled, major repairs are postponed until they can be planned and scheduled.

Periodically, a preventive maintenance inspection is made. If the inspection reveals serious wear, some maintenance operation is performed to restore the component or subassembly to a good state of repair, and reduce the probability of failure. A PM system increases the probability that the equipment will perform as expected without failure until the next inspection due date.

Determining the interval between inspections requires considering the history of maintenance for the equipment in each unique operation. Time between PM inspections (intervals) will ultimately be guided by a number of resources. These include manufacturer's recommendations, feedback information from repair history of breakdowns, and the subjective knowledge of the maintenance craftsmen and supervisors who daily maintain the asset. Equipment operators may also be a good source of information in some operations.

A central characteristic of preventive maintenance is that in most major preventive maintenance applications, the asset must be shut down for inspection. For example, a heat exchanger must be shut down and isolated when a nondestructive eddy current inspection is made on its tubes. The inspection process would require a discrete amount of downtime for the unit, which is typical for preventive maintenance.

The loss of operational time when significant preventive maintenance inspections are made is one of the reasons PM programs are often less than successful. This is especially true in applications where there are few redundant units and equipment must operate at 100% of capacity. In some situations the loss form shutdown is considered too high a penalty and preventive maintenance inspections are resisted.

7.3. Predictive Maintenance (PdM)

In contrast to preventive maintenance, predictive maintenance is a condition-based system. PdM measures some output from the equipment that is related to the degeneration of the component or subsystem. An example might be metal fatigue on the race of a rolling element bearing. The vibration amplitude produced by the rolling element as it passes over the degenerating surface is an indicator of the degree of severity of wear. As deterioration progresses, the amplitude of vibration increases. At some critical value, the vibration analyst concludes that corrective action should be taken if catastrophic failure is to be avoided.

Predictive maintenance usually permits discrete measurements that may be trended compared to some predefined limit (baseline) or tracked using statistical control charting. When an anomaly is observed, warning is provided in sufficient time to analyze the nature of the problem and take corrective action to avoid failure. Thus, predictive maintenance accomplishes the same central objective as preventive maintenance (Higgins 1995).

Confidence in the continued on-specification operation of the asset is increased. By early detection of wear, you can plan for and take corrective action to retard the rate of wear or prevent or minimize the impact of failure. The corrective maintenance work restores the component or subassembly to a good state of repair. Thus, the equipment operates with a greater probability of trouble-free performance.

The enhanced ability to trend and plot numbers collected from PdM measurement gives this method greater sensitivity than traditional preventive maintenance methods. The technique yields earlier warning of severe wear and thus provides greater lead-time for reaction. Corrective actions may be scheduled so that they have minimum impact on operations.

A principal advantage of predictive maintenance is the capability it offers the user to perform inspections while the equipment is operating. In particular, in order to reflect routine operating conditions, the technique requires that measurements be taken when the equipment is normally loaded in its production environment. Since the machine does not need to be removed from the production cycle, there is no shutdown penalty. The ability to conduct machine inspections while equipment is running is especially important in continuous operations such as in utilities, chemical, and petrochemical manufacturing.

Another advantage of predictive maintenance is that the cost of surveillance labor is much less than the cost of preventive maintenance activities. Although the technical knowledge required for predictive maintenance inspections is usually higher than that for preventive maintenance, the inspection time required per machine is much less. With predictive maintenance, the machines do not have to be disassembled for inspection. For example, with vibration analysis, 50 to 60 machines may be inspected in a single day using modern computer data collectors. When comparing cost advantages of predictive maintenance over preventive maintenance, consider production downtime costs, maintenance labor costs, maintenance materials costs, and the cost of holding spare parts in inventory.

If predictive maintenance methods are superior to preventive maintenance, why use preventive maintenance at all? The answer is simple. The nature of your operation will determine which methods are most effective. In actual practice, some combination of preventive and predictive maintenance is required to ensure maximum reliability. The degree of application of each will vary with the type of equipment and the percent of the time these machines are operating.

Pumps, fans, gear reducers, other rotating machines, and machines with large inventories of hydraulic and lubricating oils lend themselves to PdM surveillance methods. On the other hand, machines such as those that might be involved in high-speed packaging may be better inspected using traditional preventive maintenance methods. Machines that have critical timing adjustments, which tend to loosen and require precision adjustments, or that have many cams and linkages that must be reset over time lend themselves to preventive maintenance activities.

The strategy for selecting the appropriate or predictive approach involves the following decision process:

- Consider the variety of problems (defects) that develop in your equipment.
- Use the predictive method if a predictive tool is adequate for detecting the variety of maintenance problems you normally experience. One or a combination of several predictive maintenance methods may be required.
- Use preventive maintenance if it is apparent that predictive maintenance tools do not adequately apply. Inspection tasks must be developed that reveal the defects not adequately covered by preventive maintenance.
- After you have decided the combination of inspection methods, determine the frequency at which the particular inspection tasks must be applied.

Some equipment will be satisfactorily monitored using only predictive maintenance. Other equipment will require preventive maintenance. Ultimately, some combination of methods will provide the required coverage for your operation to assure reliable performance. In most operations, it is wise to apply a combination of methods to ensure that equipment defects do not go undetected. The following provides a brief summary of some of the leading predictive maintenance technologies.

7.3.1. Vibration Analysis

Today, electronic instrumentation is available that goes far beyond the human limitations with which the old time craftsmen had to contend when trying to interpret vibration signals with a screwdriver-handle-to-the-ear method. Today's instruments can detect, with accuracy and repeatability, extremely low-amplitude vibration signals. They can assign a numerical dimension to the amplitude of vibration and can isolate the frequency at which the vibration is occurring. When measurements of both amplitude and frequency are available, diagnostic methods can be used to determine the magnitude of a problem and its probable cause. When you use electronic instruments in an organized and methodical program of vibration analysis, you are able to:

- Detect machine problems long before the onslaught of failure
- Isolate conditions causing accelerated wear
- Make conclusions concerning the nature of defects causing machine problems
- Execute advance planning and scheduling of corrective repair so that catastophic failure may be avoided
- Execute repair at a time that has minimum impact on operations

7.3.2. Shock Pulse

Shock pulse is a method of surveillance that is specific to rolling element bearings. Since rolling element bearings (sometimes referred to as antifriction bearings) are so common in machines, the method has many applications. A secondary but important feature of the shock pulse method is that it permits maintenance workers to judge the adequacy of the lubrication program applied to this type of bearing.

7.3.3. Spectrometric Oil Analysis

Oil in machines carries the products of deterioration resulting from wear and mechanical failure. Analyzing the oil resident in a machine or the debris the oil carries allows predictions to be made about the state of health of the machine. The critical measurement reflecting the condition of machine wear is the number of microscopic metal wear particles that are suspended in the oil system of the machine. The spectrometric oil analysis process is a laboratory technique that uses various instruments to analyze a used oil sample from a machine. The spectrometric result is compared to a baseline level of metal found to be typically suspended in the oil under normal operating conditions. When the wear is meaningful, the sample will show high levels (in parts per million) of wear metals compared to the baseline oil sample.

7.3.4. Standard Oil Analysis

In addition to the spectrometric analysis, the oil laboratories also check the oil using common oil-analysis techniques. For example, the oil is usually checked for viscosity. Other types of standard oil analysis might include total acid number, percent moisture, particle count (for hydraulic systems), total solids, or percent silicon (representing dirt from the atmosphere in the form of silicon dioxide or perhaps just from an additive).

7.3.5. Ferrographic Oil Analysis

Ferrography provides the maintenance manager with two critical sets of decision support information: condition monitoring, which prevents unnecessary maintenance, and precise trend information, which allows maintenance to initiate repairs before equipment failure. This is accomplished by the analysis of wear particles in lubricants to determine their size, distribution, quantities, composition, and morphology (form and structure).

Ferrography is a technique that provides microscopic examination and analysis of particles separated from fluids. Developed in the early 1970s as a predictive maintenance technique, it was initially used to precipitate ferrous wear particles from lubricating oils magnetically. This technique was used successfully to monitor the condition of military aircraft engines. That success has led to the development of other applications, including testing of fluids used in vacuum pumps within the semiconductor industry.

7.3.6. Infrared Thermography

The use of infrared thermography has grown significantly in the past 10 years. Equipment is easier than ever to use and more effective. The real power of thermography is that it allows quick location and monitoring of problems. It then presents critical decision-making information in *visual* form, making it easy for management to understand. Infrared imaging systems, as they are generally called, produce a picture, either black and white or color, of the invisible thermal patterns of a component or process. These thermal patterns, when understood, can be used to monitor actual operating conditions of equipment or processes.

For instance, viewing a thermogram (heat picture) can clearly show the heat of a failing bearing or a pitted contact on a disconnect switch. Today's sophisticated imaging equipment is capable of acquiring a thermal video. This allows us to see *dynamic* thermal patterns of a casting process or a belt wear pattern in real-time, for instance. When teamed with the power of computer-based analysis systems, it can go one step further to compare and trend the critical thermal changes that often precede equipment failure or loss of production quality.

Thermography can be used to quickly locate and prevent recurrence of many equipment and process problems, such as:

- Catastrophic electrical failures
- Unscheduled electrical outages or shutdowns
- Chronic electrical problems in a piece of equipment or process
- Excessive steam usage
- Frozen or plugged product transport lines
- An inability to predict failures accurately
- Inefficient use of downtime maintenance opportunities
- Friction failures in rotating equipment
- Poor product quality due to uneven heating or cooling or moisture content
- A fire in a wall or enclosed space
- Inability to locate or verify a level in a tank
- Replacement of refractory in a boiler, furnace, or kiln
- A leaking flat roof
- Uneven room temperatures affecting product quality or employee productivity
- Trouble locating underground water, steam, or sewer lines

7.3.7. Ultrasonic Detection

A variety of tools using airborne ultrasound technology (commonly called ultrasound) have revolutionized many maintenance programs, allowing inspectors to detect deteriorating components more accurately before they fail. Ultrasound, by definition, is beyond the limits of normal human hearing, so an inspector uses a sophisticated detector to transpose ultrasonic signals to the range of human hearing.

Fluid and gas systems and other working machinery have constant ultrasound patterns. Changes in the ‘‘sonic signatures’’ can be recognized as wear in components. An ultrasonic detector senses such subtle shifts in the signature of a component and pinpoints potential sources of failure before they can cause costly damage.

7.4. Maintenance Storeroom Operations and MRO Materials Management

Maintenance repair operations (MRO) parts, materials, and supplies are the key material resources necessary for the execution of the maintenance process. Often the physical storage, control, and procurement of MRO items is never recognized for its value to the total operation. The effective operation of a maintenance storeroom is a cornerstone for maintenance excellence, but it is often a neglected area for management attention.

Best practices for maintenance storeroom operations parallel those for a finished goods warehouse in some cases but nonetheless are distinctively different. For more information on storage and warehousing, material and information handling systems, and warehouse management, readers should also refer to the respective chapters for these best practices that have general application to the maintenance storeroom operation. The following key strategies should be applied to the maintenance storeroom operation (Newhouse 1999):

- *Professionalism:* The company needs to view the maintenance storeroom as an important activity and not as a necessary evil. Both the dollars invested in maintenance storeroom materials and the impact of downtime have highlighted the need for a more professional approach to maintenance storerooms.
- *Customer awareness:* Successful maintenance storerooms will have a high regard for the customer, will know the customer requirements, and will consistently meet these requirements. The right materials will be available at the right time, in the right quantity, at the right location.
- *Measurement:* The maintenance leader will establish storeroom standards, performance will be measured against these standards, and timely actions will be taken to overcome any problems. Performance reports will be distributed to management on a monthly basis.
- *Operations planning:* Systems and procedures will be put into effect that allow the storeroom manager to plan the storeroom operations proactively as opposed to responding reactively to external circumstances.
- *Materials planning:* Systems and procedures will be put into effect that will assure having the right materials on hand in the right quantity at the right time. The materials-planning systems will provide for good inventory rotation, a minimum of stock outs, the elimination of obsolete materials, and the addition of new items, when new equipment and systems are installed.
- *Centralization:* The trend will be towards larger, centralized storerooms with responsive material-delivery systems instead of smaller, decentralized storerooms to which people travel and where they wait for required materials.
- *Adaptability:* Maintenance storeroom facilities, operations and personnel must become more adaptable. The pace of the storeroom will continue to increase: reduction of lead times, shorter equipment lives, increased inventory turns, more SKUs, and more customer demands require that storeroom adaptability be present to satisfy customers.
- *Uncertainty:* All uncertainty must be minimized; all interactions with the maintenance storeroom must be based on meeting expectations. No surprises.
- *Integration:* The activities within the maintenance storeroom will be integrated (from storeroom item identification to item issue), and the maintenance storeroom will be more integrated with the overall organization.
- *Material control:* Maintenance storeroom facilities, procedures, and systems will be designed to provide for the control of all materials. The importance and enforcement of maintenance storeroom security will be widely understood and accepted.
- *Maintenance information system:* The maintenance information system will support exceeding customer expectations. The information system will include the maintenance catalog system and the management of inventory.
- *Inventory accuracy:* Cycle counting will be used to manage inventory accuracy, and accuracy above 98% will be the norm.
- *Space utilization:* Space will be more efficiently and effectively utilized.
- *Housekeeping:* Quality housekeeping will be a priority. There will be an acceptance of that fact that there is efficiency in order.
- *Human resources:* A priority in the maintenance storeroom will be establishing a positive culture and the training and education required to achieve quality maintenance.

- *Team players:* Everyone associated with the maintenance storeroom and MRO procurement functions within the organization must be integrated into a single service-providing activity.

7.4.1. Storeroom Inventory Management

It is also important to establish an inventory planning methodology. As part of this, a determination must be made of what to stock, what the inventory policy will be, and how the inventory will be managed. A bar coding system could be just the solution for improving accountability and accuracy. This would be just one part of a strategic plan to introduce many new technologies for a major upgrade of the maintenance storeroom.

Inventory accuracy is a must for a maintenance storeroom. It must have 98% or better accuracy. If not, the craftspeople will bypass the storeroom to order a new part. It is critical that they have confidence in the accuracy. In order to attain this level of accuracy, it is necessary to cycle count. Timely detection of errors and correction of causes for the errors are essential to good control. All storerooms must have a proper cataloging system as a permanent record of all storeroom items and as a tool for identifying and locating items.

A maintenance storeroom strategic master plan is a prerequisite for success. There is efficiency in order. We must know what is to be done and what is the proper order to do it. It all starts with a plan, and successful planning requires teamwork.

7.5. Planning and Scheduling

Planning for maintenance excellence requires planning at the strategic level and the shop-floor level. This section introduces the need for implementing the maintenance best practice for planning and scheduling at the operational, shop-floor level. Surveys consistently show that only about 30–40% of an eight-hour craft day is devoted to actual hands-on wrench time. Without effective planning and scheduling, maintenance operations continue to operate in a reactive, fire-fighting mode, wasting their most valuable resource: craft time. Gambling with maintenance costs is not an option for today's organization that wants long-term survival and profitability. Achieving world-class status requires a world-class maintenance operation. World-class maintenance requires strategic planning, especially at the shop-floor level. The effective planning and scheduling of the most valuable maintenance resource, craft skills and labor, can provide an important step forward in a strategic maintenance plan.

Effective planning and scheduling improves the overall craft effectiveness (OCE) factor, which focuses upon measuring and improving the value-added contribution that people assets make to total asset management. There is also a very real concern within many areas of the United States and the world about the availability of craft skills. Technical resources and craft skills are terrible things to waste because they are so hard to find and keep. A review of the three key elements for measuring OCE shows how they very closely align with the three elements for determining the OEE factor for equipment assets.

7.5.1. The Overall Craft Effectiveness (OCE) Factor

Measuring and improving overall craft effectiveness (OCE) is one of the key benefits from maintenance planning and scheduling. The OCE factor includes three key elements very closely related to the three elements of the OEE factor.

Overall Craft Effectiveness (OCE)	Overall Equipment Effectiveness (OEE)
Craft utilization (CU)	Availability / utilization
Craft performance (CP)	Performance
Craft methods and quality (CM&Q)	Quality

7.5.1.1. Craft Utilization The first element of the OCE Factor is craft utilization, which measures how *effective* we are in planning and scheduling craft resources so that these assets are doing value-added, productive work. Craft utilization is about wrench time. Effective planning / scheduling within a proactive maintenance process is key to increased wrench time and craft utilization. It's having the right part at the right place in time to do scheduled work with minimal nonproductive time on the part of the craftsperson or crew assigned to the job. Craft utilization is expressed simply as the ratio of:

$$\frac{\text{total productive (wrench time)}}{\text{CU\%} = \text{total craft hours available/paid} \times 100}$$

Improving craft utilization provides additional craft capacity in terms of total productive craft hours available It is gained value and additional equivalent hours that can be used to reduce overtime, devote to PM/PdM, reduce the current backlog, and attack deferred maintenance, which doesn't go away.

Even if an operation does nothing to improve the other two elements of OCE (craft performance and the craft methods and quality level), significant tangible benefits can be realized with improving wrench time and craft utilization. An improvement of from 20–30% in craft utilization can typically be expected from effective maintenance planning and scheduling.

For the 30-person craft workforce, operating at 40% craft utilization, a 10% gain in wrench time hours represents a 25% increase in craft labor capacity from baseline performance. With only a 10% increase in craft utilization for a 30-person craft workforce, more than a 5:1 return can be achieved to offset a maintenance planner position, and implementation of planning and scheduling.

7.5.1.2. Craft Performance The second key element affecting overall craft effectiveness is craft performance. This element relates to how *efficient* we are in actually doing hands-on craftwork when compared to an established planned time or performance standard. Craft performance is directly related to the level of individual craft skills and overall trades experience, as well as the personal effort of each craftsperson or crew. Effective craft skills training and technical development contribute to a high level of craft performance. Craft performance (CP) is expressed as the ratio of:

$$\frac{\text{total planned time (hours)}}{\text{CP\%} = \text{total actual craft hours required} \times 100}$$

An effective planning and scheduling function requires that reasonable estimates and planning times be established for as much maintenance work as possible. Since maintenance work is not highly repetitive, the task of developing planning times is more difficult. However, there are a number of methods for establishing planning times for maintenance work, as outlined in Section 5.

Planning times are essential. They provide a number of key benefits for the planning function. First, they provide a means to determine existing workloads for scheduling by craft areas and backlog of work in each area. Planning times allow the maintenance planner to balance repair priorities against available craft hours, and to establish repair schedules realistically that can be accomplished as promised. Secondly, planning times provide a target or goal for each planned job, that allows for measurement of craft performance. Here we are not as concerned with measuring individual craft performance but rather with the overall performance of the craft workforce as a whole. Individual craft performance can be determined by comparing the group performance to individual performance over a period of time. Training needs are normally identified when individual craft performance is consistently below the group norm.

7.5.1.3. Craft Methods and Quality Level The third element affecting overall craft effectiveness relates to the relative level of the methods being used, considering personal and shop tools, special shop equipment, shop work areas, repair methods, and so on, as compared to current state-of-the-art methods. This element can include call-backs, where the poor quality of the initial repair requires another trip to fix it right the second time. Typically, the CM&Q element is a more subjective value, and is not determined based on actual data such as craft utilization and performance. However, the craft methods and quality level do affect overall craft labor productivity. The overall craft effectiveness factor is determined by multiplying each of these three elements:

$$\begin{array}{lccccc}
\text{OCE} = & \text{CU\%} & \times & \text{CP\%} & \times & \text{CM\&Q\%} \\
 & \text{craft} & \times & \text{craft} & \times & \text{craft methods and quality} \\
 & \text{utilization} & & \text{performance} & & \text{level}
\end{array}$$

Due to the subjective nature of determining the value of craft methods and quality, this element is typically not used, but it is still an important part of effective planning and scheduling. One key part of planning is determining the scope of the repair job and the special tools or equipment that might be required. A continuing concern of the maintenance planning function should be on improving existing repair methods, whether by using better tools, repair procedures, or diagnostic equipment. Providing the best possible tools, special equipment, shop areas, and repair procedures is a key contributor to improving craft performance and craft morale. It should be recognized that the value for craft methods and quality is subjective. The overall craft effectiveness Factor is best determined by:

$$\text{OCE} = \text{craft utilization} \times \text{craft performance}$$

7.5.2. Getting Started with Planning and Scheduling

Select from within the maintenance workforce the best-qualified candidate possible. Normally 1 planner per 25–30 craft personnel is sufficient. The planner position requires knowledge of existing equipment repair needs and a strong craft skill background. It also requires new skills for planning/scheduling, estimating, parts coordination, computer use, personal relations, customer service, and so on. Sufficient time must be invested in formal training of the planner(s), and for on-the-job training to get the planning function off to a smooth start. Since the planning function is often the focal point for successful CMMS utilization, planners must understand all functions of the CMMS and be capable of helping to train others in the organization.

7.5.3. Focus on Customer Service

The ultimate success of maintenance planning and scheduling will be determined by whether or not the customer is satisfied. All preliminary work to develop a plan and coordinate the scheduled repairs is wasted if execution of the schedule does not occur as promised. The customer (operations) will determine the true success of the planning process. The entire maintenance workforce must understand their service role to operations. As a formal planning process is implemented, an increased focus on customer service must be established. Operations will expect better service and maintenance must commit to providing it.

7.5.4. Measure Effectiveness of the Planning Function

The planning function should provide the focal point for measuring overall maintenance performance. However, the measurement process should start within the planning function. Planning requires an investment in staff resources. Develop and use performance measures to evaluate the return on investment and the effectiveness of the planning function.

7.6. Reliability-Centered Maintenance

7.6.1. The Evolution of Reliability-Centered Maintenance

In the early 1960s, the developmental work for reliability-centered maintenance was done by the North American civil aviation industry. During that period, the airlines began to see that many of their maintenance philosophies were not cost effective. But most importantly, they did not achieve the best possible conditions for safety. The airline industry then put together a series of Maintenance Steering Groups (MSGs) to reexamine all aspects of their aircraft maintenance operation. Representatives from the aircraft manufacturers, the airlines, and the Federal Aviation Authority were members of the MSG team. The Air Transport Association completed the first attempt at formulating maintenance strategies in 1968. Known as MSG 1, this document was later refined to MSG 2 in 1970.

In the mid-1970s, the U.S. Department of Defense commissioned a report by Stanley Nowlan and Howard Heap of United Airlines (Nowlan and Heap 1978) to define state-of-the-art strategies for maintenance within the airlines industry. It is still one of the most important documents in the history of physical asset management.

A considerable advance in thinking, well beyond MSG 2, was achieved in Nowlan and Heap's report. It was used as a basis for MSG 3 in 1980, which was revised in 1998 (Rev 1) and in 1993 (Rev 2). It is still used to develop prior-to-service maintenance programs for new aircraft types such as the Boeing 777 and the Airbus 330/340.

Nowlan and Heap's report and MSG 3 have since been used as a basis for various military RCM standards and for nonaviation-derivative programs with acronyms such as FMECA, MSG3, TPM, RCA, RBI, and RCM2. Continuous reliability improvement (CRI), developed by Tompkins Associates, uses the best from RCM, TPM and the total operations approach to leadership-driven teams, as defined in Tompkins (1998).

7.6.2. Overview of the RCM Process

The key elements of the RCM process include the following:

- Analysis and decision on what must be done to ensure that any physical asset, system, or process continues to do whatever its users want it to do. Includes essential information gathering.
- Define what users expect from their assets in terms of primary performance parameters such as output, throughput, speed, range, and carrying capacity.

- As applicable, the RCM 2 process defines what users want in terms of risks, process and operational safety, environmental integrity, quality of the output, control, comfort, economy of operation, customer service, etc.
- Identify ways in which the system can fail to live up to these expectations (failed states) and failure consequences.
- Conduct failure modes and effects analysis (FMEA) to identify all the events that are reasonably likely to cause each failed state.
- Identify a suitable failure management policy for dealing with each failure mode in the light of its consequences and technical characteristics. Failure management policy options include:
  - Predictive maintenance
  - Preventive maintenance
  - Failure finding
  - Changing the design or configuration of the system
  - Changing the way the system is operated
  - Run-to-failure (if preventive tasks not found)

7.7. Total Productive Maintenance

Total productive maintenance (TPM) is a maintenance-improvement program concept and philosophy that resembles total quality management (TQM). The TPM movement has excellent objectives:

- Zero unplanned downtime
- Zero defects
- Zero machine capacity losses
- Zero accidents
- Minimum life-cycle asset care cost

It requires a total commitment to the program by upper-level management and employees to be empowered to initiate corrective action. It is a long-term strategy, so a long-range outlook must be accepted. TPM may take more than a year to implement and is an ongoing process. Changes in employee mindset toward their job responsibilities must take place as well. TPM brings maintenance into focus as a necessary and vitally important part of the business. It is no longer regarded as a nonprofit activity.

To successfully apply TPM concepts to plant maintenance activities, the entire workforce must first be convinced that upper-level management is committed to the program. Typically, a TPM coordinator is hired or recruited to sell the TPM concepts to the workforce through an extensive training program. To do a thorough job of educating and convincing the workforce that TPM is just not another "program of the month" will take time, perhaps a year or more. Once the coordinator is convinced that the workforce is sold on the TPM program and that they understand it and its implications, the first study and action teams are formed. These teams are usually made up of people who directly have an impact on the problem being addressed. Operators, maintenance personnel, shift supervisors, schedulers, and upper management might all be included on a team. Each person becomes a stakeholder in the process and is encouraged to do his or her best to contribute to the success of the team effort. Usually, the TPM coordinator heads the teams until others become familiar with the process and natural team leaders emerge. The action teams are charged with the responsibility of pinpointing problem areas, detailing a course of corrective action, and initiating the corrective process. Recognizing problems and initiating solutions may not come easily for some team members. They will not have had experiences in other plants where they had opportunities to see how things could be done differently. In well-run TPM programs, team members often visit cooperating plants to observe and compare TPM methods and techniques and to observe work in progress. Publicity of the program and its results is one of the secrets of making the program a success. The initial stages of TPM will include taking the machine out of service for cleaning, painting, adjustment, and replacement of worn parts, belts, hoses, and so on. As a part of this process, training in operation and maintenance of the machine will be reviewed. A daily checklist of maintenance duties to be performed by the operator will be developed. Factory representatives may be called in to assist in some phases of the process. After success has been demonstrated on one machine and records begin to show how much the process had improved production, another machine is selected, then another, until the entire production area had been brought into a world-class condition and is producing at a significantly higher rate. This is one of the basic innovations of TPM. The attitude of "I just operate it!" is no longer acceptable. Routine daily maintenance checks, minor adjustments, lubrication, and minor parts change become the responsibility of the operator. Extensive overhauls and major breakdowns are handled by plant maintenance personnel, with the operator assisting in some cases.

7.8. Operator-Based Maintenance

Pride in ownership and the process of operator-based maintenance started early in American history. It is also a key element of TPM called autonomous maintenance. Operator-based maintenance is also a key element in the very successful and effective maintenance program of the U.S. armed services.

Operator-based maintenance began with the entrepreneurial spirit of the cottage industry in the U.S. free enterprise system. In most cases, the owners fixed the sawmills, the cotton gins, the printing presses, and the wagon wheels. Owners were the skilled crafts people working within a free enterprise culture and system. For many Americans, OBM started with Henry Ford's Model T, which came with tools—a crescent wrench, slip-joint pliers, screwdriver, and hammer, and so on—for the do-it-yourself repairs.

Early American culture did not normally employ the philosophy of ''I own—you fix—you operate.'' Back then it was an attitude of ''We are all responsible for the equipment.'' However simplistic it may have been prior to the industrial revolution, it was a matter of necessity and survival, just as it is in a combat zone.

Ironically, we have returned to the world-class attitude toward maintenance in many organizations, that being ''We are all responsible for our equipment.'' The key word in the TPM world-class attitude is ''our'': *our* equipment, not *the* equipment. These objectives can and will be achieved in varying degrees for organizations that include OBM as part of the total operations/maintenance strategy. The following is a basic strategy for developing successful operator-based maintenance:

- Start with an overall strategic maintenance plan, and include defined goals/objectives for OBM within this plan.
- Understand that OBM is a deliberate process for gaining commitment by operators toward:
  - Keeping equipment clean and properly lubricated
  - Keeping fasteners tightened
  - Detecting symptoms of deterioration
  - Providing early warning of catastrophic failures
  - Making minor repairs and being trained to do them
  - Assisting maintenance in making selected repairs
- Start with the first things first. Provide the necessary communication between maintenance operations and the rest of the total operation to gain the commitment and internal cooperation needed to start OBM.
- Clean the equipment to like-new condition, make minor repairs, and develop a list of major repairs for the future.
- Utilize leadership-driven, self-managed teams with whatever team names that evolve. For example, ''equipment improvement team,'' ''SWAT Team,'' or ''continuous reliability improvement team.''
- Develop a written and specific team charter.
- Avoid use of self-directed teams with no technical leadership ''driver'' for the process.
- Have teams evaluate/determine the best methods for operator cleaning, lubrication, inspection, minor repairs, and level of support during major maintenance repairs.
- Develop standard written OBM procedures for operators and include them in the quality operations and maintenance guide.
- Evaluate your current predictive and preventive maintenance procedures and include those that the operator can do as part of OBM.
- Document start-up/operating/shut-down procedures along with set-up/changeover practices.
- Consider quality control and safety requirements.
- Document operator training requirements and what maintenance must do for support.
- Develop operator certification to validate operator-performed tasks.

There are many organizational roadblocks to effective OBM. However, the roadblocks to the world-class attitude ''We are all responsible for our equipment'' exist only by self-imposed limitations we create by our attitude toward maintenance.

8. MAINTENANCE MANAGEMENT FOR THE NEW MILLENNIUM

A number of positive trends will occur within maintenance operations of the future. Just as the last 10 years of the 20th century saw extraordinary technological advances related to the physical asset, the use of CMMS/EAM, and the use of the computer and the Internet, the new millennium will see extraordinary advances. There will be a positive revolution within maintenance management, based upon four key principles from Tompkins (1998):

1. "Revolution is never a spur-of-the-moment decision. It is a process, and in a true scenario, a continuous process." Continuous reliability improvement and leadership-driven, self-managed teams will enhance the application of emerging new technologies.
2. "Although revolution is a grassroots effort, it is characterized by its leaders." Maintenance leaders will emerge to support the continuous improvement process at the grass-roots, shop-floor level. There will be a progression from maintenance manager to true maintenance leadership.
3. "A revolution cannot be managed, it must be led." An individual who is profit-centered yet understanding of the fact that people assets are the most important assets will lead new millennium maintenance. The evolution of a chief maintenance officer (CMO) will occur in both large and small operations. Successful maintenance leaders will be profit centered, establish strategic maintenance plans that are integrated with the business plan, and validate return on investment with effective measurement processes.
4. "A Revolution cannot be carried out by individuals, it must be a collaborative effort." The maintenance leader of the new millennium will recognize needs of the customer, serve the customer, and bring together all maintenance resources to maximize the value of maintenance. The synergistic effect of teams and collaborative interactions among individuals will be an additional resource and measurable gain in productivity.

8.1. The Emergence of the Chief Maintenance Officer (CMO)

The chief maintenance officer (CMO) will emerge as a recognized corporate leadership position. The CMO (or an equivalent) will be absolutely essential for long-term profitability in the new millennium and will eventually evolve into a recognized corporate position. The CEO/COO of a multisite operation that does not have a CMO accountable for physical asset management will be gambling with stockholder's equity. The small single-site operation without a CMO equivalent will realize the high cost of bad maintenance. The CEO must understand the state of maintenance in the operation and the physical asset management process. The emerging CMO with profit ability and effective leadership and technical skills will facilitate this process in larger multisite operations. The future capable company will require a proactive, capable CMO just as it needs a CFO, CIO and CEO.

8.2. Growth of Reliability-Improvement Technologies

New reliability-improvement technologies will continue to evolve and greater use of existing tools will occur. Greater use of radio frequency (RF) technology to support condition-based monitoring will occur. RF technology will enable mobile, timely, and vital communications with the mobile workforce of craftspeople.

Process-control systems will become more integrated with condition-based monitoring systems that in turn link to CMMS/EAM systems or even back to the condition-based equipment suppliers for real-time troubleshooting. Condition-based monitoring systems will eventually link data collection directly back to suppliers such as CSI, ENTEK, SKF, and others via the Internet or RF. Reliability data analysis by off-site reliability experts providing contract services will be as close as e-mail.

Information technology for the life-cycle information loop will be available. for critical assets from the internet, Intranets, and real-time data collection. The result will be "real information" on the shop floor. Information will be available to take life-cycle costing, equipment design/redesign, reliability improvement, and execution of maintenance to new levels. The craftsperson doing the repair will be in the life-cycle information loop with the OEM, subsystem and MRO providers, the asset designers, local engineering, the local asset-documentation sources, and the customers of maintenance.

A new generation of handheld computers, personal digital assistants (PDAs), and smart phones give the craftspeople their own terminal and will link them firmly to the network. Advances in voice recognition will make data collection on the shop much easier. All the necessary components are falling into place: price, performance, and communications. Since these devices are Internet capable, they will be much easier to integrate into a Web-based EAMS/CMMS package than earlier generations of handheld computers that required proprietary software.

8.3. The Role of the Internet

The use of the Internet will expand exponentially. Client/server solutions that became main stream in the 1990s have their drawbacks. From an IT standpoint, their "thick" client software is costly to deploy and support. They are difficult to integrate with other external applications. Their client software also tends to run on only one type of device, a desktop PC. Using development tools like

Java, software vendors are starting to provide solutions featuring "thin" client software or web browsers as their front end. These solutions draw their application services from centralized servers instead of megabytes of software installed on the desktop. Since they utilize the Web as a transport medium, they are much easier to integrate with the outside world. They can also support nondesktop PC devices such as handheld PDAs.

The top-tier EAMS/CMMS vendors will complete the process of developing new versions of their software, featuring a Web-centric architecture. Most of the industry will follow their lead in the coming years. They will do so because of the benefits that the technology has to offer.

8.4. MRO Materials Management

E-commerce will continue to expand, and direct links to the overall supply chain will enhance procurement of MRO parts/materials and services. E-procurement over the Internet with electronic purchasing is just too cost effective for both buyers and suppliers for it not to become the dominant MRO procurement method.

An integrated MRO supply chain management process began with MRO vendor websites as the start of this process. The creation of large trading networks of hundreds of vendors and the provision of EAMS/CMMS packages access to these networks will complete the equation. True point-and-click MRO purchasing, from requisition generation to order fulfillment, is going to be commonplace in both large and small organizations. Top-tier EAMS/CMMS vendors are now releasing e-procurement solutions and developing or joining the trading networks (Singer 2000).

8.5. The Growth of Contract Maintenance

Contract maintenance will continue to grow. The core requirement for maintenance will be even more important in the next millennium due to technology advances. Organizations will focus on core competencies. Some forget the core requirement for maintenance and also lose their core competencies to do effective maintenance. Profit-centered contract maintenance providers will consume internal maintenance operations that continue a cost-centered approach. In-house maintenance operations will continue to lose when they are unable to replenish and/or maintain their core competencies in maintenance.

Maintenance is forever. Maintenance of our bodies, minds, souls, cars, and computers and all physical assets providing products or services in today's global economy will always be required. Some organizations today have neglected maintaining their core competencies in maintenance to the point that they have lost complete control. The core requirement for good maintenance remains (forever), but the core competency to do good maintenance may be missing. In some cases, the best, and often only, solution is value-added outsourcing.

The neglect of the past and the future will be overcome by services from a growing number of profit-centered maintenance providers that clearly understand how to provide value-added maintenance service at a profit. Neglect of the past can also be overcome internally by the emergence of an internal CMO who can lead maintenance forward to profitability as if he or she owned the internal maintenance business.

9. CONCLUSION

This chapter helps to provide a firm understanding of the physical asset management and maintenance process and its important role. The contribution of maintenance to total operations success and profitability is now being more fully recognized. Effective maintenance and physical asset management are closely linked to enterprise wide performance success and profitability.

This chapter presented the 25 key requirements for maintenance success, as well as a review some of today's best maintenance practices. It outlined how the results of continuous maintenance improvement can be measured, reviewed methods for measurement, and showed how results and ROI can be validated.

There are many IE principles and practices that can support maintenance process improvement. The new and experienced IE, the engineering manager, the operations manager, and the CEO can now better understand the maintenance operation and how to improve mission-essential maintenance operations. The next step is to take action on the journey to maintenance excellence.

REFERENCES

Dunn, R. L. (1997), "Plant/Facilities Engineering—Definitions and Descriptions of Functions and Responsibilities," *AFE Facilities Engineering Journal,* December 1997.

Higgins, L. R. (1995), *Maintenance Engineering Handbook,* 5th Ed., McGraw-Hill, New York.

Newhouse, R. (1999), *The Last Great Goldmine: How to Get Control of Your Maintenance Storeroom,* Tompkins Associates, Raleigh, NC.

Nowlan, F. S., and Heap, H. F. (1978), *Reliability-Centered Maintenance,* Office of Assistant Secretary of Defense, Washington, DC.

Peters, R. W. (1994a), *Achieving Real Maintenance Return on Investment,* Tompkins Associates, Raleigh, NC.

Peters, R. W. (1994b), *The Scoreboard for Maintenance Excellence,* Tompkins Associates, Raleigh, NC.

Peters, R. W. (1996b), *Planning for Maintenance Excellence,* Tompkins Associates, Raleigh, NC.

Peters, R. W. (1996a), ‘‘The ACE Team Benchmarking System,’’ Tompkins Associates, Raleigh, NC.

Peters, R. W. (1998), ‘‘Benchmarking Your CMMS System,’’ *Industrial Maintenance and Plant Operations,* May 1998.

Peters, R. W. (1999), *The Guide to Computerized Maintenance Management Systems,* Alexander Communications, New York.

Singer, T. (2000), ‘‘Technologies for a New Millennium,’’ *Industrial Maintenance and Plant Operations,* February 2000.

Tompkins, J. (1999), *Revolution: Take-Charge Strategies for Business Success,* Tompkins Press, Raleigh, NC.

Tompkins, J. A., and Smith, J. D., Eds. (1998), *The Warehouse Management Handbook,* 2nd Ed., Tompkins Press, Raleigh, NC.

IV.E

Planning and Control

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 60

Queueing Models of Manufacturing and Service Systems

JOHN A. BUZACOTT
York University

J. GEORGE SHANTHIKUMAR
University of California at Berkeley

1. INTRODUCTION 1628
  1.1. Models 1629
    1.1.1. Basic Approach to Modeling 1629
    1.1.2. Types of Models 1630
    1.1.3. Why Model? 1630
    1.1.4. Requirements of Models 1630
  1.2. Queueing Models 1631
    1.2.1. Using Queueing Models 1632
  1.3. Modeling Manufacturing and Service Systems 1632
    1.3.1. Manufacturing Systems 1632
    1.3.2. Service Systems 1633
    1.3.3. Supply Chains and Logistic Systems 1634
  1.4. Description of Available Models 1634
    1.4.1. Assumptions of Models 1634
2. SINGLE STAGE SYSTEMS 1635
  2.1. Make-to-Order Manufacturing or Service with No Task Done Prior to Customer Arrival 1635
    2.1.1. Single-Server System 1635
    2.1.2. Multiple Servers 1635
  2.2. Make-to-Stock Manufacturing Systems or Service Systems Where Work Is Done in Advance of Customer Arrival 1636
    2.2.1. Exponential Service Time, Poisson Demands 1636
    2.2.2. Single Machine with Interruptible Demand (Stopped Arrival Queue) 1637
    2.2.3. Multiclass Backlogged Demand 1637
    2.2.4. Multiclass Lost Sales 1637
    2.2.5. Produce to Stock with Advance Information 1638
3. FLOW LINES AND SERIES SYSTEMS 1638
  3.1. Introduction 1638
  3.2. Models of Paced Systems 1638
  3.3. Models of Unpaced Lines 1639
    3.3.1. Infinite Buffer Systems 1639
  3.4. Two-Stage Flow Lines 1639
  3.5. Exponential Service Times 1639
  3.6. General Service Times 1640
  3.7. Three-Stage Flow Lines 1640
  3.8. Multiple-Stage Flow Lines with Exponential Processing Times 1642
    3.8.1. Algorithm 1: Work-in-Process 1642
    3.8.2. Algorithm 2: Throughput 1643
  3.9. General Service Time Approximation 1643
    3.9.1. Algorithm 3: Throughput 1644
    3.9.2. Squared Coefficient of Variation Recursions 1644
4. TRANSFER LINES 1645
  4.1. Models 1645
    4.1.1. Transfer Lines with No Inventory Banks 1645
    4.1.2. Time-Dependent Failures 1645
    4.1.3. Operation-Dependent Failures 1646

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

4.1.4. Systems Separated by Infinite Inventory Banks 1646
4.1.5. Two-Stage Synchronized Line with Finite Capacity Inventory Banks 1646
4.1.6. Operation-Dependent and Time-Dependent Failures 1647
4.2. Multiple Stage Transfer Lines 1648
4.2.1. Approximation 1648
4.2.2. Algorithm 4: Multistage Transfer Line 1649
5. DYNAMIC JOB SHOPS 1650
5.1. Open Jackson Queueing Network Model 1650
5.2. Multiple-Job-Class Open Jackson Queueing Network Model 1652
5.2.1. Incorporating Transport and Material Handling in the Jackson-Type Job Shop Model 1654
5.3. General Service Times 1654
5.3.1. Approximations for $\hat{n}_i$ and $C^2_{d_i}$ 1655
6. FLEXIBLE MACHINING SYSTEMS 1656
6.1. Single-Class Closed Jackson Queueing Network Model 1656
6.1.1. Algorithm 5: Convolution Algorithm 1657
6.1.2. Algorithm 6: Marginal Distribution Analysis Algorithm 1659
6.1.3. Algorithm 7: Mean Value Analysis Algorithm 1659
6.1.4. Properties of the Throughput 1660
6.2. Modeling the Effects of Dedicated Material-Handling Systems 1660
6.2.1 Algorithm 8: MVA with Material Handling Systems 1660
6.3. General Single-Class Closed Queueing Network Model 1660
6.3.1. Algorithm 9: Extended Mean Value Analysis (EMVA) 1660
6.4. Multiple-Class Model 1661
6.4.1. Algorithm 10: Multiclass MVA 1661
6.4.2. Properties of the Throughput Rate 1661
6.4.3. General Service Time Distributions 1661
6.4.4. Algorithm 11: Extended Multiclass MVA 1662
7. PRODUCTION COORDINATION 1662
7.1. Base Stock Control 1663
7.1.1. Cell Model 1663
7.1.2. Single Server in Each Cell 1664
7.2. Kanban Control 1664
7.2.1. Store Model 1665
7.2.2. Cell Model 1666
7.2.3. Connection between Store Model and Cell Model 1666
7.2.4. Performance Measures 1667
8. CONCLUSIONS 1668
REFERENCES 1668

1. INTRODUCTION

The design and improvement of the performance of manufacturing and service systems requires that we have efficient ways by which we can (1) predict the performance of the systems and (2) identify the effects of key design parameters on the system performance. Manufacturing and service systems have to cope with a wide range of variability, uncertainty, and disturbances. Different customers require different tasks to be performed, people and machines can vary in their time to perform standardized tasks, machines can break down unexpectedly, repair can prove more complicated than anticipated. So we need approaches to predicting the performance that take into account this uncertainty and variability and also help us reduce their adverse impacts. Queueing models are particularly useful in describing variability and predicting its impact on performance. This chapter contains an overview of the key models that are relevant to analysis of manufacturing and service systems. The focus of these models is on predicting throughput, inventory levels, queue lengths, and service levels after allowing for disturbances such as machine breakdowns, human operator performance variability, and quality problems.

Queueing models can be used at the system-design stage to rapidly explore alternatives and see the sensitivity to parameter values. The models can also be of great value in assessing the performance of systems once they are installed because they enable the sources of loss of productivity to be identified. Models provide understanding and insights of complex production systems and enable one to obtain answers to a variety of "what-if" questions with little effort.

Our focus in this chapter will be on manufacturing and service systems where each job or customer is distinct. Most service systems have to deal with the requirements of individual customers. In manufacturing, each job is distinct in the mechanical, electrical, and electronics industries making products such as cars, refrigerators, electric generators, and computers. Systems that process fluids, like those found in the chemical and metallurgical industries, will not be considered, although sometimes in these industries fluids are processed in distinct batches or packets and, if such a packet is taken as the unit of manufacture, then the system can be considered to process discrete jobs. For simplicity, we will call an item, part, subassembly, or assembly processed by a machine or workstation in manufacturing a *job.* In service applications, we will call each order or person a *customer.*

While each job or customer is distinct, different jobs or customers can be in all respects identical and in particular have the same processing requirements. If the system only processes one type of job or customer, then some aspects of its design and operation will be simplified because all jobs or customers can then be handled the same way. However, if the system processes many different types of jobs or customers, instructions for each type will be required and control will tend to be more complex. Particularly in service systems, it is often not known what the processing requirements of a customer will be until after the customer arrives and some diagnosis can be carried out. Even in manufacturing, quality problems can result in the processing requirements of a job changing after processing has begun. Models have to be able to represent this evolution of knowledge about processing requirements and how the information is used to modify instructions on what has to be done.

1.1. Models

Traditionally, manufacturing and service system designers relied on experience and rules of thumb in order to identify the effect of design parameters on the performance. However, the increasing cost and complexity of modern systems, and the often lengthy time required to bring them up to their designed performance targets, have resulted in designers using formal models of the system to assess performance.

1.1.1. Basic Approach to Modeling

The process of modeling involves the following steps:

1. *Identifying the issues to be addressed:* Ascertain the needs of the user. What decisions is the model required to support? These can range from the very specific, such as how large a specific storage location should be, to the somewhat vague, such as whether it is possible to identify when a particular way of operating the system is optimal.
2. *Learning about the system:* Identify the components of the system, such as the people, machines, material handling, storage and the data collection and control system. Determine the characteristics of the jobs or customers and the target volumes, quality, and cost. This step can involve close contact with the system designers and a review of any existing models to identify their capability and their shortcomings.
3. *Choosing a modeling approach:* Various types of modeling approaches can be used, ranging from formal mathematical models through computer simulations to the development of a "toy" system in which toy parts or people physically move from one toy machine to another. The choice of modeling approach is determined by the time and cost budget for model development and the anticipated way in which the model will be used.
4. *Developing and testing the model:* This step requires obtaining data on the parameters of the model, and often the lack of desirable data forces the model to be substantially simplified.
5. *Developing a model interface for the user:* If the model is to be of value in making decisions, it has to be provided with some interface so that it can be used by managers. This requires the modeler either to embody the model in a decision support system or to present the model and its implications in a way that managers can understand.
6. *Verifying and validating the model:* The model has to be checked to see that it is a reasonably correct representation of the reality it seeks to represent. Verification is the process of ensuring that the model results are correct for the assumptions made in developing the model. Validation is the process of ensuring that the model is an accurate representation of the real system. This may also involve convincing the user that the model is adequate for the decision he or she requires it to support.
7. *Experimenting with the model:* This requires exploring the impact of changes in model parameters and developing understanding of the factors influencing performance of the system so that the manager can be confident in the decisions made using the model.
8. *Presenting the results:* Using the model, the manager should come up with a recommended course of action. This recommendation may have to be presented to higher-level management

and the role of the model in aiding the decision explained. Alternatively, the model and the results of its use will be presented in a report or paper that, apart from describing the model itself, should explain what the model can do, what it cannot do, and how accurate are its predictions.

1.1.2. Types of Models

For systems processing discrete jobs or customers, there are three types of models in common use:

Physical models represent the real system by another physical system, in which jobs or customers move from one machine or service center to another and the machines or service centers perform processing operations on the jobs or customers. The major difference to the real system is that the model uses a different dimensional scale, so a large system will occupy a table top. Physical models can use toy-sized components, but they can be provided with a control system that employs the same logic as the real system. Physical models are excellent as a means of educating management and workers about the control of the system, but they do not lend themselves for assessing the long-run behavior of the system, as it is difficult to represent the statistical properties of events such as machine failures or worker absenteeism.

Simulation models represent the events that could occur as a system operates by a sequence of steps in a computer program (see Chapters 93–96). This means that the logical relationships that exist between events can be described in detail. The probabilistic nature of many events, such as machine failure, can be represented by sampling from a distribution representing the pattern of occurrence of the event, for example, the distribution of the time between machine failures. Thus, in order to represent the typical behavior of the system, it is necessary to run the simulation model for a sufficiently long time that all events can occur a reasonable number of times. Simulation models can be provided with an interactive graphic display to demonstrate the movement of jobs or customers. This can be of great value in communicating the assumptions of the model to engineers and others.

Analytical models describe the system using mathematical or symbolic relationships. These are then used to derive a formula or define an algorithm or computational procedure by which the performance measures of the system can be calculated. Analytical models can also be used to demonstrate properties of various operating rules and control strategies. Sometimes it is not possible, within a reasonable amount of computer time or space, to obtain the performance measure from the relationships describing the system without making further assumptions that modify these relationships. The resulting model is thus approximate rather than exact. Testing the approximation may then require a simulation model, so approximate models are useful only if they are easy to use and provide insight into what determines the system behavior.

1.1.3. Why Model?

Models can be developed for a variety of reasons, in particular:

- *Understanding:* The model is used in order to explain why and how. The model is intended to convey insight. Sometimes the model just indicates the direction of influence of some variable on performance, that is, as the variable increases in value does performance improve? Alternatively, the model can be quite complex but with the major function of explaining why the system behaves in certain ways.
- *Learning:* As well as providing insight, a model may be intended to teach managers or workers about the factors that determine performance. The model may omit many features of the real system and focus on those aspects that are considered crucial for those people responsible for effective operation of the system.
- *Improvement:* The model is used to improve system design and operation. Changes in parameters and rules can be explored, and factors critical for achieving performance targets can be identified. To make sure that conclusions drawn from the model will apply to the real system, such models pay particular emphasis to the adequacy with which they describe the system behavior.
- *Optimization models:* Given a model that predicts performance as a function of various parameters, an optimization model determines the optimal combination of these parameters. This usually means that the optimization problem is formulated as a mathematical programming problem, generally with a mixture of integer and continuous variables.
- *Decision making:* The model is to be used to aid decisions about either the design or operation of the system. The model has to be able to discriminate the effects of different courses of action and project their impact over time.

1.1.4. Requirements of Models

Models are based on certain assumptions about the system and its components and how the system is going to be operated. Then there are the assumptions about the nature of the disturbances that will impact the system and the range of responses to these disturbances. The hierarchy of control and the flow of information also have to be represented in the model.

- *Complexity vs. simplicity:* Modeling involves compromises in deciding how much detail to represent. A large amount of detail means that the model should be a more precise representation of reality, but the disadvantages are that the model will be more difficult to verify and validate, be harder for users to understand, and take longer to develop. A simple model may not represent the system adequately and thus may give inaccurate predictions and omit key decisions or useful responses to disturbances.
- *Flexibility:* A model may be used to support decision making as the system evolves over time. That is, from initial concept through planning, detailed design, installation, and operation, there is a need for models to support decisions. While no one single model will support all decisions, it is desirable for a model to be useful at a number of different stages in system evolution. This means that the model should permit changes in the system modeled. Some of these changes may relate to the structure of the system, such as the number of machines or service representatives, the way in which jobs or customers move through the system, or the way in which the control hierarchy is set up. Other changes may relate to the values of parameters such as the frequency of machine failures or the demand rate. A model or a modeling approach has to be evaluated with respect to the ease of making both these sorts of changes.
- *Data requirements:* While there is often a great deal of data available in manufacturing or service, it is rare that the data are in the form required by the model. At the planning stage, there is often doubt as to the applicability of data collected from different systems in different environments, while when the system is operational, the data may well only apply over a limited range of operating conditions. Thus, a model should use the least amount of data required in order to make adequate predictions, and an important component of the validation of a model is assessing the sensitivity of the model to errors in the data.
- *Transparency:* Since the model has to be accepted by its users, it is desirable that the assumptions and procedures used in the model be reasonably transparent to others beside the model developer. The developer should be able to convince the user that the model is a reasonably accurate representation of reality.
- *Efficiency:* Models can consume significant resources, both in their development and in their use. Modeling approaches differ in their requirements on the knowledge, skill, and elapsed time required for development. Since most models will be implemented on a computer, such issues as running time and storage requirements can also be important.
- *User interface:* If a model is going to be of real value, it should be usable by managers rather than only by the model developer. A user interface is essential in order to guide the user in the correct use of the model, ensuring that it is clear what data should be provided and avoiding any ambiguity in interpreting the results.

1.2. Queueing Models

Queueing models are particularly useful for determining the following performance measures of a system:

- *Capacity or throughput:* This is the maximum rate at which the system can accept jobs or customers over some long time interval. We will use the symbol TH to denote it, and it is usually measured in jobs or customers per hour (or some other suitable time interval). Individual components of the system will, of course, have a higher short-term capacity than TH, but over the long run they will lose capacity because of machine breakdowns or worker absences. Capacity will also be lost because of interaction between the different parts of the system. For example, too few pallets will reduce work flow through a machine.
- *Flow Time or Lead Time:* The flow time, sometimes called the lead time in manufacturing, is the time from when a job or customer arrives at the system until the job or customer departs the system. It will be greater than the actual processing time because jobs are held in inventory buffers and customers wait in queues. Queueing models usually focus on determining the average flow time, but it is usually also possible to determine the variance and higher moments.
- *Inventories and queue lengths:* It is often important to know where inventories and queues are distributed through the system. The average total flow time and the average total queue length are connected by Little's law:

$$\bar{l} = \lambda \overline{w}$$

where $\bar{l}$ is the average queue length, $\overline{w}$ is the average flow time, and λ is the average rate at which jobs or customers flow through the system. So given either queue length or flow time, the other performance measure can be readily determined.

- *Service level:* The service level can be measured in a variety of different ways, such as the fraction of demands met immediately or the average time to fill a customer demand. Service level is a particularly important performance measure when finished product inventories are kept and it is necessary to trade off the inventory investment with the penalties of delay in meeting customer demand.

1.2.1. Using Queueing Models

Queueing models are particularly useful in designing and improving system performance. In particular, they are of great value for addressing the following issues:

1. *Investigating alternative configurations:* There are usually alternative ways of allocating tasks to machines or people, and each alternative will result in different patterns of work flow. Queueing models are particularly valuable in rapidly exploring a wide range of alternatives and seeing how system performance is modified.
2. *Exploring the impact of parameters set by management:* Typically, management have to choose values for such parameters as inventory buffer capacities, the number of kanbans, or base stock levels. Queueing models enable their impact on performance measures such as throughput or service level to be found. If costs are available, then it is possible to determine the values that optimize performance.
3. *Comparing alternative scheduling and work-allocation rules:* Queueing models can be used to compare different scheduling rules. They can also be used to explore the way in which performance is changed as more information about jobs or customers is acquired and that information used to modify the routing and allocation of jobs or customers to machines or servers.
4. *Understanding the impact of variability:* Typically, less variability improves performance. But since reducing variability can be costly, it is desirable to know by how much performance is improved. Variability reduction is typically the aim of quality management efforts, and queueing models can help focus that effort.

1.3. Modeling Manufacturing and Service Systems

Queueing theory is described in Chapter 83. To apply it to modeling manufacturing or service systems, it is necessary to develop an understanding of the key features of the system and how these translate into queueing models. Considerable expertise is necessary. Often, as the system becomes more complex, it is necessary to develop approximations, and these have to be tested by extensive simulations. So a number of software packages have been developed that incorporate tested queueing models and approximations. These packages have user interfaces that guide the user through the selection of the appropriate model and allow the user to specify parameters values.

Alternatively, most queueing models can be easily incorporated into spreadsheets because they are either formulas or fairly straightforward iterative calculations.

Before giving the details of typical models, it is useful to describe briefly how the main features of various manufacturing and service systems are viewed as queues.

1.3.1. Manufacturing Systems

1.3.1.1. Job Shops and Flow Lines The two traditional forms of organizing manufacturing systems are the job shop and the flow line. The *job shop* consists of a variety of different types of machines, some of which can perform operations on different types of jobs, although this may require some setup or changeover time between job types. Material handling is such that different types of jobs can visit machines in different sequences. In a queueing model, the job shop is viewed as a network of queues, with each machine or machine center regarded as a server. Job routing determines the movement of jobs between the different servers. Rather than tracking individual jobs in the queueing model, it is usual to represent all the different jobs by a small number of job types or classes. By contrast, the *flow line* requires all jobs to visit machines and work centers in the same sequence, thus simplifying material handling. The simple material handling, combined with the standard routing, makes it easier to control work flow and instruct machines and workers on their tasks and thus enables high volumes to be produced economically. The queueing model thus consists of a number of queues in series. However, buffer space between machines is often limited, resulting in blocking or starving of machines, and this has to be represented by the model.

1.3.1.2. Transfer Lines Many flow lines produce only a single type of product, and with increasing volume it becomes attractive to automate individual machines and replace human operators by automatic devices and machines. The *automatic transfer line* goes one step further. Not only are the machines and the material handling from one machine to the next automated, but all machines are linked so that they begin their tasks simultaneously and thus material movement is synchronized. In this way, the number of jobs in process can be kept small and extremely high productivity is potentially possible. If any machine should fail, then the tight linkage means that the whole line stops. To overcome this, the line may be divided up into sections, with buffers between the sections. The queueing model has to capture this linkage and the impact of buffers.

1.3.1.3. Flexible Transfer Lines Initially, when automatic transfer lines came into widespread use in the 1950s, the instructions on how to perform the tasks at a machine were embodied in physical information storage devices such as cams, jigs, and other fixtures. This has been called "hard automation," and obviously it made it difficult to change instructions. However, as electronic devices for storage and processing of information were developed, combined with numerical control (i.e., transducers that convert digital information into physical motion of tools), changing the instructions on performing tasks became simpler. This resulted in the development of automated *flexible transfer lines* and *flexible flow lines,* distinguished by whether synchronized movement of jobs was retained or not. If job movement is not synchronized, then it is essential to provide some storage space between machines, otherwise job movement will always be determined by the slowest machine. Flexible transfer lines and flexible flow lines can manufacture more variants of a product than traditional automatic transfer lines, but unless the variation between jobs is very small, there may still be a requirement to change tools between jobs. The queueing model now has to describe the movement of the different job types through the system.

1.3.1.4. Flexible Manufacturing Systems Transfer lines are essentially automated flow lines, and thus all jobs have to visit all the machines in the same sequence. If numerically controlled machines are combined with a material-handling system that enables different jobs to visit machines in different sequences, then the resulting system is known as a *flexible manufacturing system* (FMS). Such systems were first implemented in the 1970s for machining tasks, so FMS sometimes implies a flexible machining system. An FMS is essentially an automated job shop. It can produce a reasonable range of products, although initially this range was limited by the ability to store or deliver the required tools to individual machines. Usually parts are mounted on pallets and the number of pallets is limited, so the queueing model represents the system as a network of queues in which the total number of customers is limited by the number of pallets.

1.3.1.5. Flexible Assembly Systems Assembly tasks are difficult to automate economically unless the tasks have unusual characteristics such as size, weight, or temperature or chemical or radiation hazards. However, beginning in the clothing and shoe industries and then spreading to the electronic industry in the early 1980s, it was recognized that *flexible assembly systems* with automated job movement to assembly, inspection, and test stations, linked to automated job-identification systems, resulted in significant improvement in work flow and control in assembly systems, producing a variety of different jobs, even though many individual tasks are still performed by human operators. Some flexible assembly systems enable jobs to move between any pair of workstations, while others, such as those introduced in the mid-1980s by the automobile industry to replace the traditional assembly line, have a generally series structure but with paralleling of workstations and some feedback loops so that jobs can be readily reprocessed if they do not meet required quality standards. Again, the number of work carriers may be limited, and there may be requirements to maintain sequence of jobs as they go through parallel work stations. Flexible assembly systems sometimes store all work in progress centrally and assign it to work stations as they become available. Alternatively, jobs may circulate through the system on conveyors. Queueing models have to be able to represent job circulation, storage, and processing.

1.3.2. Service Systems

As yet, there is no established approach for categorizing the different configurations of service systems, although there are a variety of approaches proposed (Schmenner 1986, Silvestro et al. 1992). From the perspective of developing queueing models, perhaps the most useful approach is to focus on the range and features of the tasks assigned to the people providing service (Buzacott 2000).

1.3.2.1. Narrow-Range Tasks If the system is configured so that each individual server only performs a narrow range of tasks, then this usually means that a number of different servers are required in order to perform all the required tasks for a customer. This means that the system will often be similar to a manufacturing flow line and so can be represented by queues in series. Of course, if there is some flexibility in the sequence in which tasks are done, then the system becomes more like a manufacturing job shop and can be represented by a network of queues.

1.3.2.2. Broad-Range Tasks Alternatively, individual servers can be assigned a broad range of tasks. This means that they are often able to do all the tasks required by a specific customer. However, in order to cope with increasing volume of customers, many servers are required. The system can then be represented as a number of servers in parallel. There are a variety of ways in which customers can be allocated to servers; for example, some servers may specialize on particular types of customers; alternatively, customers can be allocated to servers according to some allocation rule, such as allocate to free servers in order of customer arrivals or allocate according to a round-robin or cyclical rule (first arrival, to server 1, next to server 2, next to server 1, next to server 2, and so on).

1.3.2.3. Specialized Diagnosis In many service situations the tasks to be done on or for a customer are not known until the customer arrives and some diagnosis is carried out. As a result, it is sometimes desirable to have a first stage of service that determines the service tasks required and then allocates customers to specialized service providers capable of performing the required tasks. The appropriate queueing model now consists of a network of queues in which customer flow out of diagnosis to a specific specialized facility is random, with probability equal to the frequency of the facility being required by the diagnosis. In some situations, multiple diagnostic steps are required, depending on whether a step is able to determine the service requirements.

1.3.3. Supply Chains and Logistic Systems

A developing field for applying queueing models is in determining service levels and inventories in supply chains. For modeling purposes, the supply chain is viewed as a network of cells, where in each cell manufacturing, transport, inspection, and other functions are performed. Cells are connected by material flow from raw material through parts manufacture through assembly, test, and distribution through warehouses and retail outlets to customers. However, cells are also connected by information flow from customer orders through retailer orders, warehouse shipment instructions, production schedules, part requisitions, and raw material requests. Because material flow occurs only as a result of requisitions and orders, it is necessary to include in the queueing model not only the representation of material flow but also the representation of information flow and how the two interact. Certain types of coordination schemes for controlling work flow use advance information—for example, forecasts of receipts of customer orders are used in Material Requirements Planning—and it is then necessary to represent this in the model. We will illustrate the combined modeling of material flow and information flow by some simple models.

1.4. Description of Available Models

The remainder of this chapter will be devoted to a description of some of the available analytical models of manufacturing and service systems and supply chains. The discussion will emphasize analytical models that yield a formula-type result, as such results are most readily implementable. When no formula exists, we will outline the general principles of the computational algorithms. We will also illustrate a number of approximate approaches for determining performance.

1.4.1. Assumptions of Models

Every model is based on certain assumptions about the composition of the system, the work flow through it, and the availability of resources, such as the number of operators, the number of repair crews, and the way in which these resources are allocated when there are competing demands for them. These assumptions describe the essential features of the system. Then there are further assumptions that relate to the nature of disturbances to the system operation, such as the distribution of time between successive failures of a machine, the distribution of time to repair a machine, the distribution of time to perform a task, and the pattern of occurrence of defective parts. In some models, the results depend critically on these distributions, while in others the mean of the distribution is all that is required. Usually it is easier to derive results for certain distributions than for others; in particular, it is possible to derive explicit formulas for the performance of some systems when the associated distributions are exponential or geometric, but not for other distributions. Fortunately, in many cases the results are not particularly sensitive to the form of the distribution or, at least, they give the right general shape of the relationship of the system design parameters to the system performance measures. Thus, the results can be used to give general insight. Almost all analytical models make the following assumptions:

1. Distributions are stationary (i.e., process parameters do not change with time or cumulative output).
2. Successive events of a particular type occur at intervals that are independent of each other (i.e., there is no serial correlation).
3. Events at one machine or service center are independent of events at another machine.

These assumptions are often not strictly correct in reality. However, it would be difficult to derive results that would relax them. In most real systems, when the assumptions fail, the system would be considered out of control and managerial action would be taken to restore control (e.g., a significant worsening in quality with time would demand managerial intervention).

2. SINGLE-STAGE SYSTEMS

In a single-stage system all work required by a job or a customer is done by one service facility. However, this facility may consist of a number of servers or machines in parallel, and system performance will depend on how jobs or customers are allocated to the machines or servers.

We distinguish between two situations. One is that where no work is done prior to a job or customer's arrival. The second is where jobs are done prior to arrival of a demand, or work is done on anticipated customer requirements prior to the customer arriving.

2.1. Make-to-Order Manufacturing or Service with No Tasks Done Prior to Customer Arrival

2.1.1. Single-Server System

The system is then equivalent to a single-server queueing system for which numerous results are available (see sections 4.2 to 4.5 of Chapter 83). With Poisson arrivals, the performance measures are usually obtainable from fairly simple formulae. However, for general arrival times and general service times, it is usually necessary to resort to bounds and approximations, although such results are often remarkably accurate when the server utilization is reasonably high (see Daley et al. (1992) for a comprehensive description of various bounds and approximations). Suppose that the mean time between arrivals is $1/\lambda$ and the squared coefficient of variation (scv = variance/mean^2) of the time between arrivals is C_a^2. The service time of a job has mean $1/\mu$ and scv of C_S^2. Let $\rho = \lambda/\mu$. Then a good general upper bound on the number of jobs in the system is

$$E[N] \leq \frac{\rho(2-\rho)C_a^2 + \rho^2 C_S^2}{2(1-\rho)} + \rho \tag{1}$$

while for interarrival times that are DMRL, i.e., decreasing mean residual life, or in other words, as the time since the last arrival increases, the expected time to the next arrival becomes less, a lower bound is

$$\frac{\rho C_a^2 - \rho(1-\rho) + \rho^2 C_S^2}{2(1-\rho)} + \rho \leq E[N] \tag{2}$$

Note that the difference between these upper and lower bounds is $\rho(C_a^2 + 1)/2 < \rho$, since with DMRL arrivals $C_a^2 < 1$.

An approximation for the mean number of jobs in the system, $\hat{n}$, that gives good results for relatively high utilizations is the following:

$$\hat{n} = \left\{\frac{\rho^2(1 + C_S^2)}{1 + \rho^2 C_S^2}\right\}\left\{\frac{(C_a^2 + \rho^2 C_S^2)}{2(1-\rho)}\right\} + \rho \tag{3}$$

At low utilizations, none of the approximations or bounds that just use information on the mean and variance of the interarrival times and service times are particularly good when the criterion is the percentage error. This is because the performance of the queue tends to be determined by the burstiness of arrivals, a property not captured by the second moment.

For some situations, where all that is required is to get insight into the impact of variability in interarrival times and service times at high utilizations, it is possible to use the heavy traffic result

$$\lim_{\rho \to 1} 2(1-\rho)E[N] = C_a^2 + C_S^2 \tag{4}$$

2.1.2. Multiple Servers

2.1.2.1. Identical Servers Suppose there are c parallel servers with identical capabilities. The mean time between arrival of jobs is $1/\lambda$ and the mean time to serve a job is $1/\mu$. Jobs wait in a single queue and the first job in the queue is allocated to the first free server. Let $\rho = \lambda/c\mu$. Then with general arrivals and general service time distributions there are no exact results. One approach is to approximate the system by a $G/G/1$ queue and then modify the performance measures by the relationship between $M/M/c$ and $M/M/1$ results. When the multiple server system is represented by a $G/G/1$ queue, the arrival process at the queueing system is unchanged but the service time of the single server equivalent to the c parallel servers is scaled by a factor of $1/c$. That is, if the service time at any one of the c servers had mean $\bar{s}$, variance σ_S^2, and squared coefficient of variation (scv) $C_S^2 = \sigma_S^2/\bar{s}^2$, the equivalent single server has service time with mean $\bar{s}/c$, variance σ_S^2/c^2, and scv $= C_S^2$. Then the number of jobs in service and waiting can be approximated by

$$\hat{n}_{G/G/c} = \frac{E[L]_{M/M/c}}{E[L]_{M/\mathbf{M}/1}} (\hat{n}_{G/G/1} - \rho) + c\rho \tag{5}$$

where $\hat{n}_{G/G/1}$ is an approximation for the number of jobs in service and waiting in the approximating $G/G/1$ system. $E[L]_{M/M/c}$ is given by a well-known queueing theory result

$$E[L]_{M/M/c} = \frac{(c\rho)^c}{c!} \left(\frac{\rho}{(1 - \rho)^2} \right) p(0)$$

where $p(0) = 1/\{\Sigma_{k=0}^{c-1} (c\rho)^k/k! + (c\rho)^c/(1 - \rho)c!\}$. Also note that $E[L]_{M/M/1} = \rho^2/(1 - \rho)^2$.

2.1.2.2. Nonidentical servers Assume that all jobs have essentially the same work content. However, machines or servers differ in the time that it takes for them to perform the required work. Suppose that there are c servers or machines and the time required to perform the required work for a job or customer by server j, $j = 1,2, \ldots, c$, is a random variable S_j with mean $\bar{s}_j$.

2.1.2.3. Throughput Then the following table shows the throughput for a number of different rules for allocating jobs to machines or customers to servers.

Allocation rule	TH
First free server	$\Sigma_{j=1}^{c} 1/\bar{s}_j$
To server j with probability p_j	$\min_j 1/p_j\bar{s}_j$
Round robin or cyclic	$\min_j c/\bar{s}_j$

Note that these results depend only on the mean of the S_j. Also note that the first free server rule gives the greatest throughput, although the random allocation can reach the same throughput if p_j is chosen, so $\bar{s}_j p_j = 1/\Sigma_{u=1}^{c} 1/\bar{s}_u$. If, however, $p_j = 1/c$ for all j, then random allocation and cyclic allocation give the same throughput. In service systems, it is usual that servers differ in their capabilities even though they perform the same tasks. It can be seen that these differences will have significant impact on throughput unless it is feasible to use the first-free-server rule.

Suppose the time between arrivals has mean $1/\lambda$ and squared coefficient of variation (scv) C_a^2. With the cyclic allocation rule, the time between arrivals at a given server will have mean c/λ and scv of C_a^2/c. With random allocation, the time between arrivals at server j has mean $1/p_j\lambda_j$ and scv of $1 - p_j + p_j C_a^2$. Thus, if $p_j = 1/c$, the mean time between arrivals at a server are the same but the random allocation has higher scv of arrivals and hence will have longer queues.

2.2. Make-to-Stock Manufacturing Systems or Service Systems where Work Is Done in Advance of Customer Arrival

Assume throughout this subsection that there is a single server.

2.2.1. Exponential Service Time, Poisson Demands

Suppose the target stock level is set as z. Then, as soon as a demand arrives, it is satisfied by an item taken from inventory. A job is then released to the machine or server to begin manufacturing or serving the replenishment. Mean service time of a job is $\bar{s} = 1/\mu$.

2.2.1.1. Backlogged Demands In a system where unmet demands are backlogged, it follows that jobs will arrive at the machine in exactly the same way as demand arrives. Hence the queue length at the machine is the same as the queue length in a make-to-order system, with the same pattern or arrivals of demands and the same service time distribution at machines. However, if N is the length of the queue, then the inventory in the output store will be $\max\{z - N, 0\}$ while the size of the backlog will be $\max\{0, N - z\}$. Hence it follows that the expected backlog $E[B]$ is given by

$$E[B] = \frac{\rho^{z+1}}{1 - \rho} \tag{6}$$

and the expected delay in meeting a demand is $\bar{s}(\rho^z/(1 - \rho))$. The probability a demand cannot be met immediately will be ρ^z, so the service level, the fraction of demand met from stock, is $1 - \rho^z$.

2.2.1.2. Lost Sales Alternatively, suppose that demands that cannot be met from stock are lost. Now SL, the fraction of demand met from stock, is given by

$$\text{SL} = \frac{1 - \rho^z}{1 - \rho^{z+1}} \tag{7}$$

2.2.2. Single Machine with Interruptible Demand (Stopped Arrival Queue)

Suppose now that as soon as the store is empty, the demand process is turned off. It is turned on again once there is an item in the store. The arrival and service of jobs at the machine is now equivalent to a *stopped arrival* queue, that is, a queue in which the arrival process turns off once the queue length reaches z. Now the service level is defined by the ratio of the number of demands met to the number that would have been generated if the arrival process were never turned off. This is equal to the fraction of time that the inventory level in the store is greater than zero. With exponential arrivals and exponential service time, the service level is the same as the lost sales case. However, when the interarrival process is general, then the results are somewhat different. Consider a $G/G/1$ queue with arrivals equal to the arrival process of demands and service equal to the machine service time. Let $\hat{n}$ be an approximation for the queue length in this queue and ρ its traffic intensity. Then define σ by $\sigma = (\hat{n} - \rho)/\hat{n}$. If $\rho > 1$, then define $\sigma = 1/\sigma_R$ where $\sigma_R = (\hat{n}_R - \rho_R)/\hat{n}_R$ and $\hat{n}_R$ is the average queue length in a $G/G/1$ queue with arrivals having a distribution equal to the service process and service distribution equal to the interarrival distribution in the original system. ρ_R is defined as the ratio of the mean interarrival time of demands to the mean service time (i.e., $\rho_R = 1/\rho$). Then a good approximation for the service level in this system is given by

$$\text{SL} = \frac{1 - \rho\sigma^{z-1}}{1 - \rho^2\sigma^{z-1}}, \qquad \rho \neq 1$$

$$\text{SL} = \frac{1 + (z - 1)\nu}{2 + (z - 1)\nu}, \qquad \rho = 1 \tag{8}$$

where $\nu = \lim_{\rho\to 1} d\sigma/d\rho$. For example, if the heavy traffic approximation (4) is used, $\nu = 2/(C_a^2 + C_S^2)$ and so when $\rho = 1$,

$$\text{SL} = \frac{C_a^2 + C_S^2 + 2(z - 1)}{2(C_a^2 + C_S^2 + z - 1)} \tag{8}$$

2.2.3. Multiclass Backlogged Demand

Suppose now that there are r types of items produced on a common single machine. The demand rate for type i, $i = 1, \ldots, r$, is λ_i. The service time distribution is identical for all types and has an exponential distribution with mean $1/\mu$. Suppose that there is a target stock level of z_i for type i. Then define $\hat{\rho}_i$ by $\hat{\rho}_i = \lambda_i/(\mu - \Sigma_{j\neq i}\lambda_j)$. The service level for type i is then given by

$$\text{SL}_i = 1 - \hat{\rho}_i^{z_i} \tag{10}$$

2.2.4. Multiclass Lost Sales

Again there are r classes with the target stock z_i for class i, $i = 1,2, \ldots, r$. Demand rate for class i is λ_i while all types are produced on a single machine with the same service time distribution. It is possible to show that the probability of observing the number of jobs of each type in the machine queue of $n_1, n_2, \ldots, n_i, \ldots, n_r$ is given by

$$p(\mathbf{n}) = p(n_1, n_2, \ldots, n_i, \ldots, n_r) = G(\mathbf{Z})^{-1} \binom{\Sigma_{i=1}^r n_i}{n_1, \ldots, n_r} \prod_{i=1}^{r} \rho_i^{n_i}, \quad 0 \le n_i \le z_i;\ i = 1, \ldots, r \tag{11}$$

where the normalizing constant $G(Z)$ is determined by

$$G(\mathbf{Z}) = G(z_1, z_2, \ldots, z_r) = \sum_{n_1=0}^{z_1} \cdots \sum_{n_r=0}^{z_r} \binom{\Sigma_{i=1}^{r} n_i}{n_1, \ldots, n_r} \prod_{i=1}^{r} \rho_i^{n_i} \tag{12}$$

Note that $(\Sigma_{i=1}^{r} n_i / n_1, \ldots, n_r)$ is the multinomial coefficient, that is the number of ways to allocate $\Sigma_{i=1}^{r} n_i$ items to r cells in which cell i, $i = 1, 2, \ldots, r$, contains n_i items.

The service level for type i items is given by

$$\begin{aligned} \mathrm{SL}_i &= \lambda_i \left(1 - \sum_{n_1=0}^{z_1} \cdots \sum_{n_i=z_i} \cdots \sum_{n_r=0}^{z_r} p(\mathbf{n})\right) \\ &= \lambda_i \frac{G(z_1, z_2, \ldots, z_{i-1}, z_i - 1, z_{i+1}, \ldots, z_r)}{G(\mathbf{Z})}, \quad i = 1, \ldots, r \end{aligned} \tag{13}$$

2.2.5. Produce to Stock with Advance Information

Suppose that it is possible to obtain information about future demands in advance. That is, the arrival of the nth demand, $n = 1, 2, \ldots$, at time t is signaled at time $t - \tau$. If the target or initial stock of finished products is zero, then the item to meet the demand at time t can be released for manufacture at time $t - \tau$. Now if a target stock z of finished products is held, release of an item to manufacture will still take place at time $t - \tau$, but the nth demand, which arrives at time t, will actually be met by an item released for manufacture prior to time $t - \tau$ and triggered by the advice of demand $n - z$. Suppose service times are exponential with parameter μ and advices about future demands arrivals are Poisson with rate λ. Let $\rho = \lambda/\mu$. Then the service level SL, i.e., the fraction of demands that are met from stock, is given by (Buzacott and Shanthikumar 1994).

$$\mathrm{SL} = 1 - \rho^z e^{-\mu\tau(1-\rho)} \tag{14}$$

and the average delay in meeting a demand is given by

$$\overline{w}/\overline{s} = e^{-\mu\tau(1-\rho)} \frac{\rho^z}{1 - \rho} \tag{15}$$

The average inventory is given by

$$\bar{i} = z + \lambda\tau - \frac{\rho}{1 - \rho}(1 - \rho^z e^{-\mu\tau(1-\rho)}) \tag{16}$$

3. FLOW LINES AND SERIES SYSTEMS

3.1. Introduction

Flow lines and series systems can be divided up into two broad classes, based on their influence on the worker: *paced* and *unpaced*. In a *paced* system the time allowed to perform a task is limited and once this time is up the job or customer can no longer be worked on, so it is possible that the task may not be completed. In an *unpaced* system there is no maximum time limit imposed on the time for the worker to perform the task.

Paced systems lose throughput because of the incomplete processing of jobs or customers, while unpaced systems lose throughput because of the variability of task times.

3.2. Models of Paced Systems

In a paced system, the tolerance time, τ, the maximum time available to perform the tasks at any workstation, is set. Thus if T_i, the time required by the worker at station i to perform their required tasks, exceeds τ, the tasks will be incomplete and defective products will result. Hence the probability $Q(\tau)$ that a product will not contain any defects is

$$Q(\tau) = P\{T_1 \leq \tau, T_2 \leq \tau, \ldots, T_m \leq \tau\} = \prod_{i=1}^{m} P\{T_i \leq \tau\} \tag{17}$$

To meet a given quality target Q, the probability that the product is nondefective should be at least Q. So the minimum tolerance time is the solution to

$$F_i(\tau) = Q^{1/m} \tag{18}$$

One of the managerial controls in a paced system is the tolerance time τ, which is related to the line speed. Too short a tolerance time means that the quality is low, while too high a tolerance time reduces the utilization of the workers and so lowers productivity. For example, if $m = 10$ and the quality target $Q = 0.98$ then $F_i(\tau)$ must be at least 0.998. This means that if task times were exponential, and with the same distribution at all stations, the tolerance time τ has to be set at 6.2 × the mean processing time of a job at a station. If task times have a normal distribution with mean θ and standard deviation σ, then the tolerance time will have to be set as $\tau = \theta + 2.9\sigma$. It is clear that this would lead to a substantial loss in labor productivity unless the variability of task times can be kept small.

Note that for Q close to 1, $F_i(\tau)$ is approximately given by

$$1 - F_i(\tau) \approx \frac{1 - Q}{m} \tag{19}$$

The gross output rate will be $1/\tau$ but the system throughput of nondefectives will be

$$TH = \frac{Q(\tau)}{\tau} \tag{20}$$

3.3. Models of Unpaced Lines

The unpaced line consists of m stages. Once a job is completed at stage i, it moves on to stage $i + 1$, although it may pass through an intermediate buffer of capacity b_i, $i = 1, 2, \ldots, m - 1$, where capacity is made up of the space for jobs in process at the stage plus the space in the buffer.

3.3.1. Infinite Buffer Systems

Suppose that there is one machine or server at each stage. The processing time of a job at stage i is exponential with mean $1/\mu_i$. Jobs or customers arrive at the system according to a Poisson process with rate λ. In an infinite buffer system, we have

$$\mathrm{TH} = \min_i \left(\frac{1}{\mu_i}\right) \tag{21}$$

and if $\lambda < \mathrm{TH}$ and $\rho_i = \lambda/\mu_i$, then the distribution of N_i, the inventory at station i, is given by

$$P(N_i = k_i) = \rho_i^{k_i}(1 - \rho_i)$$

It follows that the average inventory at stage i is given by

$$E[N_i] = \frac{\rho_i}{1 - \rho_i}$$

and hence the total inventory $\bar{n}$ is given by

$$\bar{n} = \sum_{i=1}^{m} \frac{\rho_i}{1 - \rho_i} \tag{22}$$

3.4. Two-Stage Flow Lines

Consider a flow line that consists of two stages separated by a buffer of capacity z. Each stage consists of a single server, so $b = z + 1$.

3.5. Exponential Service Times

Assume that the service time at each stage is exponentially distributed with mean $1/\mu_i$, $i = 1,2$. Let $\rho = \mu_1/\mu_2$. Then, using a Markov model, it is possible to show that the throughput of this system is given by

$$\begin{aligned} \mathrm{TH} &= \mu_1 \frac{(1 - \rho^{b+1})}{1 - \rho^{b+2})} \qquad \rho \neq 1 \\ &= \mu_2 \frac{(b + 1)}{b + 2} \qquad \rho = 1 \end{aligned} \tag{23}$$

Alternatively, write

$$\text{TH} = \mu_1(1 - B(\mu_1, \mu_2, b + 1)) = \mu_2(1 - I(\mu_1, \mu_2, b + 1))$$

where

$$\begin{aligned} B(\mu_1, \mu_2, b + 1) &= \frac{\rho^{b+1}(1 - \rho)}{1 - \rho^{b+2}} \quad \rho \neq 1 \\ &= \frac{1}{b + 2} \quad \rho = 1 \end{aligned} \tag{24}$$

$$\begin{aligned} I(\mu_1, \mu_2, b + 1) &= \frac{1 - \rho}{1 - \rho^{b+2}} \quad \rho \neq 1 \\ &= \frac{1}{b + 2} \quad \rho = 1 \end{aligned} \tag{25}$$

Note that 1/TH is the mean time between parts arriving at the line and equals the mean time between parts leaving the line. That is, the system behaves as if it were equivalent to a single machine with mean service time $1/\mu^*$ given by the two equivalent expressions

$$\frac{1}{\mu^*} = \frac{1}{\mu_1} + \frac{1}{\mu_2} B(\mu_1, \mu_2, b) \tag{26}$$

$$= \frac{1}{\mu_2} + \frac{1}{\mu_1} I(\mu_1, \mu_2, b) \tag{27}$$

3.6. General Service Times

If the service times at the two stations are general, then a good approximation can be obtained by viewing the system as a stopped arrival queue, where stage 1 corresponds to the arrival process and stage 2 corresponds to the service process (Buzacott et al. 1995). Suppose the service time at stage i is a random variable S_i, $i = 1, 2$. Define $\rho = E[S_2]/E[S_1]$, $\rho_R = 1/\rho$, and $\lambda = 1/E[S_1]$. If the buffer capacity is z, then the maximum number of jobs in the system is $z + 2$ and $b = z + 1$. Then the approximation is

$$\begin{aligned} \text{TH} &= \lambda \frac{(1 - \rho\sigma^b)}{1 - \rho^2\sigma^b} \quad \rho \neq 1 \\ &= \lambda \frac{C_{S1}^2 + C_{S2}^2 + 2b}{2(C_{S1}^2 + C_{S2}^2 + b)} \quad \rho = 1 \end{aligned} \tag{28}$$

with

$$\begin{aligned} \sigma &= (\hat{n} - \rho)/\hat{n} \quad \rho < 1 \\ &= \hat{n}_R/(\hat{n}_R - \rho_R) \quad \rho > 1 \end{aligned}$$

where $\hat{n}$ is the approximate average number of jobs in a $G/G/1$ queueing system with arrival distribution S_1 and service distribution S_2, while $\hat{n}_R$ is the approximate average number of jobs in a $G/G/1$ queueing system with arrival distribution S_2 and service distribution S_1. The two moment approximations given above can be used to find $\hat{n}$ or $\hat{n}_R$ as appropriate. Table 1 shows the accuracy of the approximation for the case where S_1 and S_2 have Erlang-3 distributions.

3.7. Three-Stage Flow Lines

Consider a three-stage flow line with finite buffer storage space. The number of spaces in the buffer i between stages $i - 1$ and i is z_i. Set $b_i = z_i + 1$. Assume that the service time at each stage is exponentially distributed with mean $1/\mu_i$, $i = 1, \ldots, m$.

Such a system is best analyzed using a Markov process model. Then the state of the system is defined by $\{n_2, n_3\}$, where n_i is the number of jobs in the system that have been processed by station $i - 1$ but have not yet completed processing by station i, $i = 2, 3$. The state space $\mathcal{S} = \{(n_2, n_3)$: $0 \leq n_2 \leq b_2 + 1, 0 \leq n_3 \leq b_3 + 1, n_2 + n_3 \leq b_2 + b_3 + 1\}$. Note that when $n_i = b_i + 1$, stage $i - 1$ is blocked by stage i. Let $\mathbf{p} = (p(n_2, n_3), (n_2, n_3) \in \mathcal{S})$ be the stationary probability vector of this Markov process. The steady-state balance equations for $\mathbf{p}$, obtained by equating the rate of leaving a state with the rate of entering the state, are given by

TABLE 1 Adequacy of Throughput Approximation $C_{S1}^2 = 1/3$, $C_{S2}^2 = 1/3$

		b				
ρ	TH/λ	1	2	3	4	5
0.5	Sim.	0.9432	0.9928	0.9989	0.9999	1.0000
	approx.	0.9462	0.9895	0.9978	0.9996	0.9999
0.8	Sim.	0.8390	0.9348	0.9704	0.9864	0.9938
	approx.	0.8658	0.9411	0.9710	0.9850	0.9920
1.0	Sim.	0.7595	0.8583	0.9002	0.9250	0.9395
	approx.	0.8000	0.8750	0.9091	0.9286	0.9412
1.25	Sim.	0.6729	0.7475	0.7744	0.7882	0.7954
	approx.	0.6926	0.7529	0.7768	0.7880	0.7937
2.0	Sim.	0.4723	0.4976	0.5013	0.4991	0.5002
	approx.	0.4731	0.4947	0.4989	0.4998	0.5000

From Buzacott and Shanthikuma, © 1993. Reprinted by permission of Prentice-Hall Inc., Upper Saddle River, NJ.

$$\begin{aligned}
\mu_1 p(0, 0) &= \mu_3 p(0, 1) \\
(\mu_1 + \mu_3)p(0, n_3) &= \mu_2 p(1, n_3 - 1) + \mu_3 p(0, n_3 + 1), \quad 1 \le n_3 \le b_3 \\
(\mu_1 + \mu_3)p(0, b_3 + 1) &= \mu_2 p(1, b_3) \\
(\mu_1 + \mu_2)p(n_2, 0) &= \mu_1 p(n_2 - 1, 0) + \mu_3 p(n_2, 1), \quad 1 \le n_2 \le b_2 \\
(\mu_1 + \mu_2 + \mu_3)p(n_2, n_3 &= \mu_1 p(n_2 - 1, n_3) + \mu_2 p(n_2 + 1, n_3 - 1) + \mu_3 p(n_2, n_3 + 1), \\
&\qquad 1 \le n_2 \le b_2,\ 1 \le n_3 \le b_3 \\
(\mu_1 + \mu_3)p(n_2, b_3 + 1) &= \mu_1 p(n_2 - 1, b_3 + 1) + \mu_2 p(n_2 + 1, b_3), \quad 1 \le n_2 \le b_2 - 1 \\
\mu_3 p(b_2, b_3 + 1) &= \mu_1 p(b_2 - 1, b_3 + 1) + \mu_2 p(b_2 + 1, b_3) \\
\mu_2 p(b_2 + 1, 0) &= \mu_1 p(b_2, 0) + \mu_3 p(b_2 + 1, 1) \\
(\mu_2 + \mu_3)p(b_2 + 1, n_3) &= \mu_1 p(b_2, n_3) + \mu_3 p(b_2 + 1, n_3 + 1), \quad 1 \le n_3 \le b_3 - 1 \\
(\mu_2 + \mu_3)p(b_2 + 1, b_3) &= \mu_1 p(b_2, b_3)
\end{aligned}$$

These $|\mathbb{S}| = (b_2 + 1)(b_3 + 1) - 1$ equations along with the normalizing equation

$$\sum_{(n_2,n_3)\in\mathbb{S}} p(n_2,n_3) = 1$$

can be solved for **p**. Unfortunately, they do not possess a formula type solution.

With present computing facilities, direct solution of the equations using a standard procedure is probably as easy a method to use as any other. However, as b_2 and b_3 increase, it would be desirable to make use of the fact that the coefficient matrix has a large number of zero entries and thus a sparse matrix solution procedure would be appropriate. Once **p** has been computed, the throughput can be obtained by

$$\mathrm{TH} = \mu_3 \left(1 - \sum_{n_2=0}^{b_2+1} p(n_2, 0)\right)$$

For the special case of no (extra) buffer capacity (i.e., $b_2 = 1$, $b_3 = 1$), an explicit formula for the throughput can be obtained. It is

$$\mathrm{TH}(\mu_1, \mu_2, \mu_3) = 1 \Bigg/ \left\{\frac{1}{\mu_2} + \frac{\mu_2(\mu_1 + \mu_3)}{\mu_1\mu_3(\mu_1 + \mu_2)(\mu_2 + \mu_3)} \left[\frac{(\mu_1^2 + \mu_3^2)(\mu_1 + \mu_2 + \mu_3)^2 - \mu_1^2\mu_3^2}{(\mu_1 + \mu_3)^3 + \mu_2(\mu_1^2 + \mu_1\mu_3 + \mu_3^2)}\right]\right\} \tag{29}$$

This approach of setting up and solving a Markov model can be used for other similar systems with exponential service times. Typically, the number of equations to be solved increases rapidly with the number of stages and the size of the buffers, so while straightforward it becomes increasingly difficult to use.

3.8. Multiple-Stage Flow Lines with Exponential Processing Times

The flow line is modeled by a tandem queueing system with m stages and with finite buffer capacities $b_2, \ldots, b_m$ with $b_i = z_i + 1$, where z_i is the number of spaces in buffer i, and exponentially distributed processing times with mean $1/\mu_i$, $i = 1,2, \ldots, m$. The jobs arrive at the flow line according to a Poisson process with rate λ.

We use this situation to illustrate a particular approach to developing approximate models of manufacturing and service system performance that is of wide applicability.

The basis of the approximate approach is to consider stage i as if it were isolated from the rest of the system, and so we model it as an $M/M/1/b_i + 1$ queue. Jobs are assumed to arrive as a Poisson process with rate $\hat{\lambda}_i$ and service is exponentially distributed with rate μ_{id}, so the probability that an arriving job is blocked is given by $B(\hat{\lambda}_i, \mu_{id}, b_i + 1)$, where $B(\hat{\lambda}_i, \mu_{id}, b_i{+}1)$ is given by Eq. (24). Observe that if $\lambda < \text{TH}$, the job departure rate for stage i is λ. Thus,

$$\hat{\lambda}_i(1 - B(\hat{\lambda}_i, \mu_{id}, b_i + 1)) = \lambda \tag{30}$$

Also, considering stage $i + 1$ as a two-stage flow line, using Eq. (26) we approximate μ_{id} by

$$\frac{1}{\mu_{id}} = \frac{1}{\mu_i} + \frac{1}{\mu_{i+1d}} B(\hat{\lambda}_{i+1}, \mu_{i+1d}, b_{i+1}), \qquad i = 1, \ldots, m - 1 \tag{31}$$

and $\mu_{md} = \mu_m$. That is, we have the recursive relations (30) and (31) to solve for $\hat{\lambda}_i$, $i = 1, \ldots, m - 1$ with the initial condition $\mu_{md} = \mu_m$. Note that for (30) to have a solution, we require $\mu_{id} > \lambda$, $i = 1, \ldots, m - 1$. The mean number of jobs in the system and at the different stages can be approximated by $\Sigma_i \overline{N}_i$ and $(\hat{N}_i, i = 1, \ldots, m)$ delivered by the following algorithm. Let $\hat{N}_{m/M/1/b}(\lambda, \mu, b)$ be the mean number of customers in an $M/M/1/b$ queueing system with arrival rate λ, service rate μ, and buffer capacity b. If $\rho = \lambda/\mu$,

$$\hat{N}_{M/M/1/b}(\lambda, \mu, b) = \frac{\rho}{1 - \rho} - \frac{(b + 1)\rho^{b+1}}{1 - \rho^{b+1}}$$

3.8.1. *Algorithm 1: Work-in-Process* (Finite Buffer Flow Lines: Single Servers, Exponential Processing Times)

```
Step 1: Set $\mu_{md} = \mu_m$.
Step 2: For $i = m, \ldots, 2$,
    if $\mu_{id} < \lambda$, then the system is unstable, go to step 4;
    else, solve $\hat{\lambda}_i(1 - B(\hat{\lambda}_i, \mu_{id}, b_i + 1)0 = \lambda$ for $\hat{\lambda}_i$
    Set $1/\mu_{i-1d} = 1/\mu_{i-1} + 1/\mu_{id} B(\hat{\lambda}_i, \mu_{id}, b_i)$.
    Compute $N_1 = \hat{N}_{M/M/1/b}(\hat{\lambda}_i, \mu_{id}, b_i + 1)$.
Step 3: Compute $N_1 = \hat{N}_{M/M/1}(\lambda/\mu_{1d})$.
Step 4: Stop.
```

In step 2 we need to solve the nonlinear equation $\hat{\lambda}(1 - B(\hat{\lambda}, \mu, b + 1)) = \lambda$ for the unknown $\hat{\lambda}$ when $\mu > \lambda$. Since $\hat{\lambda}(1 - B(\hat{\lambda}, \mu, b{+}1))$ is increasing and concave in $\hat{\lambda}$, the following iterative scheme

$$\hat{\lambda}^{(k)} = \hat{\lambda}^{(k-2)} + \frac{(\hat{\lambda}^{(k-1)} - \hat{\lambda}^{(k-2)})(\lambda - \hat{\lambda}^{(k-2)}(1 - B(\hat{\lambda}^{(k-2)}, \mu, b + 1)))}{\hat{\lambda}^{(k-1)}(1 - B(\hat{\lambda}^{(k-1)}, \mu, b + 1)) - \hat{\lambda}^{(k-2)}(1 - B(\hat{\lambda}^{)k-2)}, \mu, b + 1))}, \quad k = 2, \ldots,$$

starting with $\hat{\lambda}^{(0)} = 0$; $\hat{\lambda}^{(1)} = \lambda$, will monotonically converge to the solution.

In algorithm 1 we see that if $\lambda > \mu_{id}$ for any i, $i = 1, \ldots, m$, the system is unstable. Let $\lambda^* = \max\{\lambda : \mu_{id} \geq \lambda, i = 1, \ldots, m\}$. We will use λ^* as our approximate throughput of the flow line. It can be verified that the following iterative scheme will converge and provide λ^*. Let

$$I(\lambda, \mu, b) = \frac{1 - \rho}{1 - \rho^{b+1}}$$

be the steady-state probability that the server is idle in an $M/M/1/b$ queueing system with arrival rate λ, service rate μ, and buffer capacity b.

3.8.2. Algorithm 2: Throughput (Finite Buffer Flow Lines: Single Servers, Exponential Processing Times)

```
Step 1: Set $\mu_{1u} = \mu_1$; $\mu_{id} = \mu_i$, $i = 2, \ldots, m$.
Step 2: For $i = 2, \ldots, m$, compute
    $1/\mu_{iu} = 1/\mu_i + 1/\mu_{i-1u}\, I(\mu_{i-1u}, \mu_{id}, b_i)$
Step 3: For $i = m, \ldots, 2$, compute
    $1/\mu_{i-1d} = 1/\mu_{i-1} + 1/\mu_{id}\, B(\mu_{i-1u}, \mu_{id}, b_i)$
Step 4: If $|\mu_1(1 - B(\mu_1, \mu_{2d}, b_2 + 1)) - \mu_m(1 - I(\mu_{m-1u}, \mu_m, b_m + 1))| < \epsilon$, set $\lambda^* = \mu_1(1 - B(\mu_1, \mu_{2d}, b_2 + 1))$ and stop. Otherwise go to step 2.
```

3.9. General Service Time Approximation

Suppose that the service times at the stages are random variables S_j, $j = 1, \ldots, m$. Let $1/\mu_j$ and C^2_{Sj} be the mean and squared coefficient of variation of the service time at stage j. Suppose the buffer capacity between stage $j - 1$ and j is z_j, $j = 2, \ldots, m$.

The basis of the approximation is to analyze each stage j, $j = 2, \ldots, m$ as a $GI/GI/I/z_j + 2$ stopped arrival queue. The effective arrival process and service times at stage j, $j = 2, \ldots, m$ will be chosen so that they effectively reflect the upstream and downstream portions of the flow line, where upstream means the overall effect of stages $1, \ldots, j - 1$ and downstream means the overall effect of stages $j, \ldots, m$. Let $(1/\mu_{j-1u}, C^2_{S_{j-1u}})$ denote the mean and scv of the effective input or arrival process to stage j when turned on (i.e., when stage $j - 1$ is not blocked because the queue is full), and let $(1/\mu_{jd}, C^2_{S_{jd}})$ be the mean and scv of the effective service or output process at stage j (i.e., when stage j is not starved because the stage is empty). Also let TH_j be the throughput of the stage j stopped arrival queue, obtained using a $GI/GI/1/b_j + 1$ stopped arrival queue approximation with input parameters $(1/\mu_{j-1u}, C^2_{S_{j-1u}})$ and service parameters $(1/\mu_{jd}, C^2_{S_{jd}})$.

There are useful equations relating μ_{ju} and μ_{jd}. Consider stage j. The time between the $k - 1$th and kth job departures from stage j can be written as

$$T_k^{(j)} = I_k^{(j)} + S_k^{(j)} + H_k^{(j)} \tag{32}$$

where $I_k^{(j)}$ is the idle time of stage j waiting for the next job to arrive and $H_k^{(j)}$ is the holding or blocking time of the job while it waits for queue space in the buffer at stage $j + 1$. Next observe that $S_k^{(j)} + H_k^{(j)}$ is the effective service time experienced by a job in the stopped arrival queue corresponding to stage j, that is $E[S_k^{(j)} + H_k^{(j)}] = 1/\mu_{jd}$. Also note that $I_k^{(j)} + S_k^{(j)}$ is the effective interarrival time in the stopped arrival queue corresponding to stage $j + 1$, that is, $E[I_k^{(j)} + S_k^{(j)}] = 1/\mu_{ju}$. Also, $E[T_k^{(j)}]$ is the mean interdeparture time from the original system and $1/\mathrm{TH}_{j+1}$ is from the decomposed system. In order for $1/\mathrm{TH}_{j+1}$ to be correct, we must have $E[T_k^{(j)}] = 1/\mathrm{TH}_{j+1}$. Thus, rewriting (32),

$$S_k^{(j)} + H_k^{(j)} = S_k^{(j)} + T_k^{(j)} - (I_k^{(j)} + S_k^{(j)})]$$

and taking expectations, we obtain

$$\frac{1}{\mu_{jd}} = \frac{1}{\mu_j} + \frac{1}{TH_{j+1}} - \frac{1}{\mu_{ju}}$$

Note that the average blocking time of stage j is $1/\mathrm{TH}_{j+1} - 1/\mu_{ju}$.

Similarly, by considering the time between successive job inputs to stage $j + 1$ it can be shown that

$$\frac{1}{\mu_{j+1u}} = \frac{1}{\mu_{j+1}} + \frac{1}{\mathrm{TH}_{j+1}} - \frac{1}{\mu_{j+1d}}$$

That is, we get the following set of $2(m - 1)$ equations

$$\frac{1}{\mu_{jd}} = \frac{1}{\mu_j} + \frac{1}{\mathrm{TH}_{j+1}} - \frac{1}{\mu_{ju}}, \quad j = 2, \ldots, m-1 \tag{33}$$

$$\frac{1}{\mu_{md}} = \frac{1}{\mu_m}$$

$$\frac{1}{\mu_{1u}} = \frac{1}{\mu_1}$$

$$\frac{1}{\mu_{j+1u}} = \frac{1}{\mu_{j+1}} + \frac{1}{\mathrm{TH}_{j+1}} - \frac{1}{\mu_{j+1d}}, \quad j = 1, \ldots, m-2 \tag{34}$$

Given the $C^2_{S_{ju}}$ and the $C^2_{S_{jd}}$ (see below as to alternative approaches for determining these quantities), the above $2(m-1)$ equations plus the $m-1$ equations to determine TH_j, $j = 2, \ldots, m$, the throughput of each stopped arrival queue from the mean and scv of the (equivalent) arrival and service processes, give $3(m-1)$ equations in the $3(m-1)$ variables TH_j, μ_{j-1u} and μ_{jd}, $j = 2, \ldots, m$. These equations can be solved recursively.

3.9.1. *Algorithm 3: Throughput* (Finite Buffer Flow Lines: Single Servers, General Processing Times)

Step 0: Set $\mu^{(0)}_{jd} = \mu_j$, $j = 2, \ldots, m$.

Step 1: For $k = 1, 2, \ldots$ and for $j = 1, \ldots, m-1$, calculate (i) $\mathrm{TH}^{(k)}_{j-1}$ using the $GI/GI/1/b_{j-1}+1$ stopped arrival queue approximation with input parameters $(1/\mu^{(k)}_{ju}, C^2_{S_{ju}})$ and service parameters $(1/\mu^{(k-1)}_{j+1d}, C^2_{S_{j+1d}})$, (ii) $\mu^{(k)}_{j+1u}$ using

$$\frac{1}{\mu^{(k)}_{j+1u}} = \frac{1}{\mu_{j+1}} + \frac{1}{\mathrm{TH}^{(k)}_{j+1}} - \frac{1}{\mu^{(k-1)}_{j+1d}}, \quad j = 1, \ldots, m-2,$$

and (iii) $\mu^{(k)}_{jd}$ using

$$\frac{1}{\mu^{(k)}_{jd}} = \frac{1}{\mu_j} + \frac{1}{\mathrm{TH}^{(k)}_{j+1}} - \frac{1}{\mu^{(k)}_{ju}}, \quad j = 2, \ldots, m-1$$

Step 2: Stop once

$$|\mathrm{TH}^{(k)}_m - \mathrm{TH}^{(k-1)}_m| < \epsilon$$

It can easily be verified that at convergence $\mathrm{TH}^{(k)}_i = \mathrm{TH}^{(k)}_j$ for all $i \neq j$.

3.9.2. *Squared Coefficient of Variation Recursions*

So far it has not been specified how to calculate the scv $C^2_{S_{ju}}$ and $C^2_{S_{jd}}$. Two approaches are possible. The simplest, approximation (a), is to set

$$C^2_{S_{ju}} = C^2_{S_j}, \quad j = 1, \ldots, m-1$$

$$C^2_{S_{jd}} = C^2_{S_j}, \quad j = 2, \ldots, m$$

This approximation considers only the immediate upstream and downstream stations and ignores the impact of possible blocking or starving on the variance of the equivalent service times or arrival times.

The other approach [see Buzacott et al. 1995, approximation (b)] tries to take account of the impact of possible blocking and starving on the variance of the service times. It uses the following set of $2(m-2)$ equations by which the scv's can be determined

$$E[S^2_{jd}] = E[S^2_j] + \delta_{j+1d}(E[S^2_{j+1d}] + 2E[S_j]E[S_{j+1d}]), \quad j = 2, \ldots, m-1$$

with $E[S^2_{md}] = E[S^2_m]$ and δ_{j+1d} defined by

$$\delta_{j+1d} = \mu_{j-1d}\left(\frac{1}{\mathrm{TH}_{j+1}} - \frac{1}{\mu_{ju}}\right), \quad j = 2, \ldots, m-1.$$

and

$$E[S_{ju}^2] = E[S_j^2] + \delta_{ju}(E[S_{j-1u}^2] + 2E[S_j]E[S_{j-1u}]) \quad j = 2, \ldots, m-1$$

with $E[S_{1u}^2] = E[S_1^2]$ and

$$\delta_{ju} = \mu_{j-1u}\left(\frac{1}{\mathrm{TH}_j} - \frac{1}{\mu_{jd}}\right) \quad j = 2, \ldots, m-1$$

Table 2 illustrates the approximation for a four-stage system where the service times are exponentially distributed. Note that the case (a) approximation is now identical to algorithm 2 given above for the throughput of an exponential flow line.

4. TRANSFER LINES

In a transfer line, movement of jobs at all stations in a section of the line is synchronized. Thus, transfer can only begin when the slowest station has completed its operation. If no station fails, then the mean time between successive transfers is the *cycle time*, τ. The cycle time will consist of the maximum time to perform operations at a station in the section, plus the time required for transfer. The *gross production rate* is $1/\tau$. The *net production rate* is TH and is less than the gross production rate because of line stoppages due to station or transfer mechanism failure. Note that if any station in the section fails and thus its operation is not completed, then no transfer will occur and all stations in the section will be forced down. The extent to which this section stoppage will affect other sections of the line depends on the degree of integrated linkage and control between sections. If no inventory can be kept between sections, then the rest of the line will be forced down almost immediately. In lines with many stations, there is much equipment that can break down, so the net production rate can be much less than the gross production rate. The efficiency of the transfer line is defined by

$$\eta = \frac{\text{net production rate}}{\text{gross production rate}} = \mathrm{TH}\tau$$

One means of increasing the net production rate of a transfer line and reducing the impact of stoppages is to insert in-process storage or a bank between the two sections of the line. A bank has the effect of decoupling the two sections of the line, allowing each section to operate independently.

4.1. Models

4.1.1. Transfer Lines with No Inventory Banks

Consider a transfer line with m stages or stations. The station failures can be either time dependent (failure is possible when the station is idling) or operation dependent (failure only occurs when the station is working on a part).

4.1.2. Time-Dependent Failures

Suppose the stations can fail even when they are idling. Let the mean time to failure, and repair times of stage i, be $\overline{U}_i$ and $\overline{D}_i$ respectively, for $i = 1, 2, \ldots, m$. With no inventory bank, all stations in the line stop as soon as any station fails. Then it follows that

TABLE 2 Four Stage Throughput with Exponential Service Times

Parameters							TH		
μ_i				z_i				App.	
1	2	3	4	2	3	4	Exact	(a)	(b)
1	1.1	1.2	1.3	1	1	1	0.710	0.689	0.700
1	1.2	1.4	1.6	1	1	1	0.765	0.746	0.756
1	1.5	2	2.5	1	1	1	0.861	0.850	0.855
1	2	3	4	1	1	1	0.929	0.925	0.927

From Buzacott et al. 1995. Reproduced with permission of Kluwer Academic Publishers.

$$\eta = \prod_{i=1}^{m} A_i^T \tag{35}$$

where

$$A_i^T = \frac{\bar{U}_i}{\bar{U}_i + \bar{D}_i} \qquad i = 1, \ldots, m$$

4.1.3. Operation-Dependent Failures

Suppose that stations can fail only when they are working on a part. Let $\bar{T}_i$ be the mean number of cycles station i operates between failures, $\bar{D}_i$ be the mean downtime of station i, and τ be the average cycle time when the line is running and $x_i^o = \bar{D}_i/(\bar{T}_i\tau)$. Then the line consisting just of station i would have efficiency

$$A_i^o = \frac{\bar{T}_i}{\bar{T}_i + \bar{D}_i/\tau} = \frac{1}{1 + x_i^o} \qquad i = 1, \ldots, m$$

Then it can be shown that

$$\eta = \frac{1}{1 + \Sigma_{i=1}^{m} x_i^o} \tag{36}$$

If, however, the part in process at the instant of failure must be scrapped, then

$$\eta = \frac{\Pi_{i=1}^{m}(1 - 1/\bar{T}_i)}{1 + \Sigma_{i=1} m x_i^o \Pi_{j=1}^{i-1}(1 - 1/\bar{T}_j)} \tag{37}$$

If only a fraction q_i of the failures result in scrapping of the part, then

$$\tau = \frac{\Pi_{i=1}^{m}(1 - q_i/\bar{T}_i)}{1 + \Sigma_{i=1}^{m} x_i^o \Pi_{j=1}^{i-1}(1 - q_j/\bar{T}_j)} \tag{38}$$

4.1.4. Systems Separated by Infinite Inventory Banks

Suppose the stages in an m-stage transfer line are separated by inventory banks of infinite capacity. Let the mean time to failure and repair times of stage i be $\bar{U}_i$ and $\bar{D}_i$ respectively, for $i = 1, \ldots, m$. Define

$$x_i = \frac{\bar{D}_i}{\bar{U}_i} \qquad i = 1, \ldots, m$$

4.1.4.1. No Parts Scrapped The efficiency of the line is

$$\eta = \min \left\{ \frac{1}{1 + x_i} \qquad i = 1, \ldots, m \right\} \tag{39}$$

4.1.4.2. Scrapping of Parts Suppose when station i fails, the part being processed is scrapped, with probability q_i, $i = 1, \ldots, m$. Then

$$\eta = \min_{1 \le j \le m} \left\{ \prod_{k=j}^{m} \left(\frac{1 - q_k/\bar{T}_k}{1 + x_j} \right) \right\} \tag{40}$$

4.1.5. Two-Stage Synchronized Line with Finite Capacity Inventory Banks

Consider a series transfer line with two stations, one finite intermediate inventory bank with capacity z and synchronized part transfer. Suppose

$a_i = P(\text{station } i \text{ is failed at time } n + 1 | \text{station } i \text{ is working at time } n)$
$\hat{a}_i = P(\text{station } i \text{ is failed at time } n + 1 | \text{station } i \text{ is either blocked or idle at time } n)$
$b_i = P(\text{station } i \text{ is up at time } n + 1 | \text{station } i \text{ is down at time } n) \quad i = 1, 2$

$\hat{a}_i$ is zero if failures are operation dependent, and equal to a_i if failures are time dependent.

Then a discrete time Markov chain model with the system observed just prior to the beginning of transfer can be used to show that the line efficiency is given by

$$\eta(z) = \frac{1 - r^*\rho^z}{1 + x_1 - (1 + x_2)r^*\rho^z} \qquad a_1b_2 \neq b_1a_2 \tag{41}$$

where $x_i = a_i/b_i$, that is, $1/(1 + x_i)$ is the efficiency of station i if it operates on its own,

$$r^* = \frac{(\hat{a}_1 + b_1)(\hat{a}_2\alpha_1 + a_2\beta_1)}{(\hat{a}_2 + b_2)(\hat{a}_1\alpha_2 + a_1\beta_2)} \tag{42}$$

$$\rho = \frac{\beta_1\alpha_2}{\alpha_1\beta_2} \tag{43}$$

and

$$\begin{aligned} \alpha_1 &= a_1 + a_2 - a_1a_2 - b_1a_2 \\ \alpha_2 &= a_1 + a_2 - a_1a_2 - a_1b_2 \\ \beta_1 &= b_1 + b_2 - b_1b_2 - a_1b_2 \\ \beta_2 &= b_1 + b_2 - b_1b_2 - b_1a_2. \end{aligned} \tag{44}$$

4.1.5.1. Balanced Stages (Case Where $a_1b_2 = b_1a_2$) This means that $x_1 = x_2 = \hat{x}$, $\alpha_1 = \alpha_2 = \alpha$, $\beta_2 = \beta_2 = \beta$, $\rho = 1$ and $\hat{x} = \alpha/\beta$. Then it can be shown that

$$\eta(z) = \frac{\hat{a}_1 + b_1 + \hat{a}_2 + b_2 - (\hat{a}_1 + b_1)(\hat{a}_2 + b_2) + (\hat{a}_1 + b_1)(\hat{a}_2 + b_2)z(1 + \hat{x})}{(\hat{a}_1 + b_1 + \hat{a}_2 + b_2)(1 + \hat{x}) + (\hat{a}_1 + b_1)(\hat{a}_2 + b_2)(\hat{x}(b_1 + b_2)/b_1b_2 + (z - 1)(1 + \hat{x})^2)} \tag{45}$$

Note that if $\hat{a}_1 = \hat{a}_2 = \hat{a}$, $b_1 = b_2 = b$, that is, the stations are identical,

$$\eta(z) = \frac{2 - b - \hat{a} + (b + \hat{a})z(1 + \hat{x})}{2(1 + 2\hat{x} + \hat{x}\hat{a}/b) + (b + \hat{a})(z - 1)(1 + \hat{x})^2} \tag{46}$$

4.1.6. Operation-Dependent and Time-Dependent Failures

Two limiting cases are of interest: (1) operation-dependent failures: $\hat{a}_1 = \hat{a}_2 = 0$, and (2) time-dependent failures: $\hat{a}_1 = a_1$, $\hat{a}_2 = a_2$. For operation-dependent failures and identical stations,

$$\eta(z) = \frac{2 - b + bz(1 + \hat{x})}{2(1 + 2\hat{x}) + b(z - 1)(1 + \hat{x})^2} \tag{47}$$

and for time-dependent failures

$$\eta(z) = \frac{2 - b(1 + \hat{x}) + bz(1 + \hat{x})^2}{2(1 + \hat{x})^2 + b(z - 1)(1 + \hat{x})^3} \tag{48}$$

4.1.6.1. Failure-Transfer Coefficients The quantities $\delta(z)$ and $\delta'(z)$ defined next are the failure transfer coefficients.

$$\delta(z) = Pr\{\text{bank is empty}|\text{station 1 is failed}\} \tag{49}$$

$$\delta'(z) = Pr\{\text{bank is full}|\text{station 2 is failed}\} \tag{50}$$

These quantities will be used later to describe an approximation for the efficiency of m-stage transfer lines. Later we will use $\delta(z, a_1, b_1, a_2, b_2)$ and $\delta'(z, a_1, b_1, a_2, b_2)$ for $\delta(z)$ and $\delta'(z)$ respectively to show explicitly the failure and repair probabilities of the two stages.

It can be shown that

$$
\begin{aligned}
\delta(z) &= \frac{1 - r^*}{1 - r^*\sigma^z}, \qquad x_1 \neq x_2, \\
&= \frac{b_1 + b_2 - b_1 b_2}{b_1 + b_2 - b_1 b_2 + z b_1 b_2 (1 + \hat{x})}, \qquad x_1 = x_2 = \hat{x}
\end{aligned}
\tag{51}
$$

and

$$
\begin{aligned}
\delta'(z) &= \frac{\sigma^z (1 - r^*)}{1 - r^*\sigma^z}, \qquad x_1 \neq x_2, \\
&= \frac{b_1 + b_2 - b_1 b_2}{b_1 + b_2 - b_1 b_2 + z b_2 b_2 (1 + \hat{x})}, \qquad x_1 = x_2 = \hat{x}
\end{aligned}
\tag{52}
$$

They can be expressed in an alternative way that provides further insight into their meaning.

For other two-stage transfer line models, see Buzacott and Shanthikumar (1993).

4.2. Multiple-Stage Transfer Lines

Consider a transfer line with m stages and $m - 1$ banks of capacities $z_1, z_2, \ldots, z_{m-1}$. The number of cycles of operation before failure and the repair times all have geometric distributions with mean $1/a_i$ and $1/b_i$ respectively for stage i, $I = 1, \ldots, m$. The parameters of stage i are then (a_i, b_i).

4.2.1. Approximation

The basic concept of the approximation approach is the recognition that viewed from any inventory bank the line appears to have two stages, with the arrivals at the bank determined by the upstream stages and the departures from the bank determined by the downstream stages. If the upstream stages are replaced by a single equivalent station and the downstream stages are replaced by a single equivalent station, then, viewed from the bank, the system appears to consist of the bank and two stations. If the equivalent stations have time to failure and repair time distributions for which the two station system can be solved, then the throughput of the system can be determined. The approximation arises because, even when the individual stages have distributions that in a two-station system would be solvable, grouping the upstream or downstream stages results in the grouped stages no longer having the same form of distributions of time to failure and time to repair as their constituent stages. However, in order to use the known two-stage results, it is necessary to approximate the actual distributions by distributions for which the two-stage system has a solution.

The specifics of the approximation procedure will be described for a system where each stage j, $j = 1, \ldots, m$ has geometric operation-dependent time to failure with failure probability a_j and geometric repair time with repair probability b_j. However, the procedure can easily be adapted to the case where the stages have geometric time to failure and identical deterministic repair time $\overline{D}$ with all banks having a capacity such that $z_j = l_j \overline{D}$ where l_j is an integer. Viewed from bank j, the upstream stages will be assumed to have a geometric time to failure with failure probability a_j^U and geometric repair time with repair probability b_j^U, while the downstream stages will be assumed to have geometric time to failure with failure probability a_{j+1}^D and geometric repair time with repair probability b_{j+1}^D. These parameters are given by

$$
\begin{aligned}
a_j^U &= a_j + (1 - a_j^D) a_{j-1}^U \delta_{j-1,j}, \qquad 2 \leq j \leq m - 1 \\
b_j^U &= a_j^U \Big/ \left(\frac{a_j}{b_j} + \frac{a_{j-1}^U}{b_{j-1}^U} \delta_{j-1,j} \left(1 - a_j^D \frac{b_{j-1}^U}{b_j^D + b_{j-1}^U - b_j^D b_{j-1}^U} \right) \right), \qquad 2 \leq j \leq m
\end{aligned}
$$

$$
a_I^U = a_1
$$

$$
b_1^U = b_1
$$

$$
\begin{aligned}
a_{j+1}^D &= a_{j+1} + (1 - a_{j+1}^U) a_{j+2}^D \delta_{j+2} \delta_{j+2,j+1}, \qquad 0 \leq j \leq m - 2, \\
b_{j+1}^D &= a_{j+1}^D \Big/ \left(\frac{a_{j+1}}{b_{j+1}} + \frac{a_{j+2}^D}{b_{j+2}^D} \delta_{j+2,j+1} \left(1 - a_{j+1}^U \frac{b_{j+2}^D}{b_{j+1}^U + b_{j+2}^D - b_{j+1}^U b_{j+2}^D} \right) \right), \\
&\qquad 1 \leq j \leq m - 2 \\
a_m^D &= a_m \\
b_m^D &= b_m
\end{aligned}
$$

where

$$\delta_{j-1,j} = \delta(z_{j-1}, a_{j-1}^U, b_{j-1}^U, a_j^D, b_j^D)$$

and

$$\delta_{j+1,j} = \delta'(z_j, a_j^U, b_j^U, a_{j+1}^D, b_{j+1}^D)$$

There are thus $4m - 2$ equations in $4m - 2$ unknowns. Let $(a_j^U, b_j^U, a_j^D, b_j^D, j = 1, \ldots, m)$ be a solution to these equations. Then it can be shown that the efficiency $\eta(z_j, a_j^U, b_j^U, a_{j+1}^D, b_{j+1}^D)$ of the two stage line is the same for all $j = 1, \ldots, m - 1$. Specifically, we have

$$\begin{aligned} \eta^O := \frac{1}{1 + a_m^U / b_m^U} &= \eta(z_j, a_j^U, b_j^U, a_{j+1}^D, b_{j+1}^D) \\ &= \eta^I := \frac{1}{1 + a_1^D / b_1^D}, \quad j = 1, \ldots, m - 1 \end{aligned} \tag{53}$$

Since the equations are nonlinear, they are best solved using an iterative technique, where convergence is checked by determining whether $\eta^I = \eta^O$.

4.2.2. Algorithm 4: Multistage Transfer Line

Step 0: Set $k = 1$, $a_j^D(0) = a_j$, $b_j^D(0) = b_j$, $j = 2, \ldots, m$.

Step 1: Set $a_1^U(k) = a_1$, $b_1^U(k) = b_1$. For $j = 1, 2, \ldots, m - 1$, analyze the two stage transfer line with parameters $(a_j^U(k), b_j^U(k))$ and $(a_{j+1}^D(k - 1), b_{j+1}^D(k - 1))$ and determine $a_{j+1}^U(k)$ and $b_{j+1}^U(k)$.

Step 2: Set $a_m^D(k) = a_m$, $b_m^D(k) = b_m$. For $j = m - 1, m - 2, \ldots, 1$ analyze the two-stage transfer line with parameters $(a_{j+1}^D(k), b_{j+1}^D(k))$ and $(a_j^U(k), b_j^U(k))$ and determine $a_j^D(k)$ and $b_j^D(k)$.

Step 3: Calculate $\eta^O = 1/(1 + a_m^U(k)/b_m^U(k))$ and $\eta^I = 1/(1 + a_1^D(k)/b_1^D(k))$. If $|\eta^O - \eta^I| < \epsilon$ stop, otherwise set $k = k + 1$ and go to step 1.

The convergence of this algorithm is shown in Buzacott and Shanthikumar (1993). Convergence to a solution is usually very rapid, and the solution is usually very close to the exact line efficiency. Table 3 shows exact and approximate results for a number of three-stage lines. There are some situations where convergence of the algorithm is very slow, such as, in some three station systems where the failure probability of the middle stage is much less than that of the outer stages (case 6 in Table 3). If all the b_j are identical and equal to b and the a_j's are sufficiently small that it is unlikely more than one stage is failed at any time, then the above $4(m - 1)$ equations can be reduced to the following $2(m - 1)$ equations by seting $b_j^U = b_{j+1}^D = b$ for $1 \leq j \leq m - 1$:

$$\begin{aligned} a_j^U &= a_j + a_{j-1}^U \hat{\delta}_{j-1,j}, \qquad 2 \leq j - m \\ a_{j+1}^D &= a_{j+1} + a_{j+2}^D \hat{\delta}_{j+2,j+1}, \qquad 0 \leq j \leq m - 2 \\ a_1^U &= a_1 \\ a_m^D &= a_m \end{aligned} \tag{54}$$

Here

$$\begin{aligned} \hat{\delta}_{j-1,j} &= \hat{\delta}(z_{j-1}, a_{j-1}^U, b, a_j^D, b) \\ \hat{\delta}_{j+1,j} &= \hat{\delta}'(z_j, a_j^U, b, a_{j+1}^D, b), \\ \hat{\delta}'(z, a_1, b, a_2, b) &= \hat{\delta}(z, a_2, b, a_1, b) \end{aligned} \tag{55}$$

and

$$\hat{\delta}(z, a_1, b, a_2, b) = \frac{1 - a_2/a_1}{1 - \dfrac{a_2}{a_1}\left(\dfrac{a_1 + a_2 - a_1 b}{a_1 + a_2 - a_2 b}\right)^z}$$

TABLE 3 Throughput of Three-Station Systems

Case	a_1	a_2	a_3	b_1	b_2	b_3	z_1	z_2	η exact	η approx
1	0.03	0.05	0.02	0.20	0.20	0.20	15	15	0.777846	0.777759
2	0.01	0.02	0.005	0.20	0.10	0.15	15	3	0.814949	0.814970
3	0.7	0.9	0.6	0.3	0.4	0.9	7	5	0.285463	0.285463
4	0.9	0.05	0.6	0.4	0.4	0.4	10	10	0.307692	0.307692
5	0.001	0.0003	0.005	0.7	0.02	0.3	8	4	0.970335	0.970337
6	0.6	0.04	0.6	0.8	0.4	0.8	9	9	0.565423	0.571365

From Buzacott and Shanthikumar © 1993. Reprinted by permission of Prentice-Hall Inc., Upper Saddle River, NJ.

When $a_1 = a_2$,

$$\hat{\delta}(z, a, b, a, b) = \frac{1}{1 + \dfrac{zb}{2 - b}}$$

The efficiency of the line can be found from $\eta^O = 1/(1 + a_m^U/b)$ or $\eta^I = 1/(1 + a_1^D/b)$, or by determining the efficiency of any of the two stage lines with parameters (a_j^U, b) and (a_{j+1}^D, b) and inventory bank capacity z_j by the efficiency formula

$$\hat{\eta}(z_j, a_j^U, b, a_{j+1}^D, b) = \frac{1}{1 + (a_j^U \hat{\delta}_{j,j+1} + a_{j+1}^D)/b} \tag{56}$$

It is not difficult to show that the same efficiency is obtained for all j. That is,

$$\hat{\eta}(z_j, a_j^U, b, a_{j+1}^D, b) = \hat{\eta}(z_{j+1}, a_{j+1}^U, b, a_{j+2}^D, b), \qquad j = 1, \ldots, m - 2 \tag{57}$$

Therefore, the preceding $m - 2$ equations may be used to replace $(m - 2)$ of those specified above for a_j^U's and a_{j+1}^D's and still solve for $\{(a_j^U, a_{j+1}^D), j = 1, \ldots, m - 1\}$.

5. DYNAMIC JOB SHOPS

The models in this section were originally developed to describe job shops in manufacturing, but they are also applicable to service systems where customers take different routes through the system, depending on their service requirements or diagnosis. So while we will use the language of manufacturing job shops, the reader can easily replace this with comparable language from service applications.

5.1. Open Jackson Queueing Network Model

A single class of jobs arrive at the job shop according to a Poisson process with arrival rate λ. The fraction of jobs that will join machine center i on their arrival is γ_i, $i = 1, \ldots, m$. $(\Sigma_{i=1}^m \gamma_i = 1)$. The fraction of jobs that complete service at machine center i that will directly go to machine center j is p_{ij}. Then $1 - \Sigma_{i=1}^m p_{ij}$ is the fraction of jobs among those completing service at machine center i, that will directly leave the system. Of course, at least for one or more $i = 1, \ldots, m$, $1 - \Sigma_{j=1}^m p_{ij} > 0$ so that all jobs entering the system will eventually leave the system. The service times of jobs at machine center i are i.i.d. exponential random variables with mean $1/\mu_i$, $i = 1, \ldots, m$. All the service times and the arrival times are mutually independent.

The rate at which a job is processed at machine center i when there are n jobs is assumed to be $\mu_i r_i(n) = 0, 1, \ldots$. This allows us to represent, as special cases, single or multiple machines in parallel at the machine center i. Specifically if there are c_i machines in parallel at machine center i, we set $r_i(n) = \min\{n, c_i\}$, $n = 0, 1, \ldots$; $i = 1, \ldots, m$. Leaving this representation as a general function r_i allows one to model the effect of the number of jobs in a machine center on the worker efficiency.

Assume that the job shop uses service protocols such as first come first served or last come first served that are independent of the job service time requirements.

We model this system by a Markovian open queueing network (i.e., Jackson open queueing network) where jobs are routed from one service center to another according to a transfer probability matrix $\mathbf{P} = (p_{ij})_{,,j=1\ldots,m}$. Let $N_i(t)$ be the number of jobs at machine center i at time t; $i = 1, \ldots, m$ and $\mathbf{N}(t) = \{N_1(t), \ldots, N_m(t)\}$. Then $\mathbf{N}(t)$ is a continuous time Markov process on $\mathfrak{N}_+^m$. Define the stationary distribution of $\mathbf{N}$ by $p(\mathbf{n}) = \lim_{t\to\infty} P\{\mathbf{N}(t) = \mathbf{n}\}$, $\mathbf{n} \in \mathfrak{N}_+^m$. Then it can be shown (see Jackson (1963))

$$p(\mathbf{n}) = \prod_{i=1}^{m} p_i(n_i), \qquad \dot{\mathbf{n}} \in \mathcal{N}_+^m \tag{58}$$

where if the λ_i, $i = 1, \ldots, m$ are the solution to the set of linear equations

$$\lambda_i = \lambda\gamma_i + \sum_{j=1}^{m} \lambda_j p_{ji} \qquad j = 1, \ldots, m \tag{59}$$

then

$$p_i(n_i) = p_i(0) f_i(n_i) \qquad n_i = 0, 1, \ldots, i = 1, \ldots, m \tag{60}$$

$$f_i(0) = 1;\ f_i(n_i) = \frac{f_i(n_i - 1)\lambda_i}{\mu_i r_i(n_i)}, \qquad n_i = 1, 2, \ldots;\ i = 1, \ldots, m \tag{61}$$

$$p_i(0) = 1 \Big/ \sum_{n_i=0}^{\infty} f_i(n_i), \qquad i = 1, \ldots, m \tag{62}$$

Note that λ_j is the rate of job arrivals (including both internal and external arrivals) to the machine center j, $j = 1, \ldots, m$. Furthermore, note that $p_i(.)$ is the same as the stationary distribution of the number of jobs in an $M/M(n)/1$ queueing system, with arrival rate λ_i and state-dependent service rate $\mu_i r_i(n_i)$, when there n_i jobs in it, $n_i = 0, 1, \ldots$ Therefore, the results given in Chapter 83 for the $M/M/1$ and $M/M/c$ queueing systems can be directly applied to this queueing network model. Particularly when there is only a single machine at each machine center (i.e. $c_i = 1$, $i = 1, \ldots, m$), we have

$$p(\mathbf{n}) = \prod_{i=1}^{m} (1 - \rho_i)\rho_i^{n_i}, \qquad \mathbf{n} \in \mathcal{N}_+^m \tag{63}$$

where $\rho_i = \lambda_i/\mu_i < 1$, $i = 1, \ldots, m$. In this case, the average number of jobs in machine center i is

$$E[N_i] = \frac{\rho_i}{1 - \rho_i}, \qquad i = 1, \ldots, m \tag{64}$$

and the total number of jobs in the job shop is

$$E[N] = \sum_{i=1}^{m} \frac{\rho_i}{1 - \rho_i} \tag{65}$$

Applying Little's formula, one then obtains the average flow time of an arbitrary job as

$$E[T] = \frac{1}{\lambda} \sum_{i=1}^{m} \frac{\rho_i}{1 - \rho_i} \tag{66}$$

Since $\lambda_i = \lambda v_i$, where v_i is the expected number of visits made to machine center i by an arbitrary job before it leaves the system, (66) can be rewritten as

$$E[T] = \sum_{i=1}^{m} v_i E[T_i] \tag{67}$$

where

$$E[T_i] = \frac{1}{\lambda_i} \frac{\rho_i}{1 - \rho_i} \tag{68}$$

is the average flow time of an arbitrary job in machine center i each time it visits machine center i, $i = 1, \ldots, m$.

The independence of $N_1, \ldots, N_m$ allows one to compute the variance of the total number of jobs in the job shop. Particularly since $\mathrm{var}(N_i) = \rho_i/(1 - \rho_i)^2$, $i = 1, \ldots, m$, we have

$$\operatorname{var}(N) = \sum_{i=1}^{m} \frac{\rho_i}{(1 - \rho_i)^2} \tag{69}$$

Even though the stationary number of jobs in the machine centers is statistically independent, which we used in obtaining (69), the stationary flow time of arbitrary jobs through the different machine centers is, in general, not independent. Consequently, we cannot obtain a simple expression for the variance of the flow time of an arbitrary job. In Buzacott and Shanthikumar (1993), a good approximation for this variance is given.

5.2. Multiple-Job-Class Open Jackson Queueing Network Model

In some situations, aggregation of all job types into a single job class is unacceptable in a queueing network model. We concentrate on situations where the operation/machine sequence is different for different classes of jobs, but the service requirements of the different classes at any machine center are probabilistically almost the same. In addition, we will assume that jobs are selected for service according to a First-Come-First-Served service protocol.

Class l jobs arrive at the job shop according to Poisson process with rate $\lambda^{(l)}$, $l = 1, \ldots, r$. All these r arrival processes are mutually independent. The fraction of class l jobs that join machine i is $\gamma_i^{(l)}$ ($\Sigma_{i=1}^m \gamma_i^{(l)} = 1$) and the fraction of class l jobs among those class l jobs that complete service at machine center i, that proceed directly to machine center j as a class k job, is $p_{ij}^{(l)(k)}$, $i, j = 1, 2, \ldots; l,k = 1,2, \ldots, r$. For each l ($l = 1, \ldots, r$) we assume that there exist at least one or more i such that $\Sigma_{k=1}^r \Sigma_{i=1}^m p_{ij}^{(l,k)} < 1$ so that class l jobs that enter the system will eventually leave the system. The service time of a class l job at machine center i is exponentially distributed with mean $1/\mu_i$, $i = 1, \ldots, m$. Jobs are served at a rate $r_i(n_i)$ at machine center i when there are n_i jobs, $r_i(0) = 0$, $r_i(n_i) > 0$, $n_i = 1, 2, \ldots$ Note that this service rate is independent of the number of individual classes of jobs, but depends only on the total number. The arrival, service, and job transfers from one machine center to another are mutually independent. Assuming that each job is transferred from one machine center to another and from one class to another, according to a transfer probability matrix $(p_{ij}^{(l)(k)})_{i,j=1, \ldots, m}^{l,k=1, \ldots, r}$ we model this job shop by a multiple-job-class open Jackson queueing network.

Let $N_i(t)$ be the number of jobs at machine center i at time t and $X_{i:j}(t)$ be the class index of the job in the jth position of the queue at machine center $i,j = 1, \ldots, N_i(t)$; $i = 1, \ldots, m$. We assume that the job in the first position is in service, being served at a rate $r_i(n_i)$ when $N_i(t) = n_i$, $i = 1, \ldots, m$. Then $\{\mathbf{X}(t) = (X_{i:j}(t), j = 1, \ldots, N_i(t), i = 1, \ldots, m), t \geq 0\}$ is a continuous-time Markov process. Let $q(\mathbf{x}) = \lim_{t\to\infty} P\{\mathbf{X}(t) = \mathbf{x}\}$ be the stationary probability distribution of $\mathbf{X}$. Then it can be shown that

$$q(\mathbf{x}) = \prod_{i=1}^{m} q_i(\mathbf{x}_\mathrm{i}) \tag{70}$$

and

$$q_i(\mathbf{x}_i) = \frac{1}{\Sigma_{n=0}^{\infty} f_i(n)} \prod_{s=1}^{n_i} \frac{\lambda^{(x_{i:s})}}{\mu_i r_i(s)}, \qquad i = 1, \ldots, m \tag{71}$$

with $\lambda_j^{(l)}$, $l = 1, \ldots, r; j = 1, \ldots, m$ the solution to

$$\lambda_i^{(l)} = \lambda^{(l)}\gamma_i^{(l)} + \sum_{j=1}^{m} \sum_{k=1}^{r} \lambda_j^{(k)} p_{ji}^{(k)(l)}, \qquad l = 1, \ldots, r; \ i = 1, \ldots, m \tag{72}$$

and $\lambda_j = \Sigma_{l=1}^r \lambda_j^{(l)}$, $f_i(0) = 1$, and $f_i(n_i) = f_i(n_i - 1)\lambda_i/\mu_i r_i(n)$; $n = 1,2, \ldots, m$.

We can now use (70) to obtain the joint probability distribution of $\mathbf{N}$, the number of jobs of each class at each machine center. Let $p(\mathbf{n}) = P\{\mathbf{N} = \mathbf{n}\}$ and $p_i(\mathbf{n}_i) = P\{\mathbf{N}_i = \mathbf{n}_i\}$. Here $\mathbf{n}_i = (n_i^{(1)}, n_i^{(2)}, \ldots, n_i^{(r)})$ and $n_i^{(l)}$ is the number of class l jobs at machine center i. Then one has

$$p_i(\mathbf{n}_i) = \binom{n_i}{n_i^{(1)}, \ldots, n_i^{(r)}} \prod_{l=1}^{r} \left(\frac{\lambda_i^{(l)}}{\lambda_i}\right)^{n_i^{(l)}} \frac{f_i(n_i)}{\Sigma_{n=0}^{\infty} f_i(n)}, \qquad \mathbf{n}_i \in \mathcal{N}_+^r \tag{73}$$

where $n_i = \Sigma_{l=1}^r n_i^{(l)}$.

Therefore the joint distribution of $\mathbf{N}$ is

$$p(\mathbf{n}) = \prod_{i=1}^{m} \binom{n_i}{n_i^{(1)}, \ldots, n_i^{(r)}} \prod_{l=1}^{r} \left(\frac{\lambda_i^{(l)}}{\lambda_i}\right)^{n_i^{(l)}} \left\{\frac{f_i(n_i)}{\Sigma_{n=0}^{\infty} f_i(n)}\right\}, \qquad \mathbf{n} \in \mathcal{N}_+^{r\times m} \tag{74}$$

Observe that $f_i(n_i)/\Sigma_{n=0}^{\infty} f_i(n)$, $n_i = 0,1,\ldots$ is the probability distribution of the number of customers in an $M/M(n)/1$ queueing system with arrival rate λ_i and state-dependent service rate $\mu_i r_i(\cdot)$. On the other hand

$$\binom{n_i}{n_i^{(1)},\cdots,n_i^{(r)}} \prod_{l=1}^{r} \left(\frac{\lambda_i^{(l)}}{\lambda_i}\right)^{n_i^{(l)}}$$

is the multinomial probability of choosing $n_i^{(l)}$ of type l items according to a probability $\lambda_i^{(l)}/\lambda_i$, $l = 1, \ldots, r$. Therefore, the joint distribution of the number of jobs at the different machine centers is the same as that in a single-job-class queueing network with arrival rate λ_i to machine center $i = 1, \ldots, m$. We may therefore solve the open queueing network model with single class and obtain the joint distribution of the number of jobs of different classes at different machines using a multinomial sampling with probability $\lambda_i^{(l)}/\lambda_i$ for class l jobs ($l = 1, \ldots, r$) at machine center i ($i = 1, \ldots, m$).

For a machine center that has only one machine, we have

$$p_i(\mathbf{n}_i) = \binom{n_i}{n_i^{(1)},\cdots,n_i^{(r)}} \prod_{l=1}^{r} \left(\frac{\lambda_i^{(l)}}{\lambda_i}\right)^{n_i^{(l)}} (1-\rho_i)\rho_i^{n_i}, \qquad \mathbf{n}_i \in \mathcal{N}_+^{r} \tag{75}$$

where $\rho_i = \lambda_i/\mu_i < 1$. For $c_i = +\infty$ we have

$$p_i(\mathbf{n}_i) = \prod_{l=1}^{r} \frac{(\rho_i^{(l)})^{n_i^{(l)}} e^{-\rho_i^{(l)}}}{n_i^{(l)}!}, \qquad \mathbf{n}_i \in \mathcal{N}_+^{r} \tag{76}$$

where $\rho_i^{(l)} = \lambda_i^{(l)}/\mu_i$, $l = 1, \ldots, r$. The number of jobs of different classes at a machine center with infinitely many machines are independent. In the other cases this need not be true. As we will see next, the marginal distribution of any class of job, say l, at any machine center with a single machine, say i, is exactly the same as that in an $M/M/1$ queue with arrival rate $\lambda_i^{(l)}$ and service rate $\mu_i - \Sigma_{s=1;s\neq l}^{r}\lambda_i^{(s)}$. Then it can be shown that

$$P\{N_i^{(l)} = n_i^{(l)}\} = (1-\hat{\rho}_i^{(l)})(\hat{\rho}_i^{(l)})^{n_i^{(l)}} \qquad n_i^{(l)} = 0,1,\ldots \tag{77}$$

where

$$\hat{\rho}_i^{(l)} = \frac{\rho_i^{(l)}}{1-\rho_i+\rho_i^{(l)}} = \frac{\lambda_i^{(l)}}{\mu_i - \Sigma_{k=1,k\neq l}^{r}\lambda_i^{(k)}}$$

and $\rho_i^{(l)} = \lambda_i^{(l)}/\mu_i$. Therefore, if we are interested only in studying the flow performance of a particular class of job, say l, in a job shop with single or infinite machine centers, all we have to do is redefine the service rates of the single-machine machine center (say i) by $\mu_i - \Sigma_{k=1,k\neq l}^{r}\lambda_i^{(k)}$ and analyze a single open queueing network with only class l jobs and the modified service rates. We cannot, however, do this *job class decomposition* when there are machine centers with a limited number of parallel machines.

We have seen how the number of jobs of different classes in single- and infinite-machine machine centers decomposes. Next we will make an observation about the effect of aggregating the number of jobs at different-infinite machine machine centers. Let $I \subset \{1, \ldots, m\}$ be the set of machine centers that have infinitely many machines. Then one finds that

$$P\{N_i^{(l)} = n_i^{(l)}, l = 1, \ldots, r; i \in I\} = \prod_{i\in I}\prod_{l=1}^{r} \frac{e^{-\rho_i^{(l)}}(\rho_i^{(l)})^{n_i^{(l)}}}{n_i^{(l)}!}$$

Then it is easily verified that

$$P\left\{\sum_{i\in I} N_i^{(l)} = n^{(l)}, l = 1, \ldots, r\right\} = \prod_{l=1}^{r} \frac{e^{-\rho_I^{(l)}}(\rho_I^{(l)})^{n^{(l)}}}{n^{(l)}!} \tag{78}$$

where $\rho_I^{(l)} = \Sigma_{i\in I}\rho_i^{(l)}$, $l = 1, \ldots, r$. Therefore, if we are interested only in the total number of jobs of different classes in the set I of machine centers, we may aggregate all these $|I|$ machine centers into one infinite-machine machine center.

It should be noted that the results discussed here for the infinite-machine machine centers hold true even if the distributions of the service times are general.

5.2.1. Incorporating Transport and Material Handling in the Jackson-Type Job Shop Model

Consider the material-handling configuration where each link (i,j) that connects machine centers i and $j(i,j = 1, \dots, m; i \neq j)$ has a sufficient number of transporters or a conveyor system such that when a job completes processing at machine center i, it can immediately begin movement from machine center i to its destination. The average time taken to transfer a job from machine center i to machine center j is $1/\mu_{(i,j)}$, $i,j = 1, \dots, m$; $i \neq j$. Suppose the job transfers from one machine center to another are represented by a job transfer probability matrix $\mathbf{P} = (p_{ij})_{i,j=1,\dots,m}$. If we now incorporate the transporters on each link (i,j) as a service center, the job transfer probabilities are $p'_{ii} = p_{ii}$, $p'_{i,(i,j)} = p_{ij}$ and $p'_{(i,j),j} = 1$, $i,j = 1, \dots, m$; $i \neq j$. All other transfer probabilities are zero. Here (i,j) represents the service center corresponding to the transporters on link $(i,j), i,j = 1, \dots, m$; $i \neq j$. If we model this system by an open Jackson queueing network, as described earlier, we find that the stationary distribution of the number of jobs in the machine centers ($\mathbf{n}$), and the number of jobs in the transporters ($\mathbf{l}$), is given by

$$p(\mathbf{n},\mathbf{l}) = \left\{\prod_{i=1}^{m} p_i(n_i)\right\} \prod_{i=1}^{m} \prod_{j=1;j\neq 1}^{m} p_{(i,j)}(l_{(i,j)}), \qquad \mathbf{n} \in \mathfrak{N}_+^m, \mathbf{l} \in \mathfrak{N}_+^{m^2-m} \tag{79}$$

where $p_i(n_i) = f_i(n_i p_i(0)$, $p_i(0) = 1/\Sigma_{n_i=0}^{\infty} f_i(n_i)$, and $f_i(n_i) = \lambda_i f_i(n_i - 1)/\mu_i r_i(n_i)$, $f_i(0) = 1$; $i = 1, \dots, m$,

$$p_{(i,j)}(l_{(i,j)}) = \frac{e^{-\rho(i,j)} \rho_{(i,j)}^{l(i,j)}}{l_{(i,j)}!}, \quad i,j = 0, 1, \dots; i \neq j \tag{80}$$

and $\rho_{(i,j)} = \lambda_{(i,j)}/\mu_{(i,j)} = \lambda_i p_{ij}/\mu_{(i,j)}$, $i \neq j, i,j = 1, \dots, m$. Here n_i is the number of jobs in machine center i and $l_{(i,j)}$ is the number of jobs in transit from machine center i to j, $i \neq j$, $i,j = 1, \dots, m$. Observe that if $p_{ij} = 0$ then $\rho_{(i,j)} = 0$ and the number of jobs in transit on link (i,j) is also zero, as it should be. If we are only interested in the total number of jobs in transit, we find that

$$P\{l \text{ jobs are in transit}\} = e^{-\hat{\rho}} \hat{\rho}^l / l!, \qquad l = 0,1,\ldots \tag{81}$$

where

$$\hat{\rho} = \sum_{i=1}^{m} \lambda_i \sum_{j=1;j\neq 1}^{m} p_{ij}/\mu_{(i,j)}$$

and $\hat{\rho}$ is the mean number of jobs in transit. Applying Little's law, one sees that the additional average time spent by an arbitrary job in the shop due to material handling is

$$E[\hat{T}] = \frac{1}{\lambda} \sum_{i=1}^{m} \lambda_i \sum_{j=1;j\neq i}^{m} p_{ij}/\mu_{(i,j)} = \sum_{i=1}^{m} \sum_{j=1;j\neq 1}^{m} \frac{v_{(i,j)}}{\mu_{(i,j)}} \tag{82}$$

where $v_{(i,j)}$ is the expected number of times an arbitrary job is moved along the link (i,j), $i \neq j$; $j = 1, \dots, m$. Now observe that the distribution of the number of jobs at different machine centers given in (79) is independent of the transportation times. Hence we see that the number of jobs at different machine centers is unaffected by the actual transportation time. All it does is to add an additional amount $E[\hat{T}]$ to the average flow time of an arbitrary job.

5.3. General Service Times

The modeling of open queueing networks with general service times, is discussed in Chapter 83, Section 7.2. The basis of the approximation is to assume that the queue length distributions at the different service centers are independent, that is,

$$P\{N_i = n_i, i = 1, \ldots, m\} = \prod_{i=1}^{m} P\{\hat{N}_i = n_i\}, \qquad \mathbf{n} \in \mathfrak{N}_+^M \tag{83}$$

$$E[N] = \sum_{i=1}^{m} E[\hat{N}_i] \tag{84}$$

and

$$E[T] = \sum_{i=1}^{m} v_i E[\hat{T}_i] \tag{85}$$

where N_i is the number of jobs at machine center i, $N = \Sigma_{i=1}^{m} N_i$ is the total number of jobs in the system, and T is the flow time of an arbitrary job. v_i is the average number of visits made to a service center by an arbitrary job.

Each service center i, $i = 1, \ldots, m$ is modeled by a $GI/GI/c_i$ queue. The parameters of this queue will be $(\lambda_i, C_a^2, \mu_i, C_{s_i}^2, c_i)$, where λ_i and $C_{a_i}^2$ are the mean arrival rate and squared coefficient of variation of the time between arrivals at service center i, μ_i and $C_{S_i}^2$ are the mean service rate and squared coefficient of variation of the service time of a job at service center i. Then, using some appropriate approximation, the mean number of jobs at service center i, $\hat{n}_i(\lambda_i, C_{a_i}^2, \mu_i, C_{S_i}^2, c_i)$ and the mean flow time of a job at service center i, $\hat{T}(\lambda_i, C_{a_i}^2, \mu_i, C_{S_i}^2, c_i)$ are estimated.

Given the job transfer probability matrix, $\mathbf{P} = (p_{ij})$, then the λ_i are the solution to the equations

$$\lambda_i = \lambda\gamma_i + \sum_{j=1}^{m} \lambda_j p_{ji}, \qquad i = 1, \ldots, m$$

and $v_i = \lambda_i/\lambda$, $i = 1, \ldots, m$.

The $C_{a_i}^2$ are determined by observing that if $C_{d_j}^2$ is the squared coefficient of variation of departures from service center j, then with probabilistic routing, the stream of jobs from service center j will arrive at service center i with a squared coefficient of variation $C_{a_{j^i}}^2$, given by

$$C_{a_{ji}}^2 = 1 - p_{ji} + p_{ji} C_{d_j}^2$$

Jobs coming from outside the system will arrive in a stream with squared coefficient of variation $C_{a_{0j}}^2 = 1 - \gamma_i + \gamma_i C_a^2$. This job stream will merge with job streams coming from other service centers and from outside the system. Approximately, the squared coefficient of variation of arrivals at service center i, $C_{a_i}^2$, is given by

$$C_{a_i}^2 = \frac{1}{\lambda_i} \left\{ \sum_{j=1}^{m} \lambda_j C_{a_{ji}}^2 + \lambda\gamma_i C_{a_{0i}}^2 \right\} \tag{86}$$

5.3.1. Approximations for $\hat{n}_i$ and $C_{d_i}^2$

In implementing the decomposition approach, a variety of ways of approximating $\hat{n}_i$ and $C_{d_i}^2$ have been used. It can be shown (Buzacott and Shanthikumar 1993) that in a $GI/G/1$ queue, C_d^2 and $E[N]$ are related by

$$C_d^2 = C_a^2 + 2\rho^2 C_S^2 + 2\rho(1 - \rho) - 2(1 - \rho)E[N] \tag{87}$$

so it would seem reasonable to assume that if $c_i = 1$, the approximated $C_{d_i}^2$ and the approximation $\hat{n}_i$ are connected by the same relationship as Eq. (87).

The choice of approximation $\hat{n}_i$ is largely determined by the purpose of the modeling. If the purpose is to arrive at performance predictions that will agree closely with simulation results, then a more complex approximation is used. However, if the purpose of the approximation is to compare a number of different system designs, then simpler approximations are usually quite adequate to get the necessary insight. Whitt (1983) and Shanthikumar and Buzacott (1981) used relatively complex approximations because their purpose was to demonstrate that in many situations the approximate decomposition approach yields remarkably accurate predictions. Harrison and Nguyen (1990) suggest that, in fact, a simple heavy-traffic approximation is often quite adequate. As a general rule, the more there is some randomness in job routings the more accurate the approximations are. The QNA approximations are given in Chapter 83, so we give here are two other approaches:

1. Buzacott and Shanthikumar (1993): For $c = 1$ this sets

$$E[\hat{N}(\lambda, C_a^2, \mu, C_S^2, 1)] = \left\{\frac{\rho^2(1 + C_S^2)}{1 + \rho^2 C_S^2}\right\} \left\{\frac{C_a^2 + \rho^2 C_S^2}{2(1 - \rho)}\right\} + \rho \tag{88}$$

2. Harrison and Nguyen (1990): Assume $c = 1$. Their approximation is

$$E[\hat{N}(\lambda_i, C_{a_i}^2, \mu_i, C_{S_i}^2, 1)] = \frac{\rho_i^2(C_{a_i}^2 + C_{S_i}^2)}{2(1 - \rho_i} + \rho_i \tag{89}$$

where $\rho_i = \lambda_i/\mu_i$ and they set

$$C_{d_i}^2 = C_{S_i}^2$$

We would, however, recommend setting $C_{d_i}^2$ using Eq. (87), that is,

$$C_{d_i}^2 = (1 - \rho_i^2)C_{a_i}^2 + \rho_i^2 C_{S_i}^2 \tag{90}$$

An even simpler approximation is to set

$$E[\hat{N}(\lambda_i, C_{a_i}^2, \mu_i, C_{S_i}^2, 1)] = \frac{C_{a_i}^2 + C_{S_i}^2}{2(- \rho_i)} + \rho_i \tag{91}$$

and

$$C_{d_i}^2 = C_{S_i}^2 \tag{92}$$

This approximation is particularly useful for gaining insight into alternative ways of allocating tasks to workers in service systems, for example, should workers have a broad range of tasks in a system or should they specialize (Buzacott 1996).

6. FLEXIBLE MACHINING SYSTEMS

The key feature of a flexible machining system that must be represented in queueing models is the limited number of pallets or work holders available for holding jobs in process. This means that the number of jobs in process cannot exceed the number of pallets.

6.1. Single-Class Closed Jackson Queueing Network Model

Consider a flexible machining system consisting of m machine centers $(1, \ldots, m)$ and a load/unload station "0." There are n parts of a single (probably aggregated) class of parts that circulate from one service center to another according to a transfer probability matrix $\mathbf{P} = (p_{ij})_{i,j=0,\ldots,m}$. We will assume that all self transitions have been eliminated so that $p_{ii} = 0$, $i = 1, \ldots, m$. The service requirements of parts at service center i are i.i.d. exponential random variables with mean $1/\mu_i$, $i = 0, \ldots, m$. The sequence of service requirements at the different service centers is mutually independent. The rate at which unit service requirement of a part is processed at a machine center i when there are k parts is assumed to be $r_i(k)$, $k = 1, \ldots, n$; $i = 0, \ldots, m$. This allows us to represent, as special cases, single or multiple servers in parallel at each service center. The parts at each service center are served according to a first-come-first-served service protocol.

Let $N_i(t)$ be the number of parts at service center i at time t, $i = 0, 1, \ldots, m$ and let $\mathbf{N}(t) = \{N_0(t), \ldots, N_m(t)\}$. Then $\{\mathbf{N}(t), t \geq 0\}$ is a continuous-time Markov process on the state space S_n, where

$$\mathcal{S}_l = \{\mathbf{k} : \mathbf{k} \in \mathcal{N}_+^{m+1}, |\mathbf{k}| = l\}, \qquad l = 1, 2, \ldots$$
$$\mathcal{S}_0 = \{0, \ldots, 0\} \tag{93}$$

Define the stationary distribution of $\mathbf{N}$ by $p(\mathbf{k}) = \lim_{t\to\infty} P\{\mathbf{N}(t) = \mathbf{k}\}$, $\mathbf{k} \in S_n$. Then it can be shown that (e.g., see Gordon and Newell (1967))

$$p(\mathbf{k}) = \frac{1}{G(n)} \prod_{i=0}^{m} p_i(k_i), \qquad \mathbf{k} \in \mathcal{S}_n \tag{94}$$

where

$$p_i(k_i) = \frac{v_i^{k_i}}{\prod_{j=1}^{k_i} \mu_i r_i(j)}, \qquad k_i = 0, \ldots, n; \qquad i = 0, \ldots, m \tag{95}$$

v_i, $i = 1, \ldots, m$ is the solution to

$$v_i = \sum_{j=0}^{m} v_j p_{ji}, \; i = 1, \ldots, m \tag{96}$$

with $v_0 = 1$ and

$$G(n) = \sum_{\mathbf{k} \in \mathbb{S}_n} \prod_{i=0}^{m} \frac{v_i^{k_i}}{\prod_{j=1}^{k_i} \mu_i r_i(j)} \tag{97}$$

Let Y_i be a generic random variable representing the stationary distribution of the number of parts in an $M/M(n)/1$ queueing system with arrival rate $\lambda_i = \lambda v_i$ and state-dependent service rate $\mu_i r_i(k_i)$ when there are k_i parts in the system ($k_i = 1, 2, \ldots$). λ can be any value that guarantees the existence of a stationary distribution for the $M/M(n)/1$ queue. For example, if $r_i(k_i) = \min\{k_i, c_i\}$, that is, we have c_i parallel servers at service center i, then we require that $\lambda_i < \mu_i c_i$ or equivalently $\lambda < \mu_i c_i / v_i$. It is easily verified that

$$P\{\mathbf{N} = \mathbf{k}\} = P\{\mathbf{Y} = \mathbf{k} \mid |\mathbf{Y}| = n\}, \qquad \mathbf{k} \in \mathbb{S}_n \tag{98}$$

where $Y_0, \ldots, Y_m$ are independent. It is also clear that the distribution of $\mathbf{Y}$ is the stationary distribution of an open Jackson queueing network with a set of service centers $\{0, \ldots, m\}$, and an arbitrary part visiting service center i, on the average v_i number of times before it leaves the system ($i = 1, \ldots, m$) and the load/unload center twice ($= 2v_0$). The service rate at service center i is $\mu_i r_i(k_i)$ when there are k_i parts for $i = 1, \ldots, m$ and $2\mu_0 r_0(k_0)$ when there are k_0 parts at the load/unload service center. Note that the average load or unload times are $1/2\mu_0$ while the combined load/unload operation (incorporated as a single operation) in the closed queueing network model requires an average of $1/\mu_0$ units of time. The external part arrival rate is λ. Note that this open Jackson network could be our model for the FMS if we had free inflow of raw parts into the system at rate λ. Therefore, the distribution of the number of parts in the closed queueing network model is the same as that in an open queueing network model, provided we observe the open queueing network only when there are a total of n parts in it.

Therefore, the probability distribution of the number of parts in an open queueing network model can be used to compute the probability distribution of the number of parts in the closed queueing network. To do this, we need to compute the convolution of the probability distributions $P\{Y_i = k_i\} = P\{Y_i = 0\} \, (\lambda v_i/\mu_i)^{k_i}/\Pi_{j=1}^{k_i} r_i(j)$, $i = 0, \ldots, m$. The following algorithm will compute this convolution and the stationary distribution of the number of parts in the closed queueing network.

6.1.1. Algorithm 5: Convolution Algorithm

Step 1: Set $p_i(0) = 1$, $i = 0, \ldots, m$.

Step 2: For $i = 0, \ldots, m$ and $k = 0, \ldots, n - 1$, set

$$p_i(k + 1) = p_i(k) \frac{v_i}{\mu_i r_i(k + 1)}$$

Step 3: Set $\hat{p}(k) = p_0(k)$, $k = 0, \ldots, n$.

For $i = 1, \ldots, m$ and for $k = 0, \ldots, n$, set

$$\hat{q}(k) = \sum_{l=0}^{k} \hat{p}(l) p_i(k - l)$$

For $k = 0, \ldots, n$, set

$$\hat{p}(k) = \hat{q}(k)$$

Step 4: $p(\mathbf{k}) = 1/\hat{q}(n) \, \Pi_{i=0}^{m} p_i(k_i)$, $\mathbf{k} \in \mathbb{S}_n$.

Step 5: Stop.

Let TH(n) be the throughput rate of this closed queueing network. This is the rate at which parts are loaded/unloaded at the load/unload service center, that is, TH(n) = $E[\mu_0 r_0(N_0)]$. Using (98), we see that

$$\begin{aligned} \mathrm{TH}(n) &= E[\mu_0 r_0(Y_0) \mid |\mathbf{Y}| = n] \\ &= \sum_{\mathbf{k} \in \mathcal{S}_n} \frac{\mu_0 r_0(k_0) P\{\mathbf{Y} = \mathbf{k}\}}{P\{|\mathbf{Y}| = \boldsymbol{n}\}}. \end{aligned} \tag{99}$$

Since $\mu_0 r_0(k_0) P\{Y_0 = k_0\} = \lambda P\{Y_0 = k_0 - 1\}$, $k_0 \geq 1$ and $r_0(0) = 0$, we get from (99)

$$\begin{aligned} \mathrm{TH}(n) &= \lambda \sum_{\mathbf{k}' \in \mathcal{S}_{n-1}} \frac{P\{\mathbf{Y} = \mathbf{k}'\}}{P\{|\mathbf{Y}| = n\}} \\ &= \lambda \frac{P\{|\mathbf{Y}| = n - 1\}}{P\{|\mathbf{Y}| = n\}}. \end{aligned} \tag{100}$$

Equivalently,

$$\lambda P\{|\mathbf{Y}| = n - 1\} = \mathrm{TH}(n) P\{|\mathbf{Y}| = n\}, \qquad n = 1, 2, \ldots \tag{101}$$

Hence we see that the total number of parts in the open queueing network is the same in distribution as that in an $M/M(n)/1$ queueing system with arrival rates λ and state dependent service rates TH(n), $n = 1, 2, \ldots$ For the purpose of analyzing the aggregate behavior of the total number of parts in the system, we can aggregate the whole FMS and replace it by an equivalent single-stage server with state dependent service rates equal to the production rates.

Next we will look at another property of the closed queueing network process $\{\mathbf{N}(t), t \geq 0\}$ that will help us to efficiently compute the production capacity of the flexible machining system. Denote by $\mathbf{e}_i$ a vector that is all zeroes except for a one in position i. Let $\tau_n^{(i)}$ be the nth time epoch at which an arrival occurs at service center i. We are interested in

$$p^{(i)}(\mathbf{k}) = \lim_{n \to \infty} P\{\mathbf{N}(\tau_n^{(i)}) = \mathbf{k} + \mathbf{e}_i\}, \qquad \mathbf{k} \in \mathcal{S}_{n-1}$$

the probability distribution of number of parts at different service centers seen by an arbitrary arrival at its arrival epoch to service center i. $\mathcal{S}_{n-1}$ contains all possible states that could be seen by any part excluding itself.

The rate at which an arrival to service center i sees the system state $\mathbf{k}$ at its arrival epoch is $\sum_{j=0}^{m} p(\mathbf{k} + \mathbf{e}_j)\mu_j r_j(k_j + 1)p_{ji}$. The rate at which parts arrive at service center i is $\sum_{\mathbf{k} \in S_{n-1}} \sum_{j=0}^{m} p(\mathbf{k} + \mathbf{e}_j)\mu_j r_j(k_j + 1)p_{ji}$. Then it can be shown that

$$p^{(i)}(\mathbf{k}) = \frac{\prod_{l=0}^{m} \dfrac{v_l^{k_l}}{\prod_{j=1}^{k_l} \mu_l r_l(j)}}{\sum_{\mathbf{k}' \in \mathcal{S}_{n-1}} \prod_{l=0}^{m} \dfrac{v_l^{k_l}}{\prod_{j=1}^{k_l} \mu_l r_l(j)}}, \qquad \mathbf{k} \in \mathcal{S}_{n-1} \tag{102}$$

Observe that $p^{(i)}(\mathbf{k})$ is exactly the same as the stationary distribution of the number of parts in the closed queueing network with a total of $n - 1$ parts. Furthermore, this probability distribution is independent of the service center i. Since we will be using this relationship between the closed queueing network with n and $n - 1$ parts in it to compute the performance measures, we will use an index n to associate the total number of parts to its performance measures. Earlier, we saw how to obtain the joint probability distribution $p(\mathbf{k})$, $\mathbf{k} \in \mathcal{S}_n$. However, in most applications it may be sufficient to compute the marginal probability distribution $p_i(k_i;n)$, $k_i = 0, \ldots, n$; its average (i.e., the average number of parts at service center i), $E[N_i(n)]$, the average sojourn time of an arbitrary part in service center i, $E[T_i(n)]$, $i = 0, \ldots, m$, and the throughput TH(n). Then TH(n) = $\mu_0 E[r_0(N_0(n))]$ is the rate at which parts are being loaded/unloaded at the load/unload service center. Using a simple recursive algorithm over the values of n, we can compute these performance measures with considerably less computational effort than required by the convolution algorithm described earlier.

The rate at which parts arrive at service center i is v_iTH(n). Since the probability that an arrival at service center i sees $k_i - 1$ parts in service center i is $p_i(k_i - 1; n - 1)$, the rate of upcrossings from the compound state $\{\mathbf{k}' : \mathbf{k}' \in \mathcal{S}_n, k_i' = k_i - 1\}$ is $v_i \mathrm{TH}(n) p_i(k_i - 1; n - 1)$. Equating this to the rate of downcrossings $(\mu_i r_i(k_i) p_i(k_i;n))$ into this compound state, we get

$$p_i(k_i; n) = \frac{v_i}{\mu_i r_i(k_i)} \mathrm{TH}(n) p_i(k_i - 1; n - 1), \qquad k_i = 1, \ldots, n$$

$$p_i(0; n) = 1 - \sum_{k_i=1}^{n} p_i(k_i; n), \qquad n = 1, 2, \ldots \tag{103}$$

From (103) and Little's results we find that

$$E[N_i(n)] = \mathrm{TH}(n) \sum_{k_i=0}^{n-1} \frac{(k_i + 1)v_i}{\mu_i r_i(k_i + 1)} p_i(k_i; n - 1) \tag{104}$$

and

$$E[T_i(n)] = \sum_{k_i=0}^{n-1} \frac{k_i + 1}{\mu_i r_i(k_i + 1)} p_i(k_i; n - 1) \tag{105}$$

Since $\Sigma_{i=0}^{m} E[N_i(n)] = n$, we obtain

$$\mathrm{TH}(n) = \frac{n}{\sum_{i=0}^{m} v_i E[T_i(n)]} \tag{106}$$

Equations (103)–(106) provide a recursive relationship for $p_i(\cdot; n)$ and $\mathrm{TH}(n)$. As an initial value for this recursion observe that $p_i(0; 0) = 1$, $i = 0, \ldots, m$. Then $\mathrm{TH}(1) = 1/\Sigma_{i=0}^{m} v_i/\mu_i r_i(1)$. An algorithm to compute these performance measures using this recursion is presented next:

6.1.2. Algorithm 6: Marginal Distribution Analysis Algorithm

Step 1: Set $p_i(0; 0) = 1$, $i = 0, \ldots, m$.

Step 2: For $l = 1, \ldots, n$, compute

$$E[T_i(l)] = \sum_{k_i=0}^{l-1} \frac{k_i + 1}{\mu_i r_i(k_i + 1)} p_i(k_i; l - 1); \qquad i = 0, \ldots, m$$

$$\mathrm{TH}(l) = l \Bigg/ \left\{ \sum_{i=0}^{m} v_i E[T_i(l)] \right\}$$

$$E[N_i(l)] = v_i \mathrm{TH}(l) E[T_i(l)], \qquad i = 0, \ldots, m$$

$$p_i(k_i; l) = \frac{v_i \mathrm{TH}(l)}{\mu_i r_i(k_i)} p_i(k_i - 1; l - 1), \qquad k_i = 1, \ldots, l;\ i = 0, \ldots, m$$

$$p_i(0; l) = 1 - \sum_{k_i=1}^{l} p_i(k_i; l), \qquad i = 0, \ldots, m$$

Step 3: Stop.

When we have only a single server at a service center i (i.e., $c_i = 1$) the computational effort required for the marginal distribution analysis can be reduced by the following observation $E[T_i(n)] = \{E[N_i(n - 1)] + 1\}/\mu_i$. Particularly if each service center has only a single server (i.e. $c_i = 1$, $i = 0, \ldots, m$), the following mean value analysis algorithm provides an efficient way to compute the system performance measures.

6.1.3. Algorithm 7: Mean Value Analysis Algorithm

Step 1: Set $E[N_i(0)] = 0$, $i = 0, \ldots, m$.

Step 2: For $l = 1, \ldots, n$, compute

$$E[T_i(l)] = \{E[N_i(l - 1)] + 1\}/\mu_i, \qquad i = 0, \ldots, m$$

$$\mathrm{TH}(l) = l/\left\{ \sum_{i=0}^{m} v_i E[T_i(l)] \right\},$$

$$E[N_i(l)] = v_i \mathrm{TH}(l) E[T_i(l)], \qquad i = 0, \ldots, m$$

Step 3: Stop.

6.1.4. Properties of the Throughput

The production rate TH(n) has some useful properties in terms of the influence of the number of servers c_i at service center i, and the workload $\hat{\rho}_i = v_i/\mu_i$ assigned to service center i. To indicate this dependence, the production rate is represented by TH($\hat{\rho}$, c, n)

Property 1: TH($\hat{\rho}$, c, n) is increasing and concave in n.

Property 2: TH($\hat{\rho}$, c, n) is increasing and concave in c_i, $i = 0, \ldots, m$.

Property 3: TH($\hat{\rho}$, c, n) is decreasing and convex in each $\hat{\rho}_i$, $i = 0, \ldots, m$.

Property 4: If $c_0 = c_1 = \cdots = c_m$ then, if work load can be reallocated between service centers, the production rate is maximized by having the same work load at all service centers.

6.2. Modeling the Effects of Dedicated Material-Handling Systems

Suppose we have material-handling systems available for each link connecting any two service centers, such that no queueing delays take place during part movement. The only time expended on moving parts is the travel time. Let $1/\mu_{(i,j)}$ be the average travel time of a part on the link (i,j), $i,j = 0, \ldots, m$. Now, treating such a link (i,j) as a service center with infinitely many servers, one sees that $v_{ij} = v_i p_{ij}$ and $E[T_{ij}(n)] = 1/\mu_{(i,j)}$, $i,j = 0, \ldots, m$. Incorporating this in the MVA algorithm, we get the following algorithm to compute the system performance with material-handling effects.

6.2.1. Algorithm 8: MVA with Material Handling

Step 1: Let $E[N_i(0)] = 0$, $i = 0, \ldots, m$.

Step 2: For $l = 1, \ldots, n$, compute

$$E[T_i(l)] = \{E[N_i(l-1)] + 1\}/\mu_i, \qquad i = 0, \ldots, m$$

$$\mathrm{TH}(l) = l \Big/ \left\{ \sum_{i=0}^{m} v_i E[T_i(l)] + \sum_{i=0}^{m} \sum_{j=0}^{m} v_i p_{ij}/\mu_{(ij)} \right\}$$

$$E[N_i(l)] = v_i \mathrm{TH}(l) E[T_i(l)], \qquad i = 0, \ldots, m$$

Step 3: Stop.

6.3. General Single-Class Closed Queueing Network Model

Usually service times will not be exponential. We now present an approximation by which performance measures of closed queueing networks can be found. Because the approximation extends the mean value analysis algorithm, it can only be used if there is just one server at each machine center. For an approximate algorithm that can be used when there are multiple servers at a machine center, see Buzacott and Shanthikumar (1993). The extended MVA algorithm is based on the following: each arrival to service center i will see, at the instant of its arrival, the part in service, if any, requiring an average of $E[S_i^2]/2E[S_i]$ additional processing time to complete its service. This mimics the property of an $M/G/1$ queueing system. With this, and the assumption that an arrival with n parts in the system sees the time average behavior of a system with $n - 1$ parts, we have

$$\begin{aligned} E[T_i(n)] = {} & \{E[N_i(n-1)] - v_i \mathrm{TH}(n-1)E[S_i] + 1\}E[S_i] \\ & + v_i \mathrm{TH}(n-1)E[S_i^2]/2, \qquad i = 0, \ldots, m \end{aligned} \tag{107}$$

The above relationship is then applied in the following algorithm.

6.3.1. Algorithm 9: Extended Mean Value Analysis (EMVA)

Step 1: Set $E[N_i(0)] = 0$, $i = 0, \ldots, m$; TH(0) = 0.

Step 2: For $l = 1, \ldots, n$, compute

$$\begin{aligned} E[T_i(l)] = {} & \{E[N_i(l-1)] - v_i \mathrm{TH}(l-1)E[S_i] + 1\}E[S_i] \\ & + v_i \mathrm{TH}(l-1)E[S_i^2]/2, \qquad i = 0, \ldots, m \end{aligned}$$

$$\mathrm{TH}(l) = l \Big/ \left\{ \sum_{i=0}^{m} v_i E[T_i(l)] \right\}$$

$$E[N_i(l)] = v_i \mathrm{TH}(l) E[T_i(l)], \qquad i = 0, \ldots, m$$

Step 3: Stop.

6.4. Multiple-Class Model

An FMS is often required to process multiple part types where each part type requires a unique pallet type. Let n_s be the number of pallets for type s parts, $s = 1, \ldots, p$. Suppose the service time distribution for all part types at service center i, $i = 1, \ldots, m$ is the same for all part types and is exponential with mean $1/\mu_i$. Assume that parts are served at a service center in first-come-first-served sequence. Suppose the routing of class s parts between service centers of the FMS is described by the probability matrix $p_{ij}^{(s)}$. Let $v_i^{(s)}$ be the solution to

$$v_i^{(s)} = \sum_{j=0}^{m} v_j^{(s)} p_{ji}^{(s)} \tag{108}$$

with $v_0^{(s)} = 1$, for each $s = 1, \ldots, p$. Then $v_i^{(s)}$ is the expected number of times a class s part visits service center i during its sojourn in the system.

Then again it is possible to show that the queue length distributions are product form and is thus possible to determine the performance measures. If there is just one machine at each service center, then the mean value analysis algorithm can be adapted to find the performance measures. Define $\mathbf{n} = (n_1, n_2, \ldots, n_p)$ and $\mathbf{n} - \mathbf{e}_s = (n_1, n_2, \ldots, n_s - 1, \ldots, n_p)$. Then the expected time that a class s job spends at station i, $E[T_{i,s}(\mathbf{n})]$ and the expected number of jobs of class s at station i, $E[N_{i,s}(\mathbf{n})]$ are related by $E[T_{i,s}(\mathbf{n})] = (E[N_{i,s}(\mathbf{n} - \mathbf{e}_s)] + 1)/\mu_i$. The algorithm is as follows.

6.4.1. Algorithm 10: Multiclass MVA

Step 1: Set $E[N_{i,s}(\mathbf{l})] = 0$, $l_s \leq 0$; $s = 1, \ldots, p$; $i = 0, \ldots, m$.

Step 2: For $l_1 = 0, \ldots, n_1; \ldots; l_p = 0, \ldots, n_p$ repeat step 3.

Step 3: For $i = 0, \ldots, m$ and $s = 1, \ldots, p$, compute

$$\begin{aligned} E[T_{i,s}(\mathbf{l})] &= \{E[N_{i,s}(\mathbf{l} - \mathbf{e}_s)] + 1\}/\mu_i \\ \mathrm{TH}_s(\mathbf{l}) &= l_s \Big/ \left\{\sum_{i=0}^{m} v_i^{(s)} E[T_{i,s}(\mathbf{l})]\right\} \\ E[N_{i,s}(\mathbf{l})] &= v_i^{(s)} \mathrm{TH}_s(\mathbf{l}) E[T_{i,s}(\mathbf{l})] \end{aligned}$$

Step 4: Stop.

6.4.2. Properties of the Throughput Rate

The throughput function of the multiple-class closed Jackson queueing network possesses some properties that are, at first, somewhat surprising.

Property 1: $\mathrm{TH}_s(\mathbf{n})$ is increasing in n_s for each s ($s = 1, \ldots, p$).

Property 2: $\mathrm{TH}_l(\mathbf{n})$ *need not* increase (and may *decrease*) in n_s for $s \neq l$, $s, l = 1, \ldots, p$.

Property 3: $\mathrm{TH}(\mathbf{n}) = \Sigma_{s=1}^{p} \mathrm{TH}_s(\mathbf{n})$ *need not* increase (and may *decrease*) in n_s, $s = 1, \ldots, p$.

Properties 2 and 3 mean that the selection of parts to be processed in an FMS requires extreme care. Note that these properties are in part due to the assumption of first-come-first-served sequence at service centers, so their effect can be reduced by more sophisticated scheduling.

6.4.3. General Service Time Distributions

It is possible to heuristically adapt the multiclass MVA algorithm for situations where the service times are not exponential. Let $E[N_{i,s}(\mathbf{l})]$ be the average number of class s parts at service center i when the total number of class s parts in the network is l_s, $s = 1, \ldots, p$. Let $\mathbf{l} + \mathbf{e}_s = (l_1, \ldots, l_{s-1}, l_s + 1, l_{s+1}, \ldots, l_p)$. Also let $E[T_{i,s}(\mathbf{l})]$ be the average flow time of a class s part through service center i and $\mathrm{TH}_s(\mathbf{l})$ be the throughput rate of class s parts. Then the rate at which parts arrive at service center i is

$$\lambda_i(\mathbf{l}) = \sum_{s=1}^{p} v_i^{(s)} \mathrm{TH}_s(\mathbf{l}), \qquad i = 0, \ldots, m$$

The basic idea is to adapt the equation describing the expected time a class s part spends at station i as follows:

$$\begin{aligned} E[T_{i,s}(\mathbf{l} + \mathbf{e}_s)] &= E[S_i^{(s)}] \\ &\quad + \sum_{\hat{s}=1}^{p} (E[N_{i,\hat{s}}(\mathbf{l})] - \lambda_i^{(\hat{s})}(\mathbf{l})E[S_i^{(\hat{s})}(\mathbf{l})])E[S_i^{(\hat{s})} \\ &\quad + \lambda_i(\mathbf{l})\, \frac{E[S_i^2(\mathbf{l})]}{2}, \qquad i = 0, \ldots, m \end{aligned} \tag{109}$$

where

$$E[S_i(\mathbf{l})] = \frac{1}{\lambda_i(\mathbf{l})} \sum_{\hat{s}=1}^{p} \lambda_i^{(\hat{s})}(\mathbf{l})E[S_i^{(\hat{s})}], \qquad i = 0, \ldots, m$$

and

$$E[S_i^2(\mathbf{l})] = \frac{1}{\lambda_i(\mathbf{l})} \sum_{\hat{s})=1}^{p} \lambda_i^{(\hat{s})}(\mathbf{l})E[S_i^{(\hat{s})})^2], \qquad i = 0, \ldots, m$$

are the first and second moments of the service time of an arbitrary part at service center i. Since $\sum_{i=0}^{m} E[N_{i,s}(\mathbf{l} + \mathbf{e}_s) = \mathrm{TH}_s(\mathbf{l} + \mathbf{e}_s) \sum_{i=0}^{m} v_i^{(s)}\, E[T_{i,s}(\mathbf{l} + \mathbf{e}_s)] = l_s + 1$, one has

$$\mathrm{TH}_s(\mathbf{l} + \mathbf{e}_s) = \frac{l_s + 1}{\sum_{i=0}^{m} v_i^{(s)}E[T_{i,s}(\mathbf{l} + \mathbf{e}_s)]}. \tag{110}$$

The above equations now provide a recursive scheme to compute $\mathrm{TH}_s(\mathbf{n})$, $s = 1, \ldots, p$ starting with $\mathrm{TH}_s(\mathbf{0}) = 0$, $s = 1, \ldots, p$. The following algorithm computes the approximate performance measures.

6.4.4. Algorithm 11: Extended Multiclass MVA

Step 1: Set $E[N_{i,s}(\mathbf{l})] = 0$, $\mathrm{TH}_s(\mathbf{l}) = 0$, $-1 \le l_{\hat{s}} \le 0$, $s, \hat{s} = 1, \ldots, p$; $i = 0, \ldots, m$.

Step 2: For $l_1 = 0, \ldots, n_1$; $l_2 = 0, \ldots, n_2$; $\ldots$; $l_p = 0, \ldots, n_p$ repeat step 3.

Step 3: For $i = 0, \ldots, m$; $s = 1, \ldots, p$ and $l_s \le n_s - 1$, compute

$$\lambda_i^{(s)}(\mathbf{l}) = v_i^{(s)}\mathrm{TH}_s(\mathbf{l})$$

$$\lambda_i(\mathbf{l}) = \sum_{s=1}^{p} \lambda_i^{(s)}(\mathbf{l})$$

and $E[T_{i,s}(\mathbf{l} + \mathbf{e}_s)]$ by (109). Then compute $\mathrm{TH}_s(\mathbf{l} + \mathbf{e}_s)$ by (110) and

$$E[N_{i,s}(\mathbf{l} + \mathbf{e}_s] = v_i^{(s)}\mathrm{TH}_s(\mathbf{l} + \mathbf{e}_s)/E[T_{i,s}(\mathbf{l} + \mathbf{e}_s)]$$

Step 4: Stop.

7. PRODUCTION COORDINATION

We now consider queueing models that can be used to model the control of workflow in manufacturing and logistic systems. What is different about these models is that it is necessary to model both the flow of information and the flow of material and how information is transformed into material or parts. So although the underlying model is usually that of a multiclass open or closed queue, the models have some unusual features. We consider manufacturing and logistic systems consisting of a number m of cells or stages. We restrict our discussion to the situation where each cell produces just one type of part and where the cells are arranged in series 1, 2, . . . , m and thus part flow is $1 \rightarrow 2 \rightarrow \cdots \rightarrow m$. When cell j completes processing, it puts completed parts in store j. Store j also serves as the input store to cell $j + 1$. If cell $j + 1$ were separated geographically from cell j, then it would be necessary to add in a transport cell between j and $j + 1$. Final demand is met from store m. When store m is empty then demand can not be met immediately. We consider only the case where unmet demand is backlogged. We are interested in using queueing models to find the service level, that is, the fraction of demand met from stock p_{ND}, or the expected delay in meeting demand $E[\Delta]$. We will ignore any possibility of batching and assume that work flow and work release is controlled on a part-by-part basis.

7.1. Base Stock Control

In base stock control there are target stock levels z_j set for the store j. If a demand occurs, information about the occurrence of the demand is immediately sent to each cell. On receipt of the information, a request for a part is sent to the input store, store $j - 1$. If no parts are available, then requests would queue until a part becomes available. Once the part is removed from store $j - 1$, it is then released to the production cell for processing. On completion of processing, the completed part is put in store j. Denote by $N_j(t)$ the number of parts released to the production cell j at time t and not yet completed, $N_{j-1}^{(r)}(t)$ the number of requests for parts waiting at the input store $j - 1$, and $N_j^{(e)}(t)$ the number of parts that would be required to bring the inventory in store j up to z_j. If store j is empty and there is a backlog of waiting requests of size $N_j^{(r)}(t)$, then

$$N_j^{(e)}(t) = z_j + N_j^{(r)}(t), \qquad N_j^{(r)}(t) = 1, 2, \ldots$$

In a base stock system the immediate signaling of all demands to all cells means that

$$N_j^{(e)}(t) \equiv N_j(t) + N_{j-1}^{(r)}(t)$$

Suppose demands occur as a Poisson stream with parameter λ. Then the rate at which requests for parts will arrive at store $j - 1$, $j = 2, \ldots, m + 1$, is λ. In fact, because the requests at different stores are triggered by the same demand, these requests at different cells will be correlated. Our approximation ignores this.

The basis of the base stock model is to model the behavior of each cell and each store and to take into account the way in which they are linked.

7.1.1. Cell Model

The cell j queueing model represents the arrival of the demand at the cell, its transmittal as a request to the input store $j - 1$, the release of the part from the input store to the cell j, and the delivery of the part from the cell to the output store j. That is, we model how the arriving information on the occurrence of a demand is converted to a finished part in the output store. The queueing model consists of two queues in series, where the first queue is where the requests queue at store $j - 1$ and the second queue is where parts queue for processing in cell j. We are also interested in the arrivals of finished parts at the output store j, although parts do not queue to enter the store. Since the input stream to the first queue is Poisson with parameter λ, we will approximate its output stream of releases to the cell by a Poisson stream with parameter λ. So if the cell has single or multiple servers or some network of servers, it can be modeled as an open Jackson queueing network and its stationary queue length distribution $p\{N_j = n\}$, $n = 0, 1, \ldots$, found.

Requests that arrive at store $j - 1$ when it is not empty do not queue and the part can be immediately released to cell j. However, requests that arrive when the store is empty have to wait until a part arrives at the store from cell $j - 1$. It follows that the request queue characteristics will be determined from the inventory shortfall in the store

$$\begin{aligned} p\{N_{j-1}^{(r)} = n\} &= p\{N_{j-1}^{(e)} = n + z_{j-1}\} \qquad n = 1, 2, \ldots \\ p\{N_{j-1}^{(r)} = 0\} &= \sum_{u=0}^{z_{j-1}} p\{N_{j-1}^{(e)} = u\} \end{aligned} \tag{111}$$

Since we assume that the queue lengths at the two queues are product form, we have that

$$p\{N_j^{(e)} = n\} = \sum_{u=0}^{n} p\{N_{j-1}^{(r)} = u\}p\{N_j = n - u\}, \qquad n = 0, 1, \ldots \tag{112}$$

At store 1 we have that

$$\begin{aligned} p\{N_1^{(r)} = n\} &= p\{N_1 = n + z_1\}, \qquad n = 1, 2, \ldots \\ p\{N_1^{(r)} = 0\} &= \sum_{u=0}^{z_1} p\{N_1 = u\} \end{aligned}$$

Thus, by considering store 1, store 2, . . . , store m, in sequence using Eqs. (111) and (112) we can determine the $p\{N_j^{(r)} = n\}$, $n = 0, 1, \ldots; j = 1, 2, \ldots, m$. Hence,

$$
\begin{aligned}
p_{ND} &= p\{N_m^{(r)} = 0\} \\
E[\Delta] &= \sum_{n=0}^{\infty} np\{N_m^{(r)} = n\}
\end{aligned} \tag{113}
$$

7.1.2. *Single Server in Each Cell*

If each cell consists of a single server with exponential processing time having mean $1/\mu_j$, $j = 1, \ldots, m$ then it is possible to develop a very simple recursive calculation. Let $\rho_j = \lambda/\mu_j$, and we also assume that the ρ_j, $j = 1, \ldots, m$, are all different. We postulate then that

$$
\begin{aligned}
p\{N_j^{(r)} = n\} &= \sum_{u=1}^{j} c_u^{(j)} \rho_u^n (1 - \rho_u), \qquad n = 1, 2, \ldots; j = 1, \ldots, m \\
p\{N_j^{(r)} = 0\} &= 1 - \sum_{u=1}^{j} \rho_u c_u^{(j)}, \qquad j = 1, \ldots, m
\end{aligned}
$$

It follows that

$$
\begin{aligned}
c_u^{(j)} &= c_u^{(j-1)} \rho_u^{z_j+1} \frac{(1 - \rho_j)}{\rho_u - \rho_j}, \qquad u = 1, 2, \ldots, j - 1; j = 2, \ldots, m \\
c_j^{(j)} &= \rho_j^{z_j} \left(1 - \sum_{u=1}^{j-1} c_u^{(j-1)} \rho_u \frac{(1 - \rho_j)}{\rho_u - \rho_j}\right), \qquad j = 2, \ldots, m \\
c_1^{(1)} &= \rho_1^{z_1}
\end{aligned} \tag{114}
$$

Equations (114) define a recursive scheme by which the $c_u^{(j)}$ can be determined.

The service level measures are then given by

$$
p_{ND} = 1 - \sum_{u=1}^{m} \rho_u c_u^{(m)} \tag{115}
$$

$$
E[\Delta] = \sum_{u=1}^{m} c_u^{(m)} \frac{\rho_u}{1 - \rho_u} \tag{116}
$$

Table 4 compares the approximations with simulation results for p_{ND} and $E[\Delta]$, respectively, for a system with $z_1 = 2$ and $z_2 = 2$ for a variety of values of ρ_1 and ρ_2. The approximation tends to underestimate p_{ND} and overestimate $E[\Delta]$ because of the assumption that arrivals at the machine queue are Poisson. In fact, they are less variable than Poisson. The approximation can be extended to systems where demand is notified in advance (Buzacott et al. 1992), that is, to MRP-like systems.

7.2. Kanban Control

In kanban control, there are k_j kanban cards associated with cell j, $j = 1, \ldots, m$. A cell j kanban card waits in store j until it is triggered by the arrival at the store of a cell $j + 1$ kanban card. The cell j kanban card then moves back to store $j - 1$, where it waits until parts are available. As soon as parts are available, then the kanban card and the associated part are released to cell j for processing.

TABLE 4 p_{ND} **and** $E[\Delta]$ **for a Two-Cell Base Stock System with** $z_1 = z_2 = 2$

		ρ_1					ρ_1			
ρ_2	p_{ND}	0.3	0.5	0.7	0.9	$E[\Delta]$	0.3	0.5	0.7	0.9
0.3	Sim.	0.903	0.850	0.682	0.314	Sim.	0.045	0.122	0.657	5.63
	approx.	0.899	0.840	0.670	0.298	approx.	0.047	0.131	0.702	6.21
0.5	Sim.	0.747	0.705	0.567	0.259	Sim.	0.258	0.349	0.950	6.07
	approx.	0.739	0.687	0.544	0.240	approx.	0.264	0.375	1.016	6.66
0.7	Sim.	0.512	0.487	0.396	0.181	Sim.	1.15	1.25	1.90	7.22
	approx.	0.502	0.465	0.366	0.160	approx.	1.17	1.31	2.04	7.86
0.9	Sim.	0.195	0.188	0.155	0.071	Sim.	6.94	7.05	7.76	13.3
	approx.	0.188	0.172	0.135	0.059	approx.	7.32	7.51	8.35	14.4

From Buzacott et al. 1992. Reproduced with permission of Springer-Verlag.

On completion of processing by cell j, the finished part and the kanban move to store j, where they will wait until the next cell $j + 1$ kanban arrives. That is, movement of a cell j kanban is from store j to store $j - 1$ to cell j to store j again. Thus the queueing model of a kanban controlled system consists of a multiclass queue with $m + 1$ classes of customers. Class j customers represent the cell j kanbans, $j = 1, \ldots, m$, and class m+1 customers represent customer demands. Class j customers circulate in the closed loop: store $j \rightarrow$ store $j - 1 \rightarrow$ cell $j \rightarrow$ store j, $j = 1, \ldots, m$; while class $m + 1$ customers enter the system, go to store m, and then leave the system. Assume that customer demand is Poisson with parameter λ. It will also be assumed that the cell consists of a single machine, and service times in the cells have exponential distributions with parameter μ_j for cell j, $j = 1, \ldots, m$.

To determine the system performance, it is necessary to model the circulation of class j customers and to model the interrelationship between the service of class j customers at store j and the service of class $j + 1$ customers at store j.

7.2.1. Store Model

Consider store j. Store inventory increases when a part and its associated stage j kanban arrive from cell j. Store inventory decreases when a stage $j + 1$ kanban arrives at the store and triggers the release of a part to cell $j + 1$. The maximum backlog of unmet demands at store j is k_{j+1} so, defining $n_j^{(e)}$ as the inventory shortfall, the inventory position at store j is $k_j - n_j^{(e)}$ with $n_j^{(e)} = 0, 1, \ldots, k_j, \ldots, k_j + k_{j+1}$. For a given $n_j^{(e)}$, the quantities $n_j^{(p)}$ and $n_j^{(r)}$ are determined by

$$n_j^{(p)} = \begin{cases} k_j - n_j^{(e)} & n_j^{(e)} = 0, 1, \ldots, k_j \\ 0 & n_j^{(e)} = k_j + 1, \ldots, k_j + k_{j+1} \end{cases}$$

and

$$n_j^{(r)} = \begin{cases} 0 & n_j^{(e)} = 0, 1, \ldots, k_j \\ n_j^{(e)} - k_j & n_j^{(e)} = k_j + 1, \ldots, k_j + k_{j+1} \end{cases}$$

Note that $n_j^{(r)} \times n_j^{(p)} \equiv 0$. $n_j^{(r)}$ is the number of stage $j + 1$ kanbans waiting at store j, while $n_j^{(p)}$ is the number of stage j kanbans waiting at the store (which is identical to the number of parts in the store).

Let $\lambda_{j:u}(n_j^{(p)})$ be the rate at which stage j kanbans arrive at store j from cell j when there are $n_j^{(p)}$ stage j kanbans at store j. Similarly, let $\lambda_{j:d}(n_j^{(r)})$ be the rate at which stage $j + 1$ kanbans arrive at store j from store $j + 1$ when there are $n_j^{(r)}$ stage $j + 1$ kanbans at store j.

It follows that, for a given $n_j^{(e)}$ such that $0 \leq n_j^{(e)} \leq k_j - 1$, the rate of increase to $n_j^{(e)} + 1$ is $\lambda_{j:d}(0)$ while, for $n_j^{(e)}$ such that $1 \leq n_j^{(e)} \leq k_j$, the rate of decrease to $n_j^{(e)} - 1$ is $\lambda_{j:u}(k_j - n_j^{(e)})$. Similarly, for $n_j^{(e)}$ such that $k_j \leq n_j^{(e)} \leq k_j + k_{j+1} - 1$, the rate of increase to $n_j^{(e)} + 1$ is $\lambda_{j:d}(n_j^{(e)} - k_j)$, and, for $n_j^{(e)}$ such that $k_j + 1 \leq n_j^{(e)} \leq k_j + k_{j+1}$, the rate of decrease is $\lambda_{j:u}(0)$.

It follows that the probability distribution of $n_j^{(e)}$ is given by

$$p_j(n_j^{(e)}) = \begin{cases} \dfrac{\prod_{v=n_j^{(e)}}^{k_j-1} \lambda_{j:u}(k_j - v - 1)}{\lambda_{j:d}(0)^{k_j - n_j^{(e)}}} p_j(k_j) & n_j^{(e)} = 0, 1, \ldots, k_j - 1 \\ \dfrac{\prod_{v=0}^{n_j^{(e)} - k_j b - 1} \lambda_{j:d}(v)}{\lambda_{j:u}(0)^{n_j^{(e)} - k_j}} p_j(k_j) & n_j^{(e)} = k_j + 1, \ldots, k_j + k_{j+1} \end{cases} \tag{117}$$

and $p_j(k_j)$ is determined by $\sum_{n_j^{(e)}=0}^{k_j+k_{j+1}} p_j(n_j^{(e)}) = 1$.

When a stage $j + 1$ kanban arrives at store j from store $j + 1$ then it will find no other stage $j + 1$ kanbans waiting if, just prior to its arrival, $n_j^{(e)} \leq k_j$. However, even though it finds no stage $j + 1$ kanbans waiting, it will have to wait for a part if, just prior to its arrival, $n_j^{(e)} = k_j$. Hence $q_{j:u}$, the probability that an arriving stage $j + 1$ kanban finds no other stage $j + 1$ kanban waiting yet has to wait itself is given by

$$q_{j:u} = p_j^+\{n_j^{(e)} = k_j | 0 \leq n_j^{(e)} \leq k_j\}$$

where p_j^+ denotes the probabilities at instants of arrival of stage $j + 1$ kanbans. We assume that $p_j^+(n_j^{(e)}) = p_j(n_j^{(e)})$ and set

$$q_{j:u} = \frac{p_j(n_j^{(e)} = k_j)}{\sum_{u=0}^{k_j} p_j(n_j^{(e)} = u)}$$

Once there are waiting stage $j + 1$ kanbans $n_j^{(e)} > k_j$, so the rate at which they will be served will be given by $\mu_{j:u} = \lambda_{j:u}(0)$, irrespective of the number of waiting stage $j + 1$ kanbans.

Similarly, considering the arrivals of stage j kanbans at store j from cell j, the probability $q_{j:d}$ that an arriving stage j kanban finds no other stage j kanban waiting, yet has to wait itself, is given by

$$q_{j:d} = p_j^+\{n_j^{(e)} = k_j | k_j \leq n_j^{(e)} \leq k_{j+1}\}$$

where p_j^+ denotes the probabilities at the instants of arrival of stage j kanbans. Again we assume that $p_j^+(n_j^{(e)}) = p_j(n_j^{(e)})$ and so

$$q_{j:d} = \frac{p_j(n_j^{(e)} = k_j)}{\sum_{u=k_j}^{k_j+k_{j+1}} p_j(n_j^{(e)} = u)}$$

Since $n_j^{(e)} < k_j$ if there are waiting stage j kanbans, the rate at which waiting stage j kanbans will be served will be given by $\mu_{j:d} = \lambda_{j:d}(0)$, irrespective of the number of waiting stage j kanbans.

7.2.2. *Cell Model*

Assume that the cell consists of a single server who serves customers at rate μ_j. Alternatively, with minor change in the analysis, the cell can be any open queueing network with a product form solution. Then the closed queueing network representing the circulation of stage j kanbans (or class j customers) will have a product form solution with

$$p(n_j, n_j^{(p)}, n_{j-1}^{(r)}) = K. \left(\frac{1}{\mu_j}\right)^{n_j} \left(\frac{1}{\mu_{j:d}}\right)^{n_j^{(p)}} \left(\frac{1}{\mu_{j-1:u}}\right)^{n_{j-1}^{(r)}} f_p(n_j^{(p)}) f_r(n_{j-1}^{(r)}), \qquad n_j + n_j^{(p)} + n_{j-1}^{(r)} = k_j \tag{118}$$

and

$$f_p(\ell) = \begin{cases} 1, & \ell = 0 \\ q_{j:d} & \ell = 1, \ldots, k_j \end{cases}$$

$$f_r(\ell) = \begin{cases} 1, & \ell = 0 \\ q_{j-1:u} & \ell = 1, \ldots, k_j \end{cases}$$

and K is a normalizing constant

Class $m + 1$ customers only queue at store m, so we have

$$p(n_m^{(r)}) = K \left(\frac{1}{\mu_{m:u}}\right)^{n_m^{(r)}} f_r(n_m^{(r)}), \qquad n_m^{(r)} = 0, 1, \ldots \tag{119}$$

Also note that $\mu_{m:d} = \lambda$.

7.2.3. *Connection between Store Model and Cell Model*

Lastly, it is necessary to find the $\lambda_{j:u}(n_j^{(p)})$ and the $\lambda_{j:d}(n_j^{(r)})$. To find $\lambda_{j:u}(n_j^{(p)})$ we determine the throughput of the two service center closed queue consisting of store $j - 1$ and cell j. Hence,

$$\lambda_{jw:u}(n_j^{(\mathbf{p})}) = \frac{G_u^*(k_j - n_j^{(p)} - 1)}{G_u^*(k_j - n_j^{(p)})}, \qquad n_j^{(p)} = 0, 1, \ldots, k_j - 1 \tag{120}$$

where

$$G_u^*(\ell) = \sum_{n_{j-1}^{(r)}=0}^{\ell} \left(\frac{1}{\mu_j}\right)^{\ell - n_{j-1}^{(r)}} \left(\frac{1}{\mu_{j-1:d}}\right)^{n_{j-1}^{(r)}} f_r(n_{j-1}^{(r)})$$

Similarly, by determining the throughput in the closed queue consisting of cell $j + 1$ and store $j + 1$, we have

$$\lambda_{j:d}(n_j^{(r)}) = \frac{G_d^*(k_{j+1} - n_j^{(r)} - 1)}{G_d^*(k_{j+1} - n_j^{(r)})}, \qquad n_j^{(r)} = 0, 1, \ldots, k_{j+1} - 1 \tag{121}$$

where

TABLE 5 Comparison of Approximation and Simulation for p_{ND} for a Kanban System with $k_1 = k_2 = 2$

			ρ_1				
			0.1	0.3	0.5	0.7	0.9
ρ_2	0.1	Approx.	0.990	0.982	0.937	0.796	0.447
		sim.	0.990	0.981	0.927	0.750	0.350
	0.3	Approx.	0.910	0.898	0.846	0.707	0.378
		sim.	0.911	0.902	0.847	0.677	0.301
	0.5	Approx.	0.750	0.737	0.678	0.530	0.201
		sim.	0.752	0.744	0.692	0.532	0.182
	0.7	Approx.	0.510	0.496	0.435	0.282	***
		sim.	0.515	0.507	0.454	0.300	***
	0.9	Approx.	0.190	0.177	0.117	***	***
		sim.	0.196	0.187	0.131	***	***

***: unstable.

From Buzacott and Shantikumar © 1993. Reprinted by permission of Prentice-Hall Inc., Upper Saddle River, NJ.

$$G_d^*(\ell) = \sum_{n_{j+1}^{(p)}=0}^{\ell} \left(\frac{1}{\mu_{j+1}}\right)^{\ell-n_{j+1}^{(p)}} \left(\frac{1}{\mu_{j+1:d}}\right)^{\mathbf{n}_{j+1}^{(p)}} f_p(n_{j+1}^{(p)})$$

A recursive scheme can then be developed to solve the equations. Set all $\lambda_{j:d}(u) = \lambda$, then begin by observing that all $\lambda_{1:u}(v) = \mu_1$. Determine $q_{1:u}$, then $\lambda_{2:u}(v)$, $\mathrm{q}_{2:u}$ and so on, until stage $m + 1$ is reached. Then, for stage m, determine $q_{m:d\}\text{ and }\lambda m:d}(w)$, then $q_{m-1:d}$, $\lambda_{m-1:d}(w)$, and so on. Repeat this downstream and upstream iterative process until the parameters converge.

7.2.4. Performance Measures

Define $\rho_{mu} = \lambda/\mu_{m:u}$. Then

$$p_{ND} = \frac{(1 - q_{m:u})(1 - \rho_{mu})}{1 - (1 - q_{m:u})\rho_{mu}} \tag{122}$$

and

$$\lambda E[\Delta] = \frac{\rho_{mu}}{1 - \rho_{mu}}(1 - p_{ND}) \tag{123}$$

Tables 5 and 6 compare the performance measures calculated using this approximation for a two-stage system with $k_1 = k_2 = 2$ for a variety of different values of ρ_1 and ρ_2.

TABLE 6 Comparison of Approximation and Simulation for $E[\Delta]$ for a Kanban System with $k_1 = k_2 = 2$

			ρ_1				
			0.1	0.3	0.5	0.7	0.9
ρ_2	0.1	Approx.	0.001	0.003	0.029	0.242	2.20
		sim.	0.001	0.006	0.067	0.549	5.418
	0.3	Approx.	0.039	0.047	0.102	0.376	2.52
		sim.	0.039	0.046	0.127	0.689	6.064
	0.5	Approx.	0.251	0.276	0.419	1.02	6.73
		sim.	0.251	0.266	0.402	1.23	10.5
	0.7	Approx.	1.14	1.22	1.68	3.90	***
		sim.	1.14	1.19	1.55	3.70	***
	0.9	Approx.	7.30	7.97	13.49	***	***
		sim.	6.94	7.45	11.4	***	***

***: unstable

From Buzacott and Shantikumar © 1993. Reprinted by permission of Prentice-Hall Inc., Upper Saddle River, NJ.

CONCLUSIONS

This chapter has focused on using queueing models for performance analysis. It is largely based on the authors' book (Buzacott and Shanthikumar 1993). Other books focusing on performance models of manufacturing systems are Viswanadham and Narahari (1992) and Altiok (1997). Models can also be used for determining the optimal control and scheduling of manufacturing and service systems. See Gershwin (1994) and the papers in Yin and Zhang (1996) for numerous examples.

Acknowledgements

John Buzacott's research was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Altiok, T. (1997), *Performance Analysis of Manufacturing Systems,* Springer, New York.

Buzacott, J. A. (1996), "Commonalities in Reengineered Business Processes," *Management Science,* Vol. 42, pp. 768–782.

Buzacott, J. A. (2000), "Service System Structure," *International Journal of Production Economics*, Vol. 68, No. 1, pp. 15–27.

Buzacott, J. A., and Shanthikumar, J. G. (1994), "Safety Stock versus Safety Time in MRP Controlled Production Systems," *Management Science,* Vol. 40, pp. 1678–1689.

Buzacott, J. A., and Shanthikumar, J. G. (1993), *Stochastic Models of Manufacturing Systems,* Prentice Hall, Englewood Cliffs, NJ.

Buzacott, J. A., Liu, X.-G., and Shanthikumar, J. G. (1995), "Multistage Flow Line Analysis with the Stopped Arrival Queue Model," *IIE Transactions,* Vol. 27, pp. 444–455.

Buzacott, J. A., Price, S. M., and Shanthikumar, J. G. (1992), "Service Level in Multistage MRP and Base Stock Controlled Production Systems," in *New Directions for Operations Research in Manufacturing,* G. Fandel, T. Gulledge, and A. Jones, Eds., Springer, Berlin, pp. 445–463.

Daley, D. J., Kreinin, A. Y., and Trengove, C. D. (1992), "Inequalities Concerning the Waiting Time in Single-Server Queues: A Survey," in *Queueing and Related Models,* U. N. Bhat and I. V. Basawa, Eds., Clarendon Press, Oxford, pp. 177–223.

Gershwin, S. B. (1994), *Manufacturing Systems Engineering,* Prentice Hall, Englewood Cliffs, NJ.

Gordon, W. J., and Newell, G. F. (1967), "Closed Queueing Systems with Exponential Servers," *Operations Research,* Vol. 15, pp. 254–265.

Harrison, J. M., and Nguyen, V. (1990), "The QNET Method for Two-Moment Analysis of Open Queueing Networks," *QUESTA,* Vol. 6, pp. 1–32.

Jackson, J. R. (1963), "Job-Shop-Like Queueing Systems," *Management Science,* Vol. 10, pp. 131–142.

Schmenner, R. W. (1986), "How Can Services Businesses Survive and Prosper?," *Sloan Management Review,* Vol. 27, pp. 21–32.

Shanthikumar, J. G., and Buzacott, J. A. (1981), "Open Queueing Network Models of Dynamic Job Shops," *International Journal of Production Research,* Vol. 19, pp. 255–266.

Silvestro, R., Fitzgerald, L., Johnston, R., and Voss, C. (1992), "Towards a Classification of Service Processes," *International Journal of Service Industry Management,* Vol. 3, pp. 62–75.

Viswanadham, N., and Narahari, Y. (1992), *Performance Modeling of Automated Manufacturing Systems,* Prentice Hall, Englewood Cliffs, NJ.

Whitt, W. (1983), "The Queueing Network Analyser," *Bell System Technical Journal,* Vol. 62, pp. 2779–2815.

Yin, G. G., and Zhang, Q., Eds. (1996), *Mathematics of Stochastic Manufacturing Systems,* American Mathematical Society, Providence, RI.

CHAPTER 61

Production-Inventory Systems

DAVID D. YAO
Columbia University

1. **OVERVIEW OF CLASSICAL MODELS** **1670**
    - 1.1. The EOQ Model 1670
    - 1.2. The Newsvendor Model 1670
    - 1.3. Deterministic Multiperiod Model 1671
    - 1.4. Other Standard Models 1671
2. **BASE-STOCK CONTROL** **1672**
    - 2.1. The Inventory-Queue Model 1672
    - 2.2. Normal Approximations 1673
    - 2.3. Demand and Demand over Lead Times 1674
3. **NONSTATIONARY DEMAND: THE DRP FRAMEWORK** **1675**
4. **LOT-FOR-LOT POLICY** **1676**
5. **(S, s) POLICY** **1678**
    - 5.1. Stationary Analysis 1678
    - 5.2. Approximations 1680
6. **DYNAMIC (S, s) POLICY** **1681**
    - 6.1. Setting Safety-Stock Levels 1683
7. **MULTISTAGE MODELS** **1683**
8. **ASSEMBLE-TO-ORDER SYSTEMS** **1685**
9. **KANBAN CONTROL** **1689**
    - 9.1. The Basic Model 1689
    - 9.2. Generalized Kanban Control 1689
10. **A NETWORK OF INVENTORY QUEUES** **1690**
11. **BIBLIOGRAPHICAL NOTES** **1692**

REFERENCES **1693**

Production-inventory systems are one of the most established subjects in industrial engineering. The focus is on studying inventory dynamics, with inventory viewed as a buffer between supply (production/replenishment) and customer demand. Hence the emphasis is really more on inventory than on production, the latter being the primary focus of other IE subjects, such as scheduling and production planning.

Classical textbook models of production-inventory systems focus on the issue of optimality, in particular, the optimality of simple policies that are characterized by reorder points and order sizes. More recent studies have shifted the emphasis to the inventory–service trade-off (e.g., Cheng et al. 2000; Ettl et al. 2000; Glasserman and Wang 1998; Li 1992; Zipkin 2000) and to topics that are more relevant to industrial applications, such as kanban (e.g., Buzacott and Shanthikumar 1993; Glasserman and Yao 1994b, 1996) and supply chain management (e.g., Cheng et al. 2000; Ettl et al. 2000; Lee and Billington 1986). These topics are the emphasis of this chapter as well.

We start with an overview in Section 1 of several classical models, widely available in textbooks (e.g., Nahmias 1997; Silver et al. 1998). The rest of the chapter has two parts. The first part, Sections 2, 8, and 9, relates several familiar inventory control mechanisms—base-stock control, kanban control, and assemble-to-order systems—to queueing models, emphasizing steady-state results. The theme of the second part, Sections 3, 4, 5, 6, and 7, is nonstationarity, and we use a set of DRP (distribution resource planning)-like recursions as the basic model for inventory dynamics. Finally, in Section 10, the two parts converge into a unified, decomposition-based approach for modeling a supply network. Brief bibliographical notes are given in Section 11.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. OVERVIEW OF CLASSICAL MODELS

1.1. The EOQ Model

EOQ stands for economic order quantity. As its name suggests, the EOQ model emphasizes the role of inventory in achieving economies of scale—produce or replenish in batches, rather than single units. Demand is assumed to be deterministic, with rate λ. There is a fixed setup cost of C dollars for placing each replenishment order. Order lead times are zero, which, along with deterministic demand, implies that no order need be placed until the inventory level drops down to zero. The inventory holding cost is charged at a rate of h, that is, for each dollar of inventory that is kept for one unit of time, the charge is h dollars.

Suppose the (batch) size of each order is Q, the decision variable. Then the inventory level starts from the highest point of Q, immediately after an order is placed (and then received instantaneously), and then goes down to zero—depleted by demand, at rate λ. This cycle then repeats itself over the entire time horizon. Hence the average inventory level is $Q/2$ over each cycle and hence over the entire horizon as well.

The total cost—setup cost plus holding cost—per time unit is $\lambda C/Q + hQ/2$, where λ/Q is the number of orders placed per time unit. Note that we have ignored the variable cost, the cost for purchasing the units, since this term is independent of the decision variable Q. It is equal to $c\lambda$ per time unit, with c being the purchasing cost per unit. The Q that minimizes this cost objective is $Q := \sqrt{2\lambda C/h}$, which is easily derived from setting the derivative of the cost objective (with respect to Q) to zero. The square-root order quantity is often referred to as EOQ as well.

1.2. The Newsvendor Model

This is a single-period problem. Demand, D, is random, with a distribution function $F(x)$ that is known at the beginning of the period. The actual realization of the demand will not be known until the end of the period. The problem is to decide the order quantity Q at the beginning of the period, under the following cost assumptions: Each unit of demand supplied earns a profit (selling price minus cost) of p, each unit of unmet demand incurs a penalty of π, and each surplus (i.e., unsold) unit at the end of the period carries a net loss of ℓ (i.e., cost minus any salvage value). The objective is to maximize the expected net profit:

$$\max_Q \, [p\mathsf{E}(D \wedge Q) - \pi\mathsf{E}(D - Q)^+ - \ell\mathsf{E}(Q - D)^+]$$

where $\wedge$ denotes the min operator and $[x]^+ := \max\{x, 0\}$. The above objective simplifies to

$$\max_Q \, [(p + \pi)Q - (p + \pi + \ell)\mathsf{E}(Q - D)^+]$$

making use of the following identities:

$$D \wedge Q = Q - (Q - D)^+, \quad \text{and} \quad (D - Q)^+ = D - Q + (Q - D)^+$$

and ignoring the term $\pi\mathsf{E}(D)$, which is independent of Q. Hence, setting the derivative of the objective function (with respect to Q) to zero yields

$$F(Q^*) = \frac{p + \pi}{p + \pi + \ell}$$

noticing that

$$\begin{aligned}
&\frac{d}{dQ}\mathsf{E}(Q - D)^+ \\
&= \frac{d}{dQ}\int_0^Q (Q - x)dF(x) \\
&= \frac{d}{dQ}[QF(Q) - \int_0^Q xdF(x) \\
&= F(Q) + QF'(Q) - QF'(Q) \\
&= F(Q)
\end{aligned}$$

1.3. Deterministic Multiperiod Model

Let $k = 1, \ldots, n$ index the periods; let d_k denote the demand in period k. Demand in all n periods is deterministically known.

A production or replenishment can be ordered at the beginning of each period; the decision is the lot size—how many subsequent periods' demand this order should cover. For instance,we can decide that an order placed at the beginning of period i to have a lot size of $d_i + \ldots + d_j$, so as to cover the demand in period i through period j, for some $j > i$. That is, we run a single production at the beginning of period i, to supply the demand in all subsequent periods up to j. The cost associated with this decision is:

$$c_{ij} = C + h[d_{i+1} + 2d_{i+2} + \ldots + (j - i)d_j]$$

where C is the fixed setup cost and h is the inventory holding cost per unit of inventory, per period—applied to period-end inventory only.

Note that, as in the EOQ model, there is zero lead time and the variable cost is not included in c_{ij}, as it is independent of the lot-sizing decisions. Also, no back order is allowed, so that each period's demand must be supplied either by production at the beginning of that period or by inventory carried over from the previous periods.

This model can be solved by dynamic programming (DP). Let V_k be the optimal cost-to-go in period k, that is, the total cost to supply the demand in periods k through period n, following the optimal lot-sizing decisions. Then V_1 is the desired solution, which we can derive through the DP recursion as follows.

Clearly, in period n, where there is only a single period left, the only possible solution is to place an order of d_n units to supply the demand; hence, $V_n = C$. Recursively, for period $i = n - 1, \ldots, 2, 1$, we have

$$V_i = \min_{j \geq i} \{c_{ij} + V_{j+1}\}$$

following the DP principle of optimality. That is, the decision in period i is to pick a future period j—that is, to make a lot-sizing decision on the production so as to cover the demand in period i through period j—so as to minimize the right-hand side of the above recursion. Once j is selected, the remaining problem (i.e., the one that starts from period $j + 1$) was already solved in the earlier stages of the DP recursion, that is, to follow the optimal decision embodied in V_{j+1}.

1.4. Other Standard Models

When demand is random, a standard model is to assume that the demand stream follows a renewal process, that is, $\{D_n\}$ are independent and identically distributed (i.i.d.) random variables, where D_n denotes demand in period n. Suppose each production/replenishment order requires a lead time, denoted L, which can be constant or random.

In this context, in addition to the order size, we need a second parameter, the *reorder point.* Associated with the latter is the notion of *inventory position,* defined as the on-hand inventory minus any back orders plus any outstanding orders—orders that have been placed but have not yet arrived due to the lead time. Hence, whenever the inventory position falls below the reorder point, an order is placed. Note that since demand now fluctuates randomly and there is a lead time between placing and receiving an order, it is not desirable, in general, to order only when the inventory drops down to zero, as in the case of the EOQ model.

The above describes exactly how the so-called (Q, R) model works, with R being the reorder point and Q the order size. One way to set these parameters is to let Q take the EOQ value and let R be determined by service requirements. For instance, set the value of R sufficiently large so as to ensure that the stockout probability (i.e., the probability that the on-hand inventory is zero upon a demand arrival) is limited to, say, no more than 5%, or the fill rate (the proportion of demand that is filled from on-hand inventory) is at least 95%. (Note that in general the no-stockout probability is not equal to the fill rate.) It is also possible to set up a cost objective and then optimize it to derive the best Q and R values jointly.

Closely related, but more general, is the (S, s) model, which works as follows: whenever the inventory position falls below the lower threshold, s, place an order to bring the inventory position to the upper threshold S. When demand comes in batches, typically there will be an undershoot (of random size) when the inventory position falls below s. Hence, in the (S, s) model, the order size is random, in contrast to the constant order size in the (Q, R) model.

2. BASE-STOCK CONTROL

A widely studied inventory control mechanism is *base-stock control*, which works as follows: whenever the inventory position drops below R, a constant parameter, place an order to bring the inventory position back to R. This way, every demand will trigger the placement of an order, such that the overall inventory position is always maintained at a constant level, R, which is called the *base-stock level*.

In the case when demand always comes in unit size, then the base-stock control is a special case of the (Q, R) model, with $Q = 1$. Hence, the base-stock control is sometimes also referred to as the one-for-one replenishment rule.

2.1. The Inventory-Queue Model

The base-stock control model relates to a queueing model as follows. Consider the case of demand in single units mentioned above. (Batch demand will be discussed below in Section 2.3.) Let each unit of demand correspond to a job arrival to the queue. Each unit of demand, on arrival, is supplied by the on-hand inventory or, in the stock-out situation, joins a back-order queue. Regardless, however, a replenishment/production order is triggered.

Let the outstanding orders correspond to the jobs in the queueing system. In the inventory system, each of these orders will materialize (or ''arrive''—become available to supply demand) after a lead time. This corresponds to a queueing system with an infinite number of servers, so that any job will be served immediately on arrival. Hence, the overall cycle time of each job in the system is simply its service time, which corresponds to the lead time of orders in the inventory system.

Let N be the number of jobs in the queueing system in steady state. This, as explained above, represents the number of outstanding orders. If $N < R$, then there are $R - N$ units of on-hand inventory. If $N > R$, then we know there are $N - R$ units of back orders: R units of demand have been supplied with on-hand inventory, while the remaining $N - R$ units are back ordered. Hence, with I and B denoting the on-hand inventory and the back orders, respectively, and with $[x]^+$ denoting $\max\{x, 0\}$, we have

$$I = [R - N]^+, \quad B = [N - R]^+ \tag{1}$$

Note that the above implies

$$I - B = R - N$$

and $I \cdot B = 0$, that is, I and B cannot both be positive.

Suppose demand follows a Poisson process with rate λ, and suppose the lead time is L, the time it takes to process and finish an order. Then the queueing system in question is an $M/G/\infty$ model (see, e.g., Wolff 1989), and it is known that N follows a Poisson distribution with mean $\rho := \lambda \mathsf{E}(L)$. That is,

$$\mathsf{P}[N = n] = \frac{\rho^n}{n!} e^{-\rho}, \quad n = 0, 1, 2, \ldots$$

Combining the above with (1), we can derive the distributions of the on-hand inventory and back orders:

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = 1 - \sum_{n=0}^{R-1} \frac{\rho^n}{n!} e^{-\rho}$$

$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \frac{\rho^{R-n}}{(R - n)!} e^{-\rho}, \quad n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = \sum_{n=0}^{R} \frac{\rho^n}{\mathrm{n}!} e^{-\rho},$$

$$\mathsf{P}[B = n] = \mathsf{P}[N = R + n] = \frac{\rho^{R+n}}{(R + n)!} e^{-\rho}, \quad n = 1, 2, \ldots .$$

Next, suppose that due to limited production capacity, replenishment/production orders may have to wait in queue before they can be processed. In other words, in addition to the lead time (i.e., processing time or "service time"), there is also queueing time. In this case, a queueing model with a finite number of servers is more appropriate. Suppose we use the $M/M/1$ model for the corresponding queueing system. Then, N follows a geometric distribution:

$$\mathsf{P}[N = n] = \rho^n(1 - \rho), \qquad n = 0, 1, 2, \ldots$$

The distributions of I and B in the case are:

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \rho^R,$$
$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \rho^{R-n}(1 - \rho), \quad n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = 1 - \rho^{R+1}$$
$$\mathsf{P}[B = n] = \mathsf{P}\ [N = R + n] = \rho^{R+n}(1 - \rho), \quad n = 1, 2, \ldots$$

Note from (1) that in either case, while the on-hand inventory is limited to $I \leq R$, the base-stock level, there is no upper limit on the number of back orders, B.

2.2. Normal Approximations

We can also approximate the above Poisson distribution by a normal distribution, and write

$$N = \rho + Z\sqrt{\rho}$$

where Z denotes the standard normal variate, with density function $\phi(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ and distribution function $\Phi(x)$. Below we shall also use $\overline{\Phi}(x) := 1 - \Phi(x)$. We can also write

$$R = \rho + k\sqrt{\rho}$$

where k is known as the "safety factor" in the inventory literature.

In fact, the normal approximation applies more generally; it does not have to be restricted to N following a Poisson distribution as in the inventory-queue models of Section 2.1. We can start with writing

$$N = \mu + \sigma Z \tag{2}$$

where $\mu := \mathsf{E}(N)$ and $\sigma := \mathsf{sd}(N)$; and then write

$$R = \mu + k\sigma \tag{3}$$

accordingly.

This way, we have

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \mathsf{P}[Z \geq k] = \overline{\Phi}(k) \tag{4}$$

which is the stockout probability.

To derive $\mathsf{E}(I)$ and $\mathsf{E}(B)$, we first define the following two functions:

$$G(x) := \mathsf{E}[Z - x]^+ = \int_x^\infty (z - x)\phi(z)\, dz \tag{5}$$

and

$$H(x) := \mathsf{E}(x - Z)^+ = x + G(x) = \mathsf{E}[x - Z + (Z - x)^+] = x + \mathsf{E}(Z - x)^+ = x + G(x) \tag{6}$$

We summarize below the properties of the functions $G(x)$ and $H(x)$. These properties are easily verified from first principle.

1. $G(x) = \phi(x) - x\overline{\Phi}(x)$ (note that $\phi'(x) = -x\phi(x)$); and $H(x) = \phi(x) + x\Phi(x)$. $G'(x) = -\overline{\Phi}(x)$, $H'(x) = \Phi(x)$, and $G''(x) = H''(x) = \phi(x)$.
2. For all $x \in (-\infty, +\infty)$, $G(x) \geq 0$ and $H(x) \geq 0$; $G(x)$ is decreasing and $H(x)$ increasing, and

both are convex. (Throughout, we use “increasing” and “decreasing” in the nonstrict sense.) Note in particular that $(x)^+$ is increasing and convex in x.

3. For $x >> 1$, $G(x) \cong 0$ and $G(-x) \cong x$, whereas $H(x) \cong x$ and $H(-x) \cong 0$. In fact, these approximations work very well for $x \geq 2$; for instance, $G(-2) = 2.0085$ and $G(2) = 0.0085$.

Hence, incorporating (2) and (3) with (1) and making use of the G and H functions, we can derive:

$$\mathsf{E}(I) = \sigma \mathsf{E}[k - Z]^+ = \sigma H(k), \quad \text{and} \quad \mathsf{E}(B) = \sigma \mathsf{E}[Z - k]^+ = \sigma G(k) \tag{7}$$

2.3. Demand and Demand over Lead Time

Let $D(t)$ be the demand in period t, $t = 1, 2, \ldots$ Suppose demand (per period) over time is independent and identically distributed. Let L denote the lead time to fill each replenishment order. The number of outstanding orders, as explained in Section 2.1, is equal to the number of jobs, N, in an infinite-server queueing system. In particular, if the per-period demand follows a Poisson distribution, then N also follows a Poisson distribution with mean $\mathsf{E}(N) = \mathsf{E}(D) \cdot \mathsf{E}(L)$ ($= \rho$ in Section 2.1; here D denotes the generic per-period demand). Since N follows a Poisson distribution, we know $\text{Var}(N) = \mathsf{E}(N)$.

On the other hand, the total demand over the lead time,

$$D(1, L) := D(1) + \mathrm{D}(2) + \ldots + D(L)$$

has a mean

$$\mathsf{E}D(1, L) = \mathsf{E}(D) \cdot \mathsf{E}(L)$$

and a variance

$$\mathsf{Var}D(1, L) = \mathsf{Var}(D) \cdot \mathsf{E}(L) + \mathsf{E}^2(D) \cdot \mathsf{Var}(L)$$

Hence, while the lead time demand has the same mean as $\mathsf{E}(N)$, its variance is different, unless the lead time is deterministic and D follows a Poisson distribution. This distinction is important, since it is N, *not* demand over lead time, that determines the inventory performance, in terms of on-hand inventory (I) and back orders (B) via the relations in (1).

In applications, it is often more convenient to model the demand as a *batch* Poisson arrival process. Let λ be the demand arrival rate, as in Section 2.1, and let X denote the batch size associated with each arrived demand. Assume the batches are i.i.d. and independent of the demand arrival process, with $\mathsf{E}(X) \geq 1$. Then the per-period (or per-time unit) demand has the following mean and variance:

$$\mathsf{E}(D) = \lambda \mathsf{E}(X), \quad \text{and} \quad \mathsf{Var}(D) = \lambda \mathsf{E}(X^2)$$

The main advantage of this demand model is that it has (at least) three parameters: the arrival rate λ and the first two moments of the batch size X; whereas the Poisson demand model has only one parameter (the mean, or arrival rate).

Accordingly, the queue model $M/G/\infty$ of Section 2.1) now becomes $M^X/G/\infty$. Following the queueing result in (Liu et al. 1990), the mean and the variance of N are as follows:

$$\mathsf{E}(N) = \lambda \mathsf{E}(X)\mathsf{E}(L) = \mathsf{E}(D)\mathsf{E}(L) \tag{8}$$

$$\begin{aligned}
\mathsf{Var}(N) &= \mathsf{E}(N) + \lambda[\mathsf{E}(X^2) - \mathsf{E}(X)] \int_0^\infty \bar{F}_L^2(y)\, dy \\
&\leq \mathsf{E}(N) + \lambda[\mathsf{E}(X^2) - E(X)] \int_0^\infty \bar{F}_L(y)\, dy \\
&= \mathsf{E}(N) + \lambda[\mathsf{E}(X^2) - \mathsf{E}(X)]\mathsf{E}(L) \\
&= \lambda \mathsf{E}(X^2)\mathsf{E}(L) \\
&= \mathsf{Var}(D)\mathsf{E}(L)
\end{aligned} \tag{9}$$

where $F_L(y) = 1 - \bar{F}_L(y)$ denotes the distribution function of the lead time L, the inequality is due to $\bar{F}_L^2(y) \leq \bar{F}_L(y)$, and $\mathsf{E}(X^2) \geq \mathsf{E}^2(X) \geq \mathsf{E}(X)$ (since $\mathsf{E}(X) \geq 1$). Note again that the variance of N in (9) is different from the variance of lead time demand. In fact, the $\mathsf{Var}(N)$ is even smaller than the first of the two terms of the variance of lead time demand.

The generating function of N is known for the $M/G/\infty$ model (refer to Liu et al. 1990), from which the distribution of N can be derived, at least in principle. It is, however, much simpler to derive the mean and the variance of N and then approximate it by a normal distribution, as in Section 2.2.

The queueing quantity N, by definition, is nonnegative. Hence, to approximate it with a normal distribution is not always appropriate. Sometimes an adjustment to the normal distribution is needed. The same applies if we model demand D using a normal distribution.

Specifically, instead of writing $N = \mu + \sigma Z$, where Z is the standard normal variate, we should have $\tilde{N} = [\mu + \sigma Z]^+$. The mean of $\tilde{N}$ follows from (5):

$$\mathsf{E}[\tilde{N}] = \sigma \mathsf{E}\left[Z - \left(-\frac{\mu}{\sigma}\right)\right]^+ = \sigma G\left(-\frac{\mu}{\sigma}\right) \tag{10}$$

To derive the variance of $\tilde{N}$, note the following:

$$\begin{aligned}\mathsf{E}\{[(Z-x)^+]^2\} &= \int_x^\infty (z-x)^2\phi(z)\,dz \\ &= x\phi(x) + \overline{\Phi}(x) - 2x\phi(x) + x^2\overline{\Phi}(x) \\ &= \overline{\Phi}(x) - xG(x)\end{aligned}$$

where the last equation makes use of (5). Hence,

$$\begin{aligned}\mathsf{Var}[\tilde{N}] &= \sigma^2 \mathsf{Var}\left\{\left[Z - \left(-\frac{\mu}{\sigma}\right)\right]^+\right\} \\ &= \sigma^2 \mathsf{E}\left\{\left[\left(Z - \left(-\frac{\mu}{\sigma}\right)\right)^+\right]^2\right\} - [\mathsf{E}(\tilde{N})]^2 \\ &= \sigma^2\left[\overline{\Phi}\left(-\frac{\mu}{\sigma}\right) + \frac{\mu}{\sigma} G\left(-\frac{\mu}{\sigma}\right) - G^2\left(-\frac{\mu}{\sigma}\right)\right]\end{aligned} \tag{11}$$

For moderately large x (say, $x \geq 2$), from property (c) of the G function in Section 2.2, we have $G(-x) \cong x$, and hence

$$\mathsf{E}[\tilde{N}] \cong \mathsf{E}[N], \quad \mathsf{Var}[\tilde{N}] \cong \mathsf{Var}[N]$$

from (10) and (11). Therefore, the above adjustment is *not* needed when the coefficient of variation, σ/μ, is relatively small, say, $\sigma/\mu < 0.5$.

3. NONSTATIONARY DEMAND: THE DRP FRAMEWORK

DRP stands for distribution resource planning, a class of commercial software tools that are widely used in industry for managing a firm's production/inventory/distribution systems. When demand is nonstationary, the dynamics of inventory are best captured by a set of recursive equations, which are closely related to the logic built into DRP software. Hence, we present next an overview of DRP, or more precisely, a formal abstraction of standard DRP procedures, the latter being widely available in professional references (e.g., Martin 1990; and Stenger 1994).

Time is discrete, indexed by $t = 1, 2, \ldots, n$, and referred to as periods. Suppose we are currently at the beginning of period 1 (or, the end of period 0). We are interested in the following quantities for all future periods, $t = 1, 2, \ldots, n$:

I_t = the on-hand inventory at the end of time period t
B_t = the back-ordered demand at the end of time period t
A_t = the required quantity of product needed at the beginning of period t
Q_t = the constrained (or feasible) quantity of product needed at the beginning of period t
$\tilde{Q}_t$ = the recommended order quantity at the beginning of period t

The distinction between the requirements, A_t, and the constrained order quantity, Q_t, is important. Whereas A_t represents the quantity that is needed at the beginning of period t, Q_t reflects what is feasible, taking into account lead time constraints and order quantity restrictions. For example, if $A_t = 40$ and the maximum order quantity is 30, then Q_t would be set to 30.

The following information is assumed known at the start of time period 1, for all future periods $t = 1, \ldots, n$:

D_t = the demand, in terms of its distribution, in particular $\mathsf{E}[D_t]$ and $\mathsf{sd}[D_t]$
Σ_t = the safety stock requirement, which is the portion of on-hand inventory that should be maintained in the period to protect against demand uncertainty so as to achieve a prescribed service-level requirement
I_0 = the on-hand inventory at the beginning of time period 1
B_0 = the back-ordered demand at the beginning of time period 1

In DRP, there is also the quantity of scheduled receipts, quantities that are in transit and scheduled to arrive at (the start of) each future period. Here we simply ignore these because as they can be easily netted from the inventory/back order of each period or added to Q_t.

At the beginning of each period t, the following sequence of events takes place. First, replenishment (including any scheduled receipt) arrives, that is, Q_t, the constrained order quantity (which relates to A_t and will be specified below). These units are used first to satisfy the back orders, if any, left from the previous period. Next, demand of the period, D_t, is realized and filled, which brings us to the end of the period, when I_t (on-hand inventory) and B_t (back orders) are updated.

Hence, we have the following recursive relation:

$$A_t = [B_{t-1} - I_{t-1} + D_t + \Sigma_t]^+ \tag{12}$$

which says that the replenishment requirement in each period, along with any on-hand inventory from the last period should be able to supply the demand of the current period and any backlog from the last period and still result in a surplus that is equal to the required safety stock for this period.

Next, the constrained order quantity, Q_t, is derived from A_t by applying a set of prespecified order-size restrictions, referred to as *order policies or order rules.* Several commonly used rules are listed below. Note that all of the rules below apply if and only if $A_t > 0$; otherwise, we simply set $Q_t = A_t = 0$, as $0 \leq Q_t \leq A_t$ by definition.

1. *Lot-for-lot:* $Q_t = A_t$, i.e., there is no restriction on the order quantity.
2. *Min-max:* With $Q_{\min}$ and $Q_{\max}$ being the (given) lower and upper limit on order quantities, $Q_t = \min\{\max(Q_{\min}, A_t), Q_{\max}\}$, becomes the order quantity planned for period t
3. *DOS* (days of supply): Q_t is equal to the (projected) demand over a given number of periods.
4. *EOQ* (economic order quantity): $Q_t = \sqrt{2C\mathsf{E}(D_t)/h}$, the classical EOQ formula (refer to Section 1.1), where C is a fixed cost for placing a replenishment order, and h is the inventory holding cost per period.

Note that all the rules above, with the exception of 1, impose some restrictions on the order quantity. In general, we can assume Q_t relates to A_t (and other parameters) via some prespecified function.

Define the *net* inventory level, Y_t, at the beginning of period t:

$$Y_t = I_{t-1} - B_{t-1}, \quad t = 1, \ldots, n \tag{13}$$

Then clearly,

$$I_t = [Q_t + Y_t - D_t]^+, \quad \text{and} \quad B_t = [D_t - Q_t - Y_t]^+ \tag{14}$$

Finally, suppose the order lead time is L (periods). Then the recommended order quantity, $\tilde{Q}_t$, in period t, corresponds to the calculated (constrained) order quantity $L + t$ periods later:

$$\tilde{Q}_t = \mathrm{Q}_{t+L}, \quad t = 1, \ldots, n - L$$

So the recommended order quantity at the start of period t is whatever the constrained order quantity is for period $t + L$.

4. LOT-FOR-LOT POLICY

Assume that the demand in period t can be expressed as

$$D_t = \mu_t + \sigma_t Z_t$$

where μ_t is the mean demand in period t, σ_t is the standard deviation of demand in period t, and $\{Z_t, t = 1, \ldots, n\}$ are i.i.d. random variables following any distribution with zero mean and unit variance. For ease of discussion, we shall focus on the case in which Z_t follows a standard normal distribution. The results extend readily to more general demand distributions.

Suppose, as before, that the lead time is L and we want to determine the requirement A_t for period t. We need to make this decision at the beginning of period $t - L$ so as to place the order in time for it to arrive in period t.

Position ourselves at the beginning of period $t - L$. What we know are the following: (a) the net inventory left from the previous period, Y_{t-L} and (b) the scheduled arrivals, $A_{t-L}, A_{t-L+1}, \dots, A_{t-1}$, which had already been decided and ordered. What we do *not* know is the sequence of demand over the lead time, $D_{t-L}, D_{t-L+1}, \dots, D_t$.

Denote:

$$D(s, t) := D_s + \cdots + D_t, \quad A(s, t) := A_s + \cdots + A_t$$
$$\mu(s, t) := \mathsf{E}[D(s, t)], \quad \sigma(s, t) := \mathsf{sd}[D(s, t)]$$

We claim that the right A_t value should be:

$$A_t = [\mu(t - L, t) + k_t\sigma(t - L, t) - Y_{t-L} - A(t - L, t - 1)]^+. \tag{15}$$

First, suppose $A_t > 0$, i.e., the quantity inside $[\cdot]^+$ on the right hand side above, is positive. Then, at the beginning of period $t - L$, the net inventory is Y_{t-L}, which, along with the scheduled arrivals over the lead time—a total of $A(t - L, t)$, is going to supply all the demands over the lead time—a total of $D(t - L, t)$. Hence, at the end of period t, the expected back order is:

$$\begin{aligned}\mathsf{E}[B_t] &= \mathsf{E}[D(t - L, t) - Y_{t-L} - A(t - L, t)]^+ \\ &= \mathsf{E}[\mu(t - L, t) + Z_t\sigma(t - L, t) - \mu(t - L, t) - k_t\sigma(t - L, t)]^+ \\ &= \sigma(t - L, t)\, \mathsf{E}[Z_t - k_t]^+ \\ &= \sigma(t - L, t)\, G(k_t)\end{aligned} \tag{16}$$

where the second equality follows from (15) since $A_t > 0$ as assumed:

$$Y_{t-L} + A(t - L, t) = \mu(t - L, t) + k_t\sigma(t - L, t)$$

Similarly,

$$\begin{aligned}\mathsf{E}[I_t] &= \mathsf{E}[Y_{t-L} + A(t - L, t) - D(t - L, t)]^+ \\ &= \sigma(t - L, t) + \mathsf{E}[k_t - Z_t]^+ \\ &= \sigma(t - L, t)\, H(k_t)\end{aligned} \tag{17}$$

Next, suppose $A_t = 0$ in (15). This implies

$$Y_{t-L} + A(t - L, t - 1) \geq \mu(t - L, t) + k_t\sigma(t - L, t)$$

Hence,

$$\begin{aligned}\mathsf{E}[B_t] &= \mathsf{E}[\mu(t - L, t) + Z \cdot \sigma(t - L, t) - Y_{t-L} - A(t - L, t - 1)]^+ \\ &= \sigma(t - L, t)\, G(k_t')\end{aligned} \tag{18}$$

with

$$k_t' = \frac{Y_{t-L} + A(t - L, t - 1) - \mu(t - L, t)}{\sigma(t - L, t)} \geq k_t \tag{19}$$

Hence, $\mathsf{E}[B_t]$ in (18) is no greater than $\mathsf{E}[B_t]$ in (16), since $G(\cdot)$ is a decreasing function. (Intuitively, in this case, since the available (net) inventory Y_{t-L} is higher, the effective safety factor k_t' is also higher; hence, the (projected) back order is lower.)

Similarly, when $A_t = 0$, we have

$$\mathsf{E}[I_t] = \sigma(t - L, t)\, H(k_t') \tag{20}$$

with k_t' following (19); and $\mathsf{E}[I_t]$ above is larger than $\mathsf{E}[I_t]$ of (17).

To summarize, when the order rule is lot-for-lot, that is, $Q_t = A_t$ for all t, the required quantity needed at the beginning of period t (A_t), and hence the recommended order quantity at the beginning of period $t - L$ ($\tilde{Q}_{t-L}$), should be

$$\tilde{Q}_{t-L} = Q_t = A_t = [s_{t-L} - Y_{t-L} - A(t - L, t - 1)]^+$$

where

$$s_{t-L} = \mu(t - L, t) + k_t\sigma(t - L, t)$$

Note that in this case, DRP effectively follows a base-stock control mechanism, with s_{t-L} being the base-stock level (or, reorder point) for period $t - L$. The estimates for expected on-hand inventory and back orders (for period t) follow (17, 20) and (16, 18), respectively.

Also note that the time index of the reorder point, s_{t-L}, is consistent with the time when the order is placed. On the other hand, the safety factor k_t is indexed by t since it relates more closely to the on-hand inventory and back orders in period t.

We should point out that the expression in (15) that specifies the requirement A_t can also be derived directly from the DRP logic as follows. From (13) and (14), with $Q_t = A_t$, we have

$$\begin{aligned} Y_t &= I_{t-1} - B_{t-1} \\ &= Y_{t-1} + A_{t-1} - D_{t-1} \\ &= Y_{t-2} + A_{t-2} - D_{t-2} + A_{t-1} - D_{t-1} \\ &= \dots \\ &= Y_{t-L} + A(t - L, t - 1) - D(t - L, t - 1) \end{aligned} \tag{21}$$

Now, substituting (13) into (12), we can write the latter as:

$$A_t - [D_t + \Sigma_t - Y_t]^+$$

Substituting (21) into the above and treating demand as deterministic, which is what DRP does, we have

$$D_t + D(t - L, t - 1) = \mu(t - L, t)$$

and hence recovering (15). In other words, if DRP uses (15) directly to generate the requirements A_t, instead of going through the recursions that involve I_t and B_t, then the estimates on these requirements will be protected from the in accuracies involved in estimating I_t and B_t.

Following the expressions derived above for A_t, B_t, and I_t, the recursion can be carried over to future periods through $Y_{t+1} = I_t - B_t$. Specifically, starting from period 1, we first derive A_{L+1}, followed by B_{L+1} and I_{L+1}, and then Y_{L+1}. Continue this procedure for $t = L + 2, \dots, 2L$, assuming that $Y_1, Y_2, \dots, Y_L$ have all been prespecified, i.e., given at beginning of period 1. (Note that nothing in periods 1 through L can be affected by any replenishment decision, at the beginning of period 1, due to the lead time.) In return, $Y_{L+2}, \dots, Y_{2L}$ are derived. Next, we will derive A_{2L+1}, at the beginning of period $L + 1$. At that point, we will need the value of Y_{L+1}, which has already been derived; and in return, we will derive Y_{2L+1}. The recursion then continues.

5. (*S*, *s*) POLICY

Having analyzed the lot-for-lot rule, we next want to study other rules that impose restrictions on the order quantities. It will become evident below (Section 6) that all these rules can be unified into a dynamic (S, s) control scheme—dynamic in the sense that the control parameters, S and s, in general change over time. In order to analyze the dynamic (S, s) control rule, here we first review the standard (S, s) inventory model under stationary demand and develop some simple approximations for its performance.

The (S, s) policy is known to be optimal in a quite general setting, in terms of minimizing the total costs of placing orders, keeping inventory, and paying penalty of back orders (refer to Clark and Scarf 1960; Scarf 1960). Our focus here, however, is on performance evaluation rather than cost minimization. In particular, we want to derive certain simple approximate formulas that are not only easy to evaluate but also useful in suggesting approximations in the non-stationary (i.e., dynamic) setting. More refined approximations for the stationary (S, s) systems are available in the literature (e.g., Tijms 1994), although it is not clear how they can be adapted to the nonstationary setting.

5.1. Stationary Analysis

Consider an (S, s) inventory model with periodic review. A review takes place at the beginning of each period t, at which time the inventory position is updated and a replenishment decision is made following the (S, s) rule—order up to S if and only if the inventory position (on-hand plus on-order minus back order) falls below s.

Observe that the inventory position, X_t, as defined above, always takes values between s and S. In particular, if an arriving demand brings the on-hand inventory level to *below* s, the inventory position is immediately brought up to S through placing a replenishment order.

Suppose, without loss of generality, that we start with $X_0 = S$. Then, from standard renewal theory (Ross 1996), we know the time between two consecutive replenishment orders forms a regenerative cycle. Given x $(x = s, s + 1, \ldots, S)$, let

$$T_x = \min \{t\colon D_1 + \ldots + D_t > S - x\}$$

be the time, in a cycle, for the inventory position to drop from S to below x. In particular, T_s is the cycle length.

Since demands are i.i.d., we can view $T_x - 1$ as the number of renewals by time $S - x$ of a renewal process with interarrival time D_t. (Refer to Ross 1996, Section 3.4, in particular pp. 69–70.) Hence,

$$\mathsf{E}[T_x] = M(S - x) + 1, \quad \mathsf{E}[T_s] = M(S - s) + 1$$

where

$$M(x) = \sum_{n=1}^{\infty} F_{*n}(x)$$

is the renewal function, with $F_{*n}(\cdot)$ denoting the n-fold convolution of the (per-period) demand distribution, $F(\cdot)$. Furthermore, the limiting (stationary) inventory position, X, has the following distribution:

$$\mathsf{P}[X \geq x] = \lim_{t\to\infty} \mathsf{P}[X_t \geq x] = \frac{\mathsf{E}[T_x]}{\mathsf{E}[T_s]} = \frac{1 + M(S - x)}{1 + M(S - s)}$$

for $x \in [s, S]$. Alternatively,

$$\mathsf{P}[X = x] = \frac{M(S - x) - M(S - x - 1)}{1 + M(S - s)}, \quad x = s, \ldots, S - 1$$

$$\mathsf{P}[X = S] = \frac{1}{1 + M(S - s)}$$

The mean of X then follows:

$$\mathsf{E}[X] = \frac{S + sM(S - s) + \sum_{x=1}^{S-s-1} M(x)}{1 + M(S - s)}$$

To characterize the on-hand inventory and the back orders, the key is to observe that by time t all orders that are placed before or at $t - L$ will have arrived. In other words, the on-order quantities included in the inventory position at $t - L$ will all have arrived by t. As before, let $D(t - L, t)$ denote the total demand over the $L + 1$ periods: $t - L, t - L + 1, \ldots, t$. We have

$$I_t = [X_{t-L} - D(t - L, t)]^+ \tag{22}$$

where $[x] + = \max\{x, 0\}$. Note that in (22), $D(t - L, t)$ is independent of X_{t-L}. Similarly, we have

$$B_t = [D(t - L, t) - X_{t-L}]^+ \tag{23}$$

Letting $t \to \infty$, and denoting the limits by omitting the time index, we have

$$I = [X - D(L + 1)]^+, \quad \text{and} \quad B = [D(L + 1) - X]^+ \tag{24}$$

where $D(L + 1)$ denotes a random variable that is equal in distribution to $D(t - L, t)$ and is independent of X.

5.2. Approximations

The renewal function is known to have a linear asymptote:

$$M(x) \sim \frac{x}{\mu} + \frac{m_2}{2\mu^2} - 1 \tag{25}$$

for large x, where $\mu = \mathsf{E}(D)$ and $m_2 = \mathsf{E}(D^2)$ are the first two moments of the (per-period) demand distribution. (More precisely, the difference between $M(x)$ and the linear function on the right hand side above goes to zero when $x \to \infty$).

Since in most applications the (S, s) values are large or moderately large, we can use the linear asymptote in (25) as an approximation for the renewal function. This way, the probability distribution of the inventory position X in Section 5.1 becomes:

$$\mathsf{P}[X = x] = \frac{1}{\dfrac{m_2}{2\mu} + S - s}, \quad x = s, \ldots, S - 1$$

$$\mathsf{P}[X = S] = \frac{\dfrac{m_2}{2\mu}}{\dfrac{m_2}{2\mu} + S - s}$$

where $m_2/2\mu$ is the average "undershoot"—the gap between the inventory position when an order is placed (which is necessarily smaller than s) and the reorder point s. Hence,

$$\mathsf{E}[X] = \frac{(1/2)(S + s - 1)(S - s) + (m_2/2\mu)\, S}{m_2/2\mu + S - s} = \frac{\mu(S - s)(S + s - 1) + m_2 S}{m_2 + 2\mu(S - s)} \tag{26}$$

In some special cases, the above reduces to a uniform distribution over the integers $\{s, \ldots, S\}$ and $\mathsf{E}[X] = (S + s)/2$. For instance, this is the case when the (per-period) demand follows a truncated normal distribution with unit mean and unit coefficient of variation (hence, $\mu = 1$ and $m_2 = 2$). Note that for truncated normal distributions, just as for normal distributions, assuming a unit coefficient of variation loses no generality; hence, unit mean is the only significant assumption here.

Next, making use of (24), along with the normal approximation of $D(L + 1)$, we get the following approximation:

$$\mathsf{E}[B] = \sum_{x=s}^{S} \sigma\sqrt{L+1}\, G\left(\frac{x - (L+1)\mu}{\sigma\sqrt{L+1}}\right) \mathsf{P}[X = x]$$

Substituting the uniform distribution of X into the above, we have

$$\begin{aligned}\mathsf{E}[B] &= \frac{1}{1 + S - s} \sum_{x=s}^{S} \sigma\sqrt{L+1}\, G\left(\frac{x - (L+1)\mu}{\sigma\sqrt{L+1}}\right) \\ &\le \sigma\sqrt{L+1}\, G\left(\frac{s - (L+1)\mu}{\sigma\sqrt{L+1}}\right)\end{aligned} \tag{27}$$

where the inequality follows from the fact that $G(x)$ is decreasing in x.

Once $\mathsf{E}[B]$ is derived, $\mathsf{E}[I]$ follows from (1), we have

$$\mathsf{E}[I] = \mathsf{E}[B] + \mathsf{E}[X] - \mu(L + 1)$$

For instance, from (27), taking into account $G(k) = H(k) - k$, and approximating X with a uniform distribution, we have

$$\mathsf{E}[I] = \sigma\sqrt{L+1}\, H\left(\frac{s - (L+1)\mu}{\sigma\sqrt{L+1}}\right) + \frac{S - s}{2} \tag{28}$$

Alternatively, $\mathsf{E}[X]$ can follow the approximation in (26).

6. DYNAMIC (*S*, *s*) POLICY

Observe that when following any policy other than lot-for-lot, the equation in (15) governing the requirements should be modified as follows:

$$A_t = [\mu(t-L, t) + k_t\sigma(t-L, t) - Y_{t-L} - Q(t-L, t-1)]^+ \tag{29}$$

where

$$Q(t-L, t-1) = Q_{t-L} + \ldots + Q_{t-1}$$

replaces $A(t-L, t-1)$ in (15). Hence, when $A_t > 0$, we have

$$\begin{aligned} A_t &= \mu(t-L, t) + k_t\sigma(t-L, t) - Y_{t-L} - Q(t-L, t-1) \\ &= s_{t-L} - Y_{t-L} - Q(t-L, t-1). \end{aligned}$$

That is, s_{t-L} is the reorder point, since following the DRP logic, an order ($\tilde{Q}_{t-L}$ is placed (at $t-L$) if and only if $A_t > 0$. From the above expression, A_t is the required quantity to bring the inventory position at $t-L$ back to s_{t-L}. (This is consistent with the base-stock mechanism when there are no order-size restrictions.)

Now, suppose the replenishment policy is a dynamic (*S*; *s*) rule. Specifically, when $A_t > 0$, we want to bring the inventory position to S_{t-L} ($> s_{t-L}$). Hence, an additional amount, $S_{t-L} - s_{t-L}$, is needed, and the order quantity is:

$$Q_t = A_t + S_{t-L} - s_{t-L} = S_{t-L} - Y_{t-L} - Q(t-L, t-1) \tag{30}$$

when $A_t > 0$. (As before, $Q_t = 0$ when $A_t = 0$).

On the other hand, given any replenishment policy with a prespecified Q_t (as a function of $A_t > 0$), we can implement this policy by setting [cf. (30)]:

$$S_{t-L} = Q_t + s_{t-L} - A_t = Q_t + Q(t-L, t-1) + Y_{t-L}$$

when $A_t > 0$. When $A_t = 0$, we set $Q_t = 0$, which results in $S_{t-L} = s_{t-L}$, from the first equation above.

We now turn to the performance evaluation under the dynamic (*S*, *s*) rule. Note that the inventory position at $t-L$ is:

$$X_{t-L} = Y_{t-L} + Q(t-L, t) \tag{31}$$

that is, the net inventory plus the on-order quantities (orders that have been placed but have yet to arrive). In particular, when $Q_t > 0$, from (30) and (31), we have

$$S_{t-L} = Y_{t-L} + Q(t-L, t) = X_{t-L}$$

that is, after the order is placed, the inventory position is brought up to S_{t-L}.

Just as in (21), we can iterate on (13) and (14) to obtain

$$Y_t = Y_{t-L} + Q(t-L, t-1) - D(t-L, t-1)$$

which, upon substitution back into (14), yields:

$$I_t = [Y_{t-L} + Q(t-L, t) - D(t-L, t)]^+ = [X_{t-L} - D(t-L, t)]^+$$

and

$$B_t = [D(t-L, t) - Y_{t-L} - Q(t-L, t)]^+ = [D(t-L, t) - X_{t-L}]^+$$

These are exactly the same formulas as in (22) and (23).

Therefore, we can adapt the approximations in the stationary (*S*, *s*) model. For instance, based on the back order approximation in (27), and taking into account the formulas in (16) and (18) for the lot-for-lot case, we have the following approximation:

$$\mathsf{E}[B_t] = \sigma(t - L, t) \cdot G(k_t), \quad \text{or} \quad \mathsf{E}[B_t] = \sigma(t - L, t) \cdot G(k_t') \tag{32}$$

according to whether $A_t > 0$ or $A_t = 0$, where

$$k_t' = \frac{Y_{t-L} + Q(t - L, t - 1) - \mu(t - L, t)}{\sigma(t - L, t)}$$

which, again, can be verified as dominating k_t.

To approximate $\mathsf{E}[I_t]$, just as in the stationary case, we make use of the identity [from (22)]:

$$I_t - B_t = X_{t-L} - D(t - L, t)$$

Hence,

$$\mathsf{E}[I_t] = \mathsf{E}[B_t] + X_{t-L} - \mu(t - L, t)$$

where $\mathsf{E}[B_t]$ follows the approximation in (32), and X_{t-L} follows the expression in (31). According to the two cases in (32), and making use of the expression in (31) and the relation $H(k) = k + G(k)$, we have:

$$\mathsf{E}[I_t] = \sigma(t - L, t) \cdot H(k_t) + S_{t-L} - s_{t-L} \tag{33}$$

when $A_t > 0$ (note in this case $X_{t-L} = S_{t-L}$); and when $A_t = 0$:

$$\begin{aligned}\mathsf{E}[I_t] &= \sigma(t - L, t) \cdot H(k_t') - k_t') - k_t' \sigma(t - L, t) + X_{t-L} - \mu(t - L, t) \\ &= \sigma(t - L, t) \cdot H(k_t') \end{aligned} \tag{34}$$

Approximating $H(k_t')$ by k_t' in the above equation, we have

$$\mathsf{E}[I_t] = k_t' \sigma(t - L, t) = X_{t-L} - \mu(t - L, t)$$

The above has the intuitive interpretation that if no order is placed at $t - L$ (i.e., $R_{t_L} = Q_t = A_t = 0$), then the expected on-hand inventory at t is simply the inventory position—of which all the on-order quantities will have arrived by t—minus the demand over the lead time, period $t - L$ through period t.

To implement the above approximations requires the derivation of Y_t, which is involved in both X_t and A_t. Recursively, Y_t can be approximated by its mean:

$$\mathsf{E}[Y_t] = \mathsf{E}[I_{t-1}] - \mathsf{E}[B_{t-1}]$$

Alternatively, a cruder approximation is to forgo the distinction between the two cases $A_t > 0$ and $A_t = 0$. (Observe that this distinction is not present in the stationary case; it is averaged out in each regenerative cycle.) Specifically, ignore the k_t' case in (32), and approximate X_{t-L} by $(S_{t-L} + s_{t-L})/2$. This way, we have

$$\mathsf{E}[B_t] = \sigma(t - L, t) \cdot G(k_t) \tag{35}$$

and

$$\begin{aligned}\mathsf{E}[I_t] &= \sigma(t - L, t) \cdot H(k_t) - k_t \sigma(t - L, t) + (S_{t-L} + s_{t-L})/2 - \mu(t - L, t) \\ &= \sigma(t - L, t) \cdot H(k_t) + (S_{t-L} - s_{t-L})/2 \end{aligned} \tag{36}$$

Both are consistent with the stationary approximations in (27) and (28).

To summarize, with order-quantity restrictions, the implementation of DRP can be unified into a dynamic (S, s) control rule. Under this rule, the constrained order quantity needed at the beginning of period t (Q_t), and hence the recommended order quantity at the beginning of period $t - L$ ($\tilde{Q}_{t-L}$), should be

$$\tilde{Q}_{t-L} = Q_t = [A_t + S_{t-L} - s_{t-L}] \cdot 1[A_t > 0]$$

where 1[·] denotes the indicator function, and

$$A_t = [s_{t-L} - Y_{t-L} - Q(t-L, t-1)]^+$$

following (29), with

$$s_{t-L} = \mu(t-L, t) + k_t \sigma(t-L, t)$$

6.1. Setting Safety Stock Levels

In the preceding discussions, the safety stock level, Σ_t, is assumed as given. Here we discuss one approach to set the safety stock levels, which is of particular importance because of its wide usage. It sets safety stock levels based on achieving a target fill rate.

For ease of discussion, here we assume stationary demand and omit the time index t wherever possible. Suppose the service requirement is that the fraction of demand back ordered should be limited to $1 - \beta$, where β is the required fill rate.

To characterize the fraction of back ordered demand, we need to pick a "typical" time frame. In the standard inventory literature, this is taken to be the time between two consecutive orders, known as a regenerative cycle (when demands are independent and identically distributed). The required fraction is then the ratio of the average number of back orders to the average number of demand units, both over the regenerative cycle.

In the stationary (S, s) model, the average number of back orders, following (27), is approximated by

$$\sigma\sqrt{L+1}G\left(\frac{s-(L+1)\mu}{\sigma(L+1}\right)$$

whereas the average demand per cycle is

$$\left(\frac{S-s}{\mu} + \frac{m_2}{2\mu^2}\right)\mathsf{E}(D) = S - s + \frac{m_2}{2\mu}$$

where the first factor on the left hand side is the expected cycle length $\mathsf{E}[T_s] = 1 + M(S-s)$ (refer to Section 5.1) with $M(S-s)$ approximated by the linear asymptote in (25). Note that the right-hand side above is nothing but the expected order quantity; in particular, $m_2/2\mu$ is the expected undershoot.

Therefore, based on the above, under the dynamic (S, s) rule, k_t should be the solution to:

$$\sigma(t-L, t)G(k_t) = (1-\beta)\left[\Delta_{t-L} + \frac{\mathsf{E}(D_{t-L}^2)}{2\mathsf{E}(D_{t-L})}\right] \tag{37}$$

where $\Delta_{t-L} = S_{t-L} - s_{t-L}$ is assumed given.

The lot-for-lot rule is equivalent to $S_{t-L} = s_{t-L}$, or $\Delta_{t-L} = 0$. Also, since in every period an order is placed, the order quantity is simply equal to the demand. Hence, the equation in (37) is reduced to:

$$\sigma(t-L, t)\, G(k_t) = (1-\beta)\mathsf{E}(D_{t-L}) \tag{38}$$

Note that the equations in (37) and (38) are easily solved through Newton's method (e.g., Press et al. 1994). Write the equations in the form of $G(k) = c$. The Newton's iteration, indexed by the superscript (n), is as follows:

$$k^{(n+1)} = k^{(n)} - \frac{G(k^{(n)}) - c}{G'(k^{(n)})}$$

When demand follows a normal distribution, $G'(k) = -1 + \Phi(k)$, where Φ denotes the distribution function of the standard normal variate.

7. MULTI-STAGE MODELS

The results presented so far for the single-stage model extend to a more general distribution network. For ease of exposition, we consider a model that consists of a central stocking facility (e.g., a warehouse or depot) supplying a set of local stocking facilities (e.g., retailers or outlets), each of which, in turn, supplies its own customer demand. As in the single-stage models, our focus here is on performance evaluation via simple approximations. We do not address the issue of cost optimization, for which many models can be found in the survey article by Federgruen (1993).

Suppose the central warehouse is numbered as stage 0 and the set of retailers, numbered as stages $1, \ldots, c$. Quantities that relate to these stages will be subscripted or superscripted (if there is already a subscript for time) by their stage indices. For instance, the demand stream to retailer i, $i = 1, \ldots, c$, is $\{D_t^i, t = 1, 2, \ldots\}$. Assume the demands are independent among the retailers, and independent over the time periods.

Suppose the lead time at each retailer i is L_i periods, a deterministic constant, for $i = 1, \ldots, c$. For instance, this can be the transportation time for a replenishment order to travel from the warehouse to the retailer; or, in the case where stage i is a plant, the cycle time to build a customer order. Suppose the lead time at the warehouse is L_0.

The analysis of each retailer i follows the same discussion as before. In particular, write

$$s_t^i = \mu_i(t, t + L_i) + k_{t+L_i}^i \sigma_i(t, t + L_i)$$

and if a set of target fill rates at each of the retailers, β_i, $i = 1, \ldots, c$, has been specified, then the safety factor $k_{t+L_i}^i$ is obtained from solving the following equation. (Refer to Eqs. (37) and (38)).

$$\sigma_i(t, t + L_i)G(k_{t+L_i}^i) = (1 - \beta_i)\left[\Delta_t^i + \frac{\mathsf{E}(D_t^i)^2)}{2\mathsf{E}(D_t^i)}\right]$$

(where $\Delta_t^i := S_t^i$ is assumed given); or, from

$$\sigma_i(t, t + L_i)G(k_{t+L_i}^i) = (1 - \beta_i)\mathsf{E}(D_t^i)$$

in the special case of i following a lot-for-lot rule.

The expected number of back orders is:

$$\mathsf{E}[B_t^i] = \sigma_i(t - L_i, t)G\left(\frac{s_{t-L}^i - \mu_i(t - L_i, t)}{\sigma_i(t - L_i, t)}\right)$$

corresponding to the k_t case in (32), that is, when $A_t^i > 0$. For the other case ($A_t^i = 0$), the expression is similar, corresponding to the k_t' case in (32). The expected on-hand inventory is where $X_{t-L_i}^i$ denotes the inventory position.

To analyze the warehouse, we follow the standard notion of echelon stock. That is, we aggregate stage 0, along with stages $i = 1, \ldots, c$, into a single system, indexed by E (for echelon). This aggregated system supplies the superposition of all c demand streams. Note that each demand process D_t^i still has its own lead time L_i. Let

$$L_{\max} = \max\{L_i, i = 1, \ldots, c\},$$

Then (22) and (23) should be modified, with L replaced by $L_E = L_0 + L_{\max}$.

Therefore, the lower threshold value for this aggregated system should be set, in period t, as follows:

$$s_t^E = \mu_E(t, t + L_E) + k_{t+L_E}^E \sigma_E(t, t + L_E)$$

where

$$\mu_E(t, t + L_E) = \sum_{i=1}^{c} \mu_i(t, t + L_E)$$

$$\sigma_E(t, t + L_E) = \left[\sum_{i=1}^{c} \sigma_i^2(t, t + L_E)\right]^{1/2}$$

As for the echelon safety factor, $k_{t+L_E}^E$, it can either be prespecified by the user or, if the safety stock levels are to be set automatically based on target fill rates specified at the retail level, it can be obtained from the solution to:

$$\sigma_E(t, t + L_E)G(k_{t+L_E}^E) = (1 - \beta_{\max})(\Delta_t^E + \gamma_t^E)$$

with $\Delta_t^E = S_t^E - s_t^E$ assumed given, $\beta_{\max} = \max\{\beta_i, i = 1, \ldots, c\}$, where β_i is the target fill rate at retailer i, and

$$\gamma_t^E = \frac{\mathsf{E}\left[\left(\sum_{i=1}^{c} D_t^i\right)^2\right]}{2\mathsf{E}\left(\sum_{i=1}^{c} D_t^i\right)} = \frac{\sum_{i=1}^{c} (\sigma_t^i)^2 + \left(\sum_{i=1}^{c} \mu_t^i\right)^2}{2\sum_{i=1}^{c} \mu_t^i}$$

That is, stage 0 should place an order at the beginning of period t if and only if the *echelon* inventory position falls below s_t^E. This echelon inventory position is the inventory position with the aggregated system considered as a single stage. It includes all the on-hand inventory at both the warehouse and all the retailers, plus any replenishment orders placed by the warehouse that have not yet arrived, minus any customer demand that is back ordered at all the retailers. Hence, this echelon inventory position will only change whenever there is a customer order arriving at a retailer and whenever the warehouse places a replenishment order. It will not change when a retailer takes supply from or registers a back order with the warehouse. When stage 0 issues a replenishment order, say at the beginning of period t, the order quantity is to bring the echelon stock position to $S_t^E = \Delta_t^E + s_t^E$. Hence, $\Delta_t^E + \gamma_t^E$ is the estimate of this order quantity.

Estimates of $\mathsf{E}[B_t^E]$ and $\mathsf{E}[I_t^E]$, like those at the retailers, follow in the same vein as those in Section 6 (refer to, in particular, the summary at the end of that section).

The relationship between Q_t and S_{t-L} follows (30); in particular, when $A_t > 0$, $Q_t = S_{t-L} - Y_{t-L} - Q(t-L, t-1)$; and when $A_t = 0$, $Q_t = 0$ and $S_{t-L} = s_{t-L}$. The estimates on the expected on-hand inventory and back orders (for period t) follow (33), (34) and (32), respectively; or alternatively, follow (36) and (35), respectively.

8. ASSEMBLE-TO-ORDER SYSTEMS

An assemble-to-order (ATO) system is a hybrid model of build-to-stock at the component (subassembly) level and assemble-to-order for the end product. In an ATO system, typically, the components take a substantial lead time to build, whereas the time it takes to assemble all the components into the final product is often negligible. Hence, keeping stock at the component level improves response-time performance, whereas not keeping any end-product inventory reduces inventory cost and maximizes the flexibility for customization. A good example of an ATO system is the production of a PC (personal computer). Other examples include fast-food operations and many mail-order or e-commerce services.

We consider an ATO system of m different items (components) and a single end product. Let $\mathcal{I} = \{1, 2, \ldots, m\}$ denote the set of all items. Without loss of generality, assume each end product consists of exactly one unit of each item in $\mathcal{I}$. (If the end product requires multiple units from some item, simply redefine the unit for that component and adjust inventory and lead time accordingly.) Customer demand for the end product follows a stationary Poisson process, denoted $\{A(t), t \geq 0\}$, with rate λ.

Demands are filled on a first-come-first-served (FCFS) basis. If there is positive on-hand inventory for all components upon a demand arrival, the demand is filled immediately (since the time to assemble the components into the end-product is negligible). On the other hand, if there is a stockout at one or more of the component inventory, then the demand is backlogged until the stockout inventory is replenished.

For each unit of item i, we assume the production lead times are i.i.d. random variables with a common distribution function G_i, and with L_i denoting the associated random variable, and $\mathsf{E}[L_i] = \ell_i$. Assume the lead times are independent among the items.

The inventory of each item i is controlled by a base-stock policy, with R_i being the base-stock level. Since the end product consists of a single unit of each component, the base-stock policy implies that every demand will trigger *simultaneously* the production/replenishment of one unit of each component. Therefore, as in the model of Section 2.1, for each item i, the number of units in process at any time t, denoted $X_i(t)$, is equal to the number of jobs in service in an $M/G_i/\infty$ queue, for $i = 1, \ldots, m$. Note, however, that these m queues are driven by a common Poisson arrival process $\{A(t)\}$, and hence are *not* independent.

For any time t, the performance measures of interest are the on-hand inventory and the number of back orders:

$$I_i(t) = [R_i - X_i(t)]^+, \quad \text{and} \quad B_i(t) = [X_i(t) - R_i]^+ \tag{39}$$

Note that from the FCFS rule, the number of back ordered demand (for the end product) is:

$$B(t) = \max_{1 \leq i \leq m} B_i(t) = \max_{1 \leq i \leq m} [X_i(t) - s_i]^+ \tag{40}$$

Let X_i, I_i, B_i, and B denote the corresponding steady-state limits of the above quantities.

Let f_i and f denote the fill rates for item i and for the end product, respectively. Then, due to the property that Poisson arrivals see time average (PASTA) (e.g., Wolff 1989), we have

$$f_i = \mathsf{P}(I_i > 0), \quad \text{and} \quad f = \mathsf{P}(I_1 > 0, \ldots, I_m > 0)$$

Note that the item-based performance measures f_i, I_i, and B_i depend only on the marginal distribution of X_i for each i, while the product-based performance measures f and B depend on the joint distribution of X_i, $i = 1, \ldots, m$.

From standard queuing results, the marginal distribution of X_i follows a Poisson distribution with mean $\ell_i = \lambda\mathsf{E}[L_i]$, which depends on the lead time distribution G_i only through its mean. This implies that the higher moments of the lead time (variance in particular) do *not* affect the item-based performance. This is no longer true, however, for the joint distribution of $(X_i, i = 1, \ldots, m)$ as we shall see below.

Let $N(a)$ denote a Poisson random variable with parameter (mean) a, along with the following notation:

$$p(n|a) = \mathsf{P}[N(a) = n] = \frac{a^n}{n!} e^{-a}$$

$$P(n|a) = \mathsf{P}[N(a) \leq n] = \sum_{k=0}^{n} p(k|a)$$

$$\overline{P}(n|a) = \mathsf{P}[N(a) > n] = \sum_{k=n+1}^{\infty} p(k|a) = 1 - P(n|a)$$

for $n = 0, 1, \ldots$ Given $R \geq 1$, an integer, we have

$$b(R|a) = \mathsf{E}[N(a) - R]^+ = \sum_{n=1}^{\infty} np(n + R|a) = \sum_{n=R}^{\infty} \overline{P}(n|a) = a - \sum_{n=0}^{R-1} \overline{P}(n|a)$$

The third equality above follows from rearranging the terms, and the fourth equality applies the identity $\Sigma_{n=0}^{\infty} \overline{P}(n|a) = a$. Note that $b(0|a) = \mathsf{E}[N(a)] = a$ and that $b(R|a)$ is decreasing in R, with $b(+\infty|a) = 0$.

Since X_i is equal in distribution to $N(\lambda\ell_i)$, we can express the item-based performance measures as follows:

$$f_i = \mathsf{P}[X_i \leq R_i - 1] = \mathsf{P}(R_i - 1|\lambda\ell_i)$$

$$\mathsf{E}[B_i] = b(R_i|\lambda\ell_i)$$

$$\mathsf{E}[I_i] = R_i - \mathsf{E}[X_i] + \mathsf{E}[B_i] = R_i - \lambda\ell_i + \mathsf{E}[B_i]$$

The last equality above follows from $I_i - B_i = R_i - X_i$, which in turn follows from (39).

Next, consider the joint distribution of $\{X_1(t), \ldots, X_m(t)\}$. For ease of exposition, we start with a system of two components, that is, $m = 2$. Let $N_0(t)$ denote the number of those jobs (orders) that have arrived in $[0, t]$ and are still in service at both queues. Let $N_1(t)$ [resp. $N_1(t)$] denote the number of those jobs that have arrived in $[0; t]$ and are still in service at queue 1 [resp. queue 2], but not at queue 2 [resp. queue 1]. Hence, at time t, there are

$$X_i(t) = N_0(t) + N_i(t)$$

jobs (outstanding orders) in queue i, $i = 1, 2$.

Consider a given $t > 0$. Suppose $A(\mathrm{t}) = n$. Then, it is well known (e.g., Ross 1996) (20) that the n (unordered) arrivals follow an i.i.d. uniform distribution in $[0, t]$ and that $(N_0(t), N_1(t), N_2(t))$ follows a multinomial distribution (of n objects, into four categories), with the following probabilities:

$$p_0(t) = \int_0^t \overline{G}_1(t - x)\overline{G}_2(t - x)(dx/t) = \int_0^t \overline{G}_1(x)\, \overline{G}_2(x)(dx/t)$$

$$p_1(t) = \int_0^t \overline{G}_1(t - x)G_2(t - x)(dx/t) = \int_0^t \overline{G}_1(x)G_2(x)(dx/t)$$

$$p_2(t) = \int_0^t G_1(t - x)\overline{G}_2(t - x)(dx/t) = \int_0^t G_1(x)\overline{G}_2(x)(dx/t)$$

(Note that $1/t$ is the uniform density over $[0, t]$.) Hence, we have

$$
\begin{aligned}
&\mathsf{P}[\mathrm{N}_0(t) = n_0, N_1(t) = n_1, N_2(t) = n_2] \\
&\quad = \sum_{n \ge n_0+n_1+n_2} \frac{n!}{n_0! n_1! n_2! (n - n_0 - n_1 - n_2)!} [p_0(t)]^{n_0} [p_1(t)]^{n_1} [p_2(t)]^{n_2} \\
&\qquad \cdot [1 - p_0(t) - p_1(t) - p_2(t)]^{n - n_0 - n_1 - n_2} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\
&\quad = \frac{[\lambda t p_0(t)]^{n_0} [\lambda t p_1(t)]^{n_1} [\lambda t p_2(t)]^{n_2}}{n_0! n_1! n_2!} \cdot \exp[-\lambda t (p_0(t) + p_1(t) + p_2(t))] \qquad (41)
\end{aligned}
$$

This result indicates that although $X_1(t)$ and $X_2(t)$ are driven by a common arrival process, the three underlying random variables $N_i(t)$ $i = 0, 1, 2$ have independent Poisson distributions with parameters $\lambda t p_i(t)$, $i = 0, 1, 2$, respectively. Thus, $X_1(t)$ and $X_2(t)$ are correlated only because they share a common $N_0(t)$.

Now, the joint distribution of $X_1(t)$ and $X_2(,t)$ can be expressed through the distributions of $N_i(t)$, $i = 0, 1, 2$. Let $x_1 \wedge x_2 = \min\{x_1, x_2\}$. Then

$$
\begin{aligned}
&\mathsf{P}[X_1(t) = x_1, X_2(t) = x_2] \\
&\quad = \sum_{n_0=0}^{x_1 \wedge x_2} \mathsf{P}[N_0(t) + N_1(t) = x_1, N_0(t) + N_2(t) = x_2] \\
&\quad = \sum_{n_0=0}^{x_1 \wedge x_2} \mathsf{P}[N_0(t) = n_0, N_1(t) = x_1 - n_0, N_2(t) = x_2 = n_0] \\
&\quad = \sum_{n_0=0}^{x_1 \wedge x_2} p(n_0 | \lambda t p_0(t)) p(x_1 - n_0 | \lambda t p_1(t)) p(x_2 - n_0 | \lambda t p_2(t))
\end{aligned}
$$

From (40), we have

$$
\begin{aligned}
&\mathsf{P}[B(t) \le x] \\
&\quad = \mathsf{P}[X_1(t) \le x + s_1, X_2(t) \le x + s_2] \\
&\quad = \mathsf{P}[N_0(t) + N_1(t) \le x + s_1, N_0(t) + N_2(t) \le x + s_2] \\
&\quad = \sum_{n_0=0}^{(s_1 \wedge s_2)+x} \mathsf{P}[N_0(t) = n_0] \mathsf{P}[N_1(t) \le x + s_1 - n_0, N_2(t) \le x + s_2 - n_0] \\
&\quad = \sum_{n_0=0}^{(s_1 \wedge s_2)+x} p(n_0 | \lambda t p_0(t)) \mathsf{P}(x + s_1 - n_0 | \lambda t p_1(t)) \mathsf{P}(x + s_2 - n_0 | \lambda t p_2(t))
\end{aligned}
$$

Thus, due to the special relationship between $X_i(t)$ and $(N_i(t), N_0(t))$, for $i = 1, 2$, all the performance measures of interest can be calculated by first conditioning on $N_0(t)$ and then making use of the independence of $N_1(t)$ and $N_2(t)$.

The above analysis extends readily to $m > 2$. In particular, there will be a total of $2^m - 1$ independent Poisson random variables involved: $N_{\mathbb{S}}(t)$, for $\mathbb{S} \subset \mathcal{I}$, representing the number of jobs that are still in process at time t with the queues $i \in \mathbb{S}$, but have been completed with the queues $j \in \mathcal{I} \backslash \mathbb{S}$. For each $i = 1, \ldots, m$, we can write

$$
X_i(t) = \sum_{\mathbb{S}: i \in \mathbb{S}} N_{\mathbb{S}}(t)
$$

That is, $X_i(t)$ can be expressed as the sum of $2^m - 1$ independent Poisson random variables. Hence, in principle, all the performance measures at time t can be exactly evaluated based on the distributions of the independent Poisson random variables via $X_i(t)$'s. The exponential growth (w.r.t. m), however, makes this impractical, even for systems with a moderately large number of components.

An exception is the special case of deterministic lead times, that is, $L_1 \equiv \ell_1$, $L_2 \equiv \ell_2$. Without loss of generality, assume $\ell_1 < \ell_2$. For any fixed t, there are three cases.

Case 1: $t > \ell_2$. In this case,

$$p_0(t) = \int_0^t 1[\ell_1 \geq s]1[\ell_2 \geq s](ds/t)$$
$$= \int_0^t 1[\ell_1 \geq s](ds/t) = \ell_1/t$$

Similarly,

$$p_1(t) = \int_0^t 1[\ell_1 \geq s]1[\ell_2 \leq s](ds/t) = 0$$

and

$$p_2(t) = \int_0^t 1[\ell_1 \leq s]1[\ell_2 \geq s](ds/t)$$
$$= \int_0^t 1[\ell_1 \leq s \leq \ell_2](ds/t) = (\ell_2 - L_1)/t$$

Since $p_1(t) = 0$, $N_1(t) \equiv 0$, we have

$$X_1(t) = N_0(t) \tag{42}$$
$$X_2(t) = N_2(t) + N_0(t) \tag{43}$$

Case 2: $l_1 < t \leq \ell_2$. In this case, we have $p_2(t) = (t - \ell_1)/t$, while $p_0(t) = \ell_1/t$ and $p_1(t) = 0$ stay the same as in Case 1; and (42) and (43) hold.

Case 3: $t \leq \ell_1$. In this case, we have $p_0(t) = 1$, and $p_1(t) = p_2(t) = 0$. So none of the arrived jobs is completed at either queue at time t. Therefore,

$$X_1(t) = X_2(t) = N_0(t).$$

Thus, in the special case of deterministic lead times, there are only $m - 1$ independent Poisson random variables involved, as opposed to $2^m - 1$ in the case of random lead times.

The steady-state performance evaluation can be conducted in a similar manner, as indicated in the next section. Denote

$$\theta_0 = \lim_{t \to \infty} tp_0(t) = \int_0^\infty \overline{G}_1(x)\overline{G}_2(x)\, dx$$

$$\theta_1 = \lim_{t \to \infty} tp_1(t) = \int_0^\infty \overline{G}_1(x)G_2(x)\, dx$$

$$\theta_2 = \lim_{t \to \infty} tp_2(t) = \int_0^\infty G_1(x)\overline{G}_2(x)\, dx$$

Let $t \to \infty$ (41), and write $N_i = N_i(\infty)$ for $i = 0, 1, 2$. We have

$$\begin{aligned} &\mathsf{P}[N_0 = n_0, N_1 = n_1, N_2 = n_2] \\ &= \frac{(\lambda\theta_0)^{n_0}(\lambda\theta_1)^{n_1}(\lambda\theta_2)^{n_2}}{n_0!n_1!n_2!} \cdot \exp[-\lambda(\theta_0 + \theta_1 + \theta_2)] \end{aligned}$$

Thus, N_i, $i = 0, 1, 2$, are independent Poisson random variables with parameters $\lambda\theta_i$, $i = 0, 1, 2$, respectively.

Now, consider the steady-state limit of $(X_1(t), X_2(t))$, denoted (X_1, X_2). First notice the following:

$$\theta_0 + \theta_1 = \ell_1, \quad \theta_0 + \theta_2 = \ell_2$$

From the infinitely divisible property of the Poisson distribution, we can write

$$X_1 = N_0 + N_1 = N(\lambda\theta_0) + N(\lambda\theta_1)$$
$$X_2 = N_0 + N_2 = N(\lambda\theta_0) + N(\lambda\theta_2)$$

The three Poisson random variables, N_0, N_1N_2, involved in the above expressions are independent of one another. Also, X_1 and X_2 are correlated through a common random component $N(\lambda\theta_0)$.

9. KANBAN CONTROL

9.1. The Basic Model

Kanban control refers to a control rule that limits the inventory level, including both WIP and finished goods to K units, with K a constant parameter. Specifically, there are K cards in the system; each card is attached to a job that is in waiting or in process, or to a job that is completed (waiting to supply demand). In other words, a card is needed to admit an order into the system. When all K cards are exhausted, there are three possibilities:

1. If all K cards are attached to completed jobs, then production is suspended.
2. If all K cards are attached to jobs in waiting or in process (i.e., there is no completed job), then any further arrival (demand) will be blocked, i.e., no more jobs are admitted into the system.
3. The situation in between is that some of the K cards are attached to completed jobs while others are attached to jobs in waiting or in process. Then, any arriving demand will be supplied by one of the completed jobs, with its card detached and given to another new job (representing the outstanding order).

This clearly relates to the base-stock control mechanism. In particular, $K = R$—any unit of demand will trigger production (or replenishment), whereas when the finished goods inventory reaches K, production will be suspended. Kanban, however, has the additional feature of blocking arrivals (of demand) when the on-hand inventory drops down to zero, that is, when all K cards are associated with outstanding orders, which is the situation in (2) above. Hence, kanban corresponds to a *finite* queueing system with K being the buffer capacity—the upper limit on the total number of jobs allowed in the system.

With this in mind, in particular, with $R = K$, the models in Section 2.1 should be modified to $M/G/\infty/K$ and $M/M/1/K$ queues (e.g., Wolff 1989). In terms of the distributions of N, I, and B, this amounts to a renormalization: dividing those probabilities in the last section by $\mathsf{P}[N \geq K]$. The inventory and back order expressions in (1) remain valid, with $R = K$. For instance, in the $M/G/\infty/K$ case, we have

$$\mathsf{P}[I = 0] = \mathsf{P}[N = K] = \frac{\rho^K}{K!} e^{-\rho} \left[\sum_{n=0}^{K} \frac{\rho^n}{n!} e^{-\rho} \right]^{-1}$$
$$\mathsf{P}[I = n] = \frac{\rho^{R-n}}{(R-n)!} e^{-\rho} \left[\sum_{n=0}^{K} \frac{\rho^n}{n!} e^{-\rho} \right]^{-1}, \; n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = 1,$$

since $N \leq K$.

9.2. Generalized Kanban Control

Note that in the above kanban control mechanism, the cards control both admission (of demand arrivals) into the system and the upper limit of finished goods inventory. A more general kanban control mechanism is one that uses two types of cards. In addition to the usual type, which now (only) controls admission, there is a second type of cards controlling finished goods inventory. Let K and R, with $K \geq R$, denote, respectively, the number of cards for these two types, referred to, respectively, as kanbans and production cards. Specifically, every job within the system has a kanban attached to it. In addition, to be processed by the server, a job also requires a production card. Both the kanban and the production card stay with the completed job until it supplies a demand. At that point, the released kanban admits a new order into the system, while the released production card authorizes the service (production) of another waiting order. (Note that the above kanban control with two types of cards can be shown to be equivalent to the kanban control with three types of cards, the so-called (a, b, k) model in Glasserman and Yao 1994b, 1996.) This way, the total number of jobs in the system is limited to K, while the total number of *completed* jobs (that are waiting for demand) is limited to R. Since now $R \leq K$, back orders are allowed, up to a limit of $K - R$.

Therefore, what is needed for this more general kanban control is the $M/G/\infty/K$ and $M/M/1/K$ models, in connection with the relations in (1)—with $R \leq K$. The earlier distributions for I and B can be modified accordingly. For instance, in the $M/M/1/K$ case, we have

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \frac{\rho^R - \rho^{K+1}}{1 - \rho^{K+1}}$$

$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \frac{\rho^{R-n}(1 - \rho)}{1 - \rho^{K+1}}, \; n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = \frac{1 - \rho^{R+1}}{1 - \rho^{K+1}}$$

$$\mathsf{P}[B = n] = \mathsf{P}[N = R + n] = \frac{\rho^{R+n}(1 - \rho)}{1 - \rho^{K+1}}, \; n = 1, \ldots, K - R$$

The associated expectations can be derived as follows:

$$\mathsf{E}[I] = \frac{R - (R + 1)\rho + \rho^{R+1}}{(1 - \rho)(1 - \rho^{K+1})}$$

$$\mathsf{E}[B] = \frac{\rho^{R+1} - (K - R + 1)\rho^{K+1} + (K - R)\rho^{K+2}}{(1 - \rho)(1 - \rho^{K+1})}$$

10. NETWORK OF INVENTORY QUEUES

Now consider an inventory/distribution network, also known as a supply chain. Each node in the network represents a stocking location. Suppose a base-stock control policy is followed at each node. With the discussion above, we can adapt the standard decomposition approach in analyzing queueing networks to study this inventory network.

To be specific, focus on a particular node, j, in the network. Suppose node j has a single upstream node i, which supplies any replenishment orders from j. Since base-stock control is followed throughout the network, each node is driven directly by the external demand process, which we continue to assume to be a Poisson process with rate λ.

Suppose j is modeled as an $M/G/\infty$ queue. Then, the only difference from the analysis in Section 2 is that the service time (lead time), L_j, needs to be prolonged whenever there is a stockout at node i. Hence, the modified service time is:

$$\tilde{L}_j = L_j + \tau_i, \text{ w.p. } \Phi(k_i) \tag{44}$$

While $\tilde{L}_j = L_j$ w.p. $\Phi(i_i)$. Here, τ_i is the extra delay at node i, until the next job (outstanding order) is completed. Following the analysis in Ettl et al. (2000), we can approximate τ_i as follows (provided node i is also modeled as an $M/G/\infty$ queue):

$$\tau_i = \frac{L_i \mathsf{E}(B_i)}{R_i \overline{\Phi}(k_i)} \tag{45}$$

Next, suppose both i and j are modeled as single-server queues, say $M/M/1$. Then, the only occasion in which the service time at i needs to be prolonged is when j has a stockout and server i is forced to become idle. Obviously, the probability is dominated by the stockout probability at node j, which is equal to $\rho_j^{R_j}$. In the single-server queue model, stability requires $\rho_j < 1$, hence this probability is quite negligible when R_j is reasonably large and ρ_j is not too close to 1. In this case, we can simply forgo the adjustment. Otherwise, an adjustment similar to the one in (44) is to add $\tau_i = L_i$ to L_j w.p. $\rho_j^{R_j}$. (However, this might result in overcompensation; in particular, the associated probability, $\rho_j^{R_j}$, could be too high.)

Suppose node j interacts directly with external customer demand, and follows a generalized kanban control as outlined in Section 9.2. Then a proportion of the demand will be blocked and lost when all K_j cards are occupied. This is equal to the probability $\mathsf{P}[N_j = K_j]$, where N_j follows the

$M/G/\infty/K_j$ or the $M/G/1/K_j$ model, with the adjusted service time $\tilde{L}_j$ as discussed above. Note that in this case we have a combination of back order and lost sales for external demand.

The simple Poisson demand arrival process assumed here can be readily extended to more involved processes, such as those that involve batches, like the model in Ettl et al. (2000), which makes use of the queueing results in Liu et al. (1990). The same applies to more general lead time distributions and multiple servers; approximate queueing models can be adapted; refer to Buzacott and Shanthikumar (1993).

Optimization models can be formulated based on these approximate inventory-queue models, Ettl et al. (2000) being one such example. As the basic relations in (1) serve as building blocks of the network, one should be able to establish structural properties of the objective function, based on properties of the functions in (1), which are convex in R and submodular in (R, N). These are useful properties in identifying efficient solution algorithms.

In this type of decomposition, each decomposed node is driven by exactly the same external demand processes, due to the base-stock control mechanism. This is quite different from the usual queueing network decomposition, where it is standard to assume that the decomposed queues are independent of each other, which is indeed the case in the special case of product-form networks. In the network of inventory queues, the dependence among the queues should be maximal since the queues are all driven by the same arrival processes. In this regard, the network behaves more like the ATO system in Section 8, where each unit of demand consists of a set of components; and hence, a demand arrival will simultaneously trigger the production at all component queues.

Therefore, to conclude this section, let us examine more closely the issue of stockout in the ATO system and its relation with customer order service. Consider multiple demand streams, indexed by $m \in \mathcal{M}$. Let $\mathcal{S}_m$ denote the set of components required to assemble one unit of end product for type m demand. Let α be the required service level, defined here as the off-shelf availability of all the components required to assemble a unit of type m product, for any m. Let E_i denote the event that component i is out of stock. Then we require, for each end product $m \in \mathcal{M}$,

$$\mathsf{P}[\cup_{i \in \mathcal{S}_m} E_i] \leq 1 - \alpha$$

From the well-known inclusion–exclusion formula (e.g., Ross 1996):

$$\mathsf{P}[\cup_{i \in \mathcal{S}_m} E_i] = \sum_i \mathsf{P}(E_i) - \sum_{i<j} \mathsf{P}(E_i \cap E_j) + \sum_{i<j<k} \mathsf{P}(E_i \cap E_j \cap E_k) - \dots$$

we have, as an approximation,

$$\mathsf{P}[\cup_{i \in \mathcal{S}_m} E_i] \cong \sum_{i \in \mathcal{S}_m} \mathsf{P}(E_i) = \sum_{i \in \mathcal{S}_m} \overline{\Phi}(k_i) \leq 1 - \alpha \tag{46}$$

Note that $\overline{\Phi}(k_i)$ is the stockout probability of component i, using a normal approximation [c.f. (4)].

There is another way to arrive at the above inequality. Suppose we express the service requirement as follows:

$$\prod_{i \in \mathcal{S}_m} \Phi(k_i) \geq \alpha, \quad m \in \mathcal{M} \tag{47}$$

Note that the left-hand side in (47) is, in fact, a *lower bound* of the no-stockout probability of the set of components in $\mathcal{S}_m$ that is required to assemble the end product m, that is, it is a lower bound of the desired off-shelf availability. This claim (of a lower bound) can be argued by using stochastic comparison techniques involving the notion of *association.* (Refer to, e.g. Ross 1996 for background materials.) Intuitively, since the component inventories are driven by a common demand stream $\{D_m(t)\}$, and hence positively correlated, the chance of missing one or several components must be less than when the component inventories are independent, which is what is assumed by the product on the left-hand side of (47).

Since

$$\prod_{i \in \mathcal{S}_m} \Phi(k_i) = \prod_{i \in \mathcal{S}_m} [1 - \overline{\Phi}(k_i)] \cong 1 - \sum_{i \in \mathcal{S}_m} \overline{\Phi}(k_i) \tag{48}$$

where the approximation works best when the stockout probability $\overline{\Phi}(k_i)$ is small, for all components $i \in \mathcal{S}_m$. [Note the approximation in (48) is analogous to the one in (46).] Combining (48) with (47), we arrive at the same inequality in (46).

We can now relate the above off-shelf availability requirement to the standard customer service requirements expressed in terms of response times, W_m, the time it takes to fill a customer order (of type $m \in \mathcal{M}$). Suppose the required service level of type m demand is

$$\mathsf{P}[W_m \leq w_m] \geq \alpha, \quad m \in \mathcal{M} \tag{49}$$

where w_m's are given data. This is the type of service requirement considered in Ettl et al. (2000).

Consider type m demand. We have the following two cases:

1. When there is no stockout at any store $i \in \mathcal{S}_m$—denoting the associated probability as $\pi_{0m}(t)$, the delay is simply L_m^{out}, the outbound lead time—time to process the order, assemble the product, and deliver it to the customer.
2. Suppose there is a stockout at a store $i \in \mathcal{S}_m$. Denote the associated probability as $\pi_{im}(t)$. Then the delay becomes $L_m^{\text{out}} + \tau_i$, where τ_i is the additional delay before the stocked-out component becomes available [cf. (45)].

Hence, we can write

$$\begin{aligned}\mathsf{P}[W_m \leq w_m] &\cong \pi_{0m}(t)\mathsf{P}[L_m^{\text{out}} \leq w_m] + \sum_{i\in\mathcal{S}_m} \pi_{im}(t)\mathsf{P}[L_m^{\text{out}} + \tau_i \leq w_m] \\ &= \left[\prod_{i\in\mathcal{S}_m} \Phi(k_i)\right]\mathsf{P}[L_m^{\text{out}} \leq w_m] + \sum_{i\in\mathcal{S}_m} \overline{\Phi}(k_i)\mathsf{P}[L_m^{\text{out}} + \tau_i \leq w_m]\end{aligned} \tag{50}$$

Note that in the approximation above, we have ignored the probability of two or more components stocking out at the same time [in the same spirit as in (46) and (48)].

In most applications, it is reasonable to expect $L_m^{\text{out}} \leq w_m$. For instance, this is the case when the outbound lead time L_m^{out} is nearly deterministic, and the delay limit w_m is set to be safely larger than L_m^{out}. Furthermore, τ_i can be estimated based on (45). Therefore, if we set the delay limit w_m such that $w_m \geq L_m^{\text{out}} + \tau_i$ with probability one, for all $i \in \mathcal{S}_m$, then the response-time serviceability in (50) can be met at nearly 100% [taking into account (48)]. In other words, aiming for a high off-shelf availability (say, 90–95%) will usually enable us to set a reasonable response-time limit (w_m), and to achieve a near-100% response-time service level.

11. BIBLIOGRAPHICAL NOTES

Buzacott and Shanthikumar (1993) is a rich source of exact and approximate queueing models of production-inventory systems, including generalizations of the models in Section 2.1 and Section 9. Kanban control is studied in detail from a discrete-event systems point of view in Glasserman and Yao (1994a, b, 1996); in particular, the dynamics are modeled as generalized semi-Markov processes, and structural properties such as monotonicity, concavity, and line reversibility (symmetry) are exploited to solve the optimal allocation of kanbans among the production stages. The PAC (production authorization cards) scheme in Buzacott and Shanthikumar (1993) provides a unified modeling framework for many production control schemes, including kanban and MRP.

The materials in Sections 3 through 7 are drawn from Feigin et al. (2000), which also contains a critique of DRP, including numerical examples.

The discussion on ATO systems in Sections 8 and 10 draws materials from Cheng et al. (2000) and Song and Yao (2000). Both papers also have extensive treatments of optimization models, with the latter focusing on queueing analysis and stochastic bounds (also refer to Song 1998; Song et al. 1999), whereas the former focuses on normal approximations and industrial (PC manufacturing) applications. Glasserman and Wang (1998) study ATO systems using a large deviations-based approach, deriving asymptotic results; also refer to Glasserman (1997).

The decomposition-based approach overviewed in Section 10 was first developed in Ettl et al. (2000) for the purpose of performance evaluation and optimization of a large-scale enterprise supply chain. (Refer to Lee and Billington 1986 for an earlier, related work on modeling supply chains.) A related network model, for semiconductor fabrication, appeared in Connors et al. (1996). Also refer to Buzacott and Shanthikumar (1993) for other network models using decomposition-based approximations.

An important topic that we have not discussed in this chapter is the incorporation of quality control into the modeling of production-inventory system (refer to Chen et al. 2000; Yao and Zheng 1999a, b), where the emphasis is on coordinating production and quality control (e.g., inspection), with quality adding a new dimension to the usual inventory-service trade-off.

REFERENCES

Buzacott, J. A., and Shanthikumar, J. G. (1993), *Stochastic Models of Manufacturing Systems,* Prentice Hall, Englewood Cliffs, NJ.

Chen, J., Yao, D. D., and Zheng, S. (2000), "Optimal Replenishment and Rework with Multiple Unreliable Supply Sources," *Operations Research* (forthcoming).

Cheng, F., Ettl, M., Lin, G. Y., and Yao, D. D. (2000), "Inventory-Service Optimization Configure-to-Order Systems: From Machine Type Models to Building Blocks," IBM Research Report.

Clark, A. J., and Scarf, H. (1960), "Optimal Policies for a Multi-Echelon Inventory Problem," *Management Science*, Vol. 6, pp. 475–490.

Connors, D., Feigin, G., and Yao, D. D. (1996), "A Queueing Network Model for Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, pp. 412–427.

Ettl, M., Feigin, G., Lin, G. Y., and Yao, D. D. (2000), "A Supply Network Model with Base-Stock Control and Service Requirements," *Operations Research* (forthcoming).

Federgruen, A. (1993), "Centralized Planning Models for Multi-Echelon Inventory Systems under Uncertainty," in *Logistics of Production and Inventory,* S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Eds., North-Holland, Amsterdam, pp. 133–173.

Feigin, G. E., Katircioglu, K., and Yao, D. D. (2000), "Distribution Resource PlanningSystems: A Critique and Enhancement," IBM Research Report.

Glasserman, P. and Yao, D. D. (1994a), *Monotone Structure in Discrete-Event Systems*, John Wiley & Sons, New York.

Glasserman, P., and Yao, D. D. (1994b), "A GSMP Framework for the Analysis of Production Lines," in *Stochastic Modeling and Analysis of Manufacturing Systems,* D. D. Yao, Ed., Springer, New York.

Glasserman, P., and Yao, D. D. (1996), "Structured Buffer Allocation Problems," *Discrete Event Dynamic Systems: Theory and Applications*, Vol. 6, pp. 9–42.

Glasserman, P. (1997), "Bounds and Asymptotics for Planning Critical Safety Stocks," *Operations Research*, Vol. 45, pp. 244–257.

Glasserman, P. and Wang, Y. (1998), "Leadtime–Inventory Tradeoffs in Assemble-to-Order Systems," *Operations Research*, Vol. 46, pp. 858–871.

Li, L. (1992), "The Role of Inventory in Delivery-Time Competition," *Management Science*, Vol. 38, pp. 182–197.

Lee, H., and Billington, C. (1986), "Material Management in Decentralized Supply Chains," *Operations Research*, Vol. 41, pp. 835–847.

Liu, L., Kashyap, B. R. K., and Templeton, J. G. C. (1990), "On the $GI^X/G/\infty$ System," *Journal of Applied Probability*, Vol. 27, pp. 671–683.

Martin, A. J. (1990), *DRP Distribution Resource Planning: Distribution Management's Most Powerful Tool*, 2nd Ed. Prentice Hall, Englewood Cliffs, NJ, and Oliver Wright, Essex Junction, VT.

Nahmias, S. (1997), *Production and Operations Analysis*, 3rd Ed., Irwin, Chicago.

Press, W. H., Teukolsky, S. A., Vettering, W. T., and Flannery, B. P. (1994), *Numerical Recipes in C*, 2nd Ed., Cambridge University Press, New York.

Ross, S. M. (1996), *Stochastic Processes*, 2nd Ed., John Wiley & Sons, New York.

Scarf, H. (1960), "The Optimality of (s, S) Policies in the Dynamic Inventory Problem," in *Mathematical Methods in the Social Sciences*, K. Arrow, S. Karlin and P. Suppes, Eds., Stanford University Press, Stanford, CA.

Silver, E. A., Pyke, D. F., and Peterson, R. (1998), *Inventory Management and Production Planning and Scheduling*, John Wiley & Sons, New York.

Song, J. S. (1998), "On the Order Fill Rate in a Multi-Item, Base-Stock Inventory System," *Operations Research*, Vol. 46, pp. 831–845.

Song, J. S., Xu, S., and Liu, B. (1999), "Order Fulfillment Performance Measures in an Assemble-to-Order System with Stochastic Leadtimes," *Operations Research*, Vol. 47, pp. 131–149.

Song, J. S., and Yao, D. D. (2000), "Performance Analysis and Optimization of Assemble-to-Order Systems with Random Leadtimes," preprint.

Stenger, A. J. (1994), "Distribution Resource Planning," in *The Logistics Handbook*, J. F. Robeson and W. C. Copacino, Eds., Free Press, New York.

Tijms, H. (1994), *Stochastic Modeling and Analysis: A Computational Approach*, John Wiley & Sons, New York.

Wolff, R. (1989), *Stochastic Modeling and the Theory of Queues*, Prentice Hall, Englewood Cliffs, NJ.

Yao, D. D., and Zheng, S. (1999), Sequential Inspection under Capacity Constraints. *Operations Research*, Vol. 47, pp. 410–421.

Yao, D. D. and Zheng, S. (1999), ‘‘Coordinated Quality Control in a Two-Stage System,’’ *IEEE Transactions on Automatic Control*, Vol. 44, pp. 1166–1179.

Zipkin, P. (2000), *Foundations of Inventory Management*, Irwin/McGraw-Hill, New York.

CHAPTER 62

Process Design and Reengineering

JOHN TAYLOR
TARSHA DARGAN
BEN WANG
Florida Agricultural & Mechanical University
Florida State University

1. INTRODUCTION 1696
  1.1. Definitions 1696
  1.2. Assumptions 1697
  1.3. Steps 1697
2. THE EVOLUTION OF PDR 1699
  2.1. Backdrop 1699
  2.2. Implementation 1700
  2.3. Postimplementation 1700
  2.4. Failure Reasons 1700
  2.5. The Second Wave 1701
  2.6. State of the Art 1701
3. TOOLS FOR REENGINEERING 1702
  3.1. Benchmarking 1703
  3.2. Modeling and Analysis Tools 1703
  3.3. Simulation 1703
  3.4. Activity-Based Costing (ABC) 1704
4. IMPLEMENTATION 1704
  4.1. Preplanning 1704
  4.2. Envision 1705
    4.2.1. Establish Management Commitment and Vision 1705
    4.2.2. Discover Reengineering Opportunities 1705
    4.2.3. Identify Information Technology Levers 1706
    4.2.4. Select the Process 1706
  4.3. Initiate 1706
    4.3.1. Inform Stakeholders 1706
    4.3.2. Organize Reengineering Teams 1707
    4.3.3. Conduct Project Planning 1707
    4.3.4. Determine External Process Customer Requirements 1708
    4.3.5. Set Performance Goals 1708
  4.4. Diagnose 1708
    4.4.1. Document Existing Process 1708
    4.4.2. Analyze Existing Process 1709
  4.5. Redesign 1709
    4.5.1. Define and Analyze New Process Concepts 1709
    4.5.2. Prototype and Detailed Design of New Process 1709
    4.5.3. Design Human Resource Structure 1710
  4.6. Reconstruct 1710
    4.6.1. Reorganize 1710
    4.6.2. Implement 1711
    4.6.3. Train Users 1711
  4.7. Evaluate 1711
    4.7.1. Evaluate Process Performance 1711
    4.7.2. Link to Continuous Improvement Programs 1712
5. CASE STUDIES 1712
  5.1. Corning Asahi Video 1712
  5.2. Uarco, Inc. 1713
  5.3. D2D, Ltd. 1713
  5.4. Fortune 500 Insurance Company 1713

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

6. LESSONS LEARNED 1714
7. FUTURE OUTLOOK 1714
8. REFERENCES 1715
ADDITIONAL READING 1717

1. INTRODUCTION

Process design and process reengineering have received increased attention over the last decade as companies, responding to increased competition, have struggled to become more customer focused and price competitive. There has been substantial debate over precise definitions of process, business process, and process reengineering (Love et al. 1998a). There is a desire to arrive at a uniform set of definitions so that the case studies can be effectively evaluated. Aside from the debate over labels, process design and reengineering (PDR) has offered companies new ways of assessing their operations and most importantly, improving them by large orders of magnitude. While simple in concept, PDR has proven difficult to implement successfully. Some possible reasons are:

- It is not so much a scientific theory as it is a paradigm (Weicher et al. 1995).
- It "is a messy, complex process which requires strong leadership, the application of special skills and methodologies, and an ongoing commitment to process improvement" (Caudle 1995).
- Because companies and processes are different, seemingly similar PDR projects are vastly different in execution (Caudle 1995).
- It is not "universally applicable" (Coombs and Hull 1997).

However, the benefits are lucrative, and therefore companies will continue to attempt to reengineer their processes.

1.1. Definitions

A *process* may be defined as a "sequence of pre-defined activities executed to achieve a pre-specified type or range of outcomes" (Lee and Dale 1998) or a "sequence of activities, which are performed across time and place" (Bal 1998). It is a set of activities that transform a set of inputs into a set of outputs. Furthermore, a process may be classified as "(1) the sort that starts when necessary and finished some time in the future; or (2) the sort that is constantly running" (Bal 1998).

A *business process* is "a collection of related, structured activities—a chain of events—that produces a specific service or product for a particular customer or customers" (Caudle 1995). They may be classified as "mission or external customer-facing processes, support processes, and management processes" (Caudle 1995). There are also two critical characteristics of a business process: (1) it has external or internal customers, and (2) it crosses organizational boundaries (Malhotra 1998). A representative illustration of a business process is shown in Figure 1. Note that it spans across functions and that it has internal and external customers.

Business process redesign and *business process reengineering* are synonymous with PDR. The terms were introduced, respectively, by Davenport and Short (1990) and Hammer (1990). Each refers to "a systematic, disciplined approach for achieving dramatic, measurable performance improvements by fundamentally reexamining, rethinking, and redesigning the processes that an organization uses

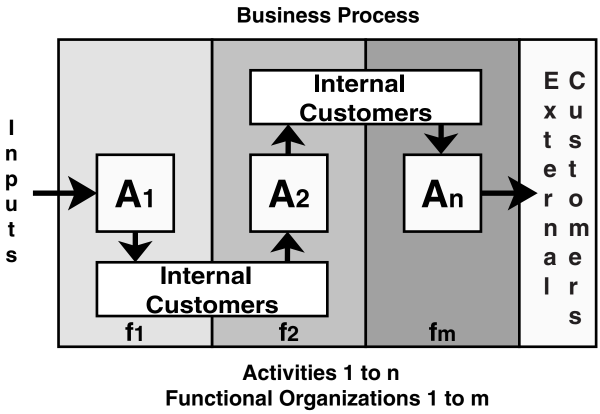


Figure 1 Business Process.

to carry out its mission'' (GAO 1995). Since that time, several additional definitions have been contributed by various authors. Because of the multiple definitions in circulation, it has been suggested that ''the reengineering label appears simply to be related to the notion of taking an organization apart and putting it back together again, more efficiently, less expensively and with fewer personnel'' (Buchanan 1998). While this statement is essentially correct, it oversimplifies the activity and challenges it involves as well as the results realized. To illustrate the different thoughts involved with PDR, consider Figure 1. Before PDR, if a person were queried about the processes present in the company, he or she would invariably answer along the lines of f1, f2, or marketing, sales, and production. Note that these are descriptions of functions within the organization and not necessarily processes. Reengineered processes are usually described in terms of the beginning and end states—for example, purchasing, which is not really a process but a function within the organization. The process would be aptly described as ''requisition to delivery.'' The difference may seem trivial, but as a starting point for PDR it forces thought towards the activities taking place between the endpoints.

In summary, the following characteristic definition regarding redesign or reengineering emerges: ''[PDR] focuses on business process, adoption of information technology, taking fundamental analysis and radical redesign, and aims to achieve dramatic improvement in a short time delivery'' (Choi and Chan 1997). It has been contrasted with TQM in that, while both center on the concept of improvement, PDR stresses radical improvement in a short time while TQM relies on incremental improvement over an indefinite period (Jarrar and Aspinwall 1999).

Finally, business process management (BPM) is the natural resultant of PDR. While some definitions include PDR as a subset of BPM, it is obvious that if processes have been radically altered, the method of managing them must be altered as well, including performance metrics, monitoring, training, and, in some cases, the organizational structure itself (Pritchard and Armistead 1999). To this end, BPM is defined as ''a systematic, structured approach to analyze, improve, control and manage processes with the aim of improving the quality of products and services'' (Lee and Dale 1998).

1.2. Assumptions

The fundamental assumptions underlying PDR are (Biazzo 1998):

1. The organization is a collection of processes that can be reengineered ''scientifically'' and systematically.
2. The nature of change is revolutionary and consists of:
   - The passage from functional units to process teams
   - A move from simple tasks to multidimensional work
   - Changes in power relations towards worker empowerment
   - Change from a ''bureaucratic'' culture to one based on customer satisfaction
   - Changes in managerial behavior from supervisors to trainers
3. Planning for this change is top-down.

1.3. Steps

There are a variety of opinions regarding the actual steps involved in process reengineering. A typical outline of the phases is listed below. It represents a consensus view from a survey of practitioners (Kettinger et al. 1997).

1. Envision:
   - Establish management commitment and vision.
   - Discover reengineering opportunities.
   - Identify information technology (IT) levers.
   - Select process.
2. Initiate:
   - Inform stakeholders.
   - Organize reengineering teams.
   - Conduct project planning.
   - Determine external process customer requirements.
   - Set performance goals.
3. Diagnose:
   - Document existing process.
   - Analyze existing process.

4. Redesign:
   - Define and analyze new process concepts.
   - Create prototype and detailed design of a new process.
   - Design human resource structure.
   - Analyze and design information systems (IS).
5. Reconstruct:
   - Reorganize.
   - Implement.
   - Train users.
   - Process cut-over.
6. Evaluate:
   - Evaluate process performance.
   - Link to continuous improvement programs.

There are critical aspects to each of the phases as well as the elements represented within. These will be discussed in the implementation section of this chapter. The six-phase process represents a logical approach to a reengineering project and is illustrated in Figure 2.

Deceptively simple, this framework implies that PDR is simply a matter of going through a sequence of steps. However, beneath this framework lies the question of how to implement the process of change. In that respect, PDR "still remains a mystery with few rules and methods to guide firms through their endeavor" (Crowe and Rolfes 1998). There is a great deal to be learned from the experience of others, but that experience must be continually evaluated against the culture of the firm that is considering reengineering.

Finally, the following general observations may be made about reengineering (Caudle 1995):

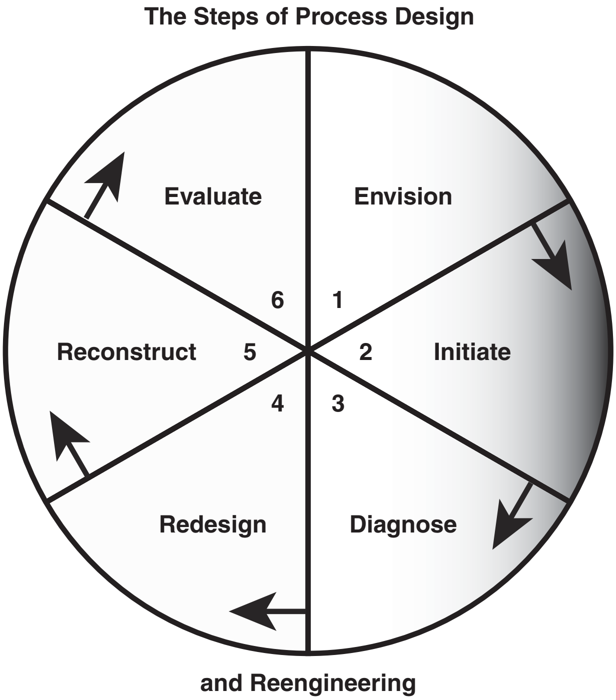


Figure 2 The Steps of Process Design and Reengineering.

1. Be aggressive about ambitions.
2. Recognize that while a lot can be done within single departments and agencies, a lot more can be accomplished through cross-departmental integration of effort.
3. Get the policy—which in private enterprise equals strategy—clear at the top.
4. Finally, always keep in mind the hidden reefs of politics.
5. Get it done fast.

2. THE EVOLUTION OF PDR

2.1. Backdrop

A variety of factors led companies to begin to experiment with changes that eventually developed into the process reengineering paradigm. Half of the Fortune 500 companies listed 20 years ago are not on that list today. The missing ones failed to see the future in terms of customer needs, evolving technology impacts, and competition (Dickie 1995). Most, if not all, were based on the management and operational structure that developed at the beginning of the twentieth century and thus had been developed for companies in stable, expanding marketplaces.

That structure performed remarkably well through the first two-thirds of the century. Unfortunately, task-oriented work required the development of functional hierarchies for supervision and control in the workplace. As the hierarchies grew, a silo mentality grew as well. Management focused on the performance of their particular function alone, making it difficult or impossible to include internal and external customers into the decision-making process. Also, efforts directed at suboptimizing the performance of a function often had little effect on the performance of the entire system. Increased competition forced companies to improve their productivity, but that effort was frustrated by the very organizational structures that once worked so well (Lucier and Torsilieri 1999). Figure 3 illustrates the inherent problems with the hierarchical structure. Note that the communication barriers that form between the functional organizations and that optimization attempts are performed vertically and do not necessarily enhance the product flow. Furthermore, as the hierarchy grows, the command-and-control nature inherent to the organization means that the distance between the decision maker and the decision point increases.

The pyramid hierarchical organization had been rendered invalid or inoperable in most cases (Love et al. 1998a). The division of labor no longer worked either (Lucier and Torsilieri 1999). TQM programs, which helped companies rediscover the customer, also introduced an incremental change paradigm, which likewise was "no longer considered to be adequate for acquiring a competitive advantage in the marketplace" (Love et al. 1998a). The internal workings of a company had to be redefined.

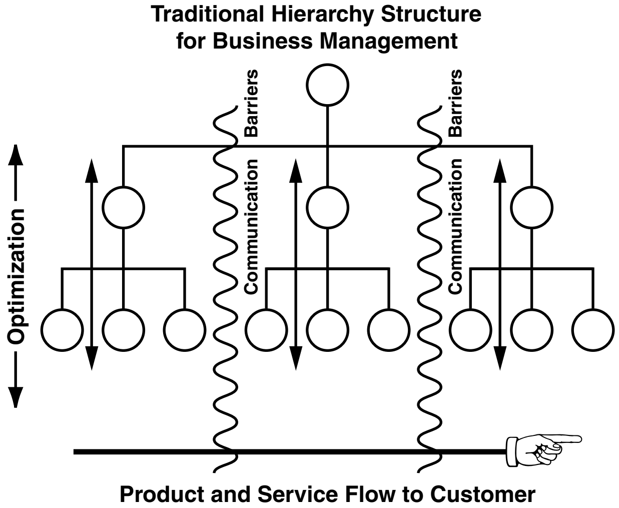

Traditional Hierarchy Structure for Business Management
Barriers
Barriers
Communication
Communication
Optimization
Product and Service Flow to Customer

Figure 3 Hierarchical Organization Schematic.

There is now a "perceived linkage between strategy and organizational structures" (Coombs and Hull 1997). The advent of the concept of a business process has permitted companies to examine how they do their work, what they do, and why they do it. "Critical business processes are the basic means to support a company's strategic objectives according to set priorities" (Dervitsiotis 1999). The new process focus facilitates improvements that allow businesses to determine which activities are value-adding ones (Guimaraes and Bond 1996).

2.2. Implementation

The concept of process reengineering went mainstream in 1993 when Michael Hammer and James Champy's *Reengineering the Corporation* was published. The concept was eagerly embraced because it offered breakthrough changes in performance. Process reengineering held promise because it encompassed three previously unconnected components: information technology, business processes, and a greenfield or blank sheet approach to change (Davenport 1996). While none of these components was new, there were inherent dangers in using them individually. These are shown in Figure 4.

The information technology piece allowed companies to imagine previously impossible scenarios of executing work. This, in concert with a process view of work, formed the foundation for the improvement of process reengineering. The last component, which is sometimes referred to as the blank sheet concept of change, promised that there would be no legacy components from an old system unless they were critical to the new process. In theory, designers would be able to design it right the first time and overcome the inertia of status quo. Figure 5 illustrates the results of using these components in tandem.

2.3. Postimplementation

As companies began to implement process reengineering programs and data were gathered, it became apparent that breakthrough performance gains were more elusive than previously thought. Estimates of failed "process reengineering" programs range anywhere from 50–70%. The failures were not merely isolated to flat responses in productivity. One of the direct consequences of a reengineered process without an increase in market demand is surplus labor. Displaced workers soon began to be referred to as "reengineered" workers. By 1996, process reengineering had come under attack as a damaging fad with "ruthless objectives and unclear methodology" (Buchanan 1998). Debate continues, and four main lines of thought have emerged. The prophets continue to recommend process reengineering in its original form. The disciples strive to achieve the gains of classical reengineering with more practical implementation strategies. The revisionists declare the shortcomings of classical process reengineering to be a lack of human focus and thus try to reengineer around those shortcomings. The last group, the skeptics, attack the concepts and methodology entirely (Buchanan 1998).

2.4. Failure Reasons

Several possible reasons for the high failure rate of process reengineering have been discussed in the literature. Harrington proposes five (Harrington 1998a):

1. The methodology was misused and the reported results were misleading.
2. Process reengineering's negative impact on the organization was poorly defined and not considered in its implementation package.
3. The creative part of process reengineering was not understood.
4. The cycle time from start of project to reaping results was too long.

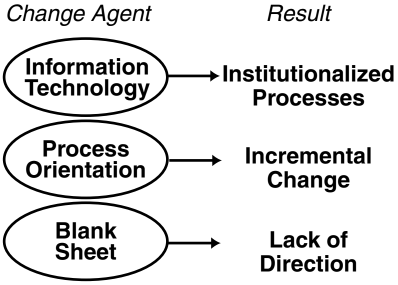


Figure 4 The Components of Reengineering: Separate.

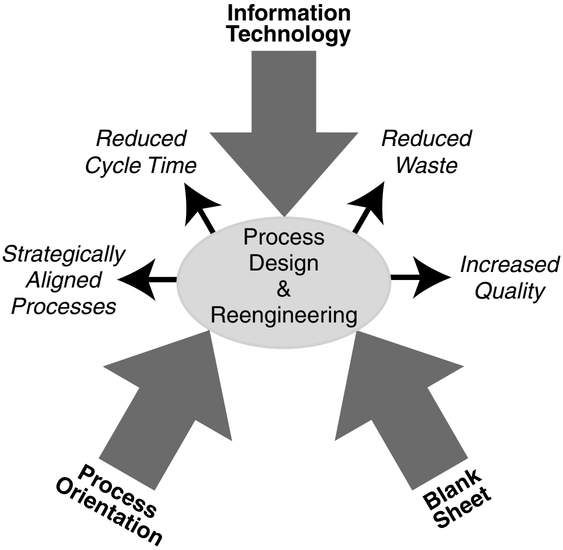


Figure 5 The Components of Reengineering: Together.

5. It produced too good a result—the methodology failed to consider anything other than performance in developing alternatives.

Another factor that contributes to less-than-desired results in process reengineering is that newly engineered processes are often implemented in traditional functional management structures (Love et al. 1998a). This can create conflict within the system (Lucier and Torsilieri 1999)—or worse, the process can revert to its old state via work-arounds (Fisher 1997). The functional management paradigm must change along with the process. The blank sheet approach has also been criticized in that it can create impractical solutions that conflict with organizational priorities (Buchanan 1998). Apparently not all of the linkage to the old process is negative. There are times when certain aspects of a given process exist for economic, political, or other good reasons, none of which may be obvious to the casual observer. These and other negative dimensions of process reengineering are changing the methodology by which it is done. The process of PDR continues to evolve as experience is gained (GAO 1995).

2.5. The Second Wave

Due in large part to the shortcomings of the original attempts at reengineering, the concept of the second wave of reengineering has developed. The companies in this wave will benefit from the experience of those who participated in the first wave. This wave will focus on ‘‘growing the top line rather than cutting costs’’ (Dickie 1995). Figure 6 illustrates this trend graphically. It has three basic principles (Moeller et al. 1996):

1. Enhance the value for cost to the customer. Rather than reengineering in order to cut or maintain the cost to the customer, attempt to increase the value to the customer.
2. Realign the processes and business systems for growth. While classical reengineering produced short-term optimization in terms of costs, the second wave will make assessments in terms of the short and long term.
3. Refocus on the soft side of capabilities development. Management vision and strategy must be transformed to reflect the values of the second wave. Measurement and reward systems will need to be revisited to ensure that they promote growth. Training and management development investments will increase, and human resource processes will need modification.

2.6. State of the Art

It appears that reengineering is poised to make a comeback (Bartholomew 1999). This wave will have the wealth of experience of the first wave to draw upon. As mentioned above, it will have significantly different goals and will attempt to shore up the weaknesses of previous attempts. While

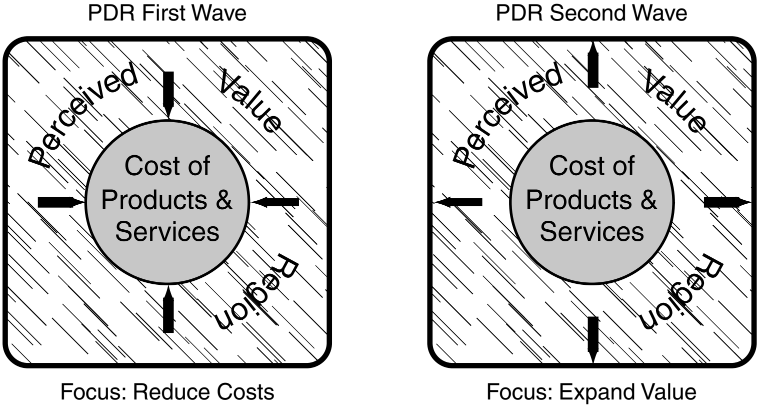


Figure 6 PDR: First to Second Wave.

IT was one of the major components of the first wave, more focus will be applied to change management in the second wave (Guha et al. 1997). The surplus problem associated with the implementation of PDR's original form drove fear into, rather than out of, the organization. As a result, people are less willing to put their heart and soul into a project that could very well cost them their jobs. As a remedy to this, some firms have implemented policies to protect personnel from layoff (Caudle 1995). The effectiveness of measures such as these remains to be seen, but the very fact that they are being attempted underscores the potential value PDR holds for companies.

3. TOOLS FOR REENGINEERING

A variety of tools are available to aid in reengineering efforts. The benefits that reengineering tools provide range from process visualization and analysis to process costing to competitive analysis. When selecting reengineering tools, it is best to give consideration to those tools that are easy to learn and easy to use. If a tool requires a long learning period, its value may be diminished. When comparing tools, proportionally discount the value of a given tool relative to the time required to learn to use it. Obtain reference data from other users about a given tool and its features. Also, consider the in-house training aspects. The rest of the reengineering team must be able to learn to use the tool effectively as well (Manganelli and Klein 1994).

A reengineering tool should have at least one of the following characteristics:

1. It must be usable by the reengineering team.
2. It must provide a return on investment (ROI).
3. It must aid in clarifying the vision.
4. It must tie top-level corporate goals to subsystem goals.
5. It must integrate into the existing infrastructure.

Reengineering tools should perform at least one of the following tasks (Roberts 1994):

1. Recording or data structuring
2. Decision support
3. Process flow diagramming
4. Flow path determination
5. Process measure determination
6. Path optimization
7. Modeling
8. Simulating
9. What-if analysis
10. Constraint isolation
11. Reporting/plotting
12. Planning/tracking

Reengineering tools include benchmarking, modeling and analysis tools, simulation, and activity-based costing.

3.1. Benchmarking

Benchmarking is defined as ". . . the search for the best practices that will lead to superior performance of a company . . . and [that will] allow a manager to compare his or her function's performance to the performance of the same function in other companies" (Camp 1989). The desired result of a benchmarking effort is a comparison to external but similar measures of system performance, such as quality and cost and productivity (Grover et al. 1995).

Typical benchmarking steps are:

1. Determine objectives and define the scope of the benchmark study.
2. Gain support from within the organization.
3. Select a benchmarking approach to use.
4. Determine who the benchmarking partners are.
5. Retrieve benchmark data via research, surveys, and benchmarking visits.
6. Using the results, determine the necessary practices or principles that are absent within the current system.
7. Choose the principles or practices to implement.
8. Implement.

There are four broad categories of benchmarks that provide the most valuable insight into the performance of a company: business results, cycle time, quality assurance, and asset. Business results are typically financial ratios, cycle time deals with task completion, quality assurance deals with customer-related measures, and asset benchmarks range from inventory turns to human asset measures (Schwartz, 1998).

The information necessary for benchmarking may be found in public domain, technical papers, and panel discussions, as well as, information exchange (benchmarking partnership). A benchmarking partnership allows each party to gather information about the other with the goal of creating better processes. Previous benchmarking studies of related industries, libraries, conferences, databases and suppliers are also key opportunities for gathering information for benchmarking (Stork 2000).

3.2. Modeling and Analysis Tools

Modeling tools allow the process to be diagrammed. Process modeling is crucial because it depicts the business, the relationships, and the flow of information. With this information, the impact or potential ripple effect of changes can be determined (Morris and Brandon 1993). In process modeling, activities and their connecting links are labeled and assigned to all necessary levels (Roberts 1994). Modeling tool-selection issues are methodology, alternative views, form of input, simulation, and standardization. Process modeling may be done with flowcharts, tree diagrams, fishbone diagrams, hierarchy charts, computer-simulated models, and mathematical models. One of the most popular methods is flowcharting. Flowcharts graphically illustrate the steps of an activity. Descriptive symbols are used for each step and arrows indicate the flow. The flowcharting steps are:

1. Define steps and sequence.
2. Identify all relationships and decisions.
3. Draw straight-line representation of all work steps.

Process analysis tools are used to enter, view, and track the process inputs. These tools, like modeling tools, must show different views of the results. Most process analysis tools are packaged with modeling tools. Process analysis tool selection considerations are basically the same as those of modeling tools. Typically, process modeling and analysis tools: (Yu and Wright 1997):

1. Provide system visualization through system diagrams
2. Are PC based
3. Are easy to learn to use and scalable
4. Provide analysis and performance measures of static system
5. Provide modeling of dynamic behaviors of system

3.3. Simulation

Simulation tools, while not specifically designed for PDR, provide additional methods to analyze the dynamic nature processes. "Fundamental to the re-engineering effort is the ability to simulate the changes that are being proposed'' (Morris and Brandon 1993). Simulation is dynamic process modeling. Simulation allows the user to predict the behavior of a system under certain circumstances without actually building the system. Simulation is just an approximation of the physical system, but it allows changes to be rapidly incorporated and tested under a variety of conditions. Through computer simulation, process fluctuations, optimal settings, and the effect of various inputs may be determined. Basic simulation inputs are task time, resources, and demand, while the outputs are cost, throughput, cycle time, utilization, and bottlenecks (Petrozzo and Stepper 1994).

3.4. Activity-Based Costing (ABC)

The need of reengineering teams for accurate information regarding the consumption of resources by processes led to renewed interest in activity-based costing (ABC). ABC is an accounting methodology in which costs are assigned to activities, as opposed to products or services. ABC more accurately distributes the applicable costs to those products and services that consume them. ''The precision of ABC is useful in establishing the true costs of the component and the compound process outputs'' (Grover et al. 1995). ABC provides necessary information as to where the money is going or where certain costs lie. An example of activity-based costing is shown in Figure 7.

In ABC, costs are assigned to activities based on resource consumption. Cost is then assigned to elements such as products or customers. This information provides the foundation for deciding on issues such as outsourcing and corporate spending. Often, when traditional accounting systems are used for costing, cost is overallocated to high-volume production and underallocated to low-volume production on a per-unit basis. To portray the value of proposed systems over current ones accurately, it is imperative that accurate methods be employed for costing.

ABC is used by reengineering teams to perform cost-benefit analysis on proposed systems including the segmented delivery of products and services. ABC will determine the cost and viability of a product/process/system, both before and after reengineering. The steps for ABC are (Maluso 1997):

1. Use a process map to identify the major business processes and key activities of the organization.
2. Attach the cost to these key activities. Financial data may be gathered from existing data, including information on labor and capital equipment expenses, budget, general ledger, and supplier invoices.
3. Link activities to processes and identify the cost drivers.
4. Summarize the total costs for each process.
5. Do this as well for the reengineering process.

4. IMPLEMENTATION

4.1. Preplanning

As mentioned in the introduction, a framework exists for conducting PDR activities. Not explicitly noted in the outline is the necessity for training and planning before any large-scale reengineering effort begins. Ultimately, the entire organization will be required to think in terms of process. This is one of the primary challenges of process management (Lee 1996). For reengineering to flow smoothly in the planning and design stages, management and staff will need to know its function,

Traditional		ABC		
Wages	$150.00		# Units	
Supplies	$ 50.00	Clean Bathroom	6	$150.00
Insurance	$ 30.24	Vacuum	3	$21.00
Transportation	$ 10.00	Dust	3	$15.00
		Change Linen	3	$12.00
Profit	$ 59.76	Empty Trash	6	$12.00
		Clean Kitchen	3	$90.00
Cost	$300.00			$300.00

Figure 7 ABC vs. Traditional Costing.

its implications, and how it fits into the scope of improvement efforts (Caudle 1995). The vision for change must be articulated over and over again. Management and staff will need to be educated in the concepts of process and reengineering. The design team should thoroughly understand the design process (Feather 1998) and be cognizant of its linkage with organizational structure (Gappmaier 1997). Spending time, effort, and money on educating people before beginning the reengineering process pays back by eliminating misunderstanding and indecision.

For most companies, the process view of work is a new one (Lee 1996). Before making the decision to reengineer along process lines, management must be firmly committed to that end as well as the necessary changes it entails (Caudle 1995). They must have in mind a target performance level that they wish the new structure to achieve (Love et al. 1998a) and a measurement system must be in place in order to let them know whether the reengineered system is, in fact, performing to expectations. This requires that employees be educated on performance measures and uses (Caudle 1995).

Lastly, management must assess the culture of the organization and determine its propensity towards change. This has proven to be a major stumbling block in a majority of the reported PDR efforts (Corrigan 1997). It is apparent from the literature that some organizational cultures have a greater capacity for accepting and processing change than others (Love et al. 1998b). Organizations that have the ability to accept change readily have been referred to as learning organizations and are "characterized by the ability to adapt and improve, to build internal and external knowledge, and to achieve higher levels of learning that may be critical to successful [PDR]" (Guha 1997). Communicating early and throughout can help overcome resistance to change (Feather 1998) and is viewed as a critical success factor to PDR projects (Andrews 1996a).

Figure 8 provides a summary of the steps involved in preplanning.

4.2. Envision

4.2.1. Establish Management Commitment and Vision

A prerequisite to successful implementation of a BPR project is unequivocal support from upper management (Ho 1999). The vision must be communicated to and embraced by the entire organization, but this communication is especially important in dealing with midlevel and functional managers who are affected by the changes (Guha 1997). As the implications of change are realized, risk aversion and uncertainty can develop, creating an atmosphere where less than optimal solutions are developed (Morris et al. 1999). In these cases, management's commitment is tested. The case for change and the vision of the future state must be coupled with a strategy for achieving change. Successful organizations require and practice hands-on senior management's aggressive ownership of process improvement through personal responsibility, involvement, and decision making. Most often this is seen in the use of executive committees, steering groups, and the assignment of process owners (Caudle 1995).

4.2.2. Discover Reengineering Opportunities

The breakthrough improvements available from reengineering can come in one of two forms. Reengineering a process can take away the inefficiencies present in terms of cost and time to process but leave the general value adding sequences the same. This is accomplished by removing organizational

Preplanning Checklist

☑ **Ramifications of Process Viewpoint Understood**
☑ **PDR as Continuous Improvement Understood**
☑ **Upper Management Commitment**
☑ **Location and Magnitude of Expected Gains Known**
☑ **Current Measures for Targeted Processes**
☑ **Corporate Culture Capable of Required Change**

Figure 8 Preplanning Step Checklist.

issues, archaic rules, and legacy procedures that are counterproductive to the overall goal (Bartholomew 1999). Companies gain from examining their processes in this way, but not every instance will yield the same ''opportunity for dramatic breakthrough'' (Bal 1998). The most significant improvements are realized when entirely new processes are developed for bringing existing products and services to customers. The following three characteristics describe efforts of this sort (Feather 1998):

1. New opportunities in markets, products, and services where value is created in the business process are discovered and exploited.
2. New ways of thinking about the process using design principles and process benchmarks are utilized.
3. Innovative core processes that encourage people to behave differently are developed.

Being creative is one of the most demanding aspects of PDR. One of the features of PDR that has allowed companies to be more creative with their process thoughts is information technology.

4.2.3. Identify Information Technology Levers

The strategic importance of information was always known, but until recent times nothing could be done to exploit it. Modern technology, however, has served to erase the geographic barriers that once stood in the way of information flow. Likewise, it has permitted the lag time that existed between information collection, data entry, and report to be reduced. Real-time information processing is a reality. Realizing this in process design permits processes to be developed that behave and perform drastically different than their predecessors. ''Successful [PDR] involves the coalescence of 'IT' and business best practice, whereby IT plays a supportive, but not always commanding role that is linked to the business case for [PDR]'' (Guha et al. 1997).

The specific roles played by IT in PDR are many and vary from case to case. It may be used to enhance communications at the customer interface through electronic data interchange, software training systems, and distributed product/service specifications and performance data. It can facilitate the distribution of information across and outside of the organization through centralized, shared databases, networks, and wireless communications. It enables companies to provide services economically at remote locations with expert systems. The rapid increase in computing power has allowed companies to eliminate the batch mentality altogether. No longer are strategic decisions made in concentrated groups at specific points in time. Plans can update in real time and data models can provide revised forecasts instantly. Accurate snapshots concerning production and distribution data are accessible with electronic tracking and identification. Customer orders can be monitored and accurate lead times can be quoted. Internal organizational performance can be monitored as well. All of these innovations typically provide the backbone of redesigned processes. They all have the common trait of providing the necessary information to the necessary destination in a usable form at the time it is needed.

While it is generally agreed that IT is the enabler for paradigm shifts required for reengineering, it has also been pointed out that its misuse can prevent PDR projects from succeeding by institutionalizing old concepts and processes (Hammer and Champy 1993). Simply automating the old systems is not reengineering and should not be confused with it. The fundamental concept in successfully utilizing IT in PDR is that it is used to change the way things are currently done or even eliminate the need for them, not merely to optimize them.

4.2.4. Select the Process

There is much to be said about selecting a process for reengineering. To achieve the greatest gain, processes selected for reengineering must have strategic importance. ''A fact often overlooked is that processes exist to support specific strategic goals considered vital for survival and success'' (Dervitsiotis 1999). Companies should begin by evaluating their business strategy and then determining which processes are critical to the strategy (Bartholomew 1999). Failure to do this has been cited as a contributor to implementation failure (Peltu et al. 1996). All processes must be reviewed so that the interdependence among them may be fully understood (Guha et al. 1997).

Figure 9 provides a summary of the envision step activities.

4.3. Initiate

4.3.1. Inform Stakeholders

Everyone associated with the process must be informed of the need for and intent of the reengineering effort. Management, customers, suppliers, employees, and labor unions are included in this group. The sharing of information from the outset decreases the possibility of misunderstanding and increases the possibility of success (Lee 1996). From the customer's perspective, this enhances the channels of communication regarding their service level needs. From the employee's perspective, this

Envision Step Checklist

☑ **Upper Management Support Fostered Through**
  - **Process Ownership**
  - **Steering Committee Representation**

☑ **Overall Vision Articulated and Embraced by Organization**
  - **Competitive Forces**
  - **Desire to be "Best" in Class**
  - **Customer Demand**

☑ **Reengineering Opportunities Isolated Through**
  - **Process Benchmarks**
  - **Focus on/Development of Value-Added Services**

☑ **Information Technology Levers**
  - **Networked, Centralized Databases**
  - **Realtime Communication**
  - **Automation**

☑ **Potential Processes Selected and Ranked from among**
  - **Processes that Serve/Impact External Customers**
  - **Strategy Supporting Processes**

Figure 9 Envision Step Checklist.

allows them the opportunity to buy into the process as opposed to having predetermined solutions forced upon them.

4.3.2. Organize Reengineering Teams

The actual size and number of teams for a given project may vary. It is essential that the team members, especially the design members, be trained to work in a team environment. The members responsible for design must also have a background of experience or training in process design, as well as some knowledge of the current technologies available for use in reengineered processes. This is where consultants are sometimes used to add missing skills to a team. Some process reengineering experts recommend that the PDR team be split into two parts: a design team and an execution team (Weicher et al. 1995). Others have suggested that the team be split into a documentation and an analysis team. The purpose of splitting the documentation and analysis functions is to maintain objectivity in the analysis. It has been reported that documenting "as-is" conditions causes the practitioners to become emotionally invested in the current system, thereby reducing their ability to see opportunities for improvement (Andrews 1996b). Separating the two neutralizes the effects of this phenomenon.

4.3.3. Conduct Project Planning

The challenges facing the project planning team are similar to those facing other large-scale project management teams. Beginning and end dates must be established. Aggressive but realistic timeframes must be incorporated. Levels of measurable accomplishment must be defined for all stages. Individuals should be made accountable for specific outcomes (Feather 1998). Resources must be made available to the team as well as the project. These may be financial but will most certainly take the form of personnel and time. Inadequate staffing at critical points in the project can lead to unsuccessful attempts at reengineering (Ho 1999). Scope creep is an ever-present danger in any project. This is doubly so in reengineering efforts. Processes are complex and difficult to visualize (Bal 1998). This can lead to the inclusion of unrelated processes into the design of the process of interest. The hazard that is peculiar to reengineering efforts is cultural change. In some cases, the process concept is so new that time to assimilate and understand the concept must be factored in. In all cases, the change itself must "proceed at a pace which can be accommodated by the organization" (Peltu et al. 1996).

4.3.4. Determine External Process Customer Requirements

There are a variety of methods for determining customer requirements. Quality function deployment models use the voice of the customer as an integral component in process design. Focus groups and benchmark data also provide valuable insight. Placing customer representatives on the actual process design team has been done with great success. Surveys can also derive information from customers. The informal nature of surveys, however, places them in a ‘‘general’’ data-gathering category.

4.3.5. Set Performance Goals

Performance goals are frequently established using the gap between the customer requirements and the performance of the current process. Likewise, benchmarking data can provide the performance goals for the new process. The metrics depend on the process being measured, but the firm can use measures that it is already familiar with. Examples include queue time, flow days, and backlog. A key element is making sure the measures cross-over from the old process to the new (Caudle 1995). It is imperative to be able to gauge the performance of the new process against the old.

Figure 10 summarizes the initiate step activity.

4.4. Diagnose

4.4.1. Document Existing Process

Many PDR proponents suggest skipping this step because they believe it limits the vision of the design team and that there is nothing to be gained from such a task. The problem of limited vision was addressed previously. The amount of information to be gained from this step is addressed here. Practitioners have suggested that only if the processes are laid out end to end do the impacts of the process become apparent (Lee 1996). It is equally important to document the cross-departmental aspects of the process (Fisher 1997). The dynamic relationships inherent in the process yield the greatest understanding of the organization and how it functions (Ould 1997). The level of technology used for this process depends on the team itself. There are a variety of examples in the literature, ranging from ‘‘sticky notes’’ on the wall to data flow diagrams to sophisticated computer programs (Bartholomew 1999). Essentially, all of these methods describe the task, the owner and customers, and the information inputs and outputs.

Initiate Step Checklist

☑ **All Impacted Internal/External Entities Informed**

☑ **Reengineering Teams Formed**
  - **Trained or Experienced in TEAM Environment**
  - **Trained or Experienced in Process Design**
  - **Knowledge of Current Technological Capabilities**
  - **Plan for Handling Documentation vs. Analysis**

☑ **Project Planned**
  - **Scope**
  - **Duration**
  - **Resources**
  - **Implementation**

☑ **Customer Voice**
  - **Integrate Customer into Design and Analysis Functions**

☑ **Performance Goals Established**
  - **Current Metrics and Values Captured**
  - **Specific Improvements Quantified**

Figure 10 Initiate Step Checklist.

4.4.2. Analyze Existing Process

Two aspects of the existing process must be assessed. The first is the performance of the process as compared to customer expectations (the external performance of the process). The second is in regard to the value added/non-value added components of the current process (the internal performance of the process). This step is critical in the delivery of bottom-line benefits (Ho 1999). Much insight into the external performance of the system was gained in the initiate stage, where the performance goals were set. In the internal performance evaluation, tasks are evaluated as value adding or non-value adding. Using manufacturing terms, tasks such as transportation and rework are non-value adding, while machining and painting are value adding. What is found typically ‘‘is that 65% to 70% of the tasks are non-value [tasks]. At the lowest level of detail, you find massive amounts of waste’’ (Bartholomew 1999).

Figure 11 summarizes the activity involved with the diagnose step.

4.5. Redesign

4.5.1. Define and Analyze New Process Concepts

Much has been said about the missing element in the reengineering prescription. Establishing the future-state solution is a difficult, creative process. Hammer and Champy (1993) point this out when they say, ‘‘[T]here are no seven or ten step procedures that will mechanically produce a radical new process design.’’ They do, however, provide some characteristics of reengineered processes: (1) several jobs are combined into one, (2) workers make decisions, (3) the steps in the process are performed in a natural order, (4) processes have multiple versions, (5) work is performed where it makes the most sense, (6) checks and controls are reduced, (7) reconciliation is minimized, (8) hybrid centralized/decentralized operations are prevalent. Insight is gained in the reengineering effort when the conventional thinking is challenged. Policies and rules, whether written or not, can often be the reason a process exists in its current form (Bartholomew 1999).

4.5.2. Prototype and Detailed Design of New Process

The detailed design effort can take several months. The new process will require new skills, thinking, and behavior from those who are a part of it. The process will not exist in isolation. It will exist among all of the other processes within a company, demanding input and providing output. Many of the inputs come from shared resources. ‘‘When designing a process there needs to be both technical and social inputs, as well as the infrastructure and architecture to support the process. If incremental, dramatic or step performance improvements are sought then the technical and social inputs must have a degree of congruence’’ (Love et al. 1998a).

Whenever possible, new process concepts, if not the entire process, should be prototyped first. There is valuable information to be gained from such activity (Feather 1998). Even in the event that the process does not perform as expected, lessons gained from the experience can be applied to the next iteration of prototype design.

Diagnose Step Checklist

☑ **Tools Selected for Data Capture and Modeling**
☑ **Document "As-Is" Process**
  - **Physical Form**
  - **Information Form**
  - **Performance Measures**

☑ **Analyze Current Process**
  - **Correlate Output Weakness with Internal Characteristics**
  - **Summarize Value-Added vs. Non Value-Added Content**
  - **Challenge Assumptions Implicit in Current Form**
  - **Review Product/Service Characteristics**

Figure 11 Diagnose Step Checklist.

4.5.3. Design Human Resource Structure

"Among the variety of reasons for the failure of reengineering processes, one of the most serious is not providing a human-performance system to support each and every performer in the organization" (Fisher 1997). "The significant factor that determines the performance of an organization is its people" (Frohman 1997). To get the people to work in concert with the new process, there is a need to develop and link reward and recognition systems to the process (Pritchard and Armistead 1999). The adage "Show me how I'm measured and I'll show you how I perform" still holds true. In order to drive home the message of process, these systems must be altered to support the new process. This may involve organizational changes (Love et al. 1998a).

Figure 12 summarizes the activities involved with the redesign step.

4.6. Reconstruct

4.6.1. Reorganize

One of the most hotly debated topics concerning reengineering is the surplus of employees that it can create. In a majority of the case studies, the employees no longer required by the new processes were laid off. Although cost cutting was arguably not one of the original primary goals of reengineering, it quickly became one as companies realized they had a surplus. Research performed since then indicates that wholesale workforce reduction may not be the correct choice of action for the following reasons:

- Stock market gains realized after the cuts are usually not long lasting (Harrington 1998b).
- Productivity either stayed the same or deteriorated after the layoff (Guimaraes and Bond 1996).
- The company cuts its capacity for future growth (Weicher et al. 1995).
- Those who are left are fearful and uncooperative (Strassman 1994).

Some companies have gone to the other extreme, guaranteeing at the outset of a reengineering effort that no jobs will be lost (Willets 1996). This is done in an effort to solidify the cooperation necessary

Redesign Step Checklist

☑ **New Process Concepts**
- **Jobs Combined Where Possible (Training & IT)**
- **More Decision Authority at Job**
- **Processes Capable of Custom Configurations**
- **Automated Error Checking/Review**
- **Wait Times Minimized**

☑ **Detail Design**
- **Impact on Other Processes Assessed**
- **Impact on Organization(s) Assessed**
- **Customer Requirements Satisfied**
- **Concepts Tested**
- **Education/Training Needs Identified**
- **Cost/Benefit Analysis**

☑ **Human Resource Structure Identified**
- **Workforce Estimates**
- **Process Measures and Human Performance Measures Linked**
- **Management Structures Aligned with Process**

Figure 12 Redesign Step Checklist.

for genuine innovation within the workplace. While this may not be the correct answer either, there is one guiding principle to use in the event that the surplus must be reduced: Don't eliminate the full number of surplus employees provided by estimates. First, no amount of planning can accurately gauge the headcount required by the new process (Macdonald 1998). Second, the knowledge and skill lost in the workforce reduction will naturally impede the performance of the new process until those skills are reacquired. This aspect is rarely modeled in workforce estimates.

4.6.2. Implement

Implementation problems are bound to develop. The interactions are too numerous and the complexity too great for this not to be expected. While it is impossible to foresee every issue that might develop, case history does provide some clues (Guimaraes and Bond 1996). Examples include (1) communication barriers between functional areas, (2) lack of leadership, (3) strategies being formed outside the company's ability to implement, (4) difficulty having the changes accepted by the employees affected, (5) unexpected enormity of the undertaking and disruption to the company, (6) difficulty balancing the incentives of traditional performance measures against what really needs to be done, (7) faltering of some projects because nervous corporate backers pull out at the first sign of difficulty, (8) executives many times being unsure whether it is a worthwhile undertaking, (9) information systems infrastructure in most large organizations being a major impediment to achieving immediate benefits, (10) the tough problems of elimination of positions and worker anxiety over losing jobs, (11) management reluctance to commit resources, (12) major training costs to make the transition and management frustration with slow results.

4.6.3. Train Users

As mentioned previously, process thinking is new to most organizations. Users must be given the opportunity to learn the implications of process thought. They must reorient their thinking towards the customer and learn to evaluate their job in terms of how it supports the customer. Also, the process may introduce new technologies and skill requirements. The users must be given the opportunity to acquire the tools necessary to perform their new job.

Figure 13 summarizes the activities involved with the reconstruct step.

4.7. Evaluate

4.7.1. Evaluate Process Performance

The evaluation of the new process will be in accordance with the process measures developed previously. If new measures were developed, they should have been implemented while the old process was still operating. It is important to realize that early measures of the new process may not reflect its true performance potential. This can be due to the disruption caused by cut-over. In the case where new measures are being used, it may reflect the need to refine the measures.

Reconstruct Step Checklist

☑ **Reorganization**
- **Plan for Dealing with Potential Surplus of Labor**
- **Plan for Dealing with Customer Implications of Elimination/Reduction of Products/Services**

☑ **Implementation**
- **Management Requirements**
- **Disruption Estimated and Anticipated**

☑ **Training**
- **Specific Skills Related to New Process**
- **General Process Thinking**

Figure 13 Reconstruct Step Checklist.

Several successful organizations are now creating assessment programs to evaluate process-improvement successes and failures and share lessons learned (Caudle 1995). This would be a system of measuring the reengineering process itself, enabling the organization to develop its own set of best practices and spread the knowledge across the organization.

4.7.2. Link to Continuous Improvement Programs

In the final analysis, PDR is an improvement program within the realm of TQM. As such, companies are compelled to constantly reevaluate themselves and their performance. The TQM framework makes the customer the absolute measure of performance and PDR makes the process the item to be measured. Including PDR in the category of improvement tools implies that the reengineering process is not a static, one-time event but rather another improvement tool that can be used as necessary to improve the products of an organization.

Figure 14 summarizes the evaluation step activities.

5. CASE STUDIES

5.1. Corning Asahi Video (Manganelli and Klein 1994)

Corning Asahi Video (CAV) is a business unit of Corning, Inc., and manufactures glass for televisions. The company had performed in the red during the period from 1987–1991. Customer dissatisfaction was high relating to lead times and order placement, and management was frustrated with its information system because the inventory and order status information was difficult to extract. To rectify this situation, the company set out to reengineer its order-fulfillment process. The overall goals of the project were to:

- Restore profitability by improving efficiency within the process
- Minimize rework and error
- Strengthen internal communications and provide process-wide access to customer information.

Initially, a cross-functional team was formed to redesign the process and implement the changes. The customer service manager was appointed as the head of this team. Customer interviews and internal studies were conducted to target inefficiencies and determine their root causes. The resulting information was shared with all of the 1200 employees at CAV.

Several changes were implemented based on study findings. To incorporate new technology, CAV adopted an existing Corning process for new product development that facilitated communication with top management to support the identification, selection, and adoption of process supporting technology. Disjointed systems of information flow were replaced with an integrated database-driven system. Customer service was consolidated into one location, and the role of customer service was expanded to allow better service. Throughout the project, employees were kept apprised of the project effort.

The project was completed in 15 months at a price of $570,000. As a result of reengineering, rework and cost overruns have been reduced by $1.6 million annually. Per order costs have decreased by 75%. Personnel costs have been reduced by $400,000 annually. Order fulfillment time has been reduced by 50% from 180 days to 90 days. The number of tasks required in the order-fulfillment process was reduced from 250 steps to 9.

Evaluation Step Checklist

☑ **Process Measurement at Inception and Evaluated Frequently**
☑ **Continuous Improvement Links**
  - **Review Other Processes**
  - **Periodic Benchmarks of New Process**

Figure 14 Evaluation Step Checklist.

5.2. Uarco, Inc. (Weston 1997)

Uarco, Inc., is a company that prints purchase order forms and other business documents. It has 14 manufacturing and service call centers located across the United States. Its annual revenues are in excess of $550 million. In 1995, the company analyzed its processes and identified $100 million in costs associated with its order-fulfillment process. This was its single most expensive process. Upon closer inspection, it found unexpected costs in their order management systems. An exorbitant amount of time was spent taking and quoting orders. Since each job is, in effect, a custom solution, each aspect of the customer specification had to be looked up to determine the cost component in order to deliver the quote to the customer. The specification components were contained in a 1000-page manual that had to be consulted for each order and change order. The process could take hours. Updating the manual was time consuming as well because change pages had to be mailed to all of the service outlets. It was determined that 60–80% of the sales force's time was spent administering orders.

As part of the solution, older mainframe units with isolated PCs were replaced with Windows NT servers and networked PCs. Client/server software was purchased along with application software that allowed the company to share database information among its geographically dispersed offices. Shared, replicated databases provided a more streamlined, real-time update process. Electronically stored specifications allowed quicker access and thus quicker order fulfillment.

With the new systems in place, the cost to provide 213,500 quotes annually has decreased from $16 million to $10 million. The capital investment by the company was $21 million. Their net profits have increased by $25 million annually. Much of the change is due to the changed nature of the workload of the salesperson. Under the old systems, a salesperson would spend 80% of his or her time administering orders and only 20% selling. Those figures have reversed under the new systems.

5.3. D2D, Ltd. (Gadd and Oakland 1995)

D2D, Ltd. is a manufacturing subsidiary of International Computers Ltd. (ICL), a systems integrator. In the mid-1980s, the IT industry underwent dramatic change. At the pinnacle of mainframe manufacturing, ICL had profit margins of over 80%. By 1994, margins were 3–4% (typical of the industry). Material costs typically accounted for 80% of the cost of finished goods. Traditional performance measures were no longer significant.

ICL faced the prospect of outsourcing its manufacturing operations in order to remain competitive. Instead, they chose to improve their operations with investments in automation, JIT, flexible assembly cells, MRP systems, EDI, and surface-mount lines. These investments enabled the company to remain competitive but put it in an under-capacity condition. Restructuring within ICL and establishing outside strategic alliances reduced the demand for goods from the manufacturing division. This provided the foundation for radical change within D2D. Extra capacity for ICL's manufacturing subsidiary, D2D, would have to be found outside the company. Given P&L accountability, the subsidiary had to transform itself from the inward-looking one it had always been into a market-led and customer-facing organization.

As a starting point, the division determined its core competencies and added a sales and marketing function. It also determined other services that were intrinsic to its culture that it could market. To help in the transformation, a customer care training program was developed. All existing employees went through the program, and all subsequent new employees now do as well. The investment in technology, particularly the manufacturing technology, enabled a flatter infrastructure with a more empowered workforce—one that was more devoted to the customer. EDI provided a connection among D2D, its vendors, and its customers that, because of its prevention-based nature, reduced order and material mistakes, thus making the supply chain more efficient. EDI created a process view that is pulled by the customer.

BPR has become an integral part of the culture at D2D. It has provided the perspective necessary to undertake the bold changes that were necessary because of the change in charter. By 1993, D2D's profits were 57% higher than in 1992. The cost associated with defects had dropped significantly as a percent of revenue. By 1995, non-ICL revenue was 50% of total revenue.

5.4. Fortune 500 Insurance Company (Feather 1998)

A Fortune 500 insurance company decided to reengineer its death claims process in order to reduce cycle time, cost, and customer dissatisfaction. Initially, a design team was formed and customer and process data were analyzed. Two important discoveries were made in this phase that would have significant bearing on the final solutions that were developed. First, a mismatch was discovered between the actual cycle time of the claim process and the cycle time deemed acceptable to the customer. Second, because of the inability of the company to provide customer data visibility to all units of the company, many opportunities—totaling hundreds of millions of dollars—were not being captured by the company.

In preparation for the reengineering effort, a design team was selected and trained in reengineering design principles. A four-stage methodology was introduced to help guide them through the effort. The methodology emphasized the discovery of new opportunities to create value in business processes, new ways of thinking in terms of process, the design of innovative, value-creating business processes encouraging different behavior, and attention to implementation details. The team used simulation exercises to familiarize themselves with design principles and encourage new thought.

The design team was separated into three subteams. The first team mapped the existing workflow. The results of their work revealed a death claims process made up of 200 steps, with only 18% of those steps deemed as value added. The second team was responsible for analyzing the cycle time of the process, while the third team conducted focus groups with beneficiaries and agents. The beneficiary focus groups provided priceless information in terms of acceptable time frames for contact, as well as typical needs of clients.

The insurance company saw tremendous opportunity in what it learned. It found that when a beneficiary needs investment counsel after receiving a death claim, it would be advantageous for the company to be able to provide options to the beneficiary coincident with the time of payout, thereby increasing the probability that the assets would be retained in the insurance company's accounts. This would effectively link the payout of benefits to the reinvestment of those benefits.

The team quickly realized that they would need to bring in technology expertise. A technology team was formed to work on information flow issues that developed through redesign efforts. It was also determined that agent and claim site personnel representation on the design team would be necessary in order to make the ideas and concepts that were beginning to develop coalesce. Design options were developed in several sessions involving all of the necessary personnel and classified into three categories: workflow, organization, and technology.

With the design options in hand, the team formed into two subteams to create the new design. This was done in order to facilitate creativity. The eventual design was a consensus design between the two teams. The features of the new design were more methods available to the beneficiary to initiate a claim and the ability to segment claims between those requiring high or low administration. The most significant change in the process was the way that company viewed it. No longer was the payment of the death claim viewed as the end of the process. With the incorporation of the correct data-recording processes and technology, the payment was now viewed as a follow-up opportunity to provide reinvestment options to the beneficiaries.

The results of the effort yielded a process with 50% fewer steps. The segmenting process at the front end of the claim permitted 65% of the claims to be processed within two days. Costs associated with handling claims were reduced by 28%, and cycle time was reduced by 56%.

6. LESSONS LEARNED

The literature points to several areas that can be considered providing advice for future practitioners. PDR is a reasoning process applied to a highly complex set of systems to produce a process. The systems include employees, customers, equipment, policies, and strategies, to name a few. The very nature of the systems—their variability—ensures that no two PDR projects will behave in the same fashion. Small differences on the input side lead to large differences on the output side (Ferrie 1995). Because of this, PDR will never be a one-size-fits-all technique. PDR, being highly situational, will require solutions to be developed on a case-by-case basis, requiring the imagination and intellect of the entire design team, if not the entire organization. As a heuristic approach, there is no one "best" answer. However, all actions planned or deployed must be considered in regard to their impacts and outcomes.

Another key point to remember is that the scope of reengineering should be limited. Despite the title of their book, Hammer and Champy never intended that the corporation be reengineered on all fronts simultaneously. Select the strategically important processes, rank them, and focus attention on the most critical ones. Only one or two processes should be reengineered at a time (Harrington 1998).

The people issues will continue to be a part of PDR since very few "lights-out" systems require reengineering. The critical aspect here is to remember to reengineer not in spite of the human aspects of the system, but in congruence with them. The key contribution of the process comes from its people rather than its hardware or systems (Frohman 1997). The first formal definition of industrial engineering, as established by the American Institute of Industrial Engineers in 1955, recognized the inseparability of people from the systems of interest to the profession. "Successfully integrating people within a system will always be the ultimate design challenge" (Ferrie 1995).

7. FUTURE OUTLOOK

There are diverse opinions concerning the future of reengineering. Some feel that "knowledge management, employee empowerment, adoption of new information technology, and shared vision" will lead to greater reengineering success (Kettinger and Grover 1995). Others feel that taxonomy will improve the success rate, while still others feel that the key is the integration of organizational theory and control and MIS (Malhotra 1998). It is predicted that reengineering projects will be undertaken more often to keep up with a dynamic business environment and that success will be based upon how quickly a company seizes the moment. The reengineering project time will continue to decrease. Regardless of exactly how it will evolve, it is safe to assume that reengineering will continue to be used within organizations as a tool for improvement.

REFERENCES

Andrews, D. (1996a), ‘‘BPR: Dialog, Not Just Design,’’ *Enterprise Reengineering,* July 1996.

Andrews, D. (1996b), ‘‘The Pros and Cons of As-Is Modeling,’‘ *Enterprise Reengineering,* May 1996.

Bal, J. (1998), ‘‘Process Analysis Tools for Process Improvement,’’ *TQM Magazine,* Vol. 10, No. 5, pp. 342–354.

Bartholomew, D. (1999), ‘‘Process Is Back,’’ *Industry Week,* Vol. 248, No. 20, pp. 31–34.

Biazzo, S. (1998), ‘‘A Critical Examination of the Business Process Reengineering Phenomenon,’’ *International Journal of Operations and Production Management,* Vol. 18, Nos. 9/10, pp. 1000–1016.

Buchanan, D. (1998), ‘‘Representing Process: The Contribution of a Re-Engineering Frame,’’ *International Journal of Operations and Production Management,* Vol. 18, No. 12, pp. 1163–1188.

Camp, R. (1989), *Benchmarking: The Search for Industry Best Practices That Leads to Superior Performance,* ASQC Quality Press, Milwaukee.

Caudle, S. L. (1995), *Reengineering for Results: Keys to Success from Government Experience,* National Academy of Public Administration, Washington, DC.

Choi, C. F., and Chan, S. (1997), ‘‘Business Process Re-Engineering: Evocation Elucidation and Exploration,’’ *Business Process Management Journal,* Vol. 3, No. 1, pp. 39–63.

Coombs, R., and Hull, R. (1997), *The Wider Research Context of Business Process Analysis,* Manchester School of Management, Centre for Research on Organizations, Management and Technical Change, Manchester, UK.

Corrigan, S. (1997), *Human and Organizational Aspects of Business Process Reengineering,* Institute of Work Psychology, University of Sheffield, Sheffield, UK.

Crowe, T. J., and Rolfes, J. D. (1998), ‘‘Selecting BPR Projects Based on Strategic Objectives,’’ *Business Process Management Journal,* Vol. 4, No. 2, pp. 114–136.

Davenport, T. H. (1996), ‘‘The Fad That Forgot People,’’ *Fast Company,* November 1996, pp. 70–74.

Davenport, T. H., and Short, J. (1990), ‘‘The New Industrial Engineering: Information Technology and Business Process Redesign,’’ *Sloan Management Review,* Vol. 31, No. 4, pp. 11–27.

Dervitsiotis, K. N. (1999), ‘‘How to Attain and Sustain Excellence with Performance Based Process Management,’’ *Total Quality Management*, Vol. 10, No. 3, pp. 309–326.

Dickie, B. N. (1995), ‘‘The C.E.O. Agenda,’‘ *Strategy and Business,* No. 1.

Feather, J. J. (1998), ‘‘The Upside of Process Improvement,’‘ *IIE Solutions,* Vol. 30, No. 12, pp. 39–42.

Ferrie, J. (1995), *Business Processes: A Natural Approach,* International Manufacturing Centre, University of Warwick, Warwick, UK.

Fisher, J. (1997), ‘‘Improving Human Performance in a Process Management Environment,’’ *CMA Management,* Vol. 71, No. 5, pp. 21–24.

Frohman, A. L. (1997), ‘‘Igniting Organizational Change from Below,’’ *Organizational Dynamics,* Vol. 25, No. 3, pp. 39–53.

Gadd, K. W., and Oakland, J. S. (1995), ‘‘Re-Engineering a Total Quality Organization,’’ *Business Process Re-engineering and Management Journal,* Vol. 1, No. 2, pp. 7–27.

Gappmaier, M. (1997), ‘‘Process Prototyping—A Methodology for Participatory Business Process Design,’’ in *IDIMT’97: Proceedings of the 5th Interdisciplinary Information Management Talks,* Oldenbourg, Munich, pp. 63–76.

General Accounting Office (GAO) (1995), *Business Process Reengineering Assessment Guide,* GAO, Washington, DC.

Grover, V., et al. (1995), *Business Process Change: Reengineering Concepts, Methods and Technologies,* IDEA Group, Harrisburg, PA.

Guha, S., Grover, V., Kettinger, W. J., and Teng, J. T. C. (1997), ‘‘Business Process Change and Organizational Performance: Exploring and Antecedent Model,’’ *Journal of Management Information Systems,* Vol. 14, No. 1, pp. 119–154.

Guimaraes, T., and Bond, W. (1996), ‘‘Empirically Assessing the Impact of BPR on Manufacturing Firms,’’ *International Journal of Operations and Production Management,* Vol. 16, No. 8, pp. 5–28.

Hammer, M. (1990), ‘‘Reengineering Work: Don’t Automate, Obliterate,’’ *Harvard Business Review,* Vol. 68, July–August 1990, pp. 104–112.

Hammer, M., and Champy, J. (1993), *Reengineering the Corporation: A Manifesto For Business Revolution,* HarperCollins, New York.

Harrington, H. J., (1998a), ‘‘Performance Improvement: The Rise and Fall of Reengineering,’’ *TQM Magazine,* Vol. 10, No. 2, pp. 69–71.

Harrington, H. J. (1998b), ‘‘Performance Improvement: The Downside to Quality Improvement (The Surplus People Problem),’’ *TQM Magazine,* Vol. 10, No. 3, pp. 154–160.

Ho, S. J. K., (1999), ‘‘The Implementation of Business Process Reengineering in American and Canadian Hospitals,’‘ *Health Care Management Review,* Vol. 24, No. 2, pp. 19–31.

Jarrar, Y. F., and Aspinwall, E. M. (1999), ‘‘Integrating Total Quality Management and Business Process Reengineering: Is It Enough?’’ *Total Quality Management,* Vol. 10, Nos. 4/5, pp. 584–593.

Kettinger W. J., and Grover, V. (1995), ‘‘ Special Section: Toward a Theory of Business Process Change Management,’’ *Journal of Management Information Systems,* Vol. 12, No. 1, pp. 9–30.

Kettinger, W. J., Teng, J. T. C., and Guha, S. (1997), ‘‘Business Process Change: A Study of Methodologies, Techniques and Tools,’’ *Management Information Systems Quarterly,* Vol. 21, No. 1, pp. 55–80.

Lee, C. R. (1996), ‘‘Process Re-engineering at GTE: Milestones on a Journey Not Yet Completed,’’ *Strategy and Business,* No. 5.

Lee, R. G., and Dale, B. G. (1998), ‘‘Business Process Management: A Review and Evaluation,’’ *Business Process Management Journal,* Vol. 4, No. 3, pp. 214–225.

Love, P. E. D., Gunasekaran, A., and Li, H. (1998a), ‘‘Putting an Engine into Re-engineering: Toward a Process-Oriented Organization,’’ *International Journal of Operations and Production Management,* Vol. 18, Nos. 9/10, pp. 937–939.

Love, P. E. D., Gunasekaran, A., and Li, H. (1998b), ‘‘Improving the Competitiveness of Manufacturing Companies by Continuous Incremental Change,’’ *TQM Magazine,* Vol. 10, No. 3, pp. 177–185.

Lucier, C. E., and Torsilieri, J. D. (1999), ‘‘Beyond Stupid, Slow and Expensive: Reintegrating Work to Improve Productivity,’‘ *Strategy and Business,* No. 17.

Macdonald, J. (1998), ‘‘The Quality Revolution—In Retrospect,’’ *TQM Magazine,* Vol. 10, No. 5, pp. 321–333.

Malhotra, Y. (1998), ‘‘Business Process Redesign: An Overview,’’ *IEEE Engineering Management Review,* Vol. 26, No. 3.

Maluso, N. (1997), ‘‘Activity Based Costing: What Is It and How Can Reengineering Teams Use It?’’ Reengineering Business Case Series, ProSci.

Manganelli, R. L., and Klein, M. M. (1994), *The Reengineering Handbook: A Step-by-Step Guide to Business Transformation,* AMACOM, New York.

Moeller, B., Tucker, J. S., and Devereaux, J. (1996), ‘‘The Next Wave: Re-engineering for Growth,’’ *Strategy and Business,* No. 5.

Morris, D., and Brandon, J. (1993), *Reengineering Your Business,* McGraw-Hill, New York.

Morris, J. R., Cascio, W. F., and Young, C. E. (1999), ‘‘Downsizing after All These Years: Questions and Answers About Who Did It, How Many Did It, and Who Benefited from It,’’ *Organizational Dynamics,* Vol. 27, No. 3, pp. 78–97.

Ould, M. A., ‘‘Designing a Reengineering Proof Process Architecture,’’ *Business Process Management Journal,* Vol. 3, No. 3, 1997, pp. 232–247.

Peltu, M., Clegg, C., and Sell, S. (1996), *Business Process Re-engineering: The Human Issues,* International Manufacturing Centre, University of Warwick, Warwick, UK.

Petrozzo, D. P., and Stepper, J. C. (1994), *Successful Reengineering,* Van Nostrand Reinhold, New York.

Pritchard, J. P., and Armistead, C. (1999) ‘‘Business Process Management—Lessons from European Business,’’ *Business Process Management Journal,* Vol. 5, No. 1, pp. 10–32.

Roberts, L. (1994), *Process Reengineering: The Key to Achieving Breakthrough Success*, ASQC Quality Press, Milwaukee.

Schwartz, K. D. (1998), ‘‘Benchmarking for Dollars,’’ *Datamation,* Vol. 4, No. 2, pp. 50–57.

Stork, K. (2000), ‘‘Confusion Still Lingers on How to Benchmark,’’ *Purchasing,* September 21.

Strassman, P. A. (1994), *The Politics of Information Management,* Information Economics Press, New Canaan, CT.

Weicher, M., Chu, W. W., Ching, W., Le, L. V., and Yu, D. (1995), *Business Process Reengineering,* Baruch College, City University of New York, New York.

Weston, R. (1997), ‘‘Overhaul Helps Chicago Printer,’’ *ComputerWorld,* February 17.

Willets, L. G. (1996), ‘‘The Best Ways to Survive Reengineering: Expert Tips on How to Reinvent Your Attitude,’’ *Enterprise Reengineering,* September.

Yu, B., and Wright, D. T. (1997), ‘‘Software Tools Supporting Business Process Analysis and Modeling,’’ *Business Process Management Journal,* Vol. 3, No. 2, pp. 133–150.

ADDITIONAL READING

Bhaskaran, K., and Lueng, Y. T., *Manufacturing Supply Chain Modeling and Reengineering, Sadhanna,* Vol. 22, 1997. pp. 165–187.

CHAPTER 63

Scheduling and Dispatching

Michael Pinedo and Sridhar Seshadri
New York University

- **1. INTRODUCTION** **1718**
- **2. FRAMEWORK AND MODELS** **1719**
  - 2.1. Models and Notation 1719
  - 2.2. Overview of Past Research Directions 1722
    - 2.2.1. Determining Computational Complexity 1722
    - 2.2.2. Development of (Efficient) Algorithms 1722
    - 2.2.3. Worst-Case Analysis of Heuristics 1722
- **3. DISPATCHING TECHNIQUES** **1723**
  - 3.1. Basic Dispatching Rules 1723
    - 3.1.1. The WSPT Rule 1723
    - 3.1.2. The EDD Rule 1723
    - 3.1.3. The LPT Rule 1723
    - 3.1.4. The SST Rule 1724
    - 3.1.5. The CP Rule 1724
  - 3.2. Composite Dispatching Rules 1724
    - 3.2.1. The Apparent Tardiness Cost Heuristic 1724
- **4. SCHEDULING TECHNIQUES** **1725**
  - 4.1. Mathematical Programming Formulations 1725
  - 4.2. Dynamic Programming 1726
    - 4.2.1. Numerical Example 1727
  - 4.3. Branch and Bound 1728
    - 4.3.1. Numerical Example 1728
  - 4.4. Decomposition Heuristics 1729
  - 4.5. Local Search 1731
  - 4.6. Multiple Objectives 1732
- **5. SCHEDULING PROBLEMS IN PRACTICE** **1732**
  - 5.1. Real-Life Scheduling Problems vs. Theoretical Models 1732
  - 5.2. Scheduling in the Packaging Industry 1733
  - 5.3. Scheduling in the Semiconductor Industry 1733
  - 5.4. Scheduling in the Automotive Industry 1734
  - 5.5. Scheduling in the Aviation Industry 1734
- **6. DEVELOPMENT OF SCHEDULING SYSTEMS** **1735**
  - 6.1. General Structure of Scheduling Systems 1735
  - 6.2. Schedule Generation 1736
  - 6.3. Software Development and Implementation 1737
- **7. CONCLUDING REMARKS** **1738**
- **REFERENCES** **1739**
- **ADDITIONAL READING** **1740**

1. INTRODUCTION

Detailed scheduling of the various elements of a production system is crucial to the efficiency and control of operations. Orders have to be released and have to be translated into one or more jobs with associated due dates. The jobs often have to be processed by the machines of a workcenter in a given order or sequence. Queueing may occur when jobs have to wait for processing on machines

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

that are busy; preemptions may occur when high-priority jobs arrive at busy machines and have to proceed at once.

The scheduling and dispatching process has to interface with several other functions in the organization. On the one hand, it is affected by the production planning process, which handles medium- and long-term planning for the entire organization. Production planning takes inventory levels, forecasts, and resource requirements into account in order to do some form of optimization at a higher level. Any decision taken by this planning function may have an impact on scheduling and dispatching. On the other hand, scheduling also receives input from shop-floor control. Events that happen on the shop-floor have to be taken into account because they may have a considerable impact on the schedules.

What follows mainly focuses on the detailed scheduling of the jobs. Given a collection of jobs that have to be processed in a given machine environment, the problem is to schedule the jobs, subject to given constraints, in such a way that one or more performance criteria are optimized. Various forms of uncertainties, such as random job-processing times, machines subject to breakdown, and rush orders, may have to be dealt with.

This chapter is organized as follows. Section 2 presents some general notation as well as a mathematical framework that is used in the theory of scheduling and dispatching. The subsequent sections focus first on dispatching techniques and then on scheduling procedures. There are two reasons for following this approach: first, dispatching rules are typically easier to explain than scheduling procedures, and second, scheduling procedures often use dispatching rules as subroutines within a more elaborate framework. Understanding the workings of a scheduling procedure thus typically requires a certain knowledge of dispatching rules. Section 3 therefore gives an overview of the theory that has been developed with regard to dispatching. It covers basic dispatching rules as well as composite dispatching rules. Section 4 subsequently focuses on the main techniques applied to scheduling; it also describes a number of empirical procedures that have proven to be useful in current scheduling systems. Section 5 focuses on scheduling and dispatching problems in practice; it discusses a variety of industrial environments where scheduling is of critical importance. Section 6 first discusses the general structure of scheduling systems and then gives a description of the major trends in the development of industrial scheduling systems during the last two decades. Section 7 contains a discussion of the difficulties encountered in the implementation of these systems (see Figure 1).

Sections 3 and 4 are somewhat technical; 5 and 6 are more descriptive and basically self-contained. The less technically inclined reader may prefer to read these last two sections first.

2. FRAMEWORK AND MODELS

2.1. Models and Notation

In the past four decades, a significant amount of theoretical research has been done in scheduling as well as in dispatching. Along the way, a notation has evolved that succinctly captures the structure of most machine scheduling models studied in the literature. A short description of this framework and notation is presented here (see Lawler et al. 1989; Pinedo 1995).

The number of jobs is denoted by n and the number of machines by m. The subscript j refers to a job, while the subscript i refers to a machine. The following data are associated with job j:

p_{ij} = the processing time of job j on machine i; if job j is only to be processed on one machine or if it has the same processing times on each one of the machines it has (or is allowed) to be processed on, the subscript i is omitted.
r_j = release date of job j.
d_j = the due date of job j.
w_j = the weight (importance) of job j.

A sequencing or scheduling problem is described by a triplet $\alpha|\beta|\gamma$, where α describes the machine environment, β the processing characteristics and constraints, and γ the objective to be minimized. Examples of the possible (exclusive) entries in the α-field are:

1 = a single machine.
Pm = m identical machines in parallel; a job may be processed on any one of the m machines, it does not matter which one.
Fm = a flow shop of m machines; that is, m machines in series. A job after completion at one machine joins the queue at the next machine. All queues operate under the first-in-first-out discipline, that is, a job cannot “pass” another while waiting in a queue.
Jm = a job shop of m machines; in such a shop each job has its own route through the various machines and a job may visit a machine more than once.

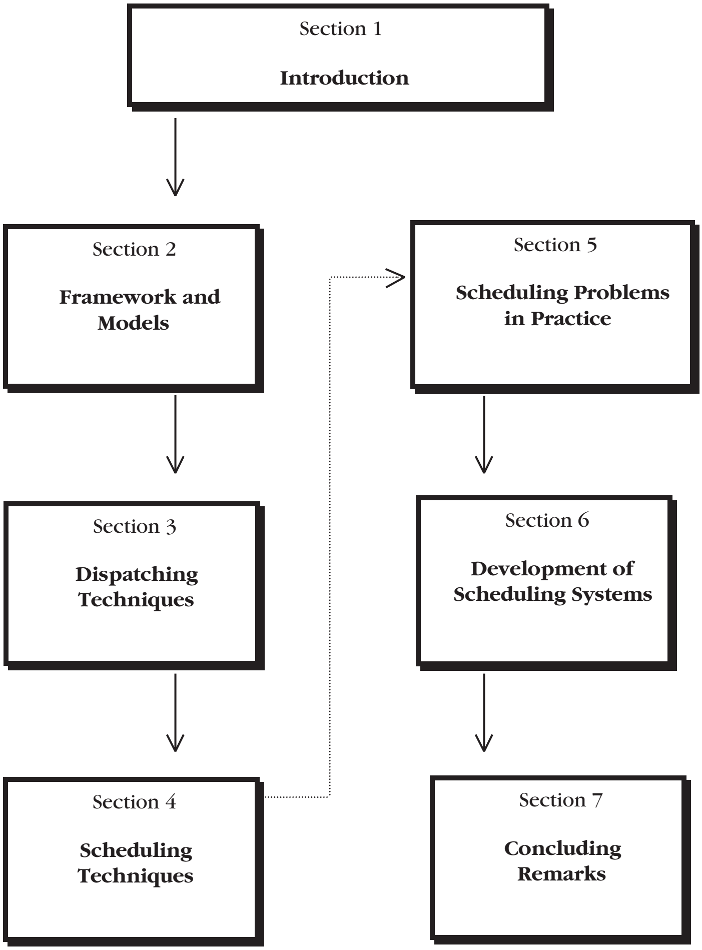

Section 1
Introduction
Section 2
Framework and Models
Section 3
Dispatching Techniques
Section 4
Scheduling Techniques
Section 5
Scheduling Problems in Practice
Section 6
Development of Scheduling Systems
Section 7
Concluding Remarks

Figure 1 Outline of This Chapter.

Examples of entries in the β-field are:

r_j = the release date of job j; job j may not start its processing before this date. If no r_j appears in this field, all r_j are assumed to be *0*.
s_{jk} = the sequence dependent setup time between jobs j and k; s_{0k} denotes the setup time for job k in case job k is first in the sequence, while s_{j0} denotes the clean-up time after job j in case job j is last in the sequence (of course, s_{0k} and s_{j0} may be zero). If no s_{jk} appears in the β-field, all setup times are assumed to be 0.
prec = the precedence constraints among the jobs. If no *prec* appears, there are no precedence constraints.
prmp = preemptions are allowed. If *prmp* does not appear, preemptions are not allowed.

Most other entries that may appear in the β-field are self-explanatory. The β-field may have multiple entries or may be completely empty. Due dates, in contrast with release dates, are not specified in this field; the ojective implies whether or not the jobs have due dates.

The objective is always a function of the completion times of the jobs, which are, of course, schedule dependent. The completion time of job j is denoted by C_j. Examples of possible objective functions to be minimized are:

C_{max} = the makespan, defined as $\max(C_1, \ldots, C_n)$, which is equivalent to the completion time of the last job to leave the system. A minimum makespan usually implies a good utilization of the machine(s).

L_{max} = the maximum lateness, which is defined as $\max(L_1, \ldots, L_n)$, where L_j equals $C_j - d_j$.

$\Sigma w_j C_j$ = the sum of the weighted completion times of the n jobs.

$\Sigma w_j T_j$ = the sum of the weighted tardinesses, where the tardiness of job j, T_j, is defined as $\max(C_j - d_j, 0)$.

$\Sigma w_j U_j$ = the weighted number of tardy jobs, with U_j being 1 if $C_j \geq d_j$ and 0 otherwise.

The following examples illustrate the notation:

$1|s_{jk}|C_{\max}$ denotes a single machine with n jobs subject to sequence-dependent setup times; the objective is to minimize the makespan. It is well known that this problem is equivalent to the so-called traveling salesman problem in which a salesman has to tour n cities in such a way that the total distance traveled is minimized.

$Pm|r_j,s_{jk}|\Sigma w_j T_j$ denotes m identical machines in parallel, n jobs with different release dates, different due dates, and different weights. The jobs are subject to sequence-dependent setup times and the objective is to find a sequence that minimizes the sum of the weighted tardinesses.

$Fm|p_{ij} = p_j|\Sigma w_j C_j$ denotes a so-called *proportionate* flow shop with m machines, i.e., m machines in series, with the processing times of job j on all m machines identical and equal to p_j (which is the reason this flow shop is called proportionate); the objective is to find the order in which the n jobs go through the system so that the sum of the weighted completion times is minimized.

$Jm||C_{max}$ denotes a job shop with m machines; the objective is to minimize the makespan.

Of course, there are many scheduling models that are not captured by this framework. Many situations in the real world have machine environments that do not fit the framework. For example, various researchers have recently begun to study hybrids between flow shops and parallel machines. In one such hybrid, there are a number of stages in series, with each stage consisting of a number of machines in parallel. Jobs progress from stage to stage, and at each stage each job has to be processed on one of the parallel machines. This machine environment has been given various names in recent years: generalized flow shop, flexible flow shop, compound flow shop. Alternatively, one can consider a number of parallel flow shops, each having a single machine at every stage. When a job is assigned to a particular flow shop, it is not allowed to switch in the middle of the process to another flow shop (see Figure 2).

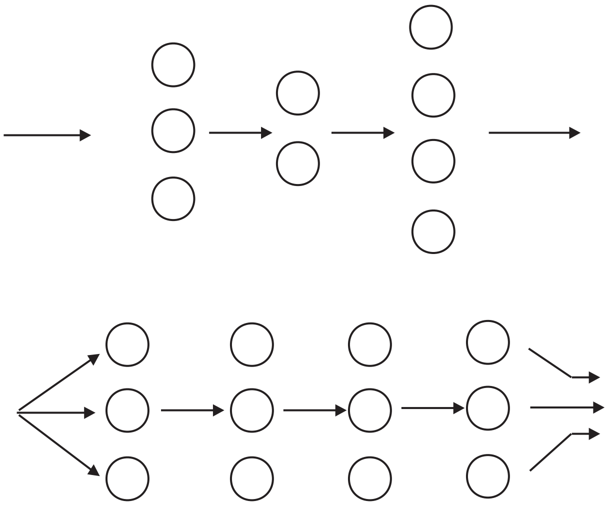


Figure 2 Hybrids of Flow Shops and Parallel Machines.

All objective functions mentioned above are so-called regular performance measures. Regular performance measures are functions that are *nondecreasing* in $C_1, \ldots, C_n$. Recently researchers have begun to pay attention to objective functions that are not regular. For example, when job j has due date d_j, there may be, besides a tardiness penalty, also an earliness penalty; the larger the deviation from the due date, in either direction, the larger the penalty. The objective may be the sum of earliness penalties and tardiness penalties.

The framework has been primarily designed for models with a single objective because most of the research has been concentrated on such models. Currently, researchers are studying models with multiple objectives as well, but a standard notation with regard to such models has not been developed yet.

Various other features in scheduling, not mentioned above, have been studied and analyzed in the past, such as finite buffers, blocking, and recirculation. A standard notation for these features still needs to be developed.

2.2. Overview of Past Research Directions

The research in deterministic scheduling problems has followed a number of different directions (see Pinedo 1995). Three areas have received considerable attention:

2.2.1. Determining Computational Complexity

For some problems so-called polynomial time algorithms are known to exist. A polynomial time algorithm implies that the number of computational steps (which is proportional to the amount of computer time) needed to find a schedule which achieves the optimum value of the objective function is a polynomial function of the parameters of the problem (e.g., the number of jobs, n and/or the number of machines, m). A polynomial time algorithm may require, for example, a number of steps that is on the order of n^3 or n^4. There are problems, however, for which no polynomial time algorithm is known to exist. These problems are the so-called NP-hard problems. The most efficient algorithms for these problems are exponential in the parameters of the problems. Such algorithms may require, for example, a number of steps that is on the order of 3^n or 4^n.

Determining whether or not a scheduling problem can be solved in polynomial time typically involves proving that the given problem is in some sense equivalent to a problem already known to be NP-hard. For example, the $1|s_{jk}|C_{max}$ problem is NP-hard because it is equivalent to the traveling salesman problem, which is known to be NP-hard.

2.2.2. Development of (Efficient) Algorithms

For problems that are solvable in polynomial time, it is usually relatively easy to find efficient algorithms. For the simplest of these problems, a simple sort is all that is required. For example, it may only be necessary to order the jobs in increasing order of their due dates (the so-called earliest due date rule [EDD]) or in decreasing order of the w_j/p_j ratios (the so-called weighted shortest processing time [WSPT] first rule. For more complicated polynomial time problems, more sophisticated techniques such as dynamic programming are required.

For the NP-hard problems, it is always significantly harder to find a good algorithm. One usually resorts to heuristics, which may give reasonably good schedules. When a more accurate solution is required, a more sophisticated technique (which also uses more computer time) such as branch and bound or lagrangean relaxation is usually employed. When only the optimal solution will do, the most likely approach would be to develop an (exponential time) algorithm based on dynamic programming or branch and bound.

2.2.3. Worst-Case Analysis of Heuristics

Solutions that are optimal for the easier problems often turn out to be good heuristics for the more complicated NP-hard problems. Examples of simple heuristics are the earliest due date (EDD) rule, the weighted shortest processing time first (WSPT) rule, the longest processing time first (LPT) rule, the shortest setup time first (SST) rule, and the critical path (CP) rule. The SST rule is of importance when the jobs are subject to sequence-dependent setup times; following the completion of a job, this rule consistently selects as the next job the one with the smallest setup time. The critical path (CP) rule is of importance when the jobs are subject to precedence constraints; following the completion of a job, this rule always selects as the next job the job that is at the head of the chain of jobs (one having to follow another) that contains the largest amount of processing.

It is of interest to know the worst-case behavior of such heuristics when applied on an NP-hard scheduling problem; that is, to have an upper bound on the ratio of the value of the objective function obtained via the heuristic divided by the true optimal value. This upper bound often, but not always, lies between 1 and 2. Besides giving an indication of how bad the result of a given heuristic may turn out to be, a worst-case analysis also gives an indication of the types of problem data for which the heuristic does not work well.

The next two sections deal, in particular, with heuristics and algorithms. At times a problem may be said to be NP-hard; however, no details will be given on how this is determined.

3. DISPATCHING TECHNIQUES

3.1. Basic Dispatching Rules

The five priority rules mentioned in the previous section, WSPT, EDD, LPT, SST, and CP, are fairly important. They provide optimal sequences in some very simple cases and serve as heuristics for more complicated scheduling models. It is useful to know the properties of these priority rules when designing a complicated computer-based scheduling system. Different modules in such a system may use at given times one of these rules to sequence a subset of the jobs. Or a composite priority rule may be constructed by combining two or more of these simple priority rules in order to minimize a mixture of various objectives. A more in-depth discussion of these five simple priority rules follows.

3.1.1. The WSPT Rule

The weighted shortest processing time first rule, which schedules the jobs in decreasing order of w_j/p_j, actually minimizes the sum of the weighted completion times on a single machine (see Smith 1956), that is, $1||\Sigma w_jC_j$. However, for slightly more complicated problems, such as $1|r_j|\Sigma w_jC_j$, or $1|r_j,prmp|\Sigma w_jC_j$, or $Pm||\Sigma w_jC_j$, the WSPT rule (or any variant of the WSPT rule that takes preemptions into account) does not necessarily result in the optimal solution. These three problems are actually NP-hard. The WSPT rule is nevertheless a very good heuristic. A worst-case analysis shows that for $Pm||\Sigma w_jC_j$:

$$\frac{\Sigma w_jC_j(\text{WSPT})}{\Sigma w_jC_j(\text{OPT})} \leq \frac{1+\sqrt{2}}{2} = 1.207$$

where $\Sigma w_jC_j(\text{WSPT})$ ($\Sigma w_jC_j(\text{OPT})$) denotes the value of the objective function under the WSPT (OPT) rule (see Kawaguchi and Kyan 1986). For the special case where all jobs have equal weights, that is $w_j = 1$ for $j = 1, \ldots, n$, the WSPT rule reduces to the shortest processing time first (SPT) rule. It is well known that the SPT rule minimizes the total completion time on parallel machines ($Pm||\Sigma C_j$).

3.1.2. The EDD Rule

The earliest due date rule, which schedules the jobs in increasing order of their due dates, minimizes the maximum lateness on a single machine (see Jackson 1955), that is, $1||L_{max}$, as well as in a proportionate flow shop, that is, $Fm|p_{ij} = p_j|L_{max}$. However, it does not provide an optimal solution for other due date-related problems, such as $1||\Sigma T_j$. Instances of $1||\Sigma T_j$ with

$$\frac{\Sigma T_j(\text{EDD})}{\Sigma T_j(\text{OPT})} = n$$

can be found easily. This implies that the EDD rule at times may result in a solution that is far from optimal. The EDD rule is usually not used as a heuristic by itself, but rather as part of a composite heuristic (such as the so-called ATC rule, which is discussed later in this section).

3.1.3. The LPT Rule

The longest processing time first rule, which schedules the jobs in decreasing order of p_j, is used as a heuristic for the $Pm||C_{\max}$ problem, which is known to be NP-hard. The importance of the makespan objective lies in the fact that a small makespan is a sign of a good job balance (partition) over the various machines. In order to find a schedule with a relatively small makespan, one orders the jobs in decreasing order of their processing times and puts the jobs on the machines, whenever one is freed, in that order. A worst-case analysis of this heuristic shows that

$$\frac{C_{\max}(\text{LPT})}{C_{\max}(\text{OPT})} \leq \frac{4}{3} - \frac{1}{3m}$$

See Graham (1969). This implies that at all times this heuristic performs reasonably well. The LPT heuristic has been used in industrial scheduling systems in order to provide a reasonable balance of the workload over the different machines. After the partition has been determined by the LPT rule, the jobs assigned to any given machine can be resequenced. Resequencing a machine clearly does not affect the balance and may be done to minimize a secondary objective.

3.1.4. The SST Rule

The shortest setup time first rule, when a machine is freed after completing job j, selects as the next job the one with the smallest setup time s_{jk}. It is used as a heuristic for the $1|s_{jk}|C_{\max}$ problem, which, as said above, is equivalent to the traveling salesman problem. The traveling salesman leaving a city for the city closest by (the nearest neighbor) is equivalent to sequencing jobs based on the smallest sequence-dependent setup time. The $1|s_{jk}|C_{\max}$ problem is known to be NP-hard even when the so-called *triangle inequality* holds. This inequality implies that $s_{jk} + s_{kl} \geq s_{jl}$ for all j, k, and l. It can be shown easily that for the $1|s_{jk}|C_{\max}$ problem, even in case the triangle inequality holds, the ratio

$$\frac{C_{\max}(\text{SST})}{C_{\max}(\text{OPT})}$$

can become arbitrarily large (see Rosenkrantz et al. 1977). The SST rule in practice is often used in composite heuristics.

3.1.5. The CP Rule

The critical path rule always selects as the next job the one that is at the head of the chain of jobs that contains the largest amount of processing. It is often used as a heuristic for the $Pm|prec|C_{\max}$ problem, which is known to be NP-hard. The heuristic actually yields the minimum makespan in case all jobs have identical processing times and the precedence constraints are in the form of a *tree* (tree-like precedence constraints imply that either all jobs have at most one sucessor or all jobs have at most one predecessor). A worst-case analysis shows that for the $Pm|prec,p_j = 1|C_{\max}$ problem with arbitrary precedence constraints:

$$\frac{C_{\max}(\text{CP})}{C_{\max}(\text{OPT})} \leq 2 - \frac{1}{m - 1}$$

See Chen and Liu (1975). The CP rule is used also when other objective functions have to be minimized. It plays an important role in composite heuristics as well.

3.2. Composite Dispatching Rules

3.2.1. The Apparent Tardiness Cost Heuristic

The problem $1||\Sigma w_j T_j$ is of importance in many practical situations. Jobs frequently have different weights (priorities) as well as different due dates (committed shipping dates) with the sum of the weighted tardinesses as the objective to be minimized. This objective is strongly correlated with loss of goodwill. It is well known that this problem is NP-hard even with a single machine. So it is important to have a heuristic that provides a reasonably good schedule with little computational effort. Some heuristics come immediately to mind, namely, the WSPT rule (which is optimal when all release dates and due dates are zero) and the EDD rule (which is optimal when all due dates are sufficiently large and spread out). It is clear that a heuristic or priority rule is needed which in one form or another combines these two heuristics. The apparent tardiness cost (ATC) heuristic is such a priority rule (Vepsalainen and Morton 1987). Under this priority rule, the jobs are scheduled one by one; that is, every time the machine is freed, a priority index is computed for all the remaining jobs that are available for processing. The job with the highest priority index is then selected to go next. This priority index is a function of the time the current job completes, say t, as well as the p_j, the w_j, and the d_j of all remaining jobs. The index is defined as

$$I_j(t) = \frac{w_j}{p_j} \exp\left(-\frac{\max(d_j - p_j - t,0)}{K\bar{p}}\right)$$

where K is a scaling parameter which is determined empirically, $\bar{p}$ is the average of the processing times of the remaining jobs, and $\max(d_j - p_j - t,0)$ is the slack of job j. If K is chosen very large, the ATC rule reduces to the WSPT rule. On the other hand, if K is chosen very close to zero, the rule reduces to the WSPT rule among the overdue jobs when there are overdue jobs; when there are no overdue jobs, it gives priority to the job with the least slack. In order to obtain good schedules, the parameter K has to be chosen very carefully. This can be done by first making a detailed analysis of the particular scheduling instance under consideration. There are several ways to characterize scheduling instances. One is through a due date tightness factor τ, which is defined as

$$\tau = 1 - \frac{\bar{d}}{C_{\max}}$$

where $\bar{d}$ is the average of the due dates. Another way is through a due date range factor R, which is defined as

$$R = \frac{d_{\max} - d_{\min}}{C_{\max}}$$

It actually pays to evaluate the τ and the R of the instance under consideration and choose the scaling parameter K based on these values. A significant amount of research has been done that establishes the relationship between the look-ahead parameter K and the τ, the R, and the machine environment.

Thus, when one wishes to minimize $\Sigma w_j T_j$ in a more complicated machine environment, one first characterizes the particular problem instance through a number of factors. Then one determines the value of the look-ahead parameter K as a function of these characterizing factors as well as of the particular machine environment. After fixing K, one applies the rule.

Several generalizations of the ATC rule have been developed in order to take into account release dates as well as sequence dependent setup times. These generalizations require the initial computation of a number of factors. These factors, together with the particular machine environment, can then be used to determine a number of the parameters (see Lee et al. 1997).

Simple heuristics such as the five described above provide adequate results only for the simplest scheduling problems. Real-world scheduling problems usually require techniques that are significantly more sophisticated than a myopic priority rule that just orders the jobs according to a function of one or two parameters. There are various ways of formulating scheduling problems as well as various types of solution procedures. These are discussed next.

4. SCHEDULING TECHNIQUES

4.1. Mathematical Programming Formulations

Many scheduling problems can be formulated as linear programs or other forms of mathematical programs.

Scheduling problems that allow preemptions are often easier than problems that do not allow preemptions. The problems that allow preemptions can often be solved in polynomial time. Consider, for example, the problem $Pm|prmp|C_{\max}$. Job j may be processed on any one of the m machines; it may be preempted and may continue its processing on another machine at another time. A schedule that minimizes the makespan can be obtained by first solving the following linear program (LP):

$$\text{minimize} \quad C_{\max}$$

subject to

$$\begin{aligned}
\sum_{i=1}^{m} x_{ij} &= p_j, && j = 1, \ldots, n \\
\sum_{i=1}^{m} x_{ij} &\leq C_{\max}, && j = 1, \ldots, n \\
\sum_{j=1}^{n} x_{ij} &\leq C_{\max}, && i = 1, \ldots, m \\
x_{ij} &\geq 0, && i = 1, \ldots, m, j = 1, \ldots, n
\end{aligned}$$

The variable x_{ij} represent the total time spent by job j on machine i. The LP can be solved in polynomial time, but the solution of the LP does not prescribe an actual schedule; it merely prescribes how much time job j should spend on machine i. However, with this information a schedule can easily (in polynomial time) be constructed. This LP formulation for the $Pm|prmp|C_{\max}$ problem is given for illustration purposes only. Actually, many schedules that minimize the makespan are easy to find and all these schedules result in a makespan

$$C_{\max} = \max(p_1, \ldots, p_n, \Sigma p_j / m).$$

One of the schedules that minimize the makespan is the preemptive longest remaining processing time first (preemptive LPT) schedule. According to this schedule, at every point in time, the m jobs with the largest remaining processing times have to be processed on the m machines. This schedule requires, of course, a large number of preemptions; there are other optimal schedules that do not require as many preemptions.

The formulation described above can be generalized to include the model where the processing time of a job may depend on the machine on which it is processed. Again, job j needs processing on only one machine and any one will do. However, if job j is processed on machine i its processing time is p_{ij}.

Linear programming formulations are often possible either when preemptions are allowed or when all processing times are identical (i.e., $p_j = 1, j = 1, \ldots, n$). Scheduling problems with all processing times identical often reduce to the so-called assignment problem, for which there exists a linear programming formulation.

When a problem is NP-hard, a linear programming formulation is, of course, not possible. However, NP-hard scheduling problems can often be formulated as integer programs or other more complicated forms of mathematical programs. Consider, for example, the $Jm||C_{\max}$ problem. Let p_{ij} denote the processing time of job j on machine i, let y_{ij} denote the starting time of this operation, and let the set N denote the set of all (i, j) operations. This set may be viewed as a set of nodes in a directed graph. Let the set A denote the set of all precedence (routing) constraints $(i, j) \prec (k, j)$ that require job j to be processed on machine i before it is processed on machine k, that is, operation (i, j) precedes operation (k, j). This set may be viewed as a set of arcs in a directed graph that has nodes N. The following mathematical program minimizes the makespan:

$$\text{minimize } C_{\max}$$

subject to

$$\begin{aligned}
y_{kj} - y_{ij} &\geq p_{ij} && \text{for all } (i,j) \prec (k,j) \in A \\
C_{\max} - y_{ij} &\geq p_{ij} && \text{for all } (i,j) \in N \\
y_{ij} - y_{il} \geq p_{il} \quad \text{or} &\quad y_{il} - y_{ij} \geq p_{ij} && \text{for all } (i,l),(i,j), \quad i = 1, \ldots, m \\
y_{ij} &\geq 0 && \text{for all } (i,j) \in N
\end{aligned}$$

The third set of constraints is often called the disjunctive arc constraints and represent the fact that some ordering must exist among operations of different jobs that are processed on the same machine. Because of these constraints, this problem is sometimes referred to as the disjunctive programming problem.

That an NP-hard scheduling problem can be formulated as an integer programming problem or a disjunctive programming problem does not imply that there is a standard solution procedure available that will work satisfactorily. For most NP-hard scheduling problems solution procedures have to be tailor made. Two general techniques are widely used, namely dynamic programming and branch and bound. In what follows, these two techniques are applied to two well-known NP-hard problems. Dynamic programming is applied to $1||\Sigma T_j$ and branch and bound to $Fm||C_{\max}$.

4.2. Dynamic Programming

Dynamic programming is an optimization technique that is particularly well suited for scheduling problems with makespans that are schedule independent, such as, single-machine problems without setups, proportionate flow shops, and problems with all processing times being identical. It can also be applied on scheduling problems with makespans that do depend on the sequence.

For the $1||\Sigma T_j$ problem the *forward* dynamic programming technique is used (see Held and Karp 1962). The following observation is crucial. If the sequence $j_1, j_2, \ldots, j_n$, a given permutation of $1, \ldots, n$, is optimal for the problem under consideration, then the subsequence $j_1, j_2, \ldots, j_k$, where $k \leq n$, has to be optimal for the smaller k-job problem, which contains only jobs $j_1, j_2, \ldots, j_k$. Let S denote a subset of the n jobs and let $G(S)$ denote the minimum value of the objective function if only the jobs in S are to be scheduled. It is easy to see that the recursive relationship

$$G(S) = \min_{j \in S} \left(G(S - \{j\}) + \max\left(\sum_{l \in S} p_l - d_j, 0\right)\right)$$

has to hold. The second term on the right-hand side of this expression represents the tardiness of job j. This recursive relationship, which in the dynamic programming literature often is referred to as the *principle of optimality,* makes it possible to solve the scheduling problem in the following manner. First, all sequencing problems that contain only a single one of the n jobs are solved. There are n

such problems. Clearly, these problems require no optimization. However, the values of the objective functions $G(S)$ still have to be computed. If such a one-job scheduling problem consists of job j, then

$$G(S) = G(\{j\}) = \max(p_j - d_j, 0)$$

After these n values have been computed, all sequencing problems containing two of the original n jobs are solved (there are $n(n - 1)/2$ such scheduling problems); the optimal order as well as the values of the objective functions have to be determined. The recursive relationship and the results for the one-job problems make it possible to solve these two-job problems efficiently. That is, if $S = \{j, k\}$, then

$$G(\{j, k\}) = \min(G(\{k\}) + \max(p_k + p_j - d_j, 0), G(\{j\}) + \max(p_j + p_k - d_k, 0)).$$

Then all sequencing problems containing three of the n jobs have to be analyzed. Again, the recursive relationship and the results for the two-job problems make it possible to solve these three-job problems efficiently. Then all four-job problems need to be analyzed, and so on.

It is easy to see that this technique is more efficient than complete enumeration. Actually, one can reduce the amount of computation even more by using *dominance rules* or *elimination criteria* (see Emmons 1969). For the $1||\Sigma T_j$ the following dominance rule exists: If there is a pair of jobs j and k such that $p_j \leq p_k$ and $d_j \leq d_k$, then there exists an optimal sequence with job j appearing in the sequence before job k. This rule makes it possible to reduce the amount of computation since it is not necessary to evaluate any sequence where job k appears before job j. (However, it still is not possible to find a polynomial time algorithm for this problem, since it is known to be NP-hard.)

4.2.1. Numerical Example

Consider three jobs with processing times $p_1 = 6$, $p_2 = 10$, $p_3 = 8$ and due dates $d_1 = 13$, $d_2 = 8$, $d_3 = 15$.

First, consider all sequencing problems that consist of a single job. If S consists of either job 1 or job 3, then $G(S) = 0$ because these jobs would be completed before their due dates. If S consists of job 2, then

$$G(S) = 10 - 8 = 2$$

Next, consider all problems that consist of two jobs. There are three possible sets S, namely $\{1, 2\}$, $\{1, 3\}$ and $\{2, 3\}$. Each one of the three problems has to be evaluated twice (because each job has to be given the chance to be the last one of the set to be completed). If $S = \{1, 2\}$, then the last job of S is completed at $p_1 + p_2 = 16$. If job 1 is the last one of the set to be completed, then $T_1 = 3$ and

$$G(S - \{1\}) = G(\{2\}) = 2$$

If job 2 is the last one to be completed, then $T_2 = 8$ and

$$G(S - \{2\}) = G(\{1\}) = 0$$

So

$$G(\{1, 2\}) = \min(2 + 3, 0 + 8) = 5$$

The optimal order is first job 2 and then job 1. In the same way it can be determined that $G(\{1, 3\}) = 0$ and that in this case job 1 should precede job 3. The computations for set $\{2,3\}$ yield $G(\{2, 3\}) = 5$ and that job 2 should precede job 3.

Now consider all problems that consist of three jobs. There is only one such set, namely $\{1, 2, 3\}$. This set has to be evaluated three times because any one of the three jobs may be the last one to finish. From the fact that $p_1 + p_2 + p_3 = 24$, it follows that if job 1 is the last one to finish, $T_1 = 11$; if job 2 is the last one, then $T_2 = 16$; if job 3 is the last one, then $T_3 = 9$. So

$$\begin{aligned} G(\{1, 2, 3\}) &= \min(G(\{1, 2\}) + T_3, G(\{2, 3\}) + T_1, G(\{1, 3\}) + T_2) \\ &= \min(5 + 9, 5 + 11, 0 + 16) = 14 \end{aligned}$$

From these computations it is concluded that the optimal order is 2,1,3 and that the minimum value of the objective function is 14.

Actually, from the elimination criteria it could have been established in advance that job 1 has to precede job 3. The use of this piece of information throughout the procedure would have reduced the number of computational steps significantly.

4.3. Branch and Bound

In the scheduling field, the branch and bound technique appears to be more widely used than dynamic programming. The technique is basically an enumeration process that attempts to eliminate, through a bounding process, as many sequences as possible from consideration. The bounding process is very problem specific. Here the technique is applied to the $Fm||C_{\max}$ problem (see Ignall and Schrage 1965).

The branching process may be described as follows: A tree is built with a number of nodes at each level of the tree. The top level, level 1, of the tree consists of a single node, called the root. At this level the sequence is completely unspecified, that is, no job has a position in the sequence yet. This root has n branches that go down to the second level. The n nodes at the second level are characterized by the first job in the sequence, that is, at each node one particular job is assigned to the first position in the sequence and the positions of the $n - 1$ remaining jobs are still unspecified. Each node at level 2 has $n - 1$ branches emanating to the third level, with each node at the third level characterized by the jobs in positions 1 and 2 of the sequence. At each subsequent level there will be more nodes with fewer branches emanating to the next level.

The bounding process attempts to develop at any given node a lower bound on the objective functions of all sequences that start with the jobs specified at this node. A node at level $k + 1$ has k jobs specified, say jobs $j_1, j_2, \ldots, j_k$; of the remaining $n - k$ jobs the positions still must be determined. Let S denote these $n - k$ jobs and let C_{i,j_k} denote the time job j_k departs machine i in the $Fm||C_{\max}$ example. In order to find one lower bound for all sequences which start with $j_1, j_2, \ldots, j_k$, observe that the first machine of the flow shop is always processing a job until the last job leaves the first machine. At best this last job goes through the remaining machines, after leaving the first machine, without having to wait for processing at any one of the remaining machines; thus,

$$C_{\max} \geq \sum_{j=1}^{n} p_{1j} + \min_{j \in S} \left(\sum_{i=2}^{m} p_{ij} \right) = LB_1$$

The first term of the lower bound represents the time the last job is completed on machine 1; the second term is the minimum time required for the last job to go through the remaining $m - 1$ machines. A second lower bound can be obtained by assuming that the second machine is continuously busy after C_{2,j_k} and that the last job to leave the second machine goes through the remaining machines without having to wait for processing at any one of these machines. Thus,

$$C_{\max} \geq C_{2,j_k} + \sum_{j \in S} p_{2j} + \min_{j \in S} \left(\sum_{i=3}^{m} p_{ij} \right) = LB_2$$

In this way it is possible to obtain m lower bounds. The last bound is

$$C_{\max} \geq C_{m,j_k} + \sum_{j \in S} p_{mj} = LB_m$$

The lower bound to use is the maximum of these m lower bounds.

The search for the optimal sequence now goes as follows. One starts out with an initial sequence, which has to be chosen in a somewhat intelligent way (e.g., through a heuristic) and computes its $C_{\max}$. The better the initial sequence, the faster the branch and bound technique works. The procedure requires branching down the tree until a node is hit with a lower bound that is higher than the $C_{\max}$ of the best sequence found so far. This node is "then fathomed" and its offspring need not be considered. The search is then resumed from another node that appears to be promising (possibly because of a low lower bound at a node somewhat close to the root). When one finds a sequence with a $C_{\max}$ that is lower than the existing sequence one already has, one retains the new sequence and discards the other. The search is terminated when all nodes of the tree have been considered explicitly or implicitly (i.e., through fathoming).

4.3.1. Numerical Example

Consider three machines and three jobs. The processing times of job 1 are $p_{11} = 8$, $p_{21} = 4$, $p_{31} = 5$, of job 2, $p_{12} = 1$, $p_{22} = 6$, $p_{32} = 3$ and of job 3, $p_{13} = 1$, $p_{23} = 9$, $p_{33} = 5$. An initial solution has to be obtained in order to have a bound to start out with. One can argue that it makes sense to let job 3 go first because it has a very small processing time on machine 1 and a very long processing time on machine 2. Choose sequence 3,2,1 as the initial sequence. Computing its makespan yields 25. This value is used as the initial bound. There are three nodes at level 2. The first node represents all sequences that start with job 1 in the first position (i.e., node (1,.,.) in Figure 3). Computing the bounds that are associated with this node yields:

$$LB_1 = 8 + 1 + 1 + \min(6 + 3, 9 + 5) = 19$$

$$LB_2 = 8 + 4 + 6 + 9 + \min(3,5) = 30$$

$$LB_3 = 8 + 4 + 5 + 3 + 5 = 25$$

So the lower bound at node (1,.,.) is 30. Because this lower bound is higher than the makespan of our initial sequence, it does not make any sense to evaluate the offspring of this node. Node (1,.,.) is thus fathomed. Evaluating node (2,.,.) and computing in a similar way a lower bound yields a value of 25. Because a sequence with a makespan of 25 is already known it is not necessary to evaluate the offspring of this node either; node (2,.,.) is therefore fathomed as well. Evaluating node (3,.,.) and computing a lower bound for this node yields a value of 23. Because this bound is lower than the makespan of the current sequence, the offspring of this node has to be evaluated; there may be a sequence with job 3 in the first position that has a makespan that is lower than the makespan of the current sequence. Evaluation of level 3, which turns out to be the last level, reveals that there are only two nodes at this level, namely nodes (3,1,2) and (3,2,1). Computing the makespan of sequence 3,1,2 yields a makespan of 23. This sequence is therefore optimal.

4.4. Decomposition Heuristics

In this section and the next, empirical techniques are presented that have proven to be useful in practice. These empirical techniques, or at least their underlying ideas, have been used in many existing industrial scheduling systems.

The technique discussed in this section has been designed for the $Jm||C_{\max}$ problem, which has received an enormous amount of attention in the literature. One of the most successful heuristic procedures developed for this problem is the shifting bottleneck heuristic (see Adams et al. 1988). A description of the procedure requires an alternative formulation of the problem.

Consider a directed graph $G = (N, A, B)$. The nodes N correspond to all the operations to be done on the n jobs. The (solid) arcs A represent the precedence relationships between the various operations of the jobs. The disjunctive (broken) arcs B form m cliques of double arcs, one clique for each machine; the operations that are connected to one another in a clique have to be done on the same machine. *All* arcs, including the disjunctive ones, emanating from a node have as length the processing time of the operation at that node. In addition, there is a source and a sink, which are dummy nodes. The source (sink) has arcs emanating to (coming from) all the first (last) operations of the jobs. The arcs emanating from the source have length zero (see Figure 4).

A feasible schedule corresponds to a *selection* of one disjunctive arc from every pair in such a way that the resulting directed graph is acyclic. This implies, then, that each selection from a clique has to be acyclic. Such a selection, thus, determines the sequence in which the operations are to be performed on that machine. The length of such a schedule is, then, determined by the longest path

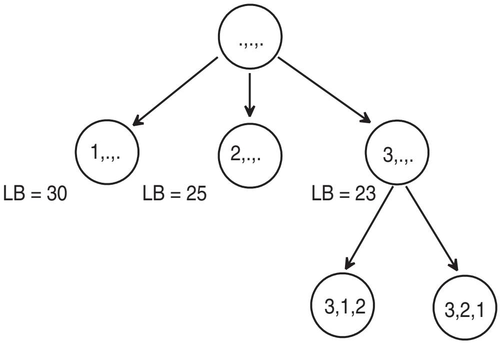


Figure 3 Branch and Bound Tree.

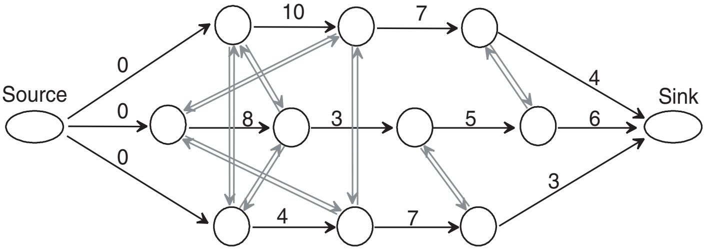


Figure 4 Disjunctive Graph of Job Shop with Three Jobs and Four Machines.

(i.e., the *critical path*) from source to sink. The problem is to find the selections of disjunctive arcs which minimize the length of the longest path. The shifting bottleneck procedure now works as follows.

Let M denote the set of all m machines. Assume that for a subset of the machines, say M_0, the selections of disjunctive arcs have been determined. That is, job sequences on these machines have been determined. An additional machine has to be added to this subset and the sequence in which the operations are to be processed on this machine needs to be specified. To determine which machine should be the next one to be included in M_0, an attempt is made to determine which machine (among the ones still to be scheduled) causes in one sense or another the severest problems. In order to do this, the original directed graph is modified by deleting all disjunctive arcs of the machines still to be scheduled (i.e., machines in the set $M - M_0$) and keeping only the relevant disjunctive arcs of the machines already in M_0 (one from every pair). Call this graph G'. Deleting all disjunctive arcs of a specific machine implies that the associated operations, which originally were supposed to be done on the machine one after another, now can be done in parallel at any point in time (as if each one of these operations has its own private machine). The graph G' has one or more critical paths as well as a makespan associated with it. Call this makespan $C_{\max}(M_0)$.

Assume now that each operation j to be scheduled on machine i, $i \in M - M_0$, has to be processed in a time window of which the release dates and due dates are determined by the critical (longest) paths in G'. Consider each one of the machines in $M - M_0$ as a separate $1|r_j|L_{\max}$ problem where the maximum lateness has to be minimized (actually, the $1|r_j|L_{\max}$ problem is NP-hard, but algorithms have been developed for this problem that perform reasonably well). After these single machine problems are solved, it has to be determined which one of these single machine problems has the *largest* maximum lateness. This machine, in a sense, is the "bottleneck" among the remaining machines still to be scheduled and therefore the one to be added next to M_0. Label this machine k, call its maximum lateness $L_{\max}(k)$, and schedule it according to the optimal solution obtained. If the corresponding disjunctive arcs, which specify the sequence of operations on machine k, are inserted in graph G', then the makespan increases by at least $L_{\max}(k)$, that is,

$$C_{\max}(M_0 \cup k) \geq C_{\max}(M_0) + L_{\max}(k).$$

Before the procedure is repeated and which machine to schedule next is decided, an additional step of resequencing each one of the machines in M_0 needs to be done. That is, a machine is taken out of the set $M_0 \cup k$, say machine l. A graph G'' has to be constructed in the same way as graph G' was constructed including now the disjunctive arcs which specify the sequence of operations on machine k and excluding the disjunctive arcs associated with machine l; machine l has to be resequenced by solving the corresponding $1|r_j|L_{\max}$ problem with the release and due dates determined by the critical paths in graph G''. After this resequencing is done for each one of the machines in the original set M_0, the entire procedure is repeated in order to add another machine to the current set $M_0 \cup k$.

The structure of this heuristic shows the relationship between the bottleneck concept and the more combinatorial concepts such as the critical (longest) path and the maximum lateness. A critical path indicates the location and timing of a bottleneck. The maximum lateness indicates the amount with which the makespan increases if a machine is added to the set of machines already scheduled.

Extensive numerical research has shown that this heuristic is extremely effective. Applied on a particular test problem with 10 machines and 10 jobs, which had remained unsolved for more than 20 years, the heuristic obtained a very good solution after only a couple of minutes of CPU time. This solution turned out to be optimal after a branch and bound approach, applied to the problem, obtained the same result and verified its optimality. The branch and bound approach, in contrast to the heuristic, needed many hours of CPU time. The disadvantage of the heuristic is, of course, that whether an optimal solution actually has been reached can never be known for sure.

4.5. Local Search

Simulated annealing and taboo search are two techniques that can be viewed as generalizations of the iterative improvement approach to combinatorial optimization problems.

Simulated annealing originated in a different field: it was first developed as a simulation model for describing a physical annealing process for condensed matter. The application of simulated annealing to scheduling requires a certain amount of structure (see Matsuo et al. 1989). At stage k of the process, there is a solution for the scheduling problem; for a single-machine problem this solution is merely a given sequence (permutation) of the jobs, say σ_p. Let $G(\sigma_p)$ denote the value of the objective function using this solution. For this solution a *neighborhood* can be defined. If the solution is just a permutation of the n jobs, then the neighborhood could be defined as all permutations that can be obtained by interchanging a pair of adjacent jobs (which implies that the neighborhood then consists of $n - 1$ different sequences). Clearly, the structure of the neighborhood can be made more complicated and is a design issue. In order to be allowed to move from solution σ_p to solution σ_q, which is an element of the neighborhood of solution σ_p an acceptance probability

$$P_{pq}(k) = \min\left\{1,\exp\left[-\frac{G(\sigma_q) - G(\sigma_p)}{\beta_k}\right]\right\}$$

is defined, where k is the so-called stage of the search and $\beta_1 \geq \beta_2 \ldots$ The stage is a level in which the same acceptance probability is used, and β_k is a control parameter. This β_k tends to zero as k increases, which implies that the acceptance probability for a move to a worse solution is lower at a later stage in the process. From the definition of the acceptance probability, it also follows that the worse a neighbouring solution is, the lower the acceptance probability is.

The simulated annealing procedure now works as follows. At each stage a series of neighborhood searches is done. A search can be done in a random way or in a organized (possibly sequential) way. A neighbor is compared with the "seed" (current) solution. When the value of the objective function of the neighbor is less than the value of the objective of the seed, the neighbor is automatically accepted and becomes the new seed. If the value of the objective of the neighbor is higher than the one of the seed, the neighbor is accepted as the new seed with a probability that is determined by the acceptance probability. However, the best solution obtained so far is always being kept in memory. In practice, several stopping criteria are used for this procedure. One way is to let the procedure run for a given (prespecified) number of iterations. Another is to let the procedure run until for a given number of iterations no improvement has been obtained.

Taboo search is in many ways similar to simulated annealing. The procedure moves again from one solution to another, with the next solution being possibly worse than the preceding solution. For each solution (or sequence) a neighborhood is defined, possibly in exactly the same way as it is defined for simulated annealing. The reason for allowing a solution to be worse than the previous one is to give the procedure the opportunity to move away from a local minimum and have a chance to find a lower minimum. The mechanism that guides the moves is different from the one in simulated annealing (see Glover 1990). At any stage of the process a so-called taboo list is being kept. This list contains the moves to the neighboring solutions the procedure is *not* allowed to make. This list has a fixed number of entries (this number usually lies between 5 and 9). Every time a move is made, the reverse move is put at the top of the taboo list; all other entries are pushed one position down and the bottom entry is deleted. The reason for putting the reverse move in the list is to avoid a move back to a local minimum that has been visited before. The search for a neighbor to which the procedure is allowed to move to is a design issue. This can, just as in simulated annealing, be done in a random way or in an organized (sequential) way.

The use of simulated annealing and taboo search has its advantages and disadvantages. One advantage is that it can be applied to a problem without having to know much about the structural properties of the problem. It can be coded very easily and it gives solutions that are usually fairly good. However, the amount of computation time needed to obtain such a solution tends to be relatively long in comparison with more rigorous problem-specific approaches.

Simulated annealing as well as taboo search are often used in the following manner. First an attempt is made to find a reasonably good initial solution for a problem via a heuristic. After this has been done, either a simulated annealing or a taboo search procedure is used as a postprocessor in order to search for an even better solution. The postprocessor is then run for a given amount of time.

4.6. Multiple Objectives

Little of the theoretical research done in the past has dealt with multiple-objective models. There are, however, several approaches for dealing with such problems. We illustrate one approach here through an example (see McCormick and Pinedo 1995).

Consider the problem $Pm|prmp|\alpha_1\Sigma C_j + \alpha_2 C_{\max}$; that is, there are m identical machines in parallel with preemptions allowed and as the objective the minimization of a weighted sum of makespan and flow time. It can be shown fairly easily that the shortest processing time first (SPT) rule minimizes ΣC_j on parallel machines, while the preemptive longest remaining processing time first (preemptive LPT) rule minimizes $C_{\max}$. Because these two rules are quite different, it is not immediately obvious what type of rule minimizes a mixture of these two objectives.

The problem can be analyzed by transforming one of the objectives into a constraint; that is, consider the problem where the flow time has to be minimized subject to the makespan being smaller than or equal to a given deadline d. This problem turns out to have a fairly simple solution. Without loss of generality, it may be assumed that $p_1 \leq p_2 \leq \ldots \leq p_n$. The scheduler schedules the jobs according to SPT until time $d - p_n$. At this point in time this largest job *must* be started on one of the machines, preempting the largest job among the ones being processed at that moment. The scheduler then continues using SPT until the second-largest job must be started, that is, at $d - p_{n-1}$. At this point in time the largest job being processed, not including job p_n, is preempted by job $n - 1$, and so on. It is easy to compute the value of the objective function, given $C_{\max} = d$. Through a parametric analysis on d, one can determine the minimum flow time as a function of the makespan. This function is decreasing convex and piecewise linear with a number of breakpoints (see Figure 5).

The values of α_1 and α_2 determine at which breakpoint the optimal solution lies.

In general, when there are a number of objectives to be minimized, the following heuristic approach can be used. Select as the first (principal) objective to minimize one that is important (that is, has a large weight) as well as sensitive to the schedule. While one optimizes this objective, one continuously keeps the other objective functions (which are of lesser importance) in mind. For example, whenever a tie needs to be broken, it is broken in a way that is beneficial for a secondary objective. After having completed the optimization procedure for the first objective, one proceeds with considering a second objective, and so on.

5. SCHEDULING PROBLEMS IN PRACTICE

This section focuses on a number of application areas where dispatching and scheduling techniques are of importance.

5.1. Real-Life Scheduling Problems vs. Theoretical Models

Real-life scheduling problems usually are very different from the mathematical models studied by researchers in academia and industrial research centers. It is difficult to categorize all differences between the real problems and the theoretical models, as each real-life scheduling problem has its own idiosyncrasies. Nevertheless, a number of these differences do stand out and are worth mentioning.

Theoretical models usually assume that there are n jobs to be scheduled and that after these n jobs are scheduled the problem is solved. In real life there may be at any point in time n jobs in the system, but every day (week or month) new jobs are added. Scheduling the current n jobs has to be

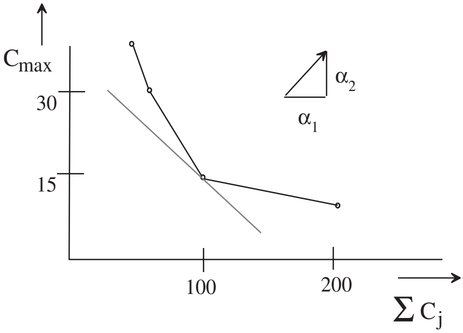


Figure 5 Trade-off Curve between Makespan and Flow Time.

done without having a perfect knowledge of what will happen in the near future. However, some provisions have to be made in order to be prepared for the unexpected. The dynamic nature of scheduling problem may therefore require, for example, slack times to be built into the schedule.

The models usually do not emphasize the *resequencing* problem. In real life the following problem often occurs. There exists a schedule that was generated based on certain assumptions; now an (unexpected) event has occurred that requires either major or minor modifications in the existing schedule. The rescheduling process, which is sometimes referred to as reactive scheduling, may have to satisfy certain constraints. For example, one may wish to keep the changes in the existing schedule at a minimum, even if an optimal schedule cannot be achieved this way.

The models usually do not consider *preferences.* In a model, a job either can or cannot be processed on a given machine. In reality, it often occurs that a job *can* be scheduled on a given machine but that there is a *preference* (for one reason or another) not to schedule it on the machine in question; scheduling it on the machine in question would only be done in case of an emergency.

Most of the theoretical research has been focused on models with a single objective. In real life there are, of course, a large number of objectives to deal with. Not only are there many objectives, but their respective weights may vary over time and may even depend on the particular person in charge.

In spite of the many differences between real-life scheduling problems and the mathematical models discussed in the previous sections, the general consensus is that the theoretical research done in the past has not been a complete waste of time. It has provided valuable insights into many scheduling problems, and these insights have proven to be useful in the development of a large number of scheduling systems.

5.2. Scheduling in the Packaging Industry

Consider a factory that produces paper bags for cement, dog food, charcoal, and so on. A scheduling system for such a factory has to be based on a flexible flow-shop model (see Adler et al. 1993). That is, there are a number of stages in series (for example, printing, glueing, and sewing) and at each stage there are a number of machines in parallel. The machines at any given stage may not be identical. Some machines may be more modern than others and may run at a higher speed or may be able to handle more complicated jobs than other machines. The main objectives are to meet the committed shipping dates as much as possible while minimizing the setup times on the machines. The algorithmic procedures adopted in such a system have to follow a number of steps.

First, the procedure goes through a bottleneck-identification process. At least one of the stages is a bottleneck. The identification of the bottleneck(s) can be done manually (specified by the user) or computed using all machine and job data. Then the procedure computes time windows during which jobs have to be processed at the bottleneck stage. The earliest time a job is allowed to start at a bottleneck stage depends on its current status upstream, while the latest time a job is allowed to be finished depends on its committed shipping date as well as on the amount of processing that is needed downstream. The current status upstream of a job may be either the stage the job currently is being processed at or the time the raw material (paper board) is expected to arrive at the facility and the job can be started at the first stage.

After this has been done, the procedure computes the various machine capacities at the bottleneck stage. For each machine it is computed how the capacity compares with the amount of processing that *has to* be done on it (that is, jobs that cannot be processed on any other machine) and with the amount of processing that *can* be done on it (that is, jobs that can be processed on other machines as well). This computation is done for various time periods (one week ahead, two weeks ahead, etc.).

Now scheduling of machines at the bottleneck stage can be done based on the information compiled in the previous steps. This schedule attempts to process all jobs within their respective windows while minimizing setup times. The procedure used here is a generalization of the ATC rule, which includes setup times. After the machines at the bottleneck stage have been scheduled, the machines at all other stages can be scheduled. The order in which the jobs go through the machines upstream and downstream of the bottleneck stage is somewhat similar to the order in which the jobs go through the bottleneck stage, with minor adjustments to improve on setup times or accommodate for machine preferences.

5.3. Scheduling in the Semiconductor Industry

In recent years a great amount of work has been done on wafer fabrication scheduling (see Hadavi and Voigt 1987). The reason for this activity is that equipment as well as products are extremely expensive. There is a great deal of randomness in the process; machines break down often and processing times are random. The machine environment can be considered as a job shop or as a flowshop with recirculation.

A system in such an environment has to create high-level production plans as well as more and more detailed scheduling plans. The timing of the generation of these detailed schedules depends on the urgency or the importance of the jobs in question.

While a large number of existing scheduling systems rely on good heuristics for solving the actual current scheduling problems (that is, *reacting* to real-time problems on the shop floor), some systems may take the more rational approach of *preventive* scheduling (that is, scheduling to avoid as many future problems as possible while at the same time optimizing the performance). In order to do so, such a system has to rely on a job-release mechanism that has a global view of the factory. Such a job-release mechanism attempts to achieve a maximal level of machine utilization and a minimal number of job delays. It may attempt to achieve this by monitoring a job parameter such as the *continuity index,* which measures how and when the job will be processed if released as proposed. The continuity index of a job is affected by bottleneck work centers, inventory shortages, tool availability, and so on. Based on these estimates, job releases are planned as much as possible to alleviate the prospective bottlenecks. This tends to minimize cycle times, which is one of the main objectives in VLSI development lines.

The sequencer and shop-floor control module of a scheduling system may track the work in progress as well as the status of the machines. It performs reactive scheduling functions based on this informtion. Such a module may be based on the axiom of *locality*, which implies that if an unexpected event occurs, an effort is made to limit, as much as possible, the number of changes in the existing schedules when correcting the problem.

The ReDS system developed by Siemens AG uses the real-time shop-floor data for another purpose as well (besides rescheduling). It uses the data in the form of a long-term feedback in order to make adjustments in the heuristics employed. This learning mechanism enables a scheduling system to mold itself over time into the shape of the factory in which it is operating.

5.4. Scheduling in the Automotive Industry

In the automotive industry, production scheduling ranges from job-shop scheduling in parts/components assembly to detailed job sequencing in automobile assembly. In automobile assembly, job sequencing and assembly line balancing are two concepts that are very closely related (see Burns and Daganzo 1987; Yano and Bolat 1990). This is largely due to two characteristics. First, most automobile assembly lines are mixed-model assembly lines with different processing time requirements for a large number of different models. In this case, different models of jobs may at times differ only in the options they carry and at other times may constitute completely different bodies. Secondly, most automobile assembly lines are paced, that is, jobs are processed at a constant rate that is determined by different model mixes and processing time requirements.

After the assembly line rate is determined, the operations with longer processing times are generally performed in sufficiently long sections of the line. This is possible because a large part of the assembly process is manually operated. However, job sequencing is now complicated because of the fact that the different operations performed on the different models require that similar jobs be adequately spaced in the sequence such that a proper balance of workload is achieved. Poor sequencing could result in reduced production and possibly quality problems. For example, suppose 10% of the jobs (cars) need a sunroof. Suppose that for a typical operation, to be done on *every* job, the processing time is p. The processing time for installing a sunroof is $3p$ since the amount of time it takes to install a sunroof is significant. The sunroof operation is designed to hold five jobs, with each job spending a total time of $6p$ in the operation (i.e., jobs are processed at a rate of $5/6p$). However, this implies that if there are three or more consecutive jobs requiring a sunroof, the operation is overloaded. An attempt is therefore made to place jobs with sunroofs as far apart as possible (i.e., approximately every tenth job) because otherwise the operation might be delayed and quality problems might occur if the workers do not have sufficient time to perform the operation properly.

Thus, the assembly line has to be sequenced in a way that more or less balances the workload at all operations, especially at those that in one form or another are critical. The operations most likely to be critical are those with a workload that tends to vary significantly from job to job.

Job sequencing is further complicated by the requirements that certain jobs be grouped together. For example, the paint process requires that each time the color of a job changes, the previous paint color must be purged from the spray gun to avoid paint overspray. Minimizing the paint purges can help reduce the paint cost, and therefore it is desirable to group jobs with the same color together in the job sequence. Also, to coordinate a better shipping schedule for finished automobiles, jobs shipped to the same destination should be grouped roughly in the same time interval.

Several large car manufacturers have developed integrated scheduling systems that take all the above requirements into account. Several systems are being used on a daily basis.

5.5. Scheduling in the Aviation Industry

In the aviation industry, many scheduling problems arise, such as crew scheduling, maintenance scheduling, and so on. In this section, a system is described that assigns airplanes to gates at an airport (see Brazile and Swigger 1988). Such scheduling systems have been developed to assign airplanes to gates at various airports in the United States, Europe, and Japan.

The problem can be viewed as a scheduling problem with machines in parallel. The gates represent the machines and the airplanes represent the jobs. The jobs have release dates that are the arrival times of the planes. These release dates are subject to random factors, such as weather and equipment failure. Any given job needs to be processed on one of the machines; the processing time is equivalent to the turnaround time of the plane. Some jobs can only be processed on a specific subset of the machines; for example, large planes can only go to specific gates while small planes may go to any one of the gates. The objective is to find a feasible assignment of jobs to machines.

Some of the systems are knowledge-based systems that attempt to find a feasible schedule through constraint satisfaction. They contain rules that guide the search and are capable of producing the trail of the rules used in reaching the conclusion. These systems may operate in one or two modes. In one mode, the static mode, it produces a daily schedule, which is followed only if everything goes as planned (which, of course, hardly ever happens). In the second mode, the dynamic mode, the system reassigns the gates in real time as information about changes in arrival and departure times becomes available.

Various different types of constraints with regard to the gate-assignment process can be formulated. There are ‘‘hard’’ constraints; for example, certain gates simply cannot accomodate 747s. There are ‘‘soft’’ constraints, which do allow exceptions; for example, domestic flights do not necessarily always have to arrive at and depart from domestic gates. There are ‘‘convenience’’ constraints; for example, planes that are scheduled to remain many hours at the terminal preferably should not be moved from one gate to another.

To arrive at an assignment, a system may search its way through two types of rules, permissive rules and conflict rules. Permissive rules determine whether it is appropiate to consider a particular gate for a particular flight. For example, when a plane is not continuing the same day, using a remote gate is acceptable. Conflict rules basically embody in the program the hard constraints mentioned before.

Systems may use a number of priority rules in their assignment of flights to gates. They first assign flights and gates that are the most constrained and the hardest to assign. Because only a small number of gates are capable of handling 747s, the wide-bodied aircraft are assigned first. Certain gates are so close to one another that planes cannot taxi in but have to be towed in. These gates have the lowest priority in the assignment.

Gate assignments are made when no rules are violated. If a feasible assignment is obtained in one pass, the process terminates. If the system does not find a feasible assignment in one pass, it automatically deactivates certain optional rules and tries again.

A system’s second mode (the reassignment mode) operates as follows. It receives the most recent data concerning the current status of arriving and departing flights from the airline’s database. When deviations are larger than given treshold values, the system checks whether the changes are creating conflicts. If there is a conflict, the system attempts to create a new schedule with a minimum number of changes. At times it may invoke the help of the scheduler.

6. DEVELOPMENT OF SCHEDULING SYSTEMS

6.1. General Structure of Scheduling Systems

During the last decade, numerous computer-based scheduling systems have been developed, many of which are currently controlling the scheduling operations in a variety of industries. For a number of reasons, the implementation of such systems usually turns out to be at least as difficult as the actual development. The development of these systems has taken place at R&D centers of industrial corporations as well as at universities (see Kanet and Adelsberger 1987; Adelsberger and Kanet 1989). Computer-based scheduling systems often consist of three modules:

1. A database management module
2. A schedule-generation module
3. A user interface module

See Figure 6.

All three parts play a crucial role in the functionality of the system. In practice, a significant amount of effort is usually required to make a factory’s database suitable for input to the system. Making the database accurate, consistent, and complete often involves the design of a series of tests the data must pass before they can be used. This module may also have capabilities of manipulating the data, performing various forms of statistical analysis, and enabling the scheduler to see data in the form of bar charts or pie charts. The schedule-generation module involves the formulation of a suitable model, the formulation of objective functions and/or constraints, and (possibly) the development of the algorithms. The user interface module is very important, especially with regard to the implementation process. Without a good user interface, there is a good chance that, regardless of its

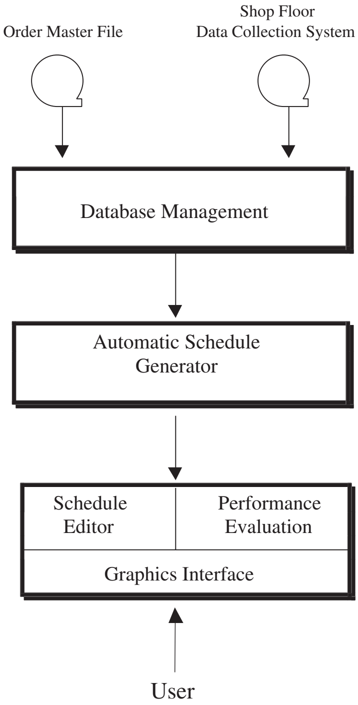


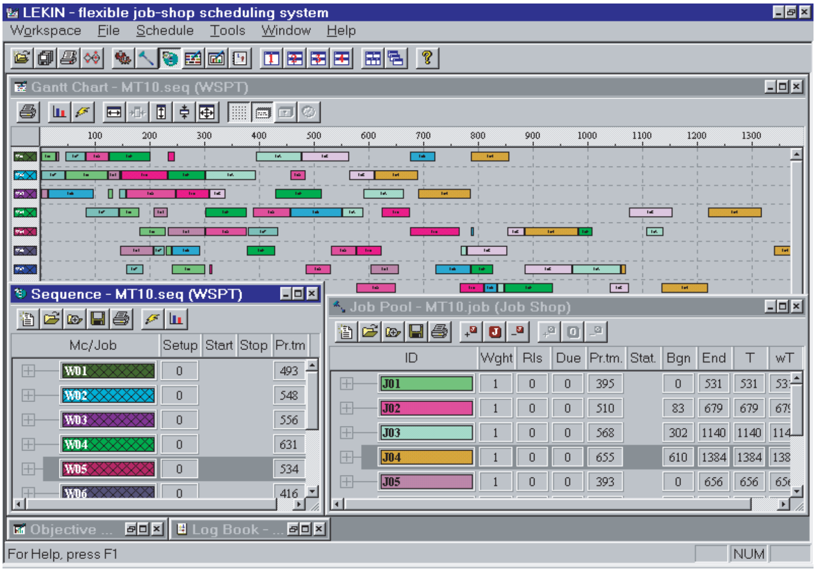
Figure 6 Configuration of a Scheduling System.

scheduling capabilities, the system will never be used. This user interface often takes the form of an electronic Gantt chart with tables and graphs that enable the scheduler to edit the schedule generated by the system and take last-minute information into account (see Figure 7). When the scheduler edits the schedule generated by the system, he or she is usually able to follow the impact of his or her changes on the various measures of performance as well as compare several schedules with one another and perform an extensive what-if analysis.

6.2. Schedule Generation

There are several schools of thought with regard to schedule generation. Two of these deserve further discussion. One, which is predominantly used by industrial engineers and operations researchers, could be called the *algorithmic* approach. The other, which is often used by computer scientists and artificial intelligence experts, is usually called the *knowledge-based* approach (see Kanet and Adelsberger 1987). Recently, these two approaches have started to converge towards one another and the differences have become more blurry. Some hybrid systems developed in the recent past use a knowledge base as well as fairly sophisticated heuristics.

The first approach usually requires a mathematical formulation of the problem, which includes objectives and constraints. The algorithm could be based on any one of the techniques or combination of techniques presented in Sections 3 and 4. The ‘‘goodness’’ of the solution is based on the values of the objectives and performance criteria under the given schedule. This form of schedule generation often may consist of three segments. In the first segment a certain amount of *preprocessing* is done. In this segment the problem instance is analyzed and a number of statistics are compiled, such as the average processing time, the maximum processing time, the due date tightness. The second segment consists of the actual algorithms and heuristics. The structure of the algorithm or heuristic may depend on the statistics compiled in the first segment (for example, in the way the look-ahead parameter K in the ATC rule may depend on the due date tightness and due date range factors). The third segment may contain a *postprocessor*. The solution that comes out of the second segment is fed into a procedure such as simulated annealing or taboo search in order to see whether any improvements can be obtained. This type of schedule generation is usually coded in a procedural language such as Fortran, Pascal, or C.


LEKIN - flexible job-shop scheduling system
Workspace
File
Schedule
Tools
Window
Help
Gantt Chart - MT10.seq (WSPT)
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
Sequence - MT10.seq (WSPT)
Job Pool - MT10.job (Job Shop)
Objective ...
Log Book - ...
For Help, press F1
NUM

Mc/Job	Setup	Start	Stop	Pr.tm
W01	0			493
W02	0			548
W03	0			556
W04	0			631
W05	0			534
W06	0			416

ID	Wght	Rls	Due	Pr.tm.	Stat.	Bgn	End	T	wT
J01	1	0	0	395		0	531	531	53
J02	1	0	0	510		83	679	679	67
J03	1	0	0	568		302	1140	1140	114
J04	1	0	0	655		610	1384	1384	138
J05	1	0	0	393		0	656	656	65

Figure 7 The Gantt Chart Interface of the Lekin System.

The second approach is different from the first in various respects. This approach is often more concerned with underlying problem structures that cannot easily be described in an analytical format. In order to incorporate the scheduler's knowledge into the system, so-called rules and objects are used. This approach is used often when it is only necessary to find a *feasible* solution given the many constraints or rules; however, because some schedules are ranked "more preferable" than others, heuristics are used at times in order to obtain a "preferred" schedule. Through an *inference engine* the approach attempts to find sequences that do not violate prescribed rules and satisfy stated preferences as much as possible. Whenever a satisfactory solution does not appear to exist or when the scheduler judges it to difficult to find, the scheduler may reformulate the problem through a relaxation of the constraints. The relaxation of constraints may actually be done automatically by the system itself. The programming style used in the development of such systems is usually different from the ones used under the first approach; systems are usually coded in languages that have so-called object oriented extensions, such as LISP and C++. These languages emphasize user-defined functions, which promote a modular programming style.

Both approaches have their advantages and disadvantages. The algorithmic approach clearly has an edge if (1) the problem allows for a crisp mathematical formulation, (2) the number of jobs involved is large, (3) the amount of randomness in the environment is minimal, and (4) some form of optimization has to be done frequently and in real time (it is very common that schedulers have neither the patience nor the time to wait more than 30 seconds for a schedule to be specified). One disadvantage of the algorithmic approach is that if the scheduling environment changes (for example, certain preferences on assignments of jobs to machines), the reprogramming effort may be substantial. The knowledge-based approach may have an edge if only a *feasible* schedule is required. Some system developers believe that changes in the scheduling environment or rules can be incorporated more easily in a knowledge-based system than in a system that is based on the algorithmic approach. Others, however, believe that the effort required to modify any system is mainly a function of how well the code is organized and written; the effort required to modify does not depend that much on the approach used. One disadvantage of the knowledge-based approach is that obtaining a reasonable schedule may take a substantial amount of computer time.

6.3. Software Development and Implementation

The last three decades have seen the design and implementation of many scheduling systems. Some of these systems were application specific and others were generic. Some were developed for research and experimentation and others were commercial.

A number of scheduling systems have been designed and developed in academia over the last three decades. Several universities developed research systems or educational systems that were often based on ideas and algorithms that were quite novel. An example of such a system is the Lekin system, which can be downloaded free of charge from the Web. Some of the academic systems have been handed over to industry and have led to the start-up of software companies.

The last two decades have witnessed the development of scores of commercial scheduling systems. There were a few major trends in the design and development of these commercial scheduling systems.

One trend started in the 1980s when a number of companies began to develop sequencing and scheduling software. Most of these companies tended to focus in the beginning only on sequencing and scheduling. They started out with the development of generic scheduling software that was designed to optimize flow lines or other types of machine environments. Some of these companies have grown significantly since their inception, such as ILOG, I2, and Manugistics.

These companies, whenever they landed a contract, had to customize their software to the specific applications. Because they realized that customization of their software customization is a way of life, they usually tried to keep their schedule generators as generic as possible. The optimization methodologies they adopted often included:

1. Shifting bottleneck procedures
2. Local search procedures
3. Mathematical programming procedures

These companies, which at the outset were focusing primarily on sequencing and scheduling, began to branch out in in the 1990s; they started to develop software for supply chain management. This diversification became necessary because clients typically had a preference for dealing with vendors that could provide a suite of software modules capable of optimizing the entire supply chain; clients did not like to have to deal with different vendors and face all kinds of integration problems.

A second major trend in the development of sequencing and scheduling software had its source in another corner of the software industry. This second trend started to take place in the beginning of the 1990s. Scheduling software started being developed by companies that at the outset specialized in ERP systems, such as SAP, Baan, J.D. Edwards, and PeopleSoft. These ERP systems basically are huge accounting systems that serve as a backbone for all the information requirements in a company. This backbone can then be used to feed information into all kinds of decision support systems, such as forecasting systems, inventory control, and sequencing and scheduling. The software vendors that specialized in ERP systems realized that it was necessary to branch out and develop decision support systems as well. A number of these companies either bought a scheduling software company (e.g., Baan bought Berclain), started their in-house scheduling software development (e.g., SAP), or established partnerships with scheduling software vendors.

Currently there are more than a hundred scheduling software vendors. Most of these are relatively small. The bigger players are I2, Cybertec, and Manugistics, all of them offering software for the entire supply chain. The main ERP vendors, such as SAP, Baan, PeopleSoft, and J.D. Edwards, all offer sequencing and scheduling packages. Some of their scheduling modules had been developed internally, whereas other modules were developed through acquisitions of smaller software companies specializing in scheduling. The algorithmic approaches differ considerably from company to company. Some companies specialize in local search procedures, while others specialize in mathematical programming techniques, and again others in decomposition techniques. Even the preferences for user interfaces may differ. However, the most popular user interface tends to be the Gantt chart.

7. CONCLUDING REMARKS

During the last two decades, with the advent of the personal computer in the factory, a large number of scheduling systems has been developed. Currently, many more scheduling systems are under development. This developmental process has made it clear that a large proportion of the theoretical research done during the decades past is of limited use in real-world applications. The system development that is currently going on in industry is fortunately encouraging theoretical researchers to tackle scheduling problems that are more relevant to the real world. At various universities in Europe, Japan, and North America, research is being focused on the development of algorithms as well as on the development of systems; significant efforts are being made in the integration of these developments (see McKay et al. 1989).

Even though during the last decade many companies have made large investments in the development and implementation of scheduling systems, not that many systems appear to be used used on a regular basis. Systems, after being implemented, often remain in use only for a limited time; after a while they are often, for one reason or another, ignored altogether.

In those situations where the systems are in use on a more or less permanent basis, the general feeling is that the operations do run more smoothly. A system in place usually does *not* reduce the time the scheduler spends on the scheduling process. However, a system usually does enable the scheduler to produce better schedules. Using an interactive schedule editor, the scheduler is able to compare different schedules and easily monitor the various performance measures. Actually, there are other reasons for smoother operations besides simply better schedules. A scheduling system imposes more "discipline" on the operations. There are compelling reasons now for keeping an accurate database. Schedules are printed out neatly and are visible on monitors. This apparently has an effect on people, encouraging them to actually even obey the schedules.

It would be interesting to know the reasons why some systems have never become implemented or are never used. In some cases databases are not sufficiently accurate. In other cases the way in which workers' productivity is measured is not in agreement with the performance criteria the system is based upon. User interfaces may not enable the scheduler to resequence quickly in the case of unexpected events. There may also be an absence of procedures that enable resequencing when the scheduler is absent (for example, if something unexpected happens during third shift). Finally, systems may not be given sufficient time to settle or stabilize in their environment (this may require many months, if not years).

Nevertheless, it appears that in the decade to come an even larger effort will be made in the development of such systems and that such systems will play an important role in computer-integrated manufacturing.

REFERENCES

Adams, J., Balas, E., and Zawack, D. (1988), "The Shifting Bottleneck Procedure for Job Shop Scheduling," *Management Science,* Vol. 34, pp. 391–401.

Adelsberger, H. H., and Kanet, J. (1989), "The Leitstand—A New Tool in Computer-Aided Manufacturing Scheduling," Technical Report, College of Commerce and Industry, Clemson University, Clemson, SC.

Adler, L., Fraiman, N. M., Kobacker, E., Pinedo, M. L., Plotnicoff, J. C., and Wu, T. P. (1993), "BPSS: A Scheduling System for the Packaging Industry," *Operations Research,* Vol. 41, pp. 641–648.

Brazile, R. P., and Swigger, K. M. (1988), "GATES: An Airline Gate Assignment and Tracking Expert System," *IEEE Expert,* Vol. 3, No. 2, pp. 33–39.

Burns, L., and Daganzo, C. F. (1987), "Assembly Line Job Sequencing Principles," *International Journal of Production Research,* Vol. 25, pp. 71–99.

Chen, N.-F., and Liu, C. L. (1975), "On a Class of Scheduling Algorithms for Multiprocessors Computing Systems," in *Parallel Processing,* Lecture Notes in Computer Science 24, T. Y. Feng, Ed., Springer, Berlin, pp. 1–16.

Emmons, H. (1969), "One-Machine Sequencing to Minimize Certain Functions of Job Tardiness," *Operations Research,* Vol. 17, pp. 701–715.

Glover, F. (1990), "Tabu Search: A Tutorial," *Interfaces,* Vol. 20, Issue 4, pp. 74–94.

Graham, R. (1969), "Bounds on Multiprocessing Anomalies," *SIAM Journal of Applied Mathematics,* Vol. 17, pp. 263–269.

Hadavi, K., and Voigt, K. (1987), "An Integrated Planning and Scheduling Environment," in *Proceedings of AI in Manufacturing Conference* (Long Beach, CA).

Held, M., and Karp, R. M. (1962), "A Dynamic Programming Approach to Sequencing Problems," *Journal of SIAM,* Vol. 10, pp. 196–210.

Ignall, E., and Schrage, L. E. (1965), "Application of the Branch and Bound Technique to Some Flow-Shop Problems," *Operations Research,* Vol. 13, pp. 400–412.

Jackson, J. R. (1955), "Scheduling a Production Line to Minimize Maximum Tardiness," Research Report 43, Management Science Research Project, University of California, Los Angeles.

Kanet, J., and Adelsberger, H. H. (1987), "Expert Systems in Production Scheduling," *European Journal of Operational Research,* Vol. 29, pp. 51–59.

Kawaguchi, T., and Kyan, S. (1986), "Worst Case Bound of an LRF Schedule for the Mean Weighted Flow Time Problem," *SIAM Journal of Computing,* Vol. 15, pp. 1119–1129.

Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., and Shmoys, D. B. (1989), "Sequencing and Scheduling: Algorithms and Complexity," Report BS-R8909, Centre for Mathematics and Computer Science, Amsterdam.

Lee, Y.-H., Bhaskaran, K., and Pinedo, M. L. (1997), "A Heuristic to Minimize the Total Weighted Tardiness with Sequence Dependent Setups," *IIE Transactions,* Vol. 29, pp. 45–52.

Matsuo, H., Suh, C. J., and Sullivan, R. S. (1989), "A Controlled Search Simulated Annealing Method for the Single Machine Weighted Tardiness Problem," *Annals of Operations Research,* Vol. 20, pp. 85–108.

McCormick, S. T., and Pinedo, M. L. (1995), "Scheduling n Independent Jobs on m Uniform Machines with Both Flow Time and Makespan Objectives: A Parametric Analysis," *ORSA Journal of Computing,* Vol. 7, pp. 63–77.

McKay, K. N., Buzacott, J. A., and Safayeni, F. (1989), "The Schedulers Information System—What is Going on? Insights for Automated Environments," in *Proceedings of the II-COM 1989 Conference* (Madrid).

Rosenkrantz, D. J., Stearns, R. E., and Lewis, P. M. (1977), "Approximate Algorithms for the Travelling Salesman Problem," *SIAM Journal of Computing,* Vol. 6, pp. 543–558.

Smith, W. E. (1956), "Various Optimizers for Single Stage Production," *Naval Research Logistics Quarterly,* Vol. 3, pp. 59–66.

Vepsalainen, A., and Morton, T. (1987), "Priority Rules for Job Shops with Weighted Tardiness Costs," *Management Science,* Vol. 33, pp. 1035–1047.

Yano, C. A., and Bolat, A. (1990), "Survey, Development and Applications of Algorithms for Sequencing Paced Assembly Lines," *Journal of Manufacturing and Operations Management,* Vol. 3, pp. 172–198.

ADDITIONAL READING

Baker, K. R., *Introduction to Sequencing and Scheduling,* John Wiley & Sons, New York, 1974.

Brucker, P., *Scheduling Algorithms,* 2nd Ed. Springer, Berlin, 1998.

Conway, R. W., Maxwell, W. L., and Miller, R. W., *Theory of Scheduling,* Addison-Wesley, Reading, MA, 1967.

Dempster, M. A. H., Lenstra, J. K., and Rinnooy Kan, A. H. G., Eds., *Deterministic and Stochastic Scheduling,* Reidel, Dordrecht, 1982.

French, S., *Sequencing and Scheduling: An Introduction to the Mathematics of the Job Shop,* Horwood, Chichester, 1982.

Pinedo, M., *Scheduling: Theory, Algorithms and Systems,* Prentice Hall, Englewood Cliffs, NJ, 1995.

Pinedo, M., and Chao, X., *Operations Scheduling with Applications in Manufacturing and Services,* Irwin/McGraw-Hill, Boston, 1999.

CHAPTER 64

Personnel Scheduling

RICHARD N. BURNS
BCW Consulting Limited

1. **INTRODUCTION** **1741**
2. **BASIC STEPS FOR PERSONNEL SCHEDULING** **1741**
   - 2.1. The Quantity of Work to Be Done 1742
   - 2.2. Determining Staffing Requirements 1742
   - 2.3. Determining Personnel Available 1742
   - 2.4. Developing Work Schedules 1742
3. **TYPES OF PERSONNEL SCHEDULING PROBLEMS** **1743**
   - 3.1. Single Shift 1743
   - 3.2. Multiple Shifts Required 1743
     - 3.2.1. Crew Scheduling 1743
     - 3.2.2. Shift Scheduling For Individuals 1744
   - 3.3 Hierarchical Workforce Requirements 1744
   - 3.4 Part-Time Workers 1744
4. **CREATING SCHEDULES** **1745**
   - 4.1. Single Shift 1746
   - 4.2. Multiple Eight-Hour Shift Crew Scheduling 1755
   - 4.3. Multiple Eight-Hour Shift Scheduling for Individuals 1757
   - 4.4. Multiple Mixed Shift Length Scheduling for Individuals 1763
   - 4.5. Hierarchical Workforce Scheduling 1764
5. **COMPUTER SYSTEMS** **1765**
   - 5.1. Personnel Interface 1765
   - 5.2. Contract Regulations 1765
   - 5.3. Payroll Interface 1765
6. **MANUAL SYSTEMS** **1765**

REFERENCES **1766**

1. INTRODUCTION

Personnel scheduling involves two stakeholders: management, who need to have the work done, and workers, who need to have working conditions that contribute positively to their quality of life. In recent years, management has recognized that the objectives of the two groups need not be mutually exclusive. Good working schedules help to make happy employees and increase the length of time a person stays with the employer.

Personnel scheduling problems can be separated into different categories, based on the types of shifts to be worked, the employment contract working conditions clauses, and the pattern of the quantity of work needing to be done. The objective of this chapter is to give an overview of the types of personnel scheduling problems that arise and an introduction to some of the solution methods available.

2. BASIC STEPS FOR PERSONNEL SCHEDULING

Generally, there is a definite set of steps to complete the process of personnel scheduling. These steps will be discussed in more detail in this section, but the first step is to determine what work needs to be done.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2.1. The Quantity of Work to Be Done

The quantity of work to be done must be defined by the appropriate time period, which could be by as little as a 15-minute interval for cashiers and as long as a day for paper mills. For continuous process industries, such as chemical plants and some mining processes, the definition of what is needed is relatively easy to determine. Even so, the work must be defined by the hour and day of the year to allow for maintenance and shutdowns for holidays. There are many other industries that also have a well-known stable demand for work profile—for example, prisons, long-term health care facilities, and many manufacturing systems, such as assembly lines. At the other end of the scale are situations such as retail outlets and telephone operators, that have a demand that varies constantly during the day, and from day to day, and from season to season. Many of the organizations with such a fluctuating demand pattern have a detailed data bank of historic data, usually by the 15-minute interval. These data can be used to predict the work requirements for future time periods.

Hospitals have an expected amount of work for each day. However, many hospitals apply a patient classification system to all patients daily to calculate the actual amount of work needed for that day. This last-minute determination of the work requirement is necessary and unavoidable, even though it makes personnel scheduling very difficult.

Managing the demand to try to smooth out the fluctuations in work required is an important part of personnel scheduling. Examples of such efforts are visible in the reduced prices for matinees at theaters, early-bird specials at restaurants, and discounted long-distance calls for off hours. In manufacturing, longer, less complicated tasks are often done at night to reduce the number of supervisory and support staff needed. It is this ability to manage or alter the work required that can make the scheduling steps iterative. If the current demand proves too difficult to schedule, then this first step can be revisited to try to alter the demand. Once the concept of managing demand is acknowledged, the organization will often find suitable ways of making the demands change to match the ability to satisfy the demand.

2.2. Determining Staffing Requirements

Even when the work required is known, the process of determining how many people are needed can be very difficult. The first step is to break down the work required into shifts, with the number of people needed for each shift. The whole process depends on several factors and will be discussed in some detail in the subsequent sections. The criteria involved are as follows: what shift lengths are to be considered, what contract conditions apply, how many part-time workers are to be used, what sickness rate can be expected, what vacation and statutory holiday conditions apply, and what are the capabilities of the staff available. This last point illustrates that the process may be iterative because the worker availability may necessitate a change in the number of workers needed.

2.3. Determine Personnel Availability

There are, of course, the obvious categories of workers, such as full-time workers, regular part-time workers, casual workers, temporary workers, and labor pools. These categories can be expanded. For example, there may be full-time workers who only work weekends. There can be other full-time workers who only work specific shifts such as nights. In the part-time worker category, there may be two workers doing job sharing so that they are scheduled as if they were one full-time worker. In most cases, regular part-time workers are associated with specific jobs and times of availability. Casual workers also are often associated with specific jobs, while others can do many jobs. Maternity leave, long-term sick leave and other leaves, as well as vacations and requests for time off that have been granted must be acknowledged in determining who is available to work. In the last few years, many industries have employed more part-time employees to reduce costs and increase flexibility in scheduling. For example, the prison systems and schools now use more part-time staff that before. The availability of all the people must also include the number of hours of work available, the shifts, and the days of work available.

2.4. Developing Work Schedules

There is one fundamental approach to scheduling that needs to be considered when developing work schedules. In the past, many of the approaches have been to determine how many workers are needed in each time period and then assign people to the jobs. The emphasis has been from the management point of view of ensuring that there are enough people available to get the work done. This is not the best point of view or the approach that should be taken. The workers are at work, on average, 40 hours per week, even if they get paid for 37½ or some other figure. This leaves 128 hours of off time in each week for the workers. To the workers, the time off is the most important time. This is the time that they have to spend with family and friends, to indulge in hobbies, and just to relax. In developing schedules, the whole approach should be to make the time off for the workers the best time possible. A secondary goal, and one that is relatively easy to achieve, is to ensure that the right number of workers is available at all times.

For example, a worker does not like to have a weekend off to spend with the family, where they work the night shift Friday night which starts at 11:30 p.m. Friday and ends at 7:30 a.m. Saturday. By the time they get home from work on Saturday morning, have something to eat, and get at least seven hours of sleep, the day is gone. Although the shift schedule might show Friday night working, Saturday and Sunday off, and return to work on the day shift Monday, it is not a desirable time off.

Once the schedule has been built, it is relatively easy to assign workers to different lines of the schedule.

3. TYPES OF PERSONNEL SCHEDULING PROBLEMS

There are many ways of categorizing shift scheduling problems. Some of the choices are based on the demand patterns, such as stable demand during a shift vs. variable demand during a shift, or demand only during a single shift or multiple shifts. Another way of considering problems is by the way workers are scheduled, such as scheduling individuals vs. crew scheduling. Within each category there are numerous subcategories, and recent research has made progress on these problems.

3.1. Single Shift

For the most basic single shift problem, where there is only demand for the day shift for five days a week, the problem does not generally create difficulties. All workers can be scheduled to work Monday to Friday with every weekend off. Recently, many companies have changed this system to have flexible hours for the workers, providing that they spend at least a given number of hours in a specified time period, such as between 10 a.m. and 4 p.m. Again, no difficulty arises, and the same holds for summer hours, where people work 10-hour shifts for four days a week.

The more difficult problems occur when the daily demand for workers extends over a longer period than eight hours. In most retail stores, banks, and service outlets, such as clinics and repair depots, the hours of operation may be as many as 14 or 16 hours per day. Even operating rooms in major hospitals are used from early in the morning to 6 or 7 in the evening. A second condition that adds a whole new level of difficulty, is if the daily demand is for more than five days per week. Now the problem involves giving workers suitable weekends off with a required frequency.

Most retail outlets have developed sophisticated databases that track the historic sales. These stores are usually open long hours, seven days a week, and have a demand for workers that varies both by the day and by the 15-minute interval within a day. Although this is still a single-shift problem, the increased difficulty is shown when building schedules for these cases. Fortunately, part-time workers are used extensively, but even so, experience has shown that there is a great deal of discontent with the final weekly schedules that are posted. Section 4 of this chapter will address how to improve schedules for the time-varying demand problem.

3.2. Multiple Shifts Required

Multiple-shift scheduling problems have many added difficulties:

- Limiting the number of shift changes in a given time period
- Having a given amount of time off between shift changes
- Preceding weekends off and vacations with a suitable shift
- Balancing the less desirable shifts among workers
- Limiting the number of successive days that some shifts are worked
- Using worker seniority when assigning shifts if required by the contract
- Choosing the mix of shifts, such as 8-hour, 10-hour, and 12-hour shifts
- Balancing the paid hours worked

3.2.1. Crew Scheduling

In many continuous process industries, such as mines, steel, and chemical companies, the workers are scheduled by crew. The author has completed many consulting contracts for such companies, but has also turned down many more contracts for the very simple reason that one recurring problem could not be solved. The problem that arises time and again, but that is not solvable, is the request to help with the scheduling to reduce the amount of overtime being paid. Consider the mathematics of the situation. A crew is needed to be working continuously, seven days a week. This means that there are 7 days of 24 hours for a total of 168 hours of work needed. The company will have four crews rotating shifts to provide the coverage. Each crew provides 40 hours of work for a total of 160 hours. The shortfall of 8 hours each and every week is made up by the same four crews, so there must be 8 hours of overtime each week, which is an average of 2 hours of overtime each week for every crew. The mathematics seems obvious, but the request for help in removing this overtime arises a startling number of times. For most continuous process companies, there is no help in reducing this overtime, but in a later section a set of suggested solutions for mining companies will be discussed.

3.2.2. Shift Scheduling for Individuals

Police, hospital workers, prison workers, all emergency systems, telephone workers, and many other seven-day-a-week operations schedule individuals rather than crews. These are the situations allowing for the most flexibility in scheduling, and also requiring complicated methods. They are also the source of most of the discontent with schedules because the workers expect better schedules than they receive, or better than may be possible. The following are some of the features not already mentioned above that make for discontent:

- Working too many days in a row too often
- Having too many single days off
- Perceiving inequities in schedules among workers
- Having too few shifts if a part-time worker
- Not having requested time off
- Not having a fair share of statutory holidays off

3.3. Hierarchical Workforce Requirements

Examples are perhaps the best way of illustrating these problems. For example, a prison system may require 20 guards for a shift, and at least one must be a qualified dog handler. In a hospital unit there may need to be eight nursing staff on the day shift, with at least six of them being four-year graduate nurses. The balance can be either nurses, nurse's aides, or nursing assistants.

Calculating the number of people needed and scheduling them is a little more difficult if there is a shortage of qualified people in some categories. Scheduling can be made easier by hiring the right mix of people or by training the people to achieve a desirable mix of worker skills. In recent years, many contract concessions have been given by unions to assist with this cross-training of workers.

3.4. Part-Time Workers

Part-time workers have become the mainstay for many industries. Various reasons exist for having part-time workers, only some of which pertain to scheduling, as can be seen in the following list:

- The full-time complement allowed is not enough to cover all the required workers, so part-time workers are scheduled into the master schedule to make up the deficit.
- People call in sick, so part-time workers are called at the last minute to fill the vacant spot in the schedule.
- A statutory holiday occurs and some full-time workers are given the day off; part-time workers are scheduled in advance for this holiday.
- Vacations, training time, business trips, long-term sickness, and other reasons for using a part-time worker are known in advance when the schedule is created.
- There is a fluctuation in demand during the day, or during the week that needs to be accounted for using part-time workers. In many cases, the need for part-time workers is known in advance, but in others, additional people will need to be called at the last minute.

Regular part-time workers are staff that are associated with the employer and used on a regular basis. In many instances these employees will fall under a union or employer contract. Some of the conditions prevailing in contracts are as follows:

- The person must be available for a given number of hours per week.
- The person must work at least a required number of weekends.
- The person must have at least a specified frequency of weekends off.
- The person cannot be required to work more than a specified number of days in a row.
- When part-time workers are needed, they are to be asked if they are available in some specified order, such as seniority, or equitably with the others.
- When called to work, the shift length must be at least as long as a specified time.

Many of the conditions above drive the way a schedule can be built. Very good records must be kept to ensure that the conditions are satisfied, as penalties often apply. The contacting of part-time people to get them to commit to a specific job assignment is very time consuming and costly. For nurse scheduling, software packages are available on the Internet that will assist in this task.

Casual part-time workers are not usually employed under a set of contract conditions. For example, in nursing, a hospital will have a list of nurses who would like some work but who are not on the regular part-time list. These nurses are called only after the list of regular part-time nurses has been exhausted. They will work a full shift, in one of a short list of units that depends on the person's skills and experience. Retail stores and fast-food outlets use many casual workers, including a large number of students. This scheduling also takes a lot of time because of the information base that must be referred to when scheduling and updated as a schedule is used. Some of the information required is similar to that of the regular part-time workers. The following is a partial list:

- The person's usual available hours
- The number of hours desired by the person
- The skills that the person has and the jobs he or she qualifies for
- The seniority of the person
- The number of hours that the person has worked in a rolling window of time
- The person's immediate past schedule record
- The means of contacting the person
- How many recent requests to work were refused or not answered
- Any short-term unavailability times requested by the person for personal reasons
- The average number of hours given to people in the same category of workers to ensure equitable assignment of work time

Many grocery stores and fast-food stores leave the scheduling of part-time workers to be done manually by administration. While they try to do a good job, it is time consuming and inequitable schedules can inadvertently arise. Such schedules can lead to discontent and employee unrest. A good system must be set up for this task.

Temporary workers are only part-time in that they are not permanent workers. They will usually be summer students or people hired to fill a job that is being reserved for someone who will be absent for a longer period of time. Long-term sickness and maternity leaves are simple examples of the need for temporary workers. Temporary workers will be scheduled just as if they are full-time workers, subject to all the rules for the full-time workers they are replacing. In seasonal situations, the temporary workers may be the bulk of the people being employed.

There is one more class of workers who appear as part-time workers to the scheduler but are in fact full-time workers. The name for this class differs from industry to industry and changes with time within an industry. "Relief labor pool" and "float pool" are names for a group of full-time people who are available to work in different jobs in the organization. The scheduler for one group of workers will use these people as needed. Essentially, the group scheduler is only assigning a job to the relief worker, who is already working to a predeveloped schedule. The person responsible for developing the relief pool schedule and for maintaining it for requests for changes for personal reasons, holidays, and vacation days is another issue. Someone must also be responsible for assigning jobs to the relief pool people who have not been requested to work by a specific unit.

4. CREATING SCHEDULES

Schedules have an important effect on employees' lives. A good schedule for one person may be considered a poor schedule by another person. There are no absolutes, but there are guidelines. If the schedules are good, worker performance and worker retention will improve, which is good for the organization as a whole. For example, in a major city hospital with 21 operating theaters, there was a 30% turnover of nurses. Changing the schedules made the turnover rate drop to the hospital average. A schedule can be considered optimal in the sense that it only reduces labor cost to the minimum, but an optimum schedule will minimize overall costs by reducing labor costs while improving performance and keeping workers satisfied with their schedules. As mentioned earlier, the goal of scheduling should be to develop optimum schedules, not just locally optimal schedules.

Much has been done in recent years to understand the personnel scheduling problem. As with most research, progress is made by making simplifying assumptions and solving the resulting problem. Insights are gained and the problem is then expanded slightly. By this iterative approach to research, problems resembling the real-world situation are finally attained. In personnel scheduling, many of the real-world problems have not yet been solved, but much of the insight obtained in the process of doing the research helps the scheduler to develop workable schedules. In the next few

TABLE 1 Week as Monday to Sunday

Line	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	X					X							X	X
2			X				X						X	X
3						X	X	X					X	
4						X	X			X				X

sections, some of the more useful facts will be presented, along with suggested ways of using them when scheduling.

4.1. Single Shift

Single-shift scheduling is still used in many situations. The shift is called a day shift, whether a person starts at 7 a.m. or 3 p.m. Schedules can be made for people and the time of starting can be determined later. The more interesting situation is a seven-day-a-week operation, with people being scheduled to work eight-hour shifts for an average of five shifts per week.

Cyclic scheduling was one of the first analytic approaches to single-shift scheduling. The situation can have a variable demand for the seven days and a requirement that all workers have two successive days off. This problem can be solved by mathematical programming, as shown in Baker (1974, 1976) and Tibrewala et al. (1972). The final schedules will have each person working five consecutive days followed by two days off. If a person has Tuesday and Wednesday off in one week, then he or she will always have the same two days off and will never have a weekend off. This type of schedule, while useful in some situations, is optimal from the short-term point of view of the employer but may in the long run cause a higher turnover of employees as they move in order to have some weekends off.

Burns (1978) introduced three very interesting facts to the process of scheduling. The first was that the definition of a work week should be Sunday to Saturday, not Monday to Sunday, as will be shown below. The second was that a simple analytic formula could be used to calculate how many workers were needed, and the third was that a simple polynomial time method could be used for developing optimal schedules without using mathematical programming. These three facts have formed the basis of many research papers and doctoral theses since that time, and they are useful in practice. Burns and Carter (1985) generalized the problem to allow for any frequency of weekends off.

Consider the problem of building a schedule that has one person working each day of the week, and the week is defined as Monday to Sunday. The schedule is to have each person working five days a week, with the maximum work stretch six days, and each person is to have every second weekend off. The difficulty arises in the week before a person has the weekend off. Once a person has Saturday and Sunday off in a week, he or she must work the five weekdays in order to have five working days per week. This means that in the week before, he or she must have either Sunday or Saturday off in order to restrict the work stretch to a maximum of six days. In Table 1, the schedule necessary to have at least one person working every day is given.

In order to have one person working each day, it is necessary to hire four people. To have n people each day, the requirement will be $4n$ workers. The schedule can be presented in a cyclic form, as shown in Table 2, with one person starting on each of the four lines and working that week before going to the next line. The person working the last line goes to the first line next. Notice that in order to have one person working on the weekend, there will be up to four people working on a weekday.

Changing the definition of the week to be from Sunday to Saturday means that the schedule can be constructed with only two people. Table 3 is an example of a cyclic schedule for a scheduling

TABLE 2 Week as Monday to Sunday, Cyclic Schedule

Line	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	X					X	
2						X	X
3			X				X
4						X	X
Totals	3	4	3	4	4	1	1

TABLE 3 Week as Sunday to Saturday, Cyclic Schedule

Line	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	X					X	
2			X				X
totals	1	2	1	2	2	1	1

problem with cyclic demand. Start one person on each line and have the person go to the next line after Saturday, where the first line is next after the second line.

In a cyclic scheduling problem, the demand on each day of the week is the same for each week. However, the demand may vary from day to day within a week. There are two basic types of schedules for the cyclic problem. The first is a cyclic schedule, such as the one in Table 3. In a cyclic schedule, the schedule itself repeats over some defined period, thereby giving a "master rotation." The second type of schedule is an iterative schedule, where an algorithm is given that schedules the first week of the schedule, and then another part of the algorithm shows how to generate week $i + 1$, given week i of the schedule. An iterative schedule might never repeat, and no "master rotation" is available.

The following is the statement of the scheduling problem that can be solved using the methods of Burns and Carter (1985):

W is the minimum number of workers required to cover a seven-day-a-week operation with the following primary constraints satisfied:

1. The demand per day n_j, $j = 1, 2, \ldots, 7$; n_1 being Sunday's demand.
2. Each worker is given at least A out of B weekends off.
3. Each worker works exactly five out of seven days each week.
4. Each worker works no more than six consecutive days.

In Burns (1978), the requirement was for every second weekend off.

This last constraint is added because a person could work the last five days of one week, followed by up to five more consecutive days of the next week, unless the restriction were in place. In Brown et al. (1976) over 50% of the schedules produced had work stretches of 10 consecutive days. Even as late as 1994, papers have appeared where the work stretch is as long as 10 days. Such schedules are not of much practical use.

The algorithm of Burns and Carter (1985) is an iterative scheduling method. The basic approach of the algorithm for the problem in this section is simple.

Ignore some of the constraints to get an absolute lower bound on the number of people required to complete the schedule. It will be a lower bound because if more constraints were incorporated, some schedules might not be feasible and more people might be needed.

Build a schedule for this lower bound number of people by first assigning the weekends off so that the required frequency of weekends off will be met for everyone.

Then assign the remaining days off for each worker to satisfy the work stretch constraint and ensure that everyone has exactly two days off per week.

Note, as outlined in the previous section, that the method does not assign workers to work specific days, but does concentrate on the workers' time off, which will produce better schedules for the individual. Following the detailed steps of the algorithm ensures that the schedule for just the lower bound number of people will, when completed, satisfy the total requirements for workers in each time period. The schedule is then optimal because no other schedule could be produced using fewer people.

Lower bounds for the number of workers needed will now be given. There are three bounds that are used:

1. *The weekend constraint:* The total number of employees' shifts available for the B weekends must be sufficient to meet the maximum weekend demand for the B weekends. In B weeks, each employee is available for $(B - A)$ weekends. Hence, $(B - A)W \geq Bn$, where $n = \max(n_1, n_7)$ is the maximum weekend demand.

$$L_1\colon W \geq \left\lceil \frac{Bn}{(B - A)} \right\rceil, \text{ where } \lceil x \rceil \text{ is the upper integer part of } x.$$

2. *The total demand constraint:* The total number of employees' days worked per week must be sufficient to meet the total weekly demand for shifts. Since each employee works exactly five days per week:

$$L_2\text{: } 5W \geq \sum_{i=1}^{7} n_{\mathrm{i}} \text{ or } W \geq \left\lceil \frac{1}{5} \sum_{i=1}^{7} n_{\mathrm{i}} \right\rceil$$

3. *The maximum daily demand constraint:* The number of employees must be sufficient to meet the maximum demand on any day.

$$L_3\text{: } W \geq \max_i(n_i),\ i = 1, 2, \ldots, 7$$

The algorithm uses a workforce equal to the maximum of the three constraints. The paper proves that if the algorithm is used, then the W employees will be sufficient to satisfy all the problem constraints, including the ones not used in calculating W.

In scheduling the weekends off, the algorithm assigns a number from 1 to W to each employee. Since n people are required to work each weekend, $(W - n)$ can have the weekend off. Assign the first weekend off to the first $(W - n)$ people. Assign the next $(W - n)$ weekends off to the next $(W - n)$ people. This process is continued cyclically with employee 1 being next after employee W.

For each week of the schedule, every employee must have exactly two days off. In any given week, there are already $(W - n)$ Sundays off at the beginning of the week and $(W - n)$ Saturdays off at the end of the week. An additional $2n$ days must be given off so that all W people have two days off and the algorithm constructs n off-day pairs for this task.

Denote the surplus workers for day j by S_j.

For Monday to Friday $S_j = W - n_j$ for $(j = 2, 3, 4, 5, 6)$
For Sunday $S_j = n - n_1$
For Saturday $S_j = n - n_7$

Iteratively, construct a list of n pairs of off days, numbered from 1 to n, as follows:

1. Choose day k such that $S_k = \max(S_j)$.
2. Choose any $i \neq k$ such that $S_i > 0$. If $S_i = 0$ for all $i \neq k$, set $i = k$.
3. Add the pair (k, i) to the list and decrease S_k and S_i by 1.
4. Repeat this procedure n times.

Pairs of the form (k, k) are called nondistinct off-day pairs and will be at the end of the list.

The next step is to assign the off-day pairs in week 1. Employees will fall into one of four categories, depending on the weekends off at the beginning and end of the week. Table 4 illustrates the categories.

There are n people working on each weekend. Therefore, $|T3| + |T4| = n$ from weekend 1 and $|T2| + |T4| = n$ from weekend 2, for $|x|$ = the cardinality of the set x. Each employee of type $T2$ is paired with one of type $T3$.

Assign the n pairs of off-days from the top of the list, first to the employees of type $T4$ and then to employees of type $T3$. The $T3$ get the earliest day of the pair, and the latest day is assigned to their associated type $T2$ employee. The paper proves that nondistinct pairs of off-days always get assigned to type $T3$ workers and then split. Hence, the algorithm will not assign two days off to a worker where the two days are the same day.

The next step is to assign the off-day pairs in week i for $i > 1$. Assume that weeks 1 up to week $(i - 1)$ have been scheduled. In week i, categorize the workers into the four types $T1$, $T2$, $T3$, $T4$. The critical problem of this step is to ensure that no employee has a work stretch longer than six days. The assignment of off-days for week i falls into two cases:

TABLE 4 Worker Categories

Category	Weekend 1	Week 1	Weekend 2
Type $T1$	off	no off-days needed	off
Type $T2$	off	one off-day needed	on
Type $T3$	on	one off-day needed	off
Type $T4$	on	two off-days needed	on

Case 1: If the list of off-day pairs contains nondistinct pairs of the form (k, k), for some day k, then week i is scheduled in precisely the same way as week 1 and is independent of week ($i - 1$).

Case 2: If all off-day pairs are distinct, of the form (j, k) where $j \neq k$, then all employees of type $T3$ and $T4$ in week i are associated with the same pair of off-days that they received in week ($i - 1$). The type $T4$ are given both days off and the type $T3$ get the earliest day of the associated pair. The $T2$ employees get the remaining days of the $T3$ pairs. In this instance, the maximum possible work stretch will be five days.

Example 1:

day j	1 Sun.	2 Mon.	3 Tue.	4 Wed.	5 Thu.	6 Fri.	7 Sat.
required:	4	4	5	5	7	7	5

where each person requires one out of three weekends off. $A = 1$; $B = 3$, and $n = 5$. Then:

$$L_1 = \left[\frac{3 * 5}{(3 - 1)}\right] = 8;\ L_2 = \left[\frac{37}{5}\right] = 8;\ L_3 = 7$$

The minimum workforce is $W = 8$. $(W - n) = 3$, which is how many workers are assigned off each weekend. The schedule after assigning the weekends off looks like Table 5.

After the weekends off are assigned, there are eight people available to work each weekday and five on the weekends. The surplus is:

day j	1	2	3	4	5	6	7
surplus S_j	1	4	3	3	1	1	0

Because of ties in the surplus, there are many different ways to make the pairs of off-days. The following is one list of pairs that would be generated using the algorithm as stated:

List 1: (Mon., Tue.), (Mon., Wed.), (Mon., Tue.), (Mon., Wed.), (Tue., Wed.)

Two features of the pair-choosing algorithm are incorporated in its design. The first is that the surplus workers on any one day, when the selection is completed, will be made as even as possible. The second feature is that, if at all possible, nondistinct pairs will be avoided. As far as the proofs of the optimality of the algorithm are concerned, only the second feature is necessary. This means that any other algorithm for choosing the pairs is acceptable as long as it does not unnecessarily create nondistinct pairs. The quality of the schedules can be improved by modifying the step of the algorithm that chooses the pairs of off-days. If list 1 above is used, the resulting schedule will have two features that are considered to be bad in schedules. The first is that there will be single-day work stretches where a person has a day off followed by a working day followed by a day off (on–off–on). Such a feature is not wanted by either management and the workers. The second poor feature is there are type 4 workers who work both weekends of a week and who will receive nonadjacent days off. If at all possible, this should be avoided.

TABLE 5 Example 1: Assigning the Weekends Off

Employee	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	_	_	_
1	X	X													X	X						
2	X	X																				
3	X	X																				
4								X	X													
5								X	X													
6								X	X													
7															X	X						
8															X	X						

Some published research papers suggest that using heuristics after the schedule is built can minimize the single-day work stretch. It is often possible to choose a different set of pairs of days off, before creating the schedule, that will result in no single day work stretches or split days off for the type 4 workers. We will improve the schedule of this example, but first it should be clear that it is not always possible to remove all the single day work stretches.

Selkirk (1999), a doctoral student of Burns, published a thesis that shows that for some problems, a new bound that increases the size of the workforce needs to be introduced to ensure that the single-day work stretches are avoided. He gives very complicated, optimal polynomial time algorithms for different subsets of the single shift scheduling problems. These algorithms are beyond the scope of this article but are mentioned here to show that heuristics cannot always be used to eliminate single-day work stretches.

The quality of the schedules can be improved if the cause of the poor schedules can be avoided when creating the off-day pairs. By applying the method for choosing pairs by Burns and Carter (1985), it is ascertained that nondistinct pairs can be avoided. The example has two type 4 workers each week. A type 4 worker who just becomes a type 4 worker in week *i* must have worked the weekend in the previous week and had one other day off, and therefore was a type 2 worker in the previous week. In the algorithm, this person gets both days of the associated pair off in week *i*. If all the pairs were adjacent pairs, then there would be no problem. Therefore, one goal in picking the pairs is to have as many pairs of adjacent days as is possible when type 4 workers are required.

The single-day work stretches are caused by two types of pairs. The first are the pairs (Mon., Tue) and (Thu., Fri.). In the algorithm the person who has Sunday at the beginning week off gets the latest day of the pair. This means that for the pair (Mon., Tue.) they would receive the Tuesday off after having the weekend off, thus creating a single-day work stretch. Similarly, the person having the weekend off at the end of the week receives the earliest day of the pair off, which for the pair (Thu., Fri.) will give him or her Thursday and Saturday off, which creates a single-day work stretch. The algorithm used for choosing the pairs of days off should avoid creating these two troublesome pairs.

The second cause of single work stretches is having a type 4 worker receive both days of a pair off, where the days in the pair are separated by only one day. This pairing should be avoided when type 4 workers exist in the schedule.

There are other quality of work schedule considerations. For example, if the demand on Saturday is different from the demand on Sunday, then it is possible to give days off on one of the weekend days. Workers prefer to have weekend days off. The pair (Mon., Fri.) is also useful because it can be exchanged in the schedule for any pair and still maintain a maximum work stretch of six days.

A new list of off-day pairs can now be prepared for example 1, taking into consideration the quality considerations just discussed.

List 2: (Sun., Mon.), (Tue., Wed.), (Tue., Wed.), (Tue., Wed.), (Mon, Fri.)

In week 1, the algorithm will assign adjacent pairs from the beginning of the list. In subsequent weeks the type 4 workers will either receive a pair of adjacent days or will receive the pair (Mon., Fri.). With a little thought, it is easy to see that this pair can be exchanged with any other pair of type 3 and type 4 workers without violating the work stretch constraint of any of the three people involved in the switch.

The final schedule is shown in Table 6.

In week 2, worker 3, a type 4 person, would have received the pair of off-days (Mon., Fri.), but this can always be switched with any paired workers of type 2 and type 3 receiving days off from

TABLE 6 Example 1: Assigning the Weekdays Off

Employee	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	_ _ _
1	X	X			X					X					X	X				
2	X	X			X						X	X								
3	X	X					X				X	X								
4				X				X	X			X								
5				X				X	X					X						
6			X					X	X	X										
7		X	X						X							X				
									X											
8				X	X						X				X	X				
surplus		1	0	0	0	1	1	0	1	0	0	0	1	1	0					

the pair (Tue., Wed.), so the first and fourth workers were used for the switch. The final schedule has no single workday stretches, and all type 4 workers have pairs of adjacent days off.

The bounds for the minimum number of workers required can be used for other purposes. For example 1 above, if a part-time worker is hired for Saturday, the variable n is reduced to 4. Bound L_2 still requires that 8 workers be hired, but the frequency of weekends off can be increased to every second weekend off without hiring more workers. For $A = 1$, $B = 2$, $n = 4$, the bound L_1 will still be 8. An interesting result will be that in the final schedule, the surplus workers on the weekdays will be increased because more days off will be given on the weekend for the same 8 people. The employer will need to find work for these surplus people, if possible.

One way of eliminating surplus workers in the final schedule is to hire part-time workers for the weekend, and for the weekdays to have the bounds given by L_1 and L_2 equal. Extra part-time workers would be hired to ensure that the bound L_2 is divisible by five.

The bounds given are very general and can be used for much more complicated scheduling problems. For example, Burns (1981) showed that the same bounds are sufficient for the problem having multiple eight-hour shifts with different demands on the shifts, and with the restriction that there must be at least 24 hours off between shift changes. As mentioned above in the reference to Selkirk (1999), enforcing no single-day work stretches may necessitate increasing the size of W by using a new bound.

Adjacent days off is another restriction that has drastic implications when the maximum work stretch is restricted to six days. The six-day work stretch restriction is the key factor. Burns and Koop (1984) explored this case and found the following:

- There are very few possible schedules.
- It is no longer possible to have each worker work exactly five days per week as is shown in Table 7. Instead, the number of days worked averages five days per week over the B week period of the cycle.
- The number of workers will need to be increased drastically from the same problem where single days off are allowed.

Consider the lower bound L_1, based on the weekend constraint. In most practical problems, the weekend constraint will be tight because it is desirable to give as many weekends off as is possible with the current work force. If the work stretch is limited to six days and all days off must be adjacent, then for each string of weekends off, a worker must have either the Saturday off in the week preceding the string of weeks or the Sunday off after the string. Table 7 illustrates this by giving the only two possible schedules, each with two work stretches of six days, and one of three days, for the example of one weekend off in three weeks.

A new weekend bound needs to be calculated. The bound is constructed as before, using the obvious fact that the weekend shifts available from W workers over B weeks must be greater than or equal to the shifts needed on weekends in B weeks. If all the weekends off are given sequentially, followed by all the weekends worked, each worker will only need one extra weekend day off. If the weekends off are separated by worked weekends in the B week cycle, then more single weekend days off will be needed, which might require more workers.

$L_{1\,\text{paired}}$: Available weekend shifts ≥ required weekend shifts

$$[2(B - A) - 1]W \geq 2Bn$$

$$W \geq \left[\frac{2Bn}{2(B - A) - 1}\right]$$

Table 8 gives the values of the new and old bounds for some of the frequencies of weekends off found in practice.

The most startling result is for every second weekend off, where the number of workers needed in the schedule must have all paired days off doubled! Even for the other two cases, the number of

TABLE 7 Paired Days off, $A = 1$, $B = 3$

S	S M T W T F S	S M T W T F S	S M T W T F S	S
X	X X X	6 days X X	6 days X	X
X	X 6 days	X X 6 days	X X X	X

TABLE 8 All Paired Days Off Compared to Some Single Days off

Frequency of Weekends Off A	Cycle Length B	L_1 paired	L_1
1	3	$2n$	$3n/2$
2	5	$2n$	$5n/3$
1	2	$4n$	$2n$

workers needed increases by a large percentage. To reduce the surplus workers that would be scheduled for weekdays, the demand for workers on weekends would need to be reduced by hiring part-time workers for weekend shifts and hiring fewer full-time workers.

Shift scheduling is like job shop scheduling in that a small change in the problem statement results in an entirely new algorithm being needed. Since Burns and Carter (1985) solve the most general problem available with variable demand, why produce special algorithms for more restricted cases? The answer is that for more restricted cases, better schedules can be guaranteed. Several cases, such as constant demand each day, or a fixed demand of N for the weekdays, and a demand of n for both weekend days, have been addressed in the research literature. The algorithms presented for the case where $N \geq n$ are easier to implement and to prove that they are optimal. One the main differences is in picking the list of paired days off. However, the Burns Carter (1985) method, with the paired-off list modified, as suggested in this article, will solve all of these cases.

A second class of single-shift scheduling problems occurs when the demand pattern changes by shortened time intervals within a day. The demand for telephone operators and cashiers in retail stores may vary by the hour or half hour or by even less time. In some cases it is also necessary to build schedules that give the workers lunch breaks and coffee breaks while ensuring that the demands are met by the workers not on breaks. Most retail stores are closed for several hours a day. It is not necessary to consider shift rotation times from one day to the next, which makes it a single-shift problem, even though there will be many different shifts during the day. There are the same three steps with this type of scheduling as with the problems where the demand is constant over a shift: determining the demand; creating the schedules that meet the demand at the lowest cost possible; and meeting the scheduling needs of the people who work the schedule.

Many of the big chain stores have a database and computer software that can predict the requirements for the days to be scheduled, based on annual history and recent history that will account for trends. Exponential smoothing, running averages, or other methods can be used to predict the daily demand. Adjustments must be made based on weather conditions in the past, special promotions, or other predictable influences.

There are several methods reported in the literature dealing with creating the schedules, given the demand. One way is to use mathematical programming methods such as linear or integer programming. Other approaches such as Segal (1974), using network flows, Henderson and Berry (1976), using heuristics, and Glover and McMillan (1986), using a taboo search, have addressed the problem of determining the shifts to be worked on one day. But why are these methods not used extensively in practice? One of the reasons is that the demand is not cyclic. The daily demand continues to change each day of the year and from year to year. The scheduling problem needs to be solved every day, for every company location, and in some instances, for different groups of workers in each location. What is done in practice? Many retail stores are given the demand pattern expected for the next scheduling period by a central computer system, while others create the demand pattern locally. Some stores are given schedules from a central location, but in many cases these have not proven to be very useful. Management in the store usually create the schedules manually. The task is time consuming and unrewarding because workers are often unhappy with their schedules.

The methods currently used for this time-varying demand problem do not address the assigning of full-time workers to have a maximum work stretch of six days, or weekends off with a given frequency, or any constraints on the pairing of off-days, or, in fact, any conditions linking one day to the next. Research for the fixed-shift requirement, cyclic demand work, and research for the time-dependent, noncyclic demand case have been developed in parallel with little if any crossover. There is no need for this separation. The following is an outline of a method for creating the shifts needed and the schedules that will satisfy the full-time workers' requirements for good schedules, followed by a method of creating shifts for the part-time workers to complete the total schedule.

The suggested methods presented below are extracted from the retail shift scheduling algorithms used in Burns (1999). In a commercial computer package, there are many small details and specifics of the particular installation built into the algorithms. For example, the third step below would be modified based on how many hours the operation is open and on the number of peaks in the demand pattern. While important to the scheduling efficiency, these specific features are not necessary to the understanding of the method.

In any particular location, there will be a core of W full-time staff and many part-time staff. The full-time staff will have a set of conditions that need to be incorporated into their schedule, such as A out of B weekends off, a maximum work stretch of six days, and five days per week of work. Schedules for the full-time workers will be considered first.

Step 1. Calculate the demand pattern to be satisfied by the W full-time workers.

Shifts available on the weekend = $(B - A)W$

Shifts required = Bn, where n is the demand to be satisfied by full-time employees

Therefore, $Bn \leq (B - A)W$, or

Weekend shifts available from full time workers is given by

$$n = \left[\frac{(B - A)W}{B} \right]$$

where $\lfloor x \rfloor$ is the largest integer contained in x.

Assuming at least N shifts per weekday are to be satisfied by full-time people, the average number of shifts per weekday from full-time workers is given by

$$N = \left[\frac{5W - 2n}{5} \right]$$

If rounding down was done to create N, then let k be the number of eight-hour shifts rounded down, which is given by $k = (5W - 2n) \bmod (5)$. k weekday shifts will need to be created in the schedule, in addition to the N shifts each weekday.

The k shifts will be added to the N shifts on weekdays, requiring the most number of workers while ensuring that the maximum required per day does not exceed W, or will be distributed evenly so that the final demand pattern will be as follows:

Sunday = $d(1) = n$. Saturday = $d(7) = n$. The weekdays have demand of either $d(i) = N$ or $d(i) = N + a$, where a = the number of extra shifts, from the k shifts, that have been allocated to day i, $i = 2, 3, 4, 5, 6$.

Step 2. Use the Burns and Carter (1985) method with the list of off-day pairs, as modified above, to schedule the W full-time workers to the next week. Since the method is iterative, the starting position, after the first time that the algorithm has been used, will be from the previous week, using the list of off-day pairs in existence. After step 1, the number of full-time shifts required per day is known, and after step 2, who will work the shifts is known; but when the full time shifts will actually start within a day has not been determined.

Step 3. Create the shifts to be worked on each day of the schedule, using the part-time workers as well as the full-time workers.

Let $s(j)$ be the number of people needed in each time period of the day being considered. The time period could be an hour, a half hour, or just 15 minutes. Let d be the requirement for eight-hour shifts on the day being considered. The algorithm will try to schedule the eight-hour shifts first and then five-hour part-time shifts. The selection of five hours is arbitrary and could be changed depending on the installation. The algorithm may need to complete the first pass at allocating part-time shifts by using shifts as short as three hours. Then the method will assign lunch breaks and coffee breaks. The final pass will be to add shifts where needed, and reduce shift lengths where needed.

Creating the eight-hour shifts: Start at period $k = 1$.

If $s(k) > 0$, let m be the minimum of $s(k)$ and d. Create m eight-hour shifts, starting at period k. Set $d = d - m$ and $s(j) = s(j) - m$ for the eight-hours, starting with period k.

If $s(k) \leq 0$, set $k = k + 1$. Repeat this step to create eight-hour shifts until $d = 0$.

Creating part-time shifts: Start with period k from the previous step.

If $s(k) > 0$, check to see if there is enough time left for a part-time shift. If the time left is at least five hours, assign the shift length L to be five hours. If the time left is greater than or equal to three hours but less than five hours, assign the shift length L to be the time left. If the time left is less than three hours, set L equal to three hours and set F equal to the period that starts three hours before the end of the day. If F was set in this step, start at period F and create $s(k)$ shifts of length L and set $s(j) = s(j) - s(k)$ for all periods from period F until the end of the day. If F was not set in this step, start at period k and create $s(k)$ shifts of length L. Set $s(j) = s(j) - s(k)$ for the L hours starting with period k. Set $k = k + 1$ and repeat this step until all $s(k) \leq 0$.

TABLE 9 Weekly Eight-Hour Shifts for Time-Dependent Workers

Worker	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S
1	X						X	X			X						X				X
2	X			X						X				X	X						X
3			X				X	X						X	X		X				

The final steps are straightforward in principal but complicated in detail. Assign all the break times needed, using the surplus workers wherever possible, that is, where $s(k) < 0$. Then either add part-time shifts where $s(k) > 0$ or increase the length of a part time shift to reduce $s(k)$ to zero. Finally, try to reduce the length of part-time shifts to increase $s(k)$ to 0 wherever $s(k) < 0$.

The entire algorithm is repeated for all the days that are to be scheduled. An important output from the algorithm is the printed reports giving the schedules and who is to work each shift. The full-time people have been assigned using the method as outlined in this section. The part-time people will be assigned using the work of Section 3.4.

The algorithm is easy to computerize or to do manually and is very fast because the amount of work is polynomial.

Example 2: Time-variant, single-shift scheduling:

Consider a retail store that is open from 9 a.m. to 9 p.m. There are three full-time employees requiring every second weekend off, a maximum work stretch of six days, and five days of work per week. The required number of people is given in Table 10 in half-hour intervals. All eight-hour shifts must have a lunch period of a half hour; coffee breaks will not be scheduled but will be given during the day as time is available. The part-time shifts must be at least three hours long and can be up to five hours long.

From the facts given, $A = 1$, $B = 2$, $W = 3$ and from step 1 of the algorithm,

$$n = 1, N = \left[\frac{15 - 2}{5}\right] = 2, k = 3$$

There will be three weekdays with three workers required and two with two required. At this point, looking ahead to step 2 of the algorithm is useful. The modification to selecting the paired off-days above suggests that (Mon., Tue.) and (Thu., Fri.) pairs should be avoided to eliminate one-day work stretches for the full time workers, and we need $n = 1$ pairs of off-days. If the three extra working days are given to Wednesday, Thursday, and Friday, the only pair of off-days would be (Mon., Tue.), which is undesirable. To help make better schedules, the three extra workdays will be assigned to Monday, Thursday and Friday leaving (Tue., Wed.) as the pair of off-days for step 2 of the algorithm. Sunday $= d(1) = 1$, $d(2) = 3$, $d(3) = d(4) = 2$, $d(5) = d(6) = 3$, $d(7) = 1$.

In step 2 of the algorithm $(W - n) = 2$ people are given the weekend off each week. Table 9 gives the schedule for the eight-hour full-time workers.

Because there are no choices of days off when creating the schedule, the schedule is cyclic, with a period of three weeks; that is, the schedule is the first week repeated. For step 3 of the algorithm, completing one day will be sufficient to illustrate the algorithm. Consider a Thursday, with the requirement for workers given in Table 10.

Step 3 first assigns the following shifts:

Eight-hour shifts starting at periods 1, 3, and 4
Five-hour shift starting at period 5
Three-and-a-half hour shift starting at period 18
Two three-hour shifts starting at period 19

TABLE 10 Example 2: Required Workers for Time-Dependent Workers at the Start

Time	9		10		11		12		1		2		3		4		5		6		7		8	
Per	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$s(j)$	1	1	2	3	4	4	4	4	4	4	4	4	3	3	3	2	2	3	3	3	3	3	3	2

TABLE 11 Required Workers for Time-Dependent Workers on Thursday in the Middle

Time	9		10		11		12		1		2		3		4		5		6		7		8	
Per	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$s(j)$	0	0	0	0	0	0	0	0	0	−1	−1	−1	−1	−1	0	−1	0	0	−1	0	0	0	0	−1

Lunch breaks can be assigned to several places. Choose the half hour lunch break for the eight-hour worker number 1 to start at period 7, worker 2 at period 8, and worker 3 at period 9 and start a new three-hour shift to start at period 7 to cover for the lunch breaks. The current status of the worker requirement table is given in Table 11.

Using the last part of the step 3, the five-hour shift starting at period 5 would be reduced to a three-hour shift, and the three-and-a-half-hour shift starting at period 18 would be reduced to a three-hour shift. The final schedule would have the following shifts:

Eight-hour shifts starting at periods 1, 3, and 4
Three-hour shift starting at period 5
Three-hour shift starting at period 7
Three-hour shift starting at period 18
Two three-hour shifts starting at period 19

The final status of the worker requirement table is given in Table 12.

For the problem requirements and scheduling restrictions, the schedule is optimal. Two of the extra shifts could be removed if the minimum part time shift is reduced to two-and-a-half hours from three hours.

4.2. Multiple Eight-Hour Shift Crew Scheduling

As discussed in Section 3.2.1, a continuous operation working four complete crews has 2 hours of overtime per week per crew. Switching to 12-hour shifts or 8-hour shifts will not remove the problem. Later in this section, some special schedules for the mining industry are presented where the work to be done is reduced for short periods of time and no overtime occurs. In continuous operations, one industry standard is to have no single days off and to allow seven-day work stretches. Because the demand is constant, 24 hours a day for every day, it is possible to have as goals the following: a good schedule will have shift rotations always moving forward, that is, day to evening to night shift; a good schedule will also have three days off after the night shift, if possible, and will not start a weekend off after working the night shift on Friday, because the worker would need to sleep most of the Saturday. Tables 13 and 14 are two different schedules for four crews, each having the desired properties. In the schedules d, e, n are 8-hour shifts for the day, evening, and night respectively. As with the single-shift schedules, a crew (person) starts on each line and when a line is finished moves to the next line, with the first line being next after the last line.

Crew schedule 1 is the more standard type of shift rotation, whereas crew schedule 2 has the fast rotation of shifts that some places like to work. Note that, as expected, each crew only has seven days off in four weeks, which means they work eight hours overtime in four weeks.

If a maximum work stretch of six days is required, then a 24-week cyclic schedule can be made. An 8-week schedule forms the basis for the schedule and is repeated three times, with the starting shift changed each time. The starting shift will be determined by just continuing on the shift that was being worked at the end of the cycle. Table 15 gives the basic schedule that is worked by all four crews, starting the cycle with one crew on each of weeks 1, 7, 13, and 19.

Over the years, the author of this article has worked on numerous consulting projects for personnel scheduling. Repeatedly, the consulting has been initiated because a company signed a new contract with the union and then found that they either could not create any schedules to match the contract conditions or could not create schedules without a considerable increase in labor costs. The motive

TABLE 12 Required Workers for Time-Dependent Workers on Thursday when Finished

Time	9		10		11		12		1		2		3		4		5		6		7		8	
Per	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$s(j)$	0	0	0	0	0	0	0	0	0	−1	0	0	0	0	0	−1	0	0	−1	0	0	0	0	0

TABLE 13 Crew Schedule 1

Crew	S	M	T	W	T	F	S
1	X	d	d	d	d	d	d
2	d	X	X	e	e	e	e
3	e	e	e	X	X	n	n
4	n	n	n	n	n	X	X

TABLE 14 Crew Schedule 2

Crew	S	M	T	W	T	F	S
1	X	d	d	e	e	n	n
2	n	X	X	d	d	e	e
3	e	n	n	X	X	d	d
4	d	e	e	n	n	X	X

TABLE 15 Crew Schedules with Six-Day Work Stretches

week 1 *	week 2	week 3 *	week 4
S M T W T F S	S M T W T F S	S M T W T F S	S M T W T F S
n n n X X d d	d d d d X X e	e e e e e X X	n n n n n n X

week 5 *	week 6	week 7 *	week 8
S M T W T F S	S M T W T F S	S M T W T F S	S M T W T F S
X d d d d d d	X X e e e e e	e X X n n n n	n n X X d d d

TABLE 16 Crew Scheduling with 12-Hour Shifts and Four 4-Hour Maintenance Shifts

Crew	S	M	T	W	T	F	S
1	X	[*d*/*D*]	[*D*/*d*]	X	X	D	D
2	D	X	X	D	D	X	X
3	X	N	N	X	X	[*n*/*N*]	[*N*/*n*]
4	N	X	X	N	N	X	X

TABLE 17 Crew Schedule with Eight-hour Shifts and One Eight-hour Maintenance Shift

Crew	S	M	T	W	T	F	S
1	X	X	d	d	d	d	d
2	d	X	X	e	e	e	e
3	e	e	e	X	X	n	n
4	n	n	n	n	n	X	X

for moving the start of the week in single-shift scheduling to be Sunday rather than Monday was one such instance, and the analysis of the cost of having paired off days, with a maximum work stretch of six days, was another. Multishift crew scheduling presented a third case of having an impossible contract signed. A mining company agreed that all overtime would be voluntary and all workers would work exactly 40 hours each week (lunch and other breaks occurred within the 40 hours). The company thought that they could change to 12-hour shifts to solve the problem. Although scheduling with extended length shift will be discussed in Section 4.4, for the sake of completeness, crew scheduling using 12-hour shifts will be included in this section. In particular, schedules for the mining industry will be discussed because it is often possible to work partial crews for a short period of time and still keep the above-ground processes working. The goal in doing this is to eliminate the overtime for the workers. When there is a partial crew working, other workers can do maintenance on the mine tunnels not in use. The methods presented are abstracted from Burns (1983).

With 12-hour schedules, the problem of working an average of 40 hours per week is difficult because 40 hours for one week, and even 80 hours for two weeks, are not divisible by 12. There are many ways of resolving this issue, and for the following examples the schedules will have six shifts of 12 hours, and one shift of 8 hours every two weeks, which has all workers average 40 hours per week every two weeks. As before, d and n are used for 8-hour day and night shifts. D and N are used for the 12-hour day and night shifts respectively. A new piece of notation is needed.

$[x/y]$ means that the crew is divided in half and the first half works shift x and the second half works shift y. If $x = d$ and $y = D$, then the first half of the crew will work an 8-hour shift and have 4 hours off while the other half of the crew is working a 12-hour shift. Table 16 is just one of many schedules that can be created for the mining industry. Each crew (person) works exactly 80 hours in each two-week schedule. On Monday, Tuesday, Friday, and Saturday there is a 4-hour period with only half a crew working.

The times of the 8-hour shifts can be moved to have two maintenance shifts of eight hours each by giving one day in every second week where there is an 8-hour shift for both halves of the crew at the same time, instead of where it is in Table 16.

The 12-hour schedule in Table 16 solves the overtime problem for the mine, but not the condition that each worker must work exactly 40 hours per week. It is possible to use 8-hour shifts and to have one full maintenance shift every Monday where no regular crew is working. In this case, everyone works exactly 40 hours per week and the overtime can be voluntary. The essence of the solution is that the reduced production happens on only one of the 21 shifts in a week. Table 17 illustrates this schedule, which is a slight modification to the schedule in Table 13. Note that on Monday there is no crew scheduled to work on the day shift. Voluntary overtime could happen on that shift.

4.3. Multiple Eight-Hour Shift Scheduling for Individuals

In addition to the requirements of the single-shift scheduling problem, there are the difficulties outlined in Sections 3.1 and 3.2.2. With the crew scheduling problems the goals included having shifts rotate forward and having three days off after completing night shifts. The methods of this section do not include these goals, because the need for workers is different for different shifts. In the research literature there are two approaches to the problem. The first, by Burns (1981), is an iterative algorithm for a very general shift scheduling problem involving eight-hour shifts. This paper presents an optimal, linear time algorithm to schedule the following problem:

- There are P shifts of length eight hours each day and the start times may overlap.
- The demand N_j per shift j, for weekdays Monday to Friday, is met and the demand n_j per shift j, for weekends Saturday and Sunday, is met, with $N_j \geq n_j \geq 0$ for all shifts $j = 1, 2, \ldots P$.
- Each employee is given at least A out of B weekends off.
- Each employee works exactly five out of seven days in each week, with the week defined as Sunday to Saturday.
- Each employee works no more than six consecutive days.
- Each employee who works two consecutive weekends has a pair of consecutive days off during the week.
- Each employee has at least one day off when changing shifts. Because shifts can have overlapping start times, the algorithm yields at least $(24 - t)$ hours off between shift changes, where, if the completion time of the last shift on one day is less than or equal to the start time of the first shift on the next day, t is the negative of the number of hours difference, otherwise t is just the difference in hours.

After the algorithm is described, ways of extending the use of the algorithm to more problems will be presented.

The first steps of the algorithm are almost identical to the modified Burns and Carter (1985) algorithm of Section 4.1. Since the demand pattern has been restricted to being the same for each weekday and the same, but of a different value, for each weekend day, rather than totally variable, the extra condition of having pairs of adjacent days off for the type $T4$ workers can be incorporated. Surprisingly, no change in the size of the workforce is required for the multiple shift problem, and the same bounds can be used.

Step 1. Compute the minimum workforce size. Set

$$n = \sum_{j=1}^{P} n_j, \text{ and } N = \sum_{j=1}^{P} N_j$$

and use n as the demand on weekends and N as the demand on weekdays to calculate the bounds on the workforce W as in the Burns and Carter (1985) algorithm.

Step 2. Schedule the weekends off. Schedule the weekends off as in the Burns and Carter (1985) algorithm.

Step 3. Determine the additional off-day pairs. Select n pairs of off-days cyclically from the list (Mon., Tue.), (Wed., Thu.), (Thu., Fri.), (Tue., Wed.), (Mon., Fri.).

Step 4. Assign off-day pairs in week 1. Use the categories of workers from the Burns and Carter algorithm. Assign $|T4|$ pairs of adjacent off-days to employees of type $T4$. Associate each employee of type $T3$ with an employee of type $T2$. From the list of pairs of off-days not yet assigned, assign the earliest day of each pair to a type $T3$ employee and the latest day to the associated type $T2$ employee.

Step 5. Assigning shifts in week i, $(i \geq 1)$. For each week of the schedule, a total of n shifts are required on the weekends and N shifts on the weekdays. These requirements are met by forming n chains [Sun.–Sat.] and N–n chains [Mon.–Fri.] that are pair-wise distinct. A [$d1$–$d2$] chain, for $d1 < d2$, is a sequence of $d2 - d1 + 1$ workdays, taken from one or more employees. Two chains are pair-wise workday disjoint if no employee workday is a member of both chains. A [$d1$–$d2$] chain is formed by starting with an employee who works on day $d1$. This employee's workdays are assigned to the chain until either day $d2$ is reached or the employee has a day dk off, $(d1 < dk < d2)$. In this latter case, the chain is transferred to a new employee who works on day dk but had day $dk - 1$ off. This new employee's workdays are then assigned to the chain. This process is continued until day $d2$ is reached.

Construct the (N–n) chains [Mon.–Fri.], first using employees of type $T1$ and pairs of type $T2$ and $T3$ employees who are associated with adjacent off-days. The type $T1$ employees form a [Mon.–Fri.] chain by themselves, while a [Mon.–Fri.] chain started with a type $T2$ employee can be completed by its associated $T3$ employee.

The n chains [Sun.–Sat.] are constructed once the current schedule has been temporarily modified so that exactly n workers are available each day, where "available" means an employee does not have the day off and has not already been assigned to a [Mon.–Fri.] chain. Sunday and Saturday already have exactly n workers available, but the weekdays may have more than n. Remove the excess workers for each weekday, starting with Monday, by temporarily giving an employee an extra day off as needed, if that employee would have started working that day. An employee would have started working on a particular day if he or she is not scheduled to have the day off and if, on the previous day, he/she had an actual day off or a designated extra day off. Then form the n chains [Sun.–Sat.], starting with any employee who is scheduled to work on Sunday, in the same way that [Mon.–Fri.] chains were formed.

The chains must now be assigned to shifts. Prior to this, however, it is necessary to associate the extra days off with chains so that they will also be assigned to shifts. Start with any extra days off that are immediately followed by a workday that is assigned to a chain. Associate this extra day off and any extra days off that precede it, back to an actual day off, with that chain. Once this is done, the only extra days off remaining are those that fall between two actual days off. Associate these extra days off with any [Sun.–Sat.] chain at random.

The (N–n) chains [Mon.–Fri.] and the n chains [Sun.–Sat.], with any extra associated workdays, must now be assigned to shifts. If it is week 1, assign the [Sun.–Sat.] chains to the P shift types according to n_j, the number of employees required for each shift of type j. If it is week i, $(i \neq 1)$, assign the [Sun.–Sat.] chains to the same shift as the Saturday preceding the Sunday. In either case, assign the [Mon.–Fri.] chains to the shifts required to make the N_i–n_i weekday chains.

Step 6. Assigning off-day pairs in week i, $(i \neq 1)$. Assume that weeks 1 up to $(i - 1)$ have already been scheduled. As in step 4, each employee can be categorized into one of four types $T1$, $T2$, $T3$, and $T4$.

For each employee who is type $T4$ in week i, who was of type $T4$ in week $(i - 1)$, assign the same pair of off-days that he or she received in week $(i - 1)$.

For each employee who is of type $T4$ in week i, who was of type $T2$ in week $(i - 1)$, and who was associated with a pair of adjacent off-days, assign both days of that pair in week i.

For each employee who is of type $T4$ in week i, who was of type $T2$ in week $(i - 1)$, and who was associated with a pair of nonadjacent off-days in week $(i - 1)$, trade pairs of off-days with an employee who was of type $T4$ in week $(i - 1)$ but is of type $T3$ in week i.

For each employee who is of type $T3$ in week i, assign the same pair of off-days that he or she received one off-day from in week $(i - 1)$, or the interchanged pair from above. The type $T3$ employees get the earliest day of the pair, while the type $T2$ employees get the remaining latest days off.

Go to step 5.

Example 3: Multiple-Shift Eight-Hour Schedules:

A company has a seven-day-a-week operation with four shifts working each day, with the following demand for workers.

Shift	Time	Weekdays	Weekends
1	Midnight to 8 a.m.	1	1
2	8 a.m. to 4 p.m.	2	2
3	1 p.m. to 9 p.m.	3	2
4	4 p.m. to midnight	1	1

Each worker is to receive at least one weekend off every three weeks; the maximum work stretch is to be six days; any worker working both weekends of a week will have a pair of adjacent days off in the week.

Step 1: $N = 7$ the daily demand on a weekday, and $n = 6$ the daily demand on the weekends.

$$L_1 = \left\lceil \frac{3 * 6}{2} \right\rceil = 9, \; L_2 = \left\lceil \frac{5 * 7 + 2 * 6}{5} \right\rceil = 10$$

Therefore, $W = 10$.

Step 2: Table 18 shows the results of step 2.

Step 3: Choosing six pairs of days from the list gives the following pairs: (Mon., Tue.), (Wed., Thu.), (Thu., Fri.), (Tue., Wed.), (Mon., Fri.), (Mon., Tue.).

Step 4: There are two type $T4$ workers every week (workers 9 and 10 in week 1). Assign the first two pairs from the list to workers 9 and 10. Associate workers 5&1, 6&2, 7&3, and 8&4, assigning the next four pairs from the list to get the weekday assignment for week i in Table 18.

Step 5: In this example, N is one bigger than n so one chain [Mon.–Fri.] needs to be constructed. Fill in this chain with a string of the letter A in Table 19 and then fill in the extra shifts with the letter Y to get the result in Table 19.

Completing step 5 using the letters B, C, D, E, F, G for the six chains, starting with workers 5–10, yields Table 20. The shifts assigned would be $B = 1$, $C = 2$, $D = 3$, $E = 4$, $F = 2$, $G = 3$, and $A = 3$, and the extra shifts would be removed to give the schedule after week 1.

In step 6 for week 2, the new type $T4$ workers are workers 3 and 4. Worker 3 is associated with worker 7 and assigned to the off-day pair (Mon., Fri.), so he or she must be interchanged with worker

TABLE 18 Multiple Eight-Hour Shifts: After Steps 1–4

	S	M	T	W	T	F	S	S	M	T	W	T	F	S
1	X					X								X
2	X			X										X
3	X					X								
4	X		X											
5					X		X	X						
6			X				X	X						
7		X					X	X						
8		X					X	X						
9		X	X											X
10				X	X									X

TABLE 19 Multiple Eight-Hour Shifts: After the First Part of Step 5

	S	M	T	W	T	F	S	S	M	T	W	T	F	S
1	X	A	A	A	A	X								X
2	X			X	Y	Y								X
3	X					X								
4	X		X	Y										
5					X	A	X	X						
6			X				X	X						
7		X					X	X						
8		X					X	X						
9		X	X											X
10				X	X									X

9, who became a type $T3$ worker. The new associations are: workers 9&7 assigned to pair (Mon., Fri.), worker 3 assigned to pair (Mon., Tue.), worker 4 assigned to pair (Mon., Tue.), and worker 8, who was associated with worker 4, joins the association 10&8 assigned to pair (Wed., Thu.). Table 21 gives the completed schedule for the first two weeks.

The algorithm works for situations where the shift demand on Monday to Friday is at least as big as on the weekends. For situations where the weekend demand is larger, the algorithm can still be used if part-time workers are hired to reduce the weekend demand to the required level. Any problem where the frequency of weekends off is greater than or equal to one off in two weeks will have no type $T4$ workers. Step 4 of the algorithm illustrates that clearly. Hence, for these cases the need for adjacent pairs disappears. One would think that the list could be modified, as was done above for the single-shift Burns and Carter (1985) algorithm, in order to reduce the number of single-day work stretches. Unfortunately, the proofs of optimality and feasibility require the existing list. All is not lost, though. Since the demand is constant for the week, it is relatively easy to modify the algorithm to ignore the list of off-day pairs and dynamically assign the days off to the type $T4$ workers as needed. Then assign the type $T3$ workers days off from Monday to Wednesday, and then assign days from Wednesday to Friday to the type $T2$ workers. A computer program can do this, along with the necessary checks to assure feasibility.

A second, completely different approach to the eight-hour scheduling problem was presented by Burns and Koop (1987). In this approach a master rotation is created rather than using an iterative algorithm. The method is optimal and works for the situation where there are three shifts a day, the weekday demand is at least as large as the weekend demand, the maximum length work stretch is six days, and each worker must have at least *A* out of *B* weekends off. A library of minischedules called modules is given, and the complete schedule is created by combining the modules. There are many situations where master shift rotations are either preferred or required; for example, in Canada, all nurse schedules and all prison guard schedules must be master rotations.

Rather than the methods of Burns and Koop (1987), a slightly different method that also uses modules to build optimal master schedules will be presented. In Burns (1985) the modules have two extra properties: the number of shift rotations is controlled and the workers receive exactly *A* out of *B* weekends off rather than the maximum number off that is possible. These modules have been used extensively to create schedules in hospitals and prisons.

TABLE 20 Multiple Eight-Hour Shifts: After Step 5

	S	M	T	W	T	F	S	S	M	T	W	T	F	S
1	X	3	3	3	3	X	5							X
2	X	2	2	X	3	3	3							X
3	X	4	4	4	4	X	2							
4	X	3	X	1	1	1	1							
5	1	1	1	1	X	3	X	X						
6	2	2	X	2	2	2	X	X						
7	3	X	2	2	2	2	X	X						
8	4	X	3	3	3	3	X	X						
9	2	X	X	3	3	3	3							X
10	3	3	3	X	X	4	4							X

TABLE 21 Multiple Eight-Hour Shifts: First Two Weeks

	S	M	T	W	T	F	S	S	M	T	W	T	F	S
1	X	3	3	3	3	X	2	2	2	2	2	X	3	X
2	X	2	2	X	3	3	3	3	3	X	4	4	4	X
3	X	4	4	4	4	X	2	2	X	X	2	2	2	2
4	X	3	X	1	1	1	1	1	X	X	2	2	2	2
5	1	1	1	1	X	3	X	X	3	3	3	3	X	4
6	2	2	X	2	2	2	X	X	2	2	X	3	3	3
7	3	X	2	2	2	2	X	X	1	1	1	1	X	3
8	4	X	3	3	3	3	X	X	3	3	3	X	1	1
9	2	X	X	3	3	3	3	3	X	3	3	3	3	X
10	3	3	3	X	X	4	4	4	4	4	X	3	3	X

The module library is separated into different categories, such as 8-hour vs. 12-hour shift. Even greater separation is created. Some modules are for people who rotate shifts, and some are for people who work a fixed shift. Within a workplace both types of schedules are often needed. In addition, there may be a group of people that want specific characteristics for their schedule, and these characteristics may differ for other groups.

The whole library of modules is too extensive to give in this work, but a sample will be given, along with enough information to show how to create modules. Having seen a sample, the reader should find it relatively easy to begin creating his or her own modules for different situations.

The reason that modules work so well in practice is that given a specific situation, there are often only a very small number of possible schedules. In addition, constructing the modules as outlined below means that the whole system works like building blocks for the complete schedule.

Consider situations such as nurse scheduling. The following is a set of basic requirements that apply in many hospitals: employees are to have one weekend off in three weeks; the maximum work stretch is six days; shift rotations to a shift that starts at an earlier time than the current one are to be done only after at least one day off; weekends off cannot start after an evening or night shift on Friday, where the night shift is the last shift of the day; employees are to work at least 50% of their time on day shift; there must be no single-day work stretches; there must be at least two days off after working a stretch of night shifts. The last restriction is necessary because if someone worked the night shift on a Tuesday, which ended, say, at 7:30 Wednesday morning, then had Wednesday off during which he or she spent most of the day sleeping, and then returned to work on the next day at 7:30 a.m. on Thursday, he or she would not really have had a very useful day off. As shown in Section 4.1, having all adjacent days off increases the number of nurses needed and reduces the possible schedules to two, which is not much choice.

The first example uses all the basic requirements, plus allowing only one single day off in the three-week cycle. The expectation is that there should be more feasible schedules than when the restriction is to have all paired off-days, but in fact there are only two possible ways of giving the days off. Table 22 shows the two basic schedules, and Table 23 shows the four possible schedules for the day (d) and evening (e) shift. Since there are only two possible ways of giving the days off, Table 22 also shows that it is not possible to give anyone a Wednesday off. This fact should be kept in mind when negotiating or promising schedules. Note that if module 1 is followed by module 2, it should be possible to have two people working on each shift every day of the week, with the extra two people on Wednesday as expected. Table 24 gives a six-week cycle with the shift assignments that give the even coverage for the two shifts.

TABLE 22 Basic Eight-Hour Shift Modules with One Single Day Off in Three Weeks

Basic Module 1							Basic Module 2						
S	M	T	W	T	F	S	S	M	T	W	T	F	S
X	X						X				X		
	X	X									X	X	
		X				X						X	X
2	1	1	3	3	3	2	2	3	3	3	1	1	2

TABLE 23 Day/Evening Modules, e Hour Shifts, One Single Day Off

d/e Module 1

	S	M	T	W	T	F	S
1	X	X	e	e	e	e	e
2	e	X	X	d	d	d	d
3	d	d	X	d	d	d	X
e	1	0	1	1	1	1	1
d	1	1	0	2	2	2	1

d/e Module 2

	S	M	T	W	T	F	S
1	X	X	d	d	d	d	d
2	d	X	X	e	e	e	e
3	e	e	X	d	d	d	X
e	1	1	0	1	1	1	1
d	1	0	1	2	2	2	1

d/e Module 3

	S	M	T	W	T	F	S
1	X	d	d	d	X	e	e
2	e	e	e	3	X	X	d
3	d	d	d	d	d	X	X
e	1	1	1	1	0	1	1
d	1	2	2	2	2	0	1

d/e Module 4

	S	M	T	W	T	F	S
1	X	d	d	d	X	d	d
2	d	d	d	d	X	X	e
3	e	e	e	e	e	X	X
e	1	1	1	1	1	0	1
d	1	2	2	2	0	1	1

For the modules with day and evening shifts, d/e modules 1, 3, and 4 can be converted to the d/n modules 1, 3, and 4, with the evening shift replaced with a night shift. Because of the one day off at a shift change, d/e module 2 cannot be converted to a d/n module. When a schedule is made, the different building blocks can be used. If the nursing unit required one nurse on night shift, and two on evening shift, with all nurses working all three shifts, then module d/e 5 would be followed by a d/n module. The nine nurses required would each start on a different line of the schedule and rotate in the usual way. Even if there were only seven or eight nurses available, the same schedule would be used because there is no other feasible schedule, and in this case, as the schedule was worked from week to week, part-time nurses would need to be used to fill in the missing shifts. Because all modules start with a Sunday off and end with a Saturday off, the modules can be put together in any order without violating the scheduling criteria. If there are more people required for the shifts, more modules will be used. If there are people who only work days and nights and others who only work days and evenings, the modules can be separated into two different master rotations for the unit.

A different module library can be built for the same basic requirements but allowing two single days off in a three-week schedule. The technique is the same as before, with the modules beginning with the Sunday off and ending with the Saturday off. In between, the four remaining off-days are placed as one adjacent pair of days off and two single days off. Each day off should be on a different day from the others to give even worker coverage, and no one-day work stretches should be allowed. As with building the chains in the iterative method, the last day of the adjacent pair off and one of the single days off should be in different weeks but adjacent days so that a chain of evening shifts is formed. There are only four basic modules that lead to only a few d/e and d/n modules for this case. A schedule can be built using modules from both the one-single-day-off library and the two-single-days-off library to give different schedules.

TABLE 24 Day/Evening Module for a Six-Week Schedule

d/e Module 5

	S	M	T	W	T	F	S
1	X	e	e	e	e	e	e
2	e	X	X	d	e	e	e
3	e	e	X	d	d	d	X
4	X	e	e	e	X	d	d
5	d	d	d	d	X	X	d
6	d	d	d	d	d	X	X
e	2	2	2	2	2	2	2
d	2	2	2	4	2	2	2

TABLE 25 Twelve-Hour Modules

D/N Module 1							D/N Module 2							D/N Module 3						
S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S
X			X	X			X	X			X	X		X			X			X
	X	X			X	X		X	X				X	X	X	X				X

The iterative schedules given before can be easily computerized. The modular scheduling is much more complicated and involves using a rule-based expert system in the computer program, but the modular system is easy to do manually.

4.4. Multiple Mixed Shift Length Scheduling For Individuals

In the 1990s there were several papers published that dealt with four day work weeks, and work weeks that had either three or four days. As is usual with research, the first set of papers gained some ground with the problem but were not of much practical use. The papers by Hung (1991, 1993, 1994a,b), ignored the length of work stretch and often had work stretches much longer than six days. In addition, some of the work by Hung applies only when the frequency of weekends off is greater than or equal to one out of two, which means that there are no workers of type $T4$ requiring pairs of adjacent days off. Burns et al. (1998) give an algorithm for the single-shift problem that restricts the work stretch to be at most five days and uses a mix of three- and four-day weeks. Burns and Narasimhan (1999) give an iterative method for optimally solving the multiple-shift scheduling problem with a restricted length of work stretch and with either four- or three-day work weeks. The algorithm is easy to understand and implement.

In this work, we will address the much more usual case of scheduling 12-hour shifts. In Burns (1985) there are many different modules given for the 12-hour scheduling problem. Only a few are needed to illustrate both the method of scheduling and the difficulties that can arise. The first difficulty is how to balance the number of hours worked. As explained in the crew section earlier, a common way of balancing the hours is to have one 8-hour shift every two weeks. Table 25 illustrates two basic modules giving every second weekend off, having a maximum work stretch of three days, having no single days off unless in a module having two weekends in a row off, where a single day off is necessary, and having a minimum work stretch of two days. The 8-hour shift replaces one of the 12-hour shifts in the module. By using two modules, with an 8-hour day shift in one and an 8-hour night shift in the other, it is possible to hire a part-time worker for an eight-hour evening shift. Many hospitals use this type of schedule for nurse schedules.

For D/N module 1, any shift can be the 12-hour N, but for D/N module 2, the Friday of the second week cannot be a N shift and still have a “good” weekend off. For this reason, many places only use the D/N module 1. Depending on the preferences of the employees, the schedule can have two weeks of nights whenever needed, followed by two or more weeks of days or the day and night shifts can be interspersed within the module.

There is an inherent problem if only module 1 is used to create a schedule. The employees working the schedule are essentially divided into two subgroups: the first working the odd-numbered weeks in the schedule and the second working the even-numbered weeks. These two groups will never work together, as they are never scheduled to work on the same day, let alone the same shift. To avoid this separation of the work force, introducing some of module 2, (or some of module 3 if needed) into the schedule will ensure that the groups see each other.

If fewer people are needed on the weekends than on the weekdays, then as many of the D/N module 3 as are required should be used.

Example 4:

The following are the requirements for workers: two people on night shifts and four people on day shift on weekdays, and two on night shift and two on day shift on the weekends. Because six people are needed each weekday, six modules will be needed. Choosing D/N1 with all the shifts D followed by D/N 1 with all the shifts N, followed by D/N module 3 on days, gives the complete schedule when two people are started on each line.

Carefully placing the 8-hour shifts allows the schedule to be created as in Table 26, with no part-time workers required. Two people are started on each line of the schedule, and each person works exactly 80 hours in every two-week period.

A second type of module can be used where no 8-hour shifts are used. In this type, an extra 12-hour shift is given off every six weeks, rather than 4 hours off every two weeks. Any of the modules can be used to do this, as long as the 12-hour shift given off is at the start or end of a three-day

TABLE 26 Example 4, Schedule 1

S	M	T	W	T	F	S
X	D	D	X	X	d	D
D	X	X	D	D	X	X
X	N	N	X	X	n	N
N	X	X	N	N	X	X
X	D	D	X	D	e	X
X	X	X	D	D	D	X

work stretch. For module 3 above, the single day off would be moved to Thursday and the Friday given as the extra day off, thus creating an off-day stretch of six days. Table 27 gives the schedule for example 4, using only 12-hour shifts. In practice, the hours worked balance over a six-week period but not over the two-week pay periods. Some agreement on overtime rules needs to be negotiated before this approach can be used.

4.5. Hierarchical Workforce Scheduling

Consider the situation where there are m different categories of workers with a definite ranking. Type i workers can do the work of any worker at a lower level j but the reverse is not true. It can be assumed that workers of type i are paid at least as much as, and probably more than, workers of type j for $i > j$. There is a demand for a minimum number of each type of worker on each shift and a cumulative demand for the number of workers of type 1 to type k for each shift. In the hierarchical problem, the requirement is to calculate the minimum number of workers in the most economical mix of types and to give each worker two days off per week and A out of B weekends off. A schedule is to be created with a maximum work stretch of six days for all values of A and B. Workers who must work two consecutive weekends should have two adjacent days off in the week.

Emmons and Burns (1991) solved the case for a single shift and a constant demand for each day of the week. Other papers were published trying extend the work, but they either did not adhere to the six-day maximum work stretch or only worked with the easier case where A of B was greater than or equal to one out of two.

Narasimhan (1993) solved several very interesting problems. He extended the work of Emmons and Burns (1987) to the case where the demand for each category of worker is constant for the weekdays and has a different constant value for the weekend. His work did not require a restriction that the weekday demand be greater than or equal to the weekend demand, and there was no restriction of the frequency of weekends off.

In the same work, Narasimhan also solved the multiple 8-hour, hierarchical workers shift problem for any number of shifts, with the weekday demand greater than or equal to the weekend demand, and he also solved the 10-hour, four-day-a-week problem for the same labor demands.

The algorithms and calculations are complicated but can be computerized. They will not be presented here because they would take too much space and because in practice they are often not applicable. For example, in the health care industry there are different categories of employees that can provide different patient needs. These categories, and the number of workers of each type required per shift, usually fall into a hierarchy that would fit the requirements of this section. However, in many cases the workers belong to different unions and have different scheduling rules in the hours of work clauses of their contracts. Since a single schedule, using one set of rules, could not be used in these instances, separate schedules are made for each category of worker. In some situations, the schedules are made sequentially, with the remaining needs for employees in the next category re-

TABLE 27 Example 4, Schedule 2

S	M	T	W	T	F	S
X	D	D	X	X	D	D
D	X	X	D	D	X	X
X	N	N	X	X	N	N
N	X	X	N	N	X	X
X	D	D	D	X	X	X
X	X	X	D	D	D	X

flecting the schedules already constructed. Generally, the schedules are done independently, using one of the methods described in this work for the single category of workers.

5. COMPUTER SYSTEMS

Despite the all-pervasive use of computers in society today, computers are not used frequently for shift-scheduling methods. There are notable exceptions in airline crew scheduling and some other industries. However, many health care companies and most retail companies still do manual scheduling. One of the reasons for this situation is the separation of institutional functions by the use of different third-party software packages, each custom-built for one function only. The merging of these companies to integrate such systems has not occurred as it has in the case of personnel records, payroll, and shift scheduling. All three systems must be integrated in a major company to have the shift scheduling work well.

5.1. Personnel Interface

Personnel systems have much of the key information about employees that is needed by a shift-scheduling computer package. The unique identifier to be used, such as a Social Security number or an employee number, is just one such item. Other needed information is a person's seniority and qualifications. Employees often take courses to upgrade their qualifications and are then eligible for a higher-paying position, and this information must be transmitted to the scheduling system on a timely basis. If someone leaves a scheduling group of employees and a new employee replaces that person, the seniority of members in the group changes and hence the schedules may need to be changed. For example, in a hospital, a change in staff may mean that a nurse may be able to move from a schedule of straight night shifts to a day shift schedule.

In industries such as pulp and paper, employees are offered chances to fill in for a higher-paying position whenever an open shift becomes available. If the person refuses the position more than once, he or she is no longer eligible for the position. The status of each employee regarding each position is required by the scheduling system.

5.2. Contract Regulations

When contracts are being negotiated, analysis using the methods of this chapter is very valuable. Calculating the required workforce size for increases in the weekend-off frequency and for any decrease in the work stretch can be done along with the related costs. If changes, such as having all pairs of days off with a work stretch maximum of six days, are contemplated, the increase in cost can be calculated and the paucity of possible scheduling modules can be illustrated.

The bounds given by Burns and Carter (1985) give the workforce size as the maximum of the bound on the total shifts required and the weekend-off bound. If a contract is being negotiated that will increase the frequency of weekends off and the new bound for this increase is higher than the bound on the total shifts required in the old contract, extra employees will be working on the weekdays. What may happen is that the employer will reduce the number of people required on the weekend until the two bounds are equal and then will hire part-time workers for the weekends. Each time the weekend requirements are reduced, the total shift requirements for full-time employees are reduced by two shifts. When the adjustments are done to balance the two bounds, the result may mean that fewer full-time employees are required under the proposed contract changes and more part-time employees will be used, which may not be what the employees want or expect.

5.3. Payroll Interface

Any computer scheduling package needs an extensive database to keep the necessary information to transmit to the payroll program. One copy of the schedule should be posted the required number of weeks in advance, usually six weeks. The employees can then trade shifts with other employees and request the scheduler to allow these changes. If the requested changes violate the contract, they may still be granted, but extra premiums, such as overtime, will not apply. The schedule, as modified with requested and granted changes, will then be saved. Up until a fixed time, say two weeks before the schedule is actually worked, management can make some types of changes without incurring a payable penalty. The final schedule, as it is to be worked, is then saved. The actual schedule worked will be saved, along with any last-minute changes made by management to satisfy unexpected needs or absenteeism. The information transferred to the payroll package will come from this final schedule, with the pay premiums calculated based on when, and by whom, the changes were made. All the information must be archived for later reference.

6. MANUAL SYSTEMS

Manual shift scheduling systems are widely used and for the most part are time consuming and frustrating. No matter how good the schedule, some employees will express discontent. This basically arises from the fact that all employees want to be home with family and friends at the same time and from the fact that different people have different ideas as to what makes a good schedule. Since many of the people creating the schedules have other tasks to do, the scheduling often gets done at home, on personal time, which is not a good situation for creating good schedules. Computer scheduling would free up the time of the person doing the scheduling and would help divert any unavoidable discontent of personnel to a machine and away from the scheduler.

An interesting result of linking the payroll and the scheduling system arose when Burns first computerized nurse scheduling in the 1980s. The head nurses were able to receive reports on the total hours worked by their full-time and part-time employees. Several nurses were discovered to be working seven days a week by doing part-time shifts in different units. This was deemed to be an unsafe work situation and changes were made to put a stop to the practice. Other work patterns can be detected once the information is put into a general payroll system rather than in a unit-by-unit shift-scheduling system.

Manual systems have a manual entry of data from the schedules worked into the payroll system. Because of the large amount of information that needs to be processed and transferred, as outlined in Section 5.2, errors are inevitable. In addition, extra people are employed just to do the data entry, an expense that can be eliminated with a computer system that is integrated with the payroll system.

For companies that will continue to use a manual system, several things can be done to improve the situation. The people doing the scheduling can be educated and trained to understand the complications of scheduling. The methods outlined in this chapter can be introduced to improve the schedules and decrease the time required to create the schedules.

REFERENCES

Baker, K. R. (1974), “Scheduling a Fulltime Workforce to Meet Cyclic Staff Requirements,” *Management Science,* Vol. 20, pp. 1561–1568.

Baker, K. R. (1976), “Workforce Allocation in Cyclical Scheduling Problems: A Survey,” *Operations Research Quarterly,* Vol. 27, pp. 156–157.

Brownell, W. D., and Lowerre, J. M. (1976), “Scheduling of Work Forces Required in Continuous Operations under Alternate Labor Policies,” *Management Science*, Vol. 22, pp. 597–605.

Burns, R. N. (1978), “Manpower Scheduling with Variable Demands and Alternate Weekends Off,” *INFOR*, Vol. 16, pp. 101–111.

Burns, R. N. (1981), “An Iterative Approach to Multiple Shift Scheduling,” Manuscript, School of Business, Queen’s University, Kingston, ON.

Burns, R. N. (1983), “Shift Scheduling at Denison Mines,” BCW Consulting Ltd., Kingston, ON.

Burns, R. N. (1985), “Shift Scheduling Modules,” BCW Consulting Ltd., Kingston, ON.

Burns, R. N. (1999), “Algorithms for Time Dependent Schedules,” BCW Consulting Ltd., Kingston, ON.

Burns, R. N., and Carter, M. W. (1985), “Work Force Size and Schedules with Variable Demands,” *Management Science,* Vol. 31, pp. 599–607.

Burns, R. N., and Koop, G. J. (1987), “A Modular Approach to Optimal Multiple Manpower Scheduling,” *Operations Research,* Vol. 35, No. 1, pp. 100–110.

Burns, R. N., and Narasimhan, R. (1999), “Multiple Shift Scheduling of Workforce on Four-Day Workweeks,” *Journal of Operational Research Society,* Vol. 50, pp. 979–981

Burns, R. N., Narasimhan., R., and Smith, L. D. (1998), “A Set-Processing Algorithm for Scheduling Staff on 4-Day of 3-Day Work Weeks,” *Naval Research Logistics,* Vol. 45, pp. 839–853.

Emmons, H., and Burns, R. N. (1991), “Off-Day Scheduling with Hierarchical Worker Categories,” *Operations Research,* Vol. 39, No. 3, pp. 484–495.

Glover, F., and McMillan, C. (1986), “The General Employee Scheduling Problem: an Integration of MS and AI,” *Computer and Operations Research,* Vol. 13, pp. 563–573.

Henderson, W. B., and Berry, W. L. (1976), “Heuristic Methods for Telephone Operator Shift Scheduling: An Experimental Analysis,” *Management Science,* Vol. 22, pp. 1372–1380.

Hung, R. (1991), “Single Shift Workforce Scheduling under a Compressed Workweek,” *OMEGA,* Vol. 19, pp. 494–497.

Hung, R. (1993), “Three-Day Workweek Multiple-Shift Scheduling Model,” *Journal of Operational Research Society,* Vol. 44, pp. 141–146.

Hung, R. (1994a), “Multiple-Shift Workforce Scheduling under the 3–4 Day Workweek,” *Management Science,* Vol. 40, No. 2, pp. 280–284.

Hung, R. (1994b), “A Multiple-Shift Workforce Scheduling Model under the 4-Day Workweek with Weekday and Weekend Labor Demand,” *Journal of the Operational Research Society,* Vol. 45, No. 9, pp. 1088–1092.

Narasimhan, R. (1993), “Optimal Workforce Shift Scheduling,” Ph. D. thesis, School of Business, Queen’s University, Kingston, ON.

Segal, M. (1974), “The Operator Scheduling Problem: Network Flow Approach,” *Operations Research,* Vol. 22, pp. 808–823.

Selkirk, C. G. (1999), “Optimal Algorithms for Single Shift Workforce Scheduling to Avoid One Day Work Stretches in a Cyclic Schedule,” Ph. D. thesis, School of Business, Queen’s University, Kingston, ON.

Tiberwala, R., Phillippe, D., and Browne, J. (1972), “Optimal Scheduling of Two Consecutive Idle Periods,” *Management Science,* Vol. 19, pp. 71–75.

CHAPTER 65

Monitoring and Controlling Operations

ALBERT JONES
National Institute of Standards and Technology

YUEHWERN YIH
Purdue University

EVAN WALLACE
National Institute of Standards and Technology

1. INTRODUCTION 1768

2. CONTROL ARCHITECTURES 1769
  - 2.1. The Purdue Enterprise Reference Architecture (PERA) 1769
    - 2.1.1. Control Hierarchy 1769
    - 2.1.2. Equipment Organization 1771
    - 2.1.3. Status 1772
  - 2.2. SEMATECH CIM Framework 1772
    - 2.2.1. CIM Framework Component Architecture 1773
    - 2.2.2. Component-Specification Methodology 1774
    - 2.2.3. Shop-Floor Application Modules 1774
    - 2.2.4. Status 1775

3. AI APPROACHES TO SHOP-FLOOR SCHEDULING AND CONTROL 1775
  - 3.1. Knowledge-Based Systems 1775
    - 3.1.1. Generating the Required Knowledge Base 1775
    - 3.1.2. Applications to Scheduling and Control 1776
    - 3.1.3. Agents to the Rescue? 1777
  - 3.2. Artificial Neural Networks 1777
    - 3.2.1. Hopfield Networks 1778
    - 3.2.2. Supervised-Learning Neural Networks 1778
    - 3.2.3. Multilayer Perceptrons 1779
    - 3.2.4. Unsupervised Neural Networks (Competition Based) 1779
    - 3.2.5. Reinforcement Learning 1780
  - 3.3. Genetic Algorithms 1780
  - 3.4. Fuzzy Logic 1781
  - 3.5. Commercial Systems 1782

4. MANUFACTURING EXECUTION SYSTEMS (MES) 1782
  - 4.1. A More Detailed Look at MES Data 1782
  - 4.2. MES Object Models 1783
  - 4.3. Market Trends and Future Directions 1787

5. SUMMARY 1787

REFERENCES 1787

1. INTRODUCTION

On the surface, companies from industry sectors such as mining, construction, manufacturing, and service would appear to be very different. Certainly, the physical activities involved and the resulting products are different. At a conceptual level, however, each of these companies can be viewed as a

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

complex system trying to utilize its resources to maximize its performance. The ability of managers and workers to control and monitor the operations within such a system has a great impact on its performance.

This chapter addresses three issues related to controlling and monitoring operations in such a system: architectures to organize those operations; artificial intelligence techniques for scheduling those operations; and commercial software to implement monitoring and control. The principal application focus of this chapter is manufacturing, but the ideas can be applied to a wide range of complex systems.

2. CONTROL ARCHITECTURES

Decisions, decisions, decisions—factory managers make them and factory workers implement them everyday. Some decisions impact events immediately; others impact events months or years into the future. Industry, academia, government agencies, and standards bodies have expended considerable effort to develop architectures that (1) organize and integrate these decisions in some meaningful way and (2) specify the information required to make those decisions monitor their execution, and control their implementation. This section describes two such architectures.

2.1. The Purdue Enterprise Reference Architecture (PERA)

A tree structure, hierarchy, is one of the most common ways of organizing functions and activities. Many such hierarchies have been proposed for decision-making and control within manufacturing systems (Jones 1990). The Purdue Enterprise Reference Architecture (PERA), which was developed by a collection of industrial and academic representatives, is one such organization of the factory that includes the control functions and information requirements (Williams 1992). Originally aimed at the process industry, it has been developed so that it can be used across all types of manufacturing. The material in the following sections is taken from Annex D and Section 5.1 of ANSI/ISA-S95.00.01-2000, *Enterprise-Control System Integration Part 1: Models and Terminology* (ANSI/ISA 2000).

2.1.1. Control Hierarchy

Figure 1 shows three levels of the PERA functional hierarchy model at which decisions are made: business planning and logistics, manufacturing operation, and control. Level 4 and level 3 deal with plant production scheduling, operation management, and plant floor coordination. Levels 2, 1, and 0 decompose control functions for three types of manufacturing: batch, continuous, and discrete. This decomposition defines the cell or line supervision functions, operations functions, and process control functions. There are several different execution methods for these functions, which are based on the actual production strategy used.

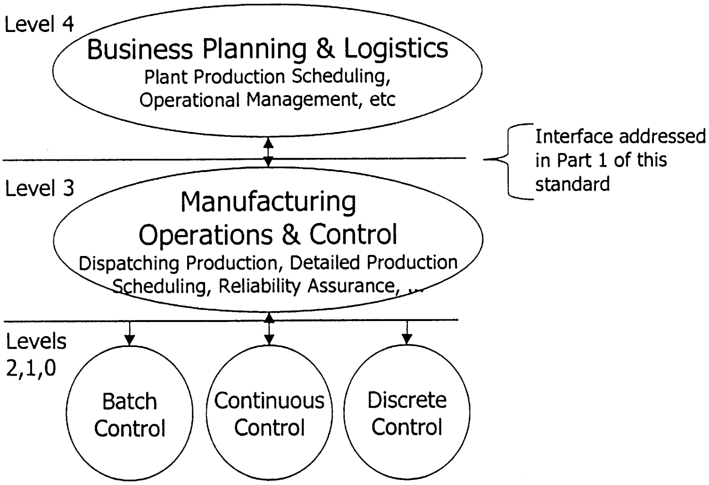

Level 4
Business Planning & Logistics
Plant Production Scheduling, Operational Management, etc
Interface addressed in Part 1 of this standard
Level 3
Manufacturing Operations & Control
Dispatching Production, Detailed Production Scheduling, Reliability Assurance, ...
Levels 2,1,0
Batch Control
Continuous Control
Discrete Control

Figure 1 Decision-Making and Control Hierarchy.

Level 4 performs the following major functions: capacity planning, plant production scheduling, materials requirements planning, and all manufacturing-related purchasing. It establishes and modifies the basic plant production schedule for orders received, based on resource availability changes, energy sources available, power demand levels, and maintenance requirements. It develops preventive maintenance and equipment renovation schedules in coordination with the basic production schedule. It determines the optimum inventory levels of raw materials, energy sources, spare parts, and in-process goods. Finally, it collects, maintains, and provides data on a large number of items. These items include raw material and spare parts usage, overall energy use, goods in process, inventory, control files as they relate to customer requirements, machinery and equipment utilization and history files, and manpower use data for transmittal to personnel and accounting.

Level 3 is the principal level of interest in this chapter. Therefore, we provide detailed descriptions of the major activities performed at this level.

- *Resource allocation and control:* Manage those resources directly associated with control and manufacturing. These resources include machines, tools, labor skills, materials, and other equipment, documents, and entities that must be available for work to start and be completed. The management of these resources may include local resource reservation to meet production-scheduling objectives, the assurance that equipment is properly set up for processing, the responsibility for providing real-time, resource status and a history of resource use.
- *Dispatching:* Manage the flow of production in the form of jobs, orders, batches, lots, and work orders by dispatching production to specific equipment and personnel. The flow is governed by the sequence of operations, which determines the order the work is done, and the time that work starts and stops. It is possible to change the sequence or times in real time as events occur on the factory floor; however, those changes are made within agreed upon limits, based on local availability and current conditions. Dispatching of production includes the ability to control the amount of work in process through buffer management and management of rework and salvage processes.
- *Data collection and acquisition:* Manage the operational production and parametric data that is associated with the production equipment and production processes, provide real-time status of the equipment and processes, and keep a history of production and parametric data.
- *Quality management:* Provide real-time measurements collected from manufacturing and analysis in order to ensure proper product quality control and identify problems requiring attention. This includes SPC/SQC tracking and management of off-line inspection operations and analysis in laboratory information management system (LIMS). This activity may recommend actions to correct the problem, including correlating the symptoms, actions, and results to determine the cause.
- *Process management:* Monitor production and provide decision support to operators who correct and improve in-process functions. These functions may be intraoperational (focusing specifically on machines or equipment being monitored and controlled) or inter-operational (tracking the process from one operation to the next). It may include alarm management to alert factory personnel of process changes that are outside of acceptable tolerances.
- *Production planning and tracking:* Provide the status of production and the disposition of work. Status information may include personnel assigned to the work, component materials used in production, current production conditions, and any alarms, rework, or other exceptions related to the product.
- *Performance analysis:* Provide up-to-the-minute reporting of actual manufacturing operations results, along with comparisons to past history and expected results. Performance results include such measurements as resource utilization, resource availability, product unit cycle time, conformance to schedule, and performance to standards. Performance analysis may include SPC/SQC analysis, and may draw from information gathered by different control functions that measure operating parameters.
- *Operations and detailed scheduling:* Generate sequences that optimize some objective (such as minimize set-up time) based on priorities, attributes, characteristics, and production rules associated with specific production equipment and specific product characteristics. This activity is carried out using the current estimate of unused capacity and recognizing alternative and overlapping/parallel operations.
- *Document control:* Control records and forms that must be maintained with the production unit. The records and forms include work instructions, recipes, drawings, standard operation procedures, part programs, batch records, engineering change notices, shift-to-shift communication, as well as the ability to edit “as-planned” and “as-built” information. This activity is responsible for providing data to operators and recipes to device controls, and for maintaining the integrity of regulatory, environmental, health and safety regulations, and SOP information such as corrective action procedures.

- *Labor management:* Provide status of personnel in real time. This activity includes time and attendance reporting, certification tracking, as well as the ability to track indirect functions such as material preparation or tool room work as a basis for activity-based costing. It may interact with resource allocation to determine optimal personnel assignments.
- *Maintenance management:* Maintain equipment and tools. This activity ensures the availability of equipment and tools for manufacturing, and manages a history of past events or problems to aid in diagnosing problems.

2.1.2. Equipment Organization

The hierarchy described above deals with the decision making, control, and information needed to manage the physical assets of a manufacturing enterprise. Those assets are usually organized in a tree structure such as the one described in Figure 2. Lower-level groupings are combined to form higher-level entities. In some cases, a grouping within one level may be incorporated into another grouping at that same level. In the following section, we define the areas of responsibility for the different levels defined in the hierarchical model and some of the objects used in the information exchanged within and across those levels.

- *Enterprise:* A collection of one or more sites. The enterprise is responsible for determining the products to be manufactured, where they will be manufactured, and in general how they will be manufactured. Level 4 functions are generally dealing at the enterprise and site levels. However, enterprise planning and scheduling may involve areas, cells, lines, or units within an area.
- *Site:* A physical, geographical or logical grouping determined by the enterprise. It may contain areas, production lines, process cells, and production units that have well defined manufacturing capabilities. The level 4 functions at a site include local site management and optimization. Sites are often used for rough-cut planning and scheduling, which may involve cells, lines, or units within the areas.
- *Area:* A physical, geographical, or logical grouping, which may contain process cells, production units, and production lines. The main production capability and geographical location within a site usually identify areas. Areas generally have well defined manufacturing capabilities and capacities that are used for planning and scheduling at levels 3 and 4. An area is made up of lower-level elements that perform continuous manufacturing operations, discrete (repetitive and nonrepetitive) manufacturing operations, and batch manufacturing operations. An area may have several of these elements in varying combinations, depending upon the manufacturing requirements. For example, a beverage manufacturer may have an area with continuous mixing equipment that feeds a batch process cell for batch processing that feeds a bottling line for discrete bottling process. Depending on the planning and scheduling strategy selected, the level 4 func-

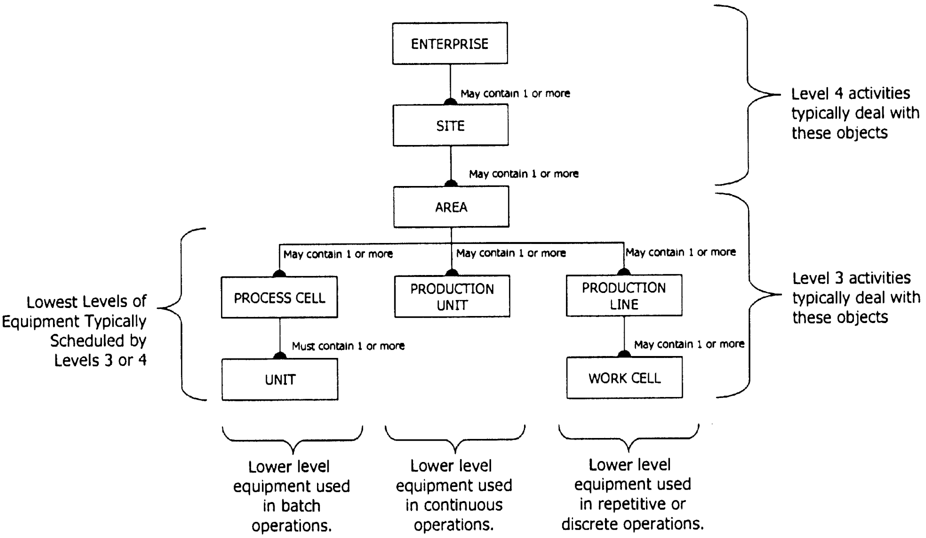

ENTERPRISE
May contain 1 or more
SITE
May contain 1 or more
AREA
May contain 1 or more
May contain 1 or more
May contain 1 or more
PROCESS CELL
PRODUCTION UNIT
PRODUCTION LINE
Must contain 1 or more
May contain 1 or more
UNIT
WORK CELL
Level 4 activities typically deal with these objects
Level 3 activities typically deal with these objects
Lowest Levels of Equipment Typically Scheduled by Levels 3 or 4
Lower level equipment used in batch operations.
Lower level equipment used in continuous operations.
Lower level equipment used in repetitive or discrete operations.

Figure 2 Typical Equipment Organization.

tions may stop at the area level, or they may schedule the functions of the lower-level elements within the areas.

- *Production units:* The lowest level of equipment typically scheduled by the level 4 or level 3 functions for continuous manufacturing processes. Production units are composed of lower level elements, such as equipment modules, sensors, and actuators. Production units have well-defined processing capabilities, and throughput capacities, and these are used for level 3 functions. The capacities and capabilities are also often used as input to level 4 scheduling, even if the production units are not scheduled by the level 4 functions.
- *Production lines and work cells:* The lowest levels of equipment typically scheduled by the level 4 or level 3 functions for discrete manufacturing processes. Work cells are usually only identified when there is flexibility in the routing of work within a production line. Production lines and work cells may be composed of lower level elements. Production line and work cells have well defined manufacturing capabilities and throughput capacities and these are used for level 3 functions. The capacities and capabilities are also often used as input to level 4 scheduling, even if the production lines and work cells are not scheduled by the level 4 functions.
- *Process cells and units:* The lowest level of equipment typically scheduled by the level 4 and level 3 functions for batch manufacturing processes. Units are usually only identified at levels 3 and 4 if there is flexibility in the routing of product within a process cell. The definitions for process cells and units are contained in the IEC 61512 and ISA S88.01 standards. Process cells and units have well-defined manufacturing capabilities and batch capacities, and these are used for level 3 functions. The capacities and capabilities may also be used as input data for level 4 scheduling, even if the process cells or units are not scheduled by the level 4 functions.

2.1.3. Status

The PERA plays a critical role in two standards, one being developed by SP95 of the ISA (Instrument Society of America) and the other by TC 184/WG 1 of the ISO (International Standards Organization). SP95 seeks to create standards for the interfaces between control functions and other enterprise functions (http://www.isa.org/sc/committee/1,1512,145,00.html) based upon the PERA. The interface initially considered is the interface between levels 3 and 4 of that model. Additional interfaces will be considered, as appropriate. WG 1 has published the PERA as an annex (http://www.nist.gov/sc5wg1/gera-std/15704fds.htm) to ISO 15704, *Requirements for Enterprise Reference Architectures and Methodologies,* which is a final draft international standard (http://www.nist.gov/sc5wg1/gera-std/15704AB.htm). From the WG1 perspective, PERA is an example of a generalized enterprise reference architecture, GERAM. A GERAM defines a toolkit of concepts for designing and maintaining enterprises for their entire life history and is meant to organize existing applications in all types of enterprises. An advantage is that previously published reference architectures can keep their own identity while identifying through GERAM their overlaps and complementing benefits compared to others.

2.2. SEMATECH CIM Framework

The CIM Framework, developed by SEMATECH, defines a standard component architecture and application component interfaces for manufacturing information and execution systems (MIES) software (Doscher 1998)—the following material is taken largely from (Hawker 1999). The CIM Framework is not hierarchical in nature; rather, it leverages distributed, object-oriented computing technology. Additionally, it uses middle-ware standards from the Object Management Group (OMG) (http://www.omg.org) to enable integration of the applications.

The CIM Framework software architecture was designed to enable the following capabilities:

- *Integration:* Applications can cooperate by exchanging data, providing services (client/server method invocation), publishing service exceptions, and publishing and subscribing to events.
- *Interoperability:* Applications from one supplier or source can be replaced easily with a functionally equivalent application (conformant to standard interface and behavior) from another source.
- *Flexibility:* Components and applications can be configured in a variety of ways that meet specific needs.
- *Reuse:* New systems can be implemented from standard components or applications more quickly, at lower cost, and with higher quality.

The major benefit of this framework is a significant reduction in the cost and time involved in building, modifying, and enhancing MIES software in response to changing business needs. Adherence to the framework allows semiconductor manufacturers to integrate applications from multiple suppliers with their legacy systems and to replace or upgrade these applications and systems over time.

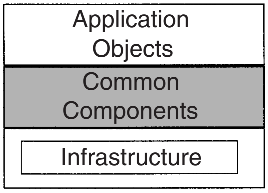

Application Objects
Common Components
Infrastructure

Figure 3 CIM Framework Architecture Layers.

2.2.1. *CIM Framework Component Architecture*

The CIM Framework architecture is a layered system that enables distributed, object-oriented applications assembled from common software components to interoperate as a single, integrated system.

The integration infrastructure is based on specifications in the Object Management Architecture (OMA) from the OMG (OMG 1995, 1996, 1997a). These specifications define standard services for distributed object communications, persistence, transactions, name services, and so forth. On top of the infrastructure, the CIM Framework architecture defines common application components. Components are software building blocks that implement some collection of functions. Typically, each MIES application will be composed of many such components. The CIM Framework common components layer defines standard models for application components that are common across MIES applications. Examples include a machine management component, a product management component, and a person management component. The machine management component includes machine resources, sensors, process capabilities, and relations to material and recipes. The product management component includes product material, lots, and relations to product and process specifications. The person management component includes persons, qualifications, and relations to skills and skill requirements. This common application model, defined in terms of common software components, is the *framework* for building integrated MIES applications.

The application objects layer of the CIM Framework architecture provides additional functionality, extending the common components to make a complete MIES. This layer, which is identified but not specified, enables MIES suppliers and users to define product-specific and site-specific application objects and components that use and extend the CIM Framework common components to implement MIES functions that meet business needs.

A given MIES application, as shown in Figure 4, implements some common components and application objects and interoperates (via the infrastructure) with other common components and

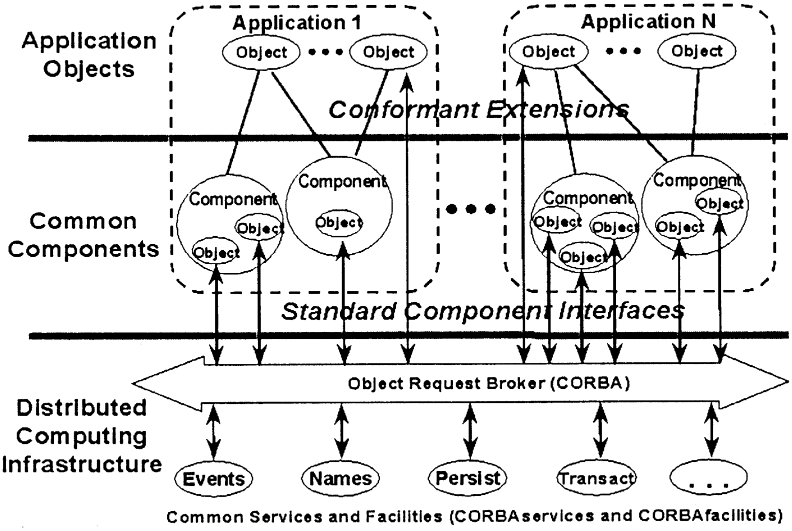

Application Objects
Application 1
Object
• • •
Object
Application N
Object
• • •
Object
Conformant Extensions
Common Components
Component
Object
Object
Component
Object
• • •
Component
Object
Object
Object
Component
Object
Object
Standard Component Interfaces
Object Request Broker (CORBA)
Distributed Computing Infrastructure
Events
Names
Persist
Transact
. . .
Common Services and Facilities (CORBAservices and CORBAfacilities)

Figure 4 CIM Framework Component Architecture.

application objects implemented in other MIES applications. The collection of interoperating MIES applications provides a complete, integrated, MIES solution.

2.2.2. *Component-Specification Methodology*

The key to the CIM Framework is the specification of components. The CIM Framework uses the following modeling methods to specify components:

- Component relationship models showing interaction between "medium-grained" components (larger than an object, smaller than an application)
- Component information models showing object interfaces and relationships in the form of OMT (object modeling technique) diagrams (Rumbaugh et al. 1991)
- Object interface definitions using OMGs Interface Definition Language (IDL)
- Published and subscribed events using an extension to OMG IDL
- Component interaction diagrams showing scenarios that trace messages and events between components
- State transition diagrams as Harel state charts (Harel 1987) and state definition tables

These modeling methods go far toward specifying components that MIES implementers can "plug-and-play" into integrated systems. SEMATECH is also working in the OMG Business Objects Domain Task Force to define and standardize additional methods for even richer semantic models, including the specification of method preconditions and postconditions, roles, rules, and dependencies (http://www.omg.org/homepages/bodtf/).

2.2.3. *Shop-Floor Application Modules*

The CIM Framework specifies application components for manufacturing information and execution systems (MIES). MIES perform factory operations functions in the context of enterprise information and control systems and systems that automate material processing, storage and movement. Figure 5 shows the MIES functional groups in the CIM Framework.

Each functional group defines a collection of related application components. Table 1 lists the CIM Framework components in each functional group. The functional groups are a convenient mechanism to organize the CIM Framework components; they are not rigid partitions, and suppliers can deliver applications that span functional groups or that implement only some of the components of a group. In contrast, the component is the smallest-grained entity that suppliers can deliver. A supplier must implement all the interfaces and behaviors of a component in order to claim conformance to that component specification.

The value and power of the CIM Framework is the application model, which specifies medium-grained components common to MIES applications. The SEMATECH CIM Framework Specification Version 2.0 (Doscher 1998) has almost 300 pages of detailed component models specified using the

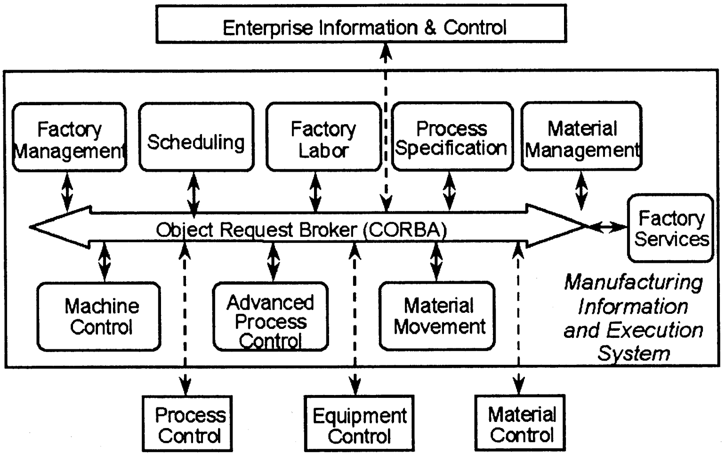

Enterprise Information & Control
Factory Management
Scheduling
Factory Labor
Process Specification
Material Management
Object Request Broker (CORBA)
Factory Services
Machine Control
Advanced Process Control
Material Movement
Manufacturing Information and Execution System
Process Control
Equipment Control
Material Control

Figure 5 Functional Groups for MIES.

TABLE 1 CIM Framework Application Components

Factory Services	Material Management	Process Specification Management
Document management	Product management	Process specification
Version management	Durable management	Process capability
History management	Consumable management	
Event broker	Product specification	*Schedule management*
	Bill of material	Dispatching
Factory management		
Factory	*Advanced process control*	*Machine control*
Product release	Plug-in management	Machine management
Factory operations	Plug-in execution	Recipe management
	Control management	Resource tracking
Factory labor	Control execution	
Person management	Control database	*Material movement*
Skill management	Data collection	Material movement

methodology of Section 2.3. Section 4 presents a portion of the CIM Framework Product Management component to illustrate the style and detail of the specification. By developing industry-wide consensus on these component definitions, SEMATECH has enabled manufacturers quickly and cost effectively to build, modify and enhance MIES by assembling standards-conformant components from multiple suppliers.

2.2.4. Status

Six of the CIM Framework specifications have become SEMI (Semiconductor Equipment and Materials International) standards (http://www.semi.org/web/wstandards.nsf/). These standards were developed with the cooperative efforts of both users and suppliers of semiconductor MES software systems. Active efforts were underway to add another standard for the CIM Framework Scheduling Component by the end of 2000, and there is interest in working on factory operations and production machine components (Hodges 2000). While there may be few instances of fully compliant MES products being shipped, we believe that vendors are using the standards in their product-development process and plans. The customers of these MES products are also using the adopted standards in integration efforts that combine best-in-class components into working factory systems.

3. AI APPROACHES TO SHOP-FLOOR SCHEDULING AND CONTROL

In the preceding section, we described two architectural approaches for organizing functions related to shop-floor scheduling control. One of the most important of these functions is scheduling. Another chapter reported on two major approaches to solving these problems: mathematical programming and heuristics. In this chapter, we describe a number of AI (artificial intelligence) techniques.

3.1. Knowledge-Based Systems

Expert and knowledge-based systems were quite prevalent in the early and mid-1980s. They have four main advantages. First, and perhaps foremost, they can use both quantitative and qualitative knowledge in the decision-making process. Second, they are capable of generating heuristics that are more complex than the simple dispatching rules described in an earlier chapter. Third, the selection of the best heuristic can be based on information about the entire shop, including current jobs, expected new jobs, and the status of all resources. Fourth, they capture complex relationships in elegant, new data structures and contain special techniques for manipulation of these data structures. They are, however, time consuming to build and verify, difficult to maintain and change, and generate only feasible solutions that can be quite far from the optimum.

3.1.1. Generating the Required Knowledge Base

Formalizations of the knowledge that human experts use—rules, procedures, heuristics, and other types of abstractions—are captured in the knowledge base. Three types of knowledge, procedural, declarative, and meta, are usually included. Procedural knowledge is domain-specific, problem-solving knowledge. Declarative knowledge provides the input data that defines the problem domain. Metaknowledge is knowledge about how to use the other knowledge to solve the actual problem. Several data structures have been proposed to represent the knowledge, including semantic nets, frames, scripts, predicate calculus, and production rules. The inference engine implements a strategy to apply to the knowledge to obtain a solution to the problem at hand. It can be forward chaining (data driven) or backward chaining (goal driven).

The first step in developing a knowledge base is knowledge acquisition, which is a two-step process: get the knowledge from knowledge sources and store that knowledge in digital form. Knowledge acquisition, such as protocol analysis, machine learning, and interactive editing (Shaw et al. 1992), has been an active area of research. Knowledge sources may be human experts, simulation data, experimental data, databases, and documents. In scheduling problems, the knowledge sources are likely to be human experts or simulation data. To extract knowledge from these two sources, any technique that learns from examples (data) becomes a promising tool. Inductive learning, which is a state classification process, is one such technique. If we view the state space as a hyperplane, training data (consisting of conditions and decisions) can be represented as points on that hyperplane. The inductive learning algorithm seeks to draw lines that divide the hyperplane into several areas within which the same decision (conclusion) will be made.

Quinlan (1986) developed an algorithm, which implements the inductive learning paradigm, called Iterative Dichotomister 3 (ID3). ID3 uses examples to induce production rules (e.g., if-then) that form a simple decision tree. Decision trees are one way to represent knowledge for the purpose of classification. The nodes in a decision tree correspond to attributes of the objects to be classified, and the arcs are alternative values for these attributes. The end nodes of the tree (leaves) indicate classes to which groups of objects belong. Each example is described by attributes and a resulting decision. To determine a good attribute, which is the basis for an object-class partitioning, entropy is employed. Entropy is a measure of the information content of each attribute. Rules are derived through a repetitive decomposition process that minimizes the overall entropy.

The attribute with the minimum entropy value will be selected as a node in the decision tree. The arcs out of this node represent different values of this attribute. If all the objects in an arc belong to one class, the partition process stops. Otherwise, another attribute will be identified using entropy values to partition further the objects that belong to this arc. This partition process continues until all the objects in an arc are in the same class. Before this algorithm is applied, all attributes that have continuous values need to be transformed to discrete values.

In the context of job shop scheduling, the attributes represent system status and the classes represent the dispatching rules. Very often the attribute values are continuous. Yih (1990) proposed a trace-driven knowledge-acquisition (TDKA) methodology to deal with continuous data and avoid the problems that occur interviewing human experts. TDKA learns scheduling knowledge from expert schedulers without resorting to an interview. There are three steps. In Step 1, an interactive simulator is developed to model the system of interest. The expert will interact with this simulator and make decisions. The entire decision-making process will be recorded in the simulator and can be repeated for later analysis. The pair, system information and scheduling decision, is called a trace. Step 2 analyzes the trace and proposes classification rules to partition the trace into groups. The partition process stops when most of the cases in each group use the same dispatching rule (error rate is below the threshold defined by the knowledge engineer). Then the decision rules are formed. The last step is to verify the generated rules. The resulting rule base is used to schedule jobs in the simulator. If it performs as well as or better than the expert, the process stops. Otherwise the threshold value is increased and the process returns to Step 2. This approach was applied in an electroplating process line and the rule base system outperforms the users. Later, Yih (1994) developed a prolog-based controller that handles the time-window problems in the same manufacturing environment.

3.1.2. Applications to Scheduling and Control

ISIS (Fox 1983) was the first expert system aimed specifically at job shop scheduling problems. ISIS used a constraint-directed-reasoning approach with three constraint categories: organizational goals, physical limitations, and causal restrictions. Organizational goals specified five objective functions based on due date and work-in-progress. Physical limitations specified the processing capability of each resource. Casual restrictions included all procedural constraints and resource requirements. Several issues related to these constraints were considered, such as conflicts among constraints, relative importance of constraints, and interactions of constraints. ISIS used a three-level, hierarchical, constraint-directed search. Orders were selected at level 1. Capacity analysis was performed at level 2 to determine the availability of the resources required by the order. Detailed scheduling was performed at level 3, to assign times to the resources identified at level 2. ISIS utilized its constraint knowledge to maintain the consistency of the schedule and identify scheduling decisions that would result in poorly satisfied constraints. It also included the capability to construct and alter schedules interactively. Chiu and Yih (1995) proposed a learning-based approach for dynamic scheduling in a distributed manufacturing system. An incremental approach to training a decision tree is proposed in this study. Each training sample consists of system attributes as inputs and a dispatching rule as its output. In their work, simulations are conducted first to collect some scenarios, and then the genetic algorithm is performed to search a good dispatching rule for each scenario. The learning algorithm is then applied to obtain a decision tree for dynamic selection of scheduling rules.

Chang (1996) proposed a fuzzy-based methodology to control the number of kanbans in a generic kanban system. In this approach, the simulated annealing algorithm is employed to find the near-optimal number of kanbans for different system status, and thereafter a training instance is generated. Then the proposed fuzzy system will be generated for dynamic kanban control. Other work in developing a control system includes Huang and Chang (1992), Gupta et al. (1989), Chandra and Talavage (1991), and Talavage and Shodhan (1992).

Adachi et al. (1988) proposed a pattern-recognition-based method for controlling a multiloop production system. In their proposed approach, a state table is constructed for the control-decision support system (CDSS) based on the simulation results. After the user indicates the desired level and importance weight for each performance measure, the one with shortest distance to the desired pattern will be selected by the control system and the associated performance level will be displayed. These procedures are repeated until the user is satisfied with the expected performance levels. The authors further constructed a rule-based decision support system (RBDSS) to control the same production system and compared the performance of CDSS and RBDSS (Adachi et al. 1989).

Several researchers have attempted to use the knowledge-based approach to model the shop-floor control problem (Farhoodi 1990; Pluym 1990; Adachi et al. 1989). Under this approach, a central database with several production rules handles scheduling and monitors system status. Each production rule consists of a condition part and an action portion with a form of an if-then clause. Typically, these rules are based on the simulation results from different scenarios or the knowledge from the experience of schedulers. When a decision-making point is encountered, the database is scanned to find the condition that could match the current situation and the associated action is then executed. However, it is not easy to generate a database consisting of every possible situation for a system. Besides, if this database is large or the production rules are complex, it will take a long time to search the database and it is impractical for real-time implementation.

O'Grady and Lee (1988) proposed a cell control system, called PLATO-Z, by using a rule-based expert system and a multiblackboard/actor model. In the proposed control system, the major functions are performed by four blackboard subsystems: scheduling, operation dispatching, monitoring, and error handling. Adequate messages are passed between blackboard subsystems in order to achieve the control requirements. This control framework was further implemented by an object-oriented programming technique (O'Grady and Seshadri 1992).

Wu and Wysk (1988, 1989) also proposed a multipass, expert control system for flexible manufacturing cells. Under their proposed system, some candidate rules are selected by a knowledge-based system and then the performance of each candidate rule is evaluated through simulation. Weighted objective values are compared in order to achieve the multicriterion objective. Cho and Wysk (1993) then refined it by using a neural network instead of the knowledge-based system for selecting the candidate rules in the initial stage.

3.1.3. Agents to the Rescue?

It is difficult to use expert and knowledge-based systems to solve large, real-world scheduling problems because of their limited knowledge and problem solving abilities. To address this, AI researchers have used the "divide and conquer" approach to develop distributed scheduling approaches (Parunak et al. 1985). This requires a technique to decompose the scheduling problem and a collection of associated knowledge-based systems that cooperate to solve the overall problem (Zhang and Zhang 1995). Cooperation is handled through an agent paradigm. Each agent is a complete knowledge-based system with its own long-term knowledge, solution-evaluation criteria, languages, algorithms, and hardware requirements. A multiagent system is created by integrating agents selected from a "library" of agents.

For example, one such multiagent system could involve two types of agents: tasks and resources. Each task agent might schedule a certain class of tasks, such as material handling, machining, or inspection, on those resources capable of performing such tasks. The schedule is generated using any task-related performance measure, such as minimize tardiness. The schedules generated by task agents become goals for the resource agents (Daouas et al. 1995). Each resource agent schedules tasks for its assigned resource(s) using resource-related performance measures, such as maximize utilization. Each resource agent will use its schedule to decide whether it can meet the scheduling goals set by the task agents. Clearly, a situation can arise where no resource will accept a given task; coordination mechanisms must be developed to avoid this situation.

While there is promise for these types of agent-based approaches, there are no general guidelines for the design and implementation of such approaches.

3.2. Artificial Neural Networks

Neural networks, also called connectionist or distributed/parallel processing models, have been studied for many years in an attempt to mirror the learning and prediction abilities of human beings.

Neural network models are distinguished by network topology, node characteristics, and training or learning rules. They are important because they can match current shop status and the desired performance measures to near-optimal scheduling strategies and they can learn (Yih and Jones 1992).

Among the many network topologies and learning algorithms, the Hopfield network and the multilayer perceptron are preferred by several researchers for scheduling problems. Therefore, in the following sections, these two networks will be briefly discussed, and the related works on scheduling problems will be reviewed.

3.2.1. Hopfield Networks

A Hopfield network consists of nodes that are fully connected to each other bidirectionally. Instead of a continuous value, this network takes only the binary or bipolar value as its input. In addition, it is also regarded as a symmetrically weighted network because the weights on the links between nodes are the same in both directions. When an input pattern is applied, the Hopfield network will adjust the weights until it converges to a stable state. This happens when the output value of each node is no longer changed. In other words, the network will reduce its "energy" until it stabilizes in a hollow of the energy landscape.

Foo and Takefuji (1988a, b) used the Hopfield network to solve job shop scheduling problems. The scheduling problem was first mapped into a two-dimensional matrix representation. Feasibility constraints and performance measures were then formulated as the energy function, named cost function. The characteristic of this energy function is that it will result in very large value when the schedule is not feasible or the performance is far from expectations. The solution is obtained by reducing the energy in the network. The authors concluded that this approach could produce near-optimal solutions, though the optimality was not guaranteed. In addition, it was claimed that the proposed approach would not be feasible in a large-scale problem.

Zhou et al. (1991) modified this approach by using a linear cost function and concluded that this modification not only produced better results but also reduced network complexity. Other works related to using the Hopfield network for the scheduling problem include Zhang et al. (1991) and Arizono et al. (1992).

3.2.2. Supervised-Learning Neural Networks

Through exposure to historical data, supervised-learning neural networks attempt to capture desired relationships between the inputs and the outputs. Back-propagation is the most popular and widely used capture procedure. Back-propagation (Rumelhart et al. 1986, Werbos 1995) applies the gradient-descent technique to change a collection of weights so that some cost function can be minimized. The cost function, which is dependent on weights and training patterns only, is defined by:

$$C(W) = \tfrac{1}{2}\Sigma\,(T_{ij} - O_{ij}) \tag{1}$$

where the T is the target value, O is the output of the network, i represents the output nodes, and j represents the training patterns.

After the network propagates from the input layer to the output layer, the error between the desired output and actual output will be back-propagated to the previous layer. In the hidden layers, the error for each node is computed by the weighted sum of errors in the next layer's nodes. In a three-layered network, the next layer means the output layer. The activation function is usually a sigmoid function with the weights modified according to (2) or (3).

$$\Delta W_{ij} = \eta\, X_j\,(1 - X_j)(T_j - X_j)\,X_i \tag{2}$$

or

$$\Delta W_{ij} = \eta\, X_j\,(1 - X_j)\,(\Sigma\,\delta_k W_{jk})\,X_i \tag{3}$$

where W_{jk} is weight from node i to node (e.g., neuron) j, η is the learning rate, X_j is the output of node j, T_j is the target value of node j, and δ_k is the error function of node k. If j is in the output layer, (2) is used. If j is the hidden layers, (3) is used. The weights are updated to reduce the cost function at each step. The process continues until the error between the predicted and the actual outputs is smaller than some predetermined tolerance.

Rabelo (1990) was the first to use back-propagation neural nets to solve job shop scheduling problems. He allowed several job types, with different arrival patterns, process plans, precedence requirements, and batch sizes. Examples were generated to train the neural network to select those characterizations of the manufacturing environments suitable for various scheduling policies and applied to the target manufacturing system. The neural networks were trained for problems involving

3, 4, 5, 8, 10, and 20 machines. To carry out this training, a special input-feature space was developed. This space contained information on both job characteristics (such as job types, number of jobs in each type, routings, due dates, and processing times) and shop characteristics (such as number of machines and their capacities). Neural networks were tested on numerous scheduling problems with a variety of performance measures. For each test, the output of the neural network represented a relative ranking of the available dispatching rules. The one with the largest ranking was selected. Rabelo showed that the same rule did not always minimize a specified performance measure under all input conditions. For example, SPT does not always minimize mean flow time. In addition, he showed that the rule selected by the neural network never performed worse than the presumed optimum.

3.2.3. Multilayer Perceptrons

A multilayer perceptron is a fully connected feed-forward network consisting of an input layer, an output layer, and several hidden layers in between. Each layer is composed of nodes that are fully connected with those in the succeeding layer by weights. Each node computes a weighted sum of the elements in the preceding layer, subtracts a threshold, and then passes the result through a nonlinear function, called an activation function. Typically, the activation function is a sigmoid energy function and the learning algorithm employed to adjust weights is the Backpropagation algorithm (Rumelhart et al. 1986).

When an input pattern in training data is fed into the network, the error between the desired output and actual output values will be back-propagated to the previous layer and the weights will be adjusted accordingly. This procedure is called training and it is done to obtain the proper weight matrices so that the total is minimized. Yih et al. (1993) conducted a three-phased experiment to quantify the benefits of training. Schedules were generated by a human expert, an untrained neural network, and a neural network with training data refined by a semi-Markov decision model. The results indicated that the untrained neural network performed worse than the human expert did. However, the trained neural network outperformed both. This implies that good training data will significantly improve network performance.

Several works have used multilayer perceptrons with the Backpropagation training algorithm in scheduling or in candidate rules selection. Potvin et al. (1992) modified the network structure but still used the Backpropagation learning algorithm to build up the dispatcher for automated vehicles. Rabelo et al. (1993) used modular neural networks to serve as a candidate rule selector. In 1996, Chen and Yih discussed the impact of the network input attributes on the performance of the resulting control system.

As mentioned above, Cho and Wysk (1993) utilized the multilayer perceptron to take the place of the knowledge-based system in selecting candidate scheduling rules. In their proposed framework, the neural network will output a "goodness" index for each rule based on the system attributes and a performance measure. Sim et al. (1994) used an expert neural network for the job shop scheduling problem. In their approach, an expert system will activate one of 16 subnetworks based on whether the attribute corresponding to the node (scheduling rules, arrival rate factor, and criterion) is applicable to the job under consideration. Then the job with the smallest output value will be selected to process.

Yih and Jones (1992) proposed using multilayer perceptrons in selecting some candidate rules for further evaluation of their performance. In their proposed approach, a multilayer perceptron will take the attributes describing the system configuration and the performance measures and will output a proper matching score for each dispatching rule. They also used this approach for multiple-criterion objectives.

Sun and Yih (1996) adopted their idea to develop a neural network-based controller for manufacturing cells. In their approach, a neural network was trained to serve as decisionmaker that will select a proper dispatching rule for its associated machine to process the next job. Based on their results, the controller performs well under multiple criterion environments. In addition, when the production objectives change, the controller can respond to such change in a short time.

3.2.4. Unsupervised Neural Networks (Competition Based)

Competition-based neural networks, which are good at classifying or clustering input data, can also be applied to scheduling problems. Since the classes or clusters are not known in advance, the network must discover them by finding correlation in the input data. Multidimensional data sets are presented to the network, which adaptively adjusts its weights. This input presentation process is repeated until the network reaches stability—each output unit is activated only for a particular subset of the input patterns. Variations of these neural networks have been used to solve scheduling problems. For example, Bourret et al. (1989) applied some of these principles to develop a neural network that was able to schedule optimal time periods of low-level satellites to one or several antennas. The neural network was able to take into account that each satellite has a given priority and several other operational constraints. Min et al. (1998) adopted this concept and developed a methodology in a multiobjective scheduling problem. Kim et al. (1998) integrated the network with inductive learning module to develop a real-time controller for flexible manufacturing systems.

3.2.5. Reinforcement Learning

We noted above that supervised learning neural networks attempt to capture desired relationships between inputs and outputs through exposure to training patterns. For some problems, the training period may be too short to find those relationships. When the desired response is obtained, changes to the neural network are performed by assessing penalties for the actions previously decided by the neural network. As summarized by Tesauro (1992), ''In the simplest form of this paradigm, the learning system passively observes a temporal sequence of input states that eventually leads to a final reinforcement or reward signal (usually a scalar). The learning system's task in this case is to predict expected reward given an observation of an input state or sequence of input states. The system may also be set up so that it can generate control signals that influence the sequence of states.'' For scheduling, the learning task is to produce a schedule that minimizes (or maximizes) the performance measure. Several procedures have been developed to train neural networks in a variety of generic cases.

One of the popularly adopted reinforcement learning algorithms is called Q-learning. In this approach, an action-value function, which assigns an expected utility to take a given action in a given state, is defined. The output of this function is called Q-values. The relation between Q-values and the utility values is as follows:

$$U(s) = \underset{a}{\text{Max}}\ Q(a, s)$$

where $U(s)$ = the utility value at state s
$Q(a, s)$ = the Q-value of taking action a in state s

The learning process in Q-learning is to find an appropriate Q-value associated with each action in each state that the decision can be based on.

Rabelo et al. (1994) utilized a procedure developed by Watkins (1989), called Q-learning, to solve dynamic scheduling problems. The procedure followed trends in the shop floor and selected a dispatching rule that provided the maximum reward according to performance measures based on tardiness and flow time. Zhang and Dietterich (1996) utilized a procedure developed by Sutton (1988) called TD(λ) to schedule payload processing of NASA's space shuttle program. The scheduling system was able to outperform an iterative repair scheduler that combined heuristics with simulated annealing.

Kim and Lee (1995) formulated the machine-scheduling problem as a reinforcement learning problem and then developed a learning-based heuristic, called EVIS, for solving the scheduling problem. The EVIS, implementing reinforcement learning with the genetic algorithm, was then applied to a few deterministic scheduling problem instances. The results show that the proposed heuristic has good average-case performances for most of the problem instances.

Another alternative in reinforcement learning involves using the CMAC network. A CMAC network can be regarded as an associative memory system which stores the appropriate output in the associated memory cells. As mentioned by Miller et al. (1990), the CMAC network is an alternative to the back-propagated, multilayer, neural network because it has the advantages of local generalization, rapid training, and output superposition. Several researches have been involved in applying the CMAC network to develop a controller. Miller et al. (1990) demonstrated the application of CMAC networks in real-time robot control without providing any initial knowledge, in character recognition, and in signal processing. Lin and Kim (1991) constructed a CMAC-based controller and demonstrated its capability in the inverted pendulum problem. Moody (1989) proposed a multiresolution CMAC (MRC) that combines some CMAC networks with different resolution to increase the accuracy and generalization ability. The proposed approach shows good performance and online learning ability in prediction of a time series.

3.3. Genetic Algorithms

Genetic algorithms (GA) provide an optimization methodology based on a direct analogy to Darwinian natural selection and mutations in biological reproduction. In principle, genetic algorithms encode a parallel search through concept space, with each process attempting coarse-grain hill climbing (Goldberg 1988). Instances of a concept correspond to individuals of a species. Induced changes and recombinations of these concepts are tested against an evaluation function to see which ones will survive to the next generation. The use of genetic algorithms requires five components:

1. A way of encoding solutions to the problem—fixed length string of symbols
2. An evaluation function that returns a rating for each solution
3. A way of initializing the population of solutions
4. Operators that may be applied to parents when they reproduce to alter their genetic composition, such as crossover (i.e., exchanging a randomly selected segment between parents), mutation (i.e., gene modification), and other domain-specific operators
5. Parameter setting for the algorithm, the operators, and so forth

A number of approaches have been utilized in the application of genetic algorithms (GA) to job shop scheduling problems (Davis 1985; Goldberg and Lingle 1985; Starkweather et al. 1992):

1. Genetic algorithms with blind recombination operators have been utilized in job shop scheduling. Their emphasis on relative ordering schema, absolute ordering schema, cycles, and edges in the offsprings will lead to differences in such blind recombination operators.
2. Sequencing problems have been addressed by mapping their constraints to a Boolean satisfiability problem using partial payoff schemes. This scheme has produced good results for very simple problems.
3. Heuristic genetic algorithms have been applied to job shop scheduling. In these genetic schemes, problem specific heuristics are incorporated in the recombination operators (such as optimization operators based).

Starkweather et al. (1992, 1993) were the first to use genetic algorithms to solve a dual-criteria job shop scheduling problem in a real production facility, a beer plant. Those criteria were the minimization of average inventory in the plant and the minimization of the average waiting time for an order to be selected. These criteria are negatively correlated: as the inventory increases (decreases), the wait decreases (increases). To represent the production/shipping optimization problem, a symbolic coding was used for each member (chromosome) of the population. In this scheme, customer orders are represented by discrete integers. Therefore, each member of the population is a permutation of customer orders. The GA used to solve this problem was based on blind recombinant operators. This operator emphasizes information about the relative order of the elements in the permutation because this impacts both inventory and waiting time. A weighted sum of the two criteria was utilized to rank each member of the population. That ranking was based on an online simulation of the plant operations. This approach generated schedules that produced inventory levels and waiting times that were acceptable to the plant manager. In addition, the integration of the genetic algorithm with the online simulation made it possible to react to plant dynamics.

These applications have emphasized the utilization of genetic algorithms as a ‘‘solo’’ technique. This limits both the complexity of the problems solved and levels of success. Recent research has demonstrated the sensitivity of genetic algorithms to the initial population. When the initial population is generated randomly, genetic algorithms are shown to be less efficient that the annealing-type algorithms but better than the heuristic methods alone. However, if the initial population is generated by a heuristic, the genetic algorithms become as good as or better than the annealing-type algorithms. In addition, integration with other search procedures (e.g., taboo search) has enhanced the capabilities of both. This result is not surprising, as it is consistent with results from nonlinear optimization.

3.4. Fuzzy Logic

Fuzzy set theory has been utilized to develop hybrid-scheduling approaches. Fuzzy set theory can be useful in modeling and solving job shop scheduling problems with uncertain processing times, constraints, and set-up times. These uncertainties can be represented by fuzzy numbers, which are described by the concept called interval of confidence. These approaches usually are integrated with other methodologies (e.g., search procedures, constraint relaxation). For example, Slany (1994) stressed the imprecision of straightforward methods presented in the mathematical approaches and introduced a method known as fuzzy constraint relaxation, which is integrated with a knowledge-based scheduling system. Chang (1996) and Chang and Yih (1998, 1999) proposed a machine learning methodology to develop a fuzzy rule-based system for controlling a kanban system. Grabot and Geneste (1994) use fuzzy logic principles to combine dispatching rules for multi-criteria problems. Krucky (1994) used fuzzy logic to minimize setup times for a production line with a medium-to-high product mix by clustering assemblies into families of products that share the same setup. The clustering was achieved by balancing a product’s placement time between multiple-high-speed placement process steps. Tsujimura et al. (1993) presented a hybrid system, which uses fuzzy set theory to model the processing times as triangular fuzzy numbers (TFNs). Each job is defined by two TFNs, a lower bound and an upper bound. A branch and bound procedure is utilized to minimize makespan through the shop based on these estimates.

3.5. Commercial Systems

A number of university software systems use these techniques to do scheduling. A few commercial software systems use expert systems and genetic algorithms. Commercial hardware and software systems are available that implement neural networks, but none have been designed specifically for scheduling.

4. MANUFACTURING EXECUTION SYSTEMS (MES)

During the 1990s a new category of software system, manufacturing execution system (MES), emerged that consolidated and automated a number of functions involved in the management and operation of a production facility. An MES is a collection of hardware/software components that enables the management and optimization of production activities from order launch to finished goods. While maintaining current and accurate data, an MES guides, initiates, responds to, and reports on plant activities as they occur. An MES provides mission-critical information about production activities to decision support processes across the enterprise (MESA 1997; Wallace 1999). The term *order launch* is to be interpreted as initiation of physical production activities, typically beginning with materials preparation or machine preparation. Activities relating to planning and scheduling physical production operations are included within the scope of MES, but activities related to defining physical operations are not. The word "component" is used in a generic way to mean a separable portion of a larger whole.

Wallace (1999) provides a list of 12 major functions, derived from the original list in MESA (1997). These functions are similar to those found in the PERA model described in Section 3.

1. Resource allocation and tracking
2. Operations/detailed scheduling
3. Production unit dispatching
4. Specification management
5. Data collection/acquisition
6. Labor management
7. Quality management
8. Process management
9. Maintenance management
10. Product tracking
11. Performance analysis
12. Material management

The relationship between these functions and software products that call themselves MES is not clear. Some MES products can be purchased prepackaged as a single unit that performs all of these functions. Many other products provide only a subset of these functions; some call themselves MES, some do not. Standard interfaces would facilitate the integration of components into an overall MES and facilitate the integration of that MES with other enterprise software applications such as ERP (enterprise resource planning). No viable standards are emerging to fulfill this need, but there are four organizations looking at the problem: ISO, ISA, SEMI, and OMG.

Within the International Organization for Standardization (ISO), the Manufacturing Management Data Exchange (MANDATE) work in TC184/SC4/WG8 is focusing MES (http://www.iso.ch/meme/TC184SC4.html). Within ISA, MES-related work is being done in SP95 (http://www.isa.org/sc/committee/1,1512,145,00.html), see **2.1**. Within SEMI, MES-related standards are being generated as part of the standardization of CIM Framework specifications, (http://www.semi.org/web/wstandards.ns), see **2.2**. Within OMG, the MES-related work is being done by the Manufacturing Execution Systems/Machine Control group (http://www.omg.org/homepages/mfg/mfgmesmc.htm). Most of the remaining material in this section is based on work done in the OMG working group.

4.1. A More Detailed Look at MES Data

MES functions create, collect, modify, analyze, react to, and manage a great deal of data. These data are summarized below:

Dispatch data: job/operation dispatch list, or commands to operators and equipment

Equipment resource data: resource state, staffing, setup, current operations and assignments, and job usage history

Labor resource data: personnel availability and tracking information, and job assignment history

Maintenance data: machine availability data, maintenance history, and usage data

Material location data: location and state of materials with respect to active resources and material-handling components

Order data: units of a particular product to be manufactured, including status and associations with particular material groups

Performance, cost, and usage data: cost of operations performed, materials and resources used, idle time

Process control data: process control parameters

Product data (WIP): the amount, state, and disposition of materials in production and their relationship with manufacturing orders

Quality analysis data: data resulting from the quality analysis function, that is, interpreted measurements of process and product

Quality data: product and process measurement data, which can include such data taken from process operations

Resource description data: characteristics of labor and equipment resources such as capabilities, skills, types, and assigned cost

Schedule data: allocation of resources to jobs and processes per time period

Shop-floor data: raw data collected by data collection systems that can be used to derive product (WIP) data, resource data, performance data, and so on

Specification data: specifications for how to perform a manufacturing process, including sequence of operations to be performed, equipment, tooling and skills requirements, and materials to be used

Tooling resource data: usage, location and allocation information, which may be used for tracking, scheduling, and maintenance of tools

Table 2 shows the nature of the data usage by the MES functions. The numbers in each cell correspond to the numbers assigned above. The designations HRM, ERP, and PPE refer to non-MES systems of the enterprise that use some of the same data. HRM refers to human resource management systems, ERP refers to enterprise resource planning systems, and PPE refers to product and process engineering systems.

4.2. MES Object Models

An analysis of the MES functions and the data in Table 2 led to the development of the two MES object models shown in Figure 6 and Figure 7. These models, which are based on the work in Ray and Wallace (1995) and OMG (1997b), organize the functions and data in a way that facilitates the implementation of MES components based on object or component middleware technology. They also provide an implementor or integrator with an organization of a distributed MES architecture that simplifies integration with other manufacturing information systems.

The partitions that we have shown with dotted boxes in Figure 6 are groupings of model entities that are closely coupled or functionally similar. Coupling between these partitions should be supported via domain specific names or keys. Each partition in the model is described in detail below. While all these partitions are important to MES, they need not all be supported directly within an MES. We make special note below of those that we consider to be core MES partitions. The tags, shown in italics above or below each partition box in the figures, are general characterizations of the resource entities within the corresponding partition. These are further elaborated in the text below:

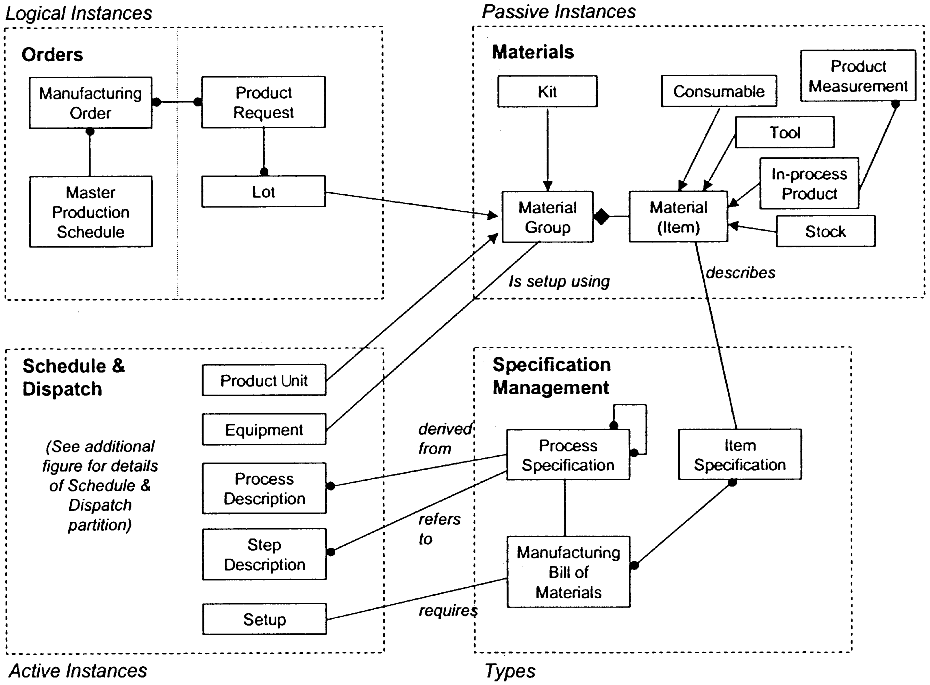
Orders: The orders partition contains two planning entity classes: those that are created by ERP systems and those that may be MRPII entities. This distinction is made here since it is not clear whether MRPII is considered an ERP or MES function. We have tagged this partition ‘‘logical instances’’ in the diagram to indicate that the entities within it are logical resources, which merely represent groups of physical resources or planned physical resources that are modeled in other partitions.

Schedule and dispatch: This partition is at the heart of MES operations, containing the primary entities involved in scheduling and dispatching. If the MES supports reactive scheduling, then the internal components must be able to share a common understanding of process description. Therefore, the model has a separate entity for process description, which is distinct from (and an instance of which is derived from a) process specification. We note with the tag ‘‘active instances’’ that unlike the entities in the other partitions, many Schedule and Dispatch entities can initiate actions.

Specification management: The entities in this partition have similar access characteristics but only a loose coupling. This means that some efficiency may be gained by putting all these entities into one component. Nevertheless, as long as all access requirements shared by these entities are met, these entities could be stored in multiple components. We note with the tag ‘‘types’’ that this partition contains information resource entities that provide type information used to instantiate or describe other entities.

TABLE 2 Relationship of Data to MES Activities

	Collects	Creates	Changes	Manages	Analyzes	Reacts To	Delivers	Displays	Derived From
Dispatch data		3	3, 8			8, 10, 12 Control	3	3, 10	
Equipment resource data	5		5, 9	1	9	2, 3			
Labor resource data	6		10	6, HRM	11	2, 3			
Maintenance data	5	9	9	9	9, 10	2, 3, 9	9		
Material location data	5, 12	12	12	12		3, 10		12	Shop-floor data
Order data	ERP		10	ERP		2, 3, 10	ERP	2, 10	Product data
Performance data		11		11		ERP, P/PE	11	11	Resource data, Schedule data, Shop-floor data, Process data, Product data
Process control data		P/PE	8			Control	8		
Product data (WIP)	5		7, Control	10	11	2, 10, ERP		10	Shop-floor data, Process data
Quality analysis data		7		7, 10		3, 9	7	7	Quality data
Quality data	5?, 7	Control		7	7, 10	?			Product data, Process data
Resource description data		ERP, HRM, P/PE		6, HRM, ERP, P/PE		2, 3			
Schedule data		2	2	2	11	2, 3			
Shop-floor data	5			5	2, 10	2, 3	5		
Specification data	P/PE	P/PE	4	4		3, 7	P/PE		
Tooling resource data	12	12	Control	1, 10		2, 3, 10			


Logical Instances
Passive Instances
Orders
Manufacturing Order
Product Request
Master Production Schedule
Lot
Materials
Kit
Consumable
Product Measurement
Tool
In-process Product
Material Group
Material (Item)
Stock
Is setup using
describes
Schedule & Dispatch
(See additional figure for details of Schedule & Dispatch partition)
Product Unit
Equipment
Process Description
Step Description
Setup
Specification Management
derived from
Process Specification
Item Specification
refers to
Manufacturing Bill of Materials
requires
Active Instances
Types

*The graphic notation used in this figure and subsequent object models, is a simplified version of the object modeling technique (OMT) notation described in Rumbaugh (1991).

Figure 6 Simplified Object Model.*

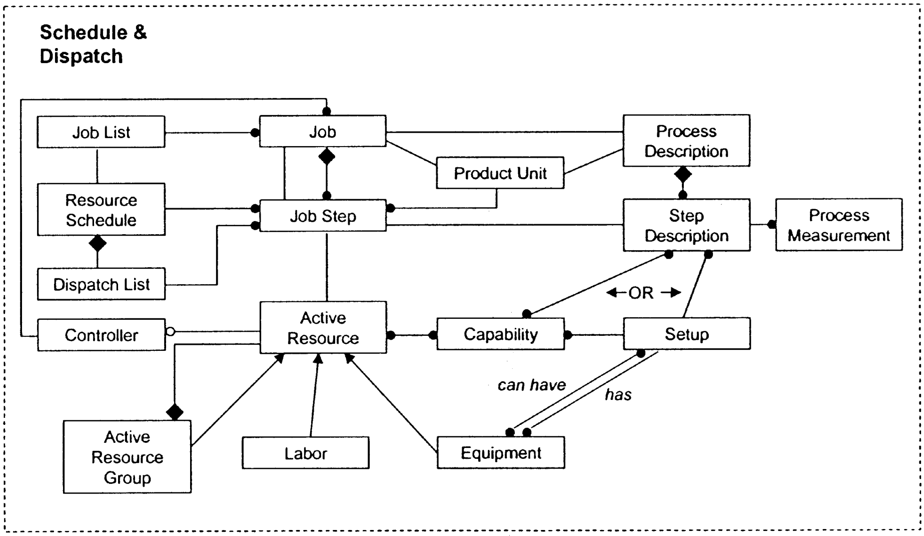

Schedule & Dispatch
Job List
Job
Process Description
Product Unit
Resource Schedule
Job Step
Step Description
Process Measurement
Dispatch List
OR
Controller
Active Resource
Capability
Setup
can have
has
Active Resource Group
Labor
Equipment

Figure 7 Detailed View of Scheduling and Dispatching.

Materials: The entities in this partition are related to material instances such as material status, history, and genealogy data. It may also provide inventory information. We note with the tag ‘‘passive instances’’ that the entities in this partition represent physical resources that are manipulated or acted upon by active resources represented in other models.

The entities of the MES Object Model shown in Figures 6 and 7 are described in detail below:

Active resource: a physical resource, or a logical grouping a resources, that can perform process operations and supports specified capabilities during manufacturing activities—people and equipment such as machines, robots, storage devices, transport devices, etc.

Capability: the ability of a resource to perform a specified operation. For human resources, a capability is sometimes called a (certified) skill. For equipment resources, this potential may be dependent on setup and on having an appropriately skilled operator.

Consumable: a unit of material that is used and expended during manufacturing or support processes.

Controller: a software system, or human agent, that provides the intelligence needed by an active resource to perform all or part of a process operation automatically.

Item specification: the complete description of a kind of material item including its shape, weight, handling characteristics, image, etc.

Job: a unit of work assigned to an active resource. At the schedule/dispatch level, the unit of work is a lot; at lower levels, the unit of work is a process operation, transport operation, or operation step.

Job step: an instantiation of a step description. Depending on its granularity, a job step becomes a job for the active resource to which the step execution is assigned.

Kit: a physical collection of material items that is handled as a unit.

Labor: a human active resource (employee or contractor or group) who performs certain jobs directly or in collaboration with a machine. The ability of a human resource to perform a specific kind of process with specific kinds of equipment is called a certified skill.

Lot: the unit of product into which a manufacturing order is decomposed for scheduling, dispatching, and tracking. A lot is a conceptual material group. The actual corresponding product units on the factory floor may be decomposed and regrouped, depending on production policies and needs.

Manufacturing bill of materials: a list of the types and quantities of all materials needed to manufacture a unit of product or to perform a particular process.

Manufacturing order: a quantity of a particular product to manufacture as specified by an ERP system.

Master production schedule: A long-term schedule created and maintained by enterprise planning systems that defines quantities of particular products to be produced in particular time frames based on customer demands, manufacturing capacity, and resource availability.

Material item: a physical resource that is acted upon or used by an active resource in the performance of a manufacturing activity. It is characteristic of most material items that they can be moved around the factory floor, and it is often the case that their location is tracked in some way.

Material group: a logical or physical collection of material instances.

Process description: a breakdown of a process into subtasks or steps, expressing the requirements for each step and its relationship to other steps in the recipe. Every process description has an associated active resource that is responsible for planning and executing the process described. Process descriptions come in many forms: NC programs, routings, operation sheets, recipes, etc.

Process specification: archival form of a process description, usually a document. An information resource is targeted for a particular type of active resource. It is copied as needed to create process descriptions for individual resources.

Product in-process: a material item, which becomes part of a final product.

Product request: a unit of product that is the basis for planning/scheduling activities at the highest level of a manufacturing execution system.

Product unit: a quantity of product (or a group of in-process product items) that undergoes manufacturing activities together.

Resource schedule: a collection of assignments of active resources to job steps at specific times.

Setup: a particular configuration of an equipment resource that results in a set of active capabilities for that resource.

Step description: a part of a process description that identifies a job step to be performed, resources necessary to perform the job step, other information needed to plan the job step, and, usually, the process specification that the active resource will use to perform the job step activity.

Stock Material: a kind of material that is the starting point for, or an ingredient in, a unit of product.

Tool: a material item used by an active resource in the performance of some manufacturing activity. A tool is a material item that is needed during the manufacturing process but is not (usually) consumed and does not become part of the finished goods. Tools are used to set up an equipment resource, or augment a workstation setup, in order to enable the workstation to perform a particular process operation.

4.3. Market Trends and Future Directions

Originally predicted to be a strong new market area, the MES product category had only grown about $218 million in 1998 (Callaway 1998). This has been attributed to such diverse factors as manufacturing management's preoccupation with higher-level enterprise systems such as enterprise resource planning (ERP) and supply chain management (SCM) and resources being diverted for Year 2000 projects. No matter what the cause, MES has not become a ubiquitous system across all manufacturing domains.

However, the functionality it targeted remains important as the changing marketplace for manufactured goods forces manufacturers to respond more quickly to market opportunities and participate in more dynamic supply chains. An MES provides the higher level of integration needed to support rapid access to data, as well as the means to respond rapidly to new production tasks or priorities. Because this need still remains, vendors in other product categories have expanded their product line to include many of the MES functions described in Section 2.2, with mixed success. This encroachment has come from above with ERP systems and from below with open control systems. While it is certain that the functionality and data described in this section will be available in the factory of the future, it is difficult to predict what systems will be managing it. This underscores the importance of developing standard interfaces and data models for this area so that manufacturers can make use of this technology no matter what the products are called or where the lines are drawn between them.

5. SUMMARY

This chapter has addressed three subjects related to the control of shop-floor operations: control architectures, AI-based scheduling, and manufacturing execution systems (MES). The ideas discussed in this chapter are applicable to many different types of manufacturing and service industries. To date, very few of the advanced scheduling techniques have been incorporated into commercial MES software packages, and little or no effort has been put into integrating these packages into an open architecture for shop-floor control. We expect this situation to change dramatically over the next few years as companies push more toward Internet-based electronic commerce.

REFERENCES

Adachi, T., Moodie, C. L., and Talavage, J. J. (1988), "A Pattern-Recognition-Based Method for Controlling a Multi-loop Production System," *International Journal of Production Research,* Vol. 26, No. 12, pp. 1943–1957.

Adachi, T., Talavage, J. J., and Moodie, C. L. (1989), "A Rule-Based Control Method for a Multi-loop Production System," *Artificial Intelligence in Engineering,* Vol. 4 , No. 3, pp. 115–125.

ANSI/ISA-S95.00.01-2000 (2000), *Enterprise-Control System Integration Part 1: Models and Terminology,* International Standards Assocation, Research Triangle Park, NC.

Arizono, I., Yamamoto, A., and Ohta, H. (1992), "Scheduling for Minimizing Total Actual Flow Time by Neural Networks," *International Journal of Production Research,* Vol. 30, No. 3, pp. 503–511.

Bourret, P., Goodall, S., and Samuelides, M. (1989), "Optimal Scheduling by Competitive Activation: Application to the Satellite-Antennae Scheduling Problem," in *Proceedings of the International Joint Conference on Neural Networks*.

Callaway (1998), "A Second Chance for MES," in *Managing Automation,* Vol. 13, No. 12, pp. 34–43.

Chandra, J., and Talavage, J. (1991), "Intelligent Dispatching for Flexible Manufacturing," *International Journal of Production Research,* Vol. 29, No. 11, pp. 2259–2278.

Chang, T. (1996), "A Fuzzy Rule-Based Methodology for Dynamic Kanban Control in a Generic Kanban System," Ph.D. dissertation, Purdue University, West Lafayette, IN.

Chang, T., and Yih, Y. (1998), "A Fuzzy Rule-Based Approach for Dynamic control of kanbans in a generic kanban system," *International Journal of Production Research,* Vol. 36, No. 8, pp. 2247–2257.

Chang, T., and Yih, Y. (1999), "Constructing a Fuzzy Rule System from Examples," *Journal of Integrated Computer-Aided Engineering,* Vol. 6, pp. 213–221.

Chen, C., Yih, Y., and Wu, Y. (1999), "Auto-bias Selection for Developing Learning-Based Scheduling Systems," *International Journal of Production Research,* Vol. 37, No. 9, pp. 1987–2002.

Chiu, C., and Yih, Y. (1995), "Learning-Based Methodology for Dynamic Scheduling in Distributed Manufacturing Systems," *International Journal of Production Research,* Vol. 33, No. 11, pp. 3217–3232.

Cho, H., and Wysk, R. A. (1993), "A Robust Adaptive Scheduler for an Intelligent Workstation Controller," *International Journal of Production Research,* Vol. 31, No.4, pp. 771–789.

Daouas, T., Ghedira, K., and Muller, J. (1995), "Distributed Flow Shop Scheduling Problem versus Local Optimization," in *Proceedings of the First International Conference on Multi-Agent Systems,* MIT Press, Cambridge, MA.

Davis, L. (1985), "Job Shop Scheduling with Genetic Algorithms," in *Proceedings of an International Conference on Genetic Algorithms and Their Applications* (Carnegie Mellon University), pp. 136–140.

Doscher, D., Ed. (1998), *Computer Integrated Manufacturing* (*CIM*) *Framework Specification Version 2.0*, Technology Transfer #93061697J-ENG, SEMATECH, Austin, Tx.

Farhoodi, F. (1990), "A Knowledge-Based Approach to Dynamic Job-Shop scheduling," *International Journal of Computer Integrated Manufacturing,* Vol. 3, No. 2, pp. 84–95.

Foo, Y. S., and Takefuji, Y. (1988a), "Stochastic Neural Networks for Solving Job-Shop Scheduling. Part 1: problem representation," in *IEEE International Joint Conference on Neural Networks,* Vol. 2, pp. 275–282.

Foo, Y. S., and Takefuji, Y. (1988b), "Stochastic Neural Networks for Solving Job-Shop Scheduling. Part 2: Architecture and Simulation," in *IEEE International Joint Conference on Neural Networks,* Vol. 2, pp. 283–290.

Fox, M. (1983), "Constraint-Directed Search: A Case Study of Job Shop Scheduling," Ph.D. dissertation, Carnegie Mellon University.

Goldberg, D. (1988), *Genetic Algorithms in Search Optimization and Machine Learning,* Addison-Wesley, Menlo Park, CA.

Goldberg, D., and Lingle, R. (1985), "Alleles, Loci, and the Traveling Salesman Problem," in *Proceedings of the of the International Conference on Genetic Algorithms and Their Applications* (Carnegie Mellon University), pp. 162–164.

Grabot, B., and Geneste, L. (1994), "Dispatching Rules in Scheduling: A Fuzzy Approach," *International Journal of Production Research,* Vol. 32, No. 4, pp. 903–915.

Gupta, M. C., Judt, C., Gupta, Y. P., and Balakrishnan, S. (1989), "Expert Scheduling System for a Prototype Flexible Manufacturing Cell: A Framework," *Computers and Operations Research,* Vol. 16, No. 4, pp. 363–378.

Harel, D. (1987), "Statecharts: A Visual Formalism for Complex Systems," *Science of Computer Programming* Vol. 8, pp. 231–274.

Hawker, J. (1999), "CIM Framework and Application Models," in *Information Infrastructure Systems for Manufacturing,* J. Mills, and F. Kimura, Eds., Kluwer, Boston.

Hodges, R. (2000), private communication.

Huang, H., and Chang, P. (1992), "Specification, Modeling and Control of a Flexible Manufacturing Cell," *International Journal of Production Research,* Vol. 30, No. 11, pp. 2515–2543.

Jones, A., Ed. (1990), *Proceedings of COMCON'90,* NIST Special Publication, National Institute of Standards and Technology, Gaithersburg, MD.

Kim, C., Min, H., and Yih, Y. (1998), "Integration of Inductive Learning and Neural network for Multi-objective FMS Scheduling," *International Journal of Production Research,* Vol. 36, No. 9, pp. 2497–2509.

Kim, G. H., and Lee, C. S. G. (1995), "Genetic Reinforcement Learning Approach to the Machine Scheduling Problem," in *Proceedings of IEEE International Conference on Robotics and Automation,* Vol. 1, pp. 196–201.

Krucky, J. (1994), "Fuzzy Family Setup Assignment and Machine Balancing," *Hewlett-Packard Journal,* June, pp. 51–64.

Lin, C., and Kim, H. (1991), "CMAC-Based Adaptive Critic Self-Learning Control," *IEEE Transactions on Neural Networks,* Vol. 2, No. 5, pp. 530–533.

MESA International (1997), "MES Explained: A High Level Vision," White paper #6, MESA International, Pittsburgh.

Miller, W. T., Glanz, F. H., and Kraft, L. G., III (1990), "CMAC: An Associative Neural Network Alternative to Backpropagation," *Proceedings of the IEEE,* Vol. 78, No. 10, pp. 1561–1567.

Min, H., Yih, Y. and Kim, C. (1998), "A Competitive Neural Network Approach to Multi-objective FMS Scheduling," *International Journal of Production Research,* Vol. 36, No. 7, pp. 1749–1765.

Moody, J. (1989), "Fast Learning in Multi-Resolution hierarchies," *Advances in Neural Information Processing,* Vol. 1, pp. 29–39.

Object Management Group (OMG) (1995), *Object Management Architecture Guide, Revision 3.0*, OMG, Framingham, MA.

Object Management Group (OMG) (1996), *CORBAservices: Common Object Services Specification,* OMG, Framingham, MA.

Object Management Group (OMG) (1997a), *Common Object Request Broker: Architecture and Specification,* Revision 2.1, OMG, Framingham, MA.

Object Management Group (OMG) (1997b), "Manufacturing Domain Task Force RFI-3 Manufacturing Execution Systems (MES)," Document mfg/97-11-01, OMG, Framingham, MA.

O'Grady, P., and Lee, K. H. (1988), "An Intelligent Cell Control System for Automated Manufacturing," *International Journal of Production Research,* Vol. 26, No. 5, pp. 845–861.

O'Grady, P., and Seshadri, R. (1992), "Operation of X-Cell—an Intelligent Cell Control System," *Computer Integrated Manufacturing Systems,* Vol. 5, No. 1, pp. 21–30.

Parunak, H., Irish, B., Kindrick, J., and Lozo, P. (1985), "Fractal Actors for Distributed Manufacturing Control," in *Proceedings of the Second IEEE Conference on Artificial Intelligence Applications,* pp. 653–660.

Pluym, B. V. D. (1990), "Knowledge-Based Decision-Making for Job Shop Scheduling," *International Journal of Computer Integrated Manufacturing,* Vol. 3, No. 6, pp. 354–363.

Potvin, J., Shen, Y., and Rousseau, J. (1992), "Neural Networks for Automated Vehicle Dispatching," *Computers and Operations Research,* Vol. 19, Nos. 3/4, pp. 267–276.

Quinlan, J. (1986), "Induction of Decision Trees," *Machine Learning,* Vol. 1, pp. 81–106.

Rabelo, L. (1990), "Hybrid Artificial Neural Networks and Knowledge-Based Expert Systems Approach to Flexible Manufacturing System Scheduling," PhD. dissertation, University of Missouri-Rolla.

Rabelo, L., Yih, Y., Jones, A., and Tsai, J. (1993), "Intelligent Scheduling for Flexible Manufacturing Systems," in *Proceedings of IEEE International Conference on Robotics and Automation,* Vol. 3, pp. 810–815.

Rabelo, L., Sahinoglu, M., and Avula, X. (1994), "Flexible Manufacturing Systems Scheduling Using Q-Learning," in *Proceedings of the World Congress on Neural Networks* (San Diego), pp. I378–I385.

Ray, S., and Wallace, S. (1995), "A Production Management Information Model for Discrete Manufacturing," in *Production Planning and Control,* Vol. 6, No. 1, pp. 65–79.

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W. et al. (1991), *Object-Oriented Modeling and Design,* Prentice-Hall, Englewood Cliffs, NJ.

Rumelhart, D., McClelland, J., and the PDP Research Group (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1: Foundations,* MIT Press, Cambridge, MA.

Shaw, M., Park, S. and Raman, N. (1992), "Intelligent Scheduling with Machine Learning Capabilities: The Induction of Scheduling Knowledge," *IEE Transactions on Design and Manufacturing,* Vol. 24, pp. 156–168.

Sim, S. K., Yeo, K. T., and Lee, W. H. (1994), "An Expert Neural Network System for Dynamic Job Shop Scheduling," *International Journal of Production Research,* Vol. 34 , No. 8, pp. 2353–2373.

Slany, W. (1994), "Scheduling as a Fuzzy Multiple Criteria Optimization Problem." CD-Technical Report 94/62, Technical University of Vienna.

Starkweather, T., Whitley, D., Mathias, K., and McDaniel, S. (1992), "Sequence Scheduling with Genetic Algorithms," in *Proceedings of the US/German Conference on New Directions for OR in Manufacturing,* pp. 130–148.

Starkweather, T., Whitley, D., and Cookson, B. (1993), "A Genetic Algorithm for scheduling with Resource Consumption," in *Proceedings of the Joint German/US Conference on Operations Research in Production Planning and Control,* pp. 567–583.

Sun, Y. and Yih, Y. (1996), "An Intelligent Controller for Manufacturing Cells," *International Journal of Production Research,* Vol. 34, No. 8, pp. 2353–2373.

Sutton, R. (1988), "Learning to Predict by the Methods of Temporal Differences," *Machine Learning,* Vol. 3, pp. 9–44.

Talavage, J. J., and Shodhan, R. (1992), "Automated Development of Design and Control Strategy for FMS," *International Journal of Computer Integrated Manufacturing,* Vol. 5, No. 6, pp. 335–348.

Tesauro, G. (1992), "Practical Issues in Temporal Difference Learning," *Machine Learning,* Vol. 8, pp. 257–277.

Tsujimura, Y., Park, S., Chang, S., and Gen, M. (1993), "An Effective Method for Solving Flow Shop Scheduling Problems with Fuzzy Processing Times," *Computers and Industrial Engineering,* Vol. 25, pp. 239–242.

Wallace E., Ed. (1999), "NIST Response to MES Request for Information," Internal Report 6397, National Institute of Standards and Technology, Gaithersburg, MD.

Watkins, C. (1989), "Learning from Delayed Rewards," Ph.D. dissertation, King's College, Cambridge.

Werbos, P. (1995), "Neurocontrol and Supervised Learning: An Overview and Evaluation," in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches,* Van Nostrand Reinhold, New York, pp. 65–89.

Williams, T. (1992), *The Purdue Enterprise Reference Architecture—A Technical Guide for CIM Planning and Implementation,* Instrument Society of America, Research Triangle Park, NC.

Wu, S. D., and Wysk, R. A. (1988), "Multi-pass Expert Control System—a Control Scheduling Structure for Flexible Manufacturing Cells," *Journal of Manufacturing Systems,* Vol. 7, No. 2, pp. 107–120.

Wu, S. D., and Wysk, R. A. (1989), "An Application of Discrete-Event Simulation to On-line Control and Scheduling in Flexible Manufacturing," *International Journal of Production Research,* Vol. 27, No. 9, pp. 1603–1623.

Yih, Y. (1988), "Trace-Driven Knowledge Acquisition (TDKA) for Rule-Based Real-Time Scheduling Systems," *Journal of Intelligent Manufacturing,* Vol. 1, No. 4, pp. 217–230.

Yih, Y. (1994), "An Algorithm for Hoist Scheduling Problems," *International Journal of Production Research,* Vol. 32, No. 3, pp. 501–516.

Yih, Y., and Jones, A. T. (1992), "Candidate Rule Selection to Develop Intelligent Scheduling Aids for Flexible Manufacturing Systems (FMS)," in *Proceedings of a Joint German/US Conference,* Hagen, Germany, pp. 201–217.

Yih, Y., Liang, T., and Moskowitz, H. (1993), "Robot Scheduling in a Circuit Board Production Line: A hybrid OR/ANN Approach," *IIE Transactions,* Vol. 25, No. 2, pp. 26–33.

Zhang, C., Yan, P., and Chang, T. (1991), "Solving Job-Shop Scheduling Problem with Priority Using Neural Network," in *IEEE International Joint Conference on Neural Networks,* pp. 1361–1366.

Zhang, M., and Zhang, C. (1995), "The Consensus of Uncertainties in Distributed Expert Systems," in *Proceedings of the First International Conference on Multi-Agent Systems,* MIT Press, Cambridge, MA.

Zhang, W., and Dietterich, T. (1996), "High-Performance Job-Shop Scheduling with a Time-delay TD(() network," *Advances in Neural Information Processing Systems,* MIT Press, Cambridge, MA, pp. 1025–1030.

Zhou, D. N., Cherkassky, V., Baldwin, T. R., and Olson, D. E. (1991), "A Neural Network Approach to Job-Shop Scheduling," *IEEE Transactions on Neural Networks,* Vol. 2, No. 1, pp. 175–179.

CHAPTER 67

Quality Tools for Learning and Improvement

Lloyd Provost
Associates in Process Improvement

1. MODEL FOR IMPROVEMENT (PDSA) 1808
2. TOOLS FOR VIEWING SYSTEMS AND PROCESSES 1809
3. TOOLS FOR GATHERING INFORMATION 1810
4. TOOLS FOR ORGANIZING INFORMATION 1813
5. TOOLS FOR UNDERSTANDING VARIATION 1821
6. TOOLS FOR UNDERSTANDING RELATIONSHIPS 1821
7. INTEGRATION OF THE TOOLS FOR IMPROVEMENT 1822
8. CASE STUDY: USING THE TOOLS TO IMPROVE 1823
ADDITIONAL READING 1827

1. MODEL FOR IMPROVEMENT

Improvement comes from the application of knowledge to relevant problems or opportunities. This knowledge may be knowledge of engineering, operations, complex theories, or simply experience of the way some activity is currently done. Generally, the more complete the appropriate knowledge, the better the improvements will be when applying the knowledge to making changes. Therefore, improvement must be based on building and applying knowledge. This view leads to the model for improvement in Figure 1 (Langley 1998). The model consists of three fundamental questions that define the opportunity and a method (the PDSA cycle) to increase the requisite knowledge.

This model provides a framework for a trial-and-learning approach to improvement. The word ‘‘trial’’ suggests a test of a change. The word ‘‘learning’’ implies that some criteria exist by which to study and learn from the trial. The focus on the questions accelerates the building of knowledge by emphasizing a framework for learning, the use of data, and the design of effective tests or trials. Learning from testing changes on a small scale is encouraged, rather than from extensive study of the problem before any changes are attempted. The PDSA (Plan, Do, Study, Act) cycle in the model provides the framework for an efficient trial-and-learning methodology. We refer to this cycle as the cycle for learning and improvement. The cycle begins with a plan and ends with action being taken based on the learning gained from the Plan, Do, and Study phases of the cycle.

Many improvement efforts can be successfully completed using only subject matter knowledge to answer the three questions of the model for improvement. In other cases, additional knowledge is needed to develop a change. What are some ways to obtain this knowledge? A number of different tools and methods are described in this Handbook and are summarized in this chapter. These methods and tools for improvement in this chapter are organized into five use categories:

1. Viewing Systems and Processes
2. Gathering Information
3. Organizing Information

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

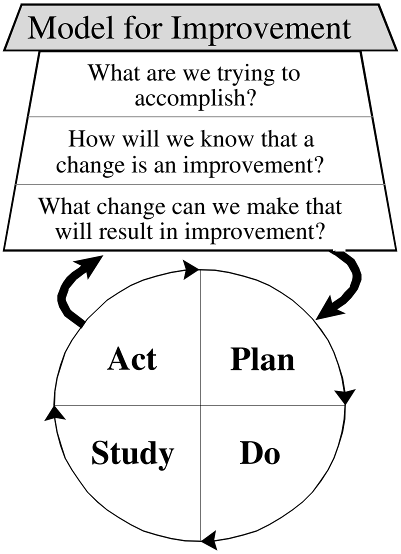

Model for Improvement
What are we trying to accomplish?
How will we know that a change is an improvement?
What change can we make that will result in improvement?
Act
Plan
Study
Do

Figure 1 The Model for Improvement. (Copyright 1980–1998 Associates in Process Improvement)

4. Understanding Variation
5. Understanding Relationships

The 19 improvement tools and methods described in this chapter are summarized in Table 1. Note that these tools are sometimes called by slightly different names in other references.

Figures 2 through 20 show pictorial examples of these tools/methods. Each tool or method is considered in light of five questions about the tool:

1. *What* is this tool?
2. *Why* would someone choose to use this tool?
3. What are important considerations when *planning* to use this tool?
4. What *other tools* might be useful in conjunction with this tool?
5. What are the basic *mechanics* of using this tool?

A case study at the end of the chapter illustrates the use of the tools in an improvement project.

2. TOOLS FOR VIEWING SYSTEMS AND PROCESSES

- *What?* The flow diagram (often called flowchart) is a graphic representation of a series of activities that define a process.
- *Why?* This tool is useful when there is a need to describe how a process is being carried out or should be carried out. In particular, the flow diagram is very useful when a team needs to understand how the process works.
- *Planning:* The scope of the process (beginning and end) must be defined. The developers of the flow diagram must agree on the use of symbols and the level of detail needed. These choices will be guided by the purpose of the flowchart.
- *Other tools:* The flow diagram can help plan where to collect data. Use of the flow diagram may help initiate using the tools for organizing information, such as the cause-and-effect diagram.
- *Mechanics:* Decide on the process stop–start points, level of detail, and symbols. Have those knowledgeable about the process construct the chart. There are a number of special versions of

TABLE 1 Overview of Methods and Tools for Improvement

Use Category	Method or Tool	Typical Use of Method or Tool
Viewing systems and processes	1. Flow diagram	Develop a picture of a process. Communicate and standardize processes.
	2. Linkage of processes	Develop a picture of a system composed of processes linked together.
Gathering information	3. Form for collecting data	Plan and organize a data-collection effort
	4. Surveys	Obtain information from people.
	5. Benchmarking	Obtain information on performance and approaches from other organizations.
	6. Creativity methods	Develop new ideas and fresh thinking.
Organizing information	7. Affinity diagram	Organize and summarize qualitative information.
	8. Force field analysis	Summarize forces supporting and hindering change.
	9. Cause-and-effect diagram	Collect and organize current knowledge about potential causes of problems or variation.
	10. Matrix diagram	Arrange information to understand relationships and make decisions.
	11. Tree diagram	Visualize the structure of a problem, plan, or any other opportunity of interest.
	12. Quality function deployment (QFD)	Communicate customer needs and requirements through the design and production processes.
Understanding variation	13. Run chart	Study variation in data over time; understand the impact of changes on measures.
	14. Control chart	Distinguish between special and common causes of variation.
	15. Pareto chart	Focus on areas of improvement with greatest impact.
	16. Frequency plot	Understand location, spread, shape, and patterns of data.
Understanding relationships	17. Scatterplot	Analyze the associations or relationship between two variables; test for possible cause and effect.
	18. Two-way table	Understand cause and effect for qualitative variables.
	19. Planned experimentation	Design studies to evaluate cause-and-effect relationships and test changes.

Copyright 1980–1998 Associates in Process Improvement.

flow diagrams (for example, complexity flow diagram, deployment flowchart, and top-down flow diagram).

- *What?* The linkage of processes (LOP) is a method to describe a system composed of processes linked together to accomplish a common purpose.
- *Why?* The LOP can be used to develop a systems view of an organization or a group of processes. A LOP can be used to explore the interdependencies and complexities in a system or subsystem.
- *Planning:* The boundaries of the system must be defined; the level of process detail should be uniform. This will be determined by the purpose of the LOP. People knowledgeable about the entire system will be needed to build a LOP.
- *Other Tools:* The linkage can be used as a data collection form. Parts of a LOP can be expanded using flow diagrams. The LOP provides a structure for using the tools to organize information.
- *Mechanics:* Decide on purpose and system boundaries; define processes at a uniform level of detail (initially target about 20-80 processes); assemble related processes and show linkages between processes.

3. TOOLS FOR GATHERING INFORMATION

- *What?* The form for collection of data provides an organized method to record observations or measurements to be used in the process of analysis.

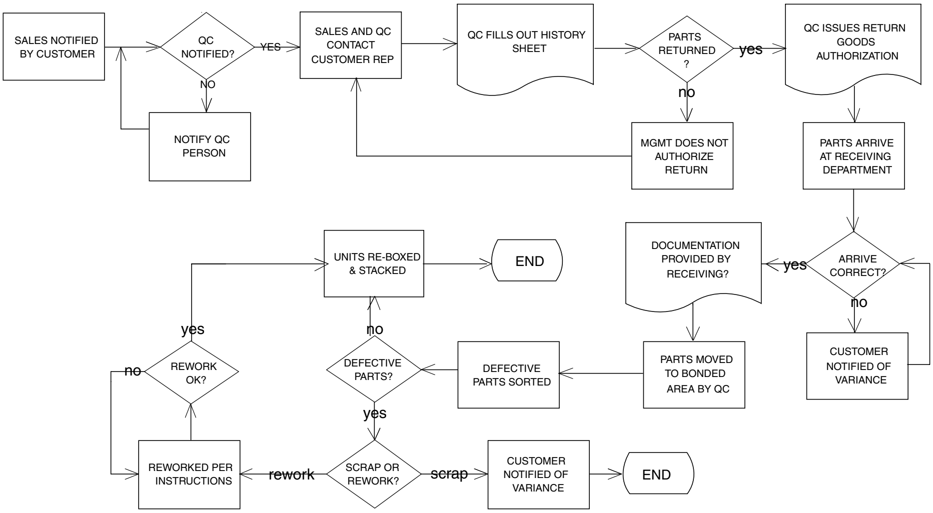

SALES NOTIFIED BY CUSTOMER
QC NOTIFIED?
YES
NO
NOTIFY QC PERSON
SALES AND QC CONTACT CUSTOMER REP
QC FILLS OUT HISTORY SHEET
PARTS RETURNED ?
yes
no
MGMT DOES NOT AUTHORIZE RETURN
QC ISSUES RETURN GOODS AUTHORIZATION
PARTS ARRIVE AT RECEIVING DEPARTMENT
ARRIVE CORRECT?
yes
no
CUSTOMER NOTIFIED OF VARIANCE
DOCUMENTATION PROVIDED BY RECEIVING?
PARTS MOVED TO BONDED AREA BY QC
DEFECTIVE PARTS SORTED
DEFECTIVE PARTS?
no
yes
UNITS RE-BOXED & STACKED
END
SCRAP OR REWORK?
scrap
rework
CUSTOMER NOTIFIED OF VARIANCE
END
REWORKED PER INSTRUCTIONS
REWORK OK?
yes
no

Figure 2 Flow Diagram. (Copyright 1980–1998 Associates in Process Improvement)

- *Why?* This tool is used to facilitate the recording of data and, in some cases, begin the analysis of data.
- *Planning:* It is important to have a clear understanding of what we want to learn from the data and who will collect and analyze it. The process of planning the form may force a more thorough consideration of the issues. Two types of forms are the check sheet and the recording form.
- *Other tools:* A form for collecting data is often built into other data analysis tools such as a control chart. All of the other data analysis tools will require some sort of data collection form for capturing observations or measurements.
- *Mechanics:* Decide on the questions to be answered. Decide on exactly what observations or data will be recorded and what format will be most useful. Operational definitions of terms can be important on a form for data collection.
- *What?* A survey is a method of collecting information directly from people about their feelings, motivations, plans, beliefs, experiences, and backgrounds.
- *Why?* As part of a PDSA cycle, a survey is a type of data-collection process that focuses on getting information from people to answer a question(s) posed in the planning phase of the cycle.
- *Planning:* Determine the objective of the survey. Consider why a survey is the most appropriate method for obtaining the desired information. Determine what questions are to be answered by the survey.
- *Other tools:* Any of the data-analysis tools may be used with data from surveys. Benchmarking is a special kind of survey. Surveys are required for a QFD analysis.
- *Mechanics:* Surveys can be administered in a number of ways: written surveys, personal interviews, group interviews, observations, and trading places.
- *What?* A process of measuring products, services, and business practices against the toughest competitors or those companies recognized as industry leaders.
- *Why?* Benchmarking, in its simplest form, is merely looking around at how others are doing things and trying to learn new approaches and possibilities. We can all benefit from doing this, and most organizations are already doing this on an informal basis. Benchmarking provides a formal method, with some structure, for making these observations and then using this information for improvement.
- *Planning:* Successful benchmarking efforts are organized using the PDSA cycle. A typical benchmarking exercise would be organized into four or five learning cycles.
- *Other tools:* Benchmarking is a type of survey. All of the other methods and tools can be used to organize and analyze data from benchmarking studies.
- *Mechanics:* The 10-step benchmarking process can be organized using the Model as a series of PDSA Cycles.

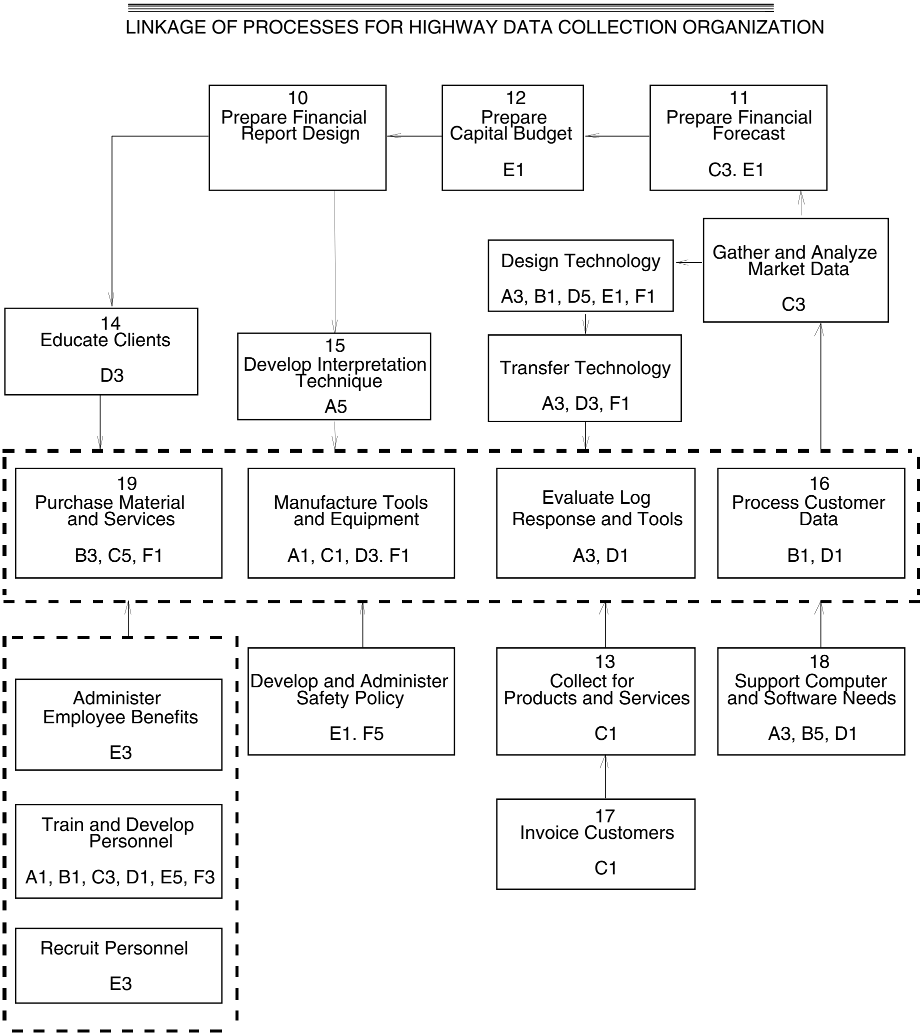

LINKAGE OF PROCESSES FOR HIGHWAY DATA COLLECTION ORGANIZATION
10 Prepare Financial Report Design
12 Prepare Capital Budget E1
11 Prepare Financial Forecast C3. E1
Design Technology A3, B1, D5, E1, F1
Gather and Analyze Market Data C3
14 Educate Clients D3
15 Develop Interpretation Technique A5
Transfer Technology A3, D3, F1
19 Purchase Material and Services B3, C5, F1
Manufacture Tools and Equipment A1, C1, D3. F1
Evaluate Log Response and Tools A3, D1
16 Process Customer Data B1, D1
Administer Employee Benefits E3
Develop and Administer Safety Policy E1. F5
13 Collect for Products and Services C1
18 Support Computer and Software Needs A3, B5, D1
Train and Develop Personnel A1, B1, C3, D1, E5, F3
17 Invoice Customers C1
Recruit Personnel E3

Figure 3 Linkage of Processes. (Copyright 1980–1998 Associates in Process Improvement)

- *What?* A collection of tools and methods (provocations, random entry, six thinking hats, concept triangle, etc.) based on creative thinking that fit with traditional quality improvement methods. The methods are serious, deliberate, and systematic; they do not rely on acting crazy or pure natural talent.
- *Why?* Based on theory and knowledge of how the brain works as an active, self-organizing information system, the lateral (creative) thinking tools are designed to allow people to deliberately produce thoughts that are outside their normal thinking patterns. This in turn greatly increases the chances of producing new ideas, new concepts, and new perceptions from old situations.
- *Planning:* The tools should be used when new ideas or concepts are required to accomplish the improvement initiative.
- *Other tools:* The creativity methods provide a refreshing alternative to the other methods to gather information. Tools for organizing information can be used with information collected using the creativity methods.

Problem	Occurrences				
	Monday	Tuesday	Wednesday	Thursday	Friday
Out of paper					
Out of toner					
Copies too light					
Sorter problems					
Document feeder					
Transparency feed					
Other: Copier stopped					
Panel won't clear					
No power					

Figure 4 Forms for Collection of Data. (Copyright 1980–1998 Associates in Process Improvement)

- *Mechanics:* Provocations seek to jolt us or start us outside of these mainstream patterns so that we can then increase the probability of connecting with other patterns to produce new ideas, concepts and perceptions.

4. TOOLS FOR ORGANIZING INFORMATION

- *What?* The affinity diagram is a method to summarize qualitative data into groups with common themes.
- *Why?* The affinity diagram should be used when there are large amounts of qualitative data that must be summarized before proceeding. An example is the answers to open ended questions on a survey.
- *Planning:* Much will depend on the type of qualitative data. There may be many useful ways to group the data, so being creative and open-minded is helpful.

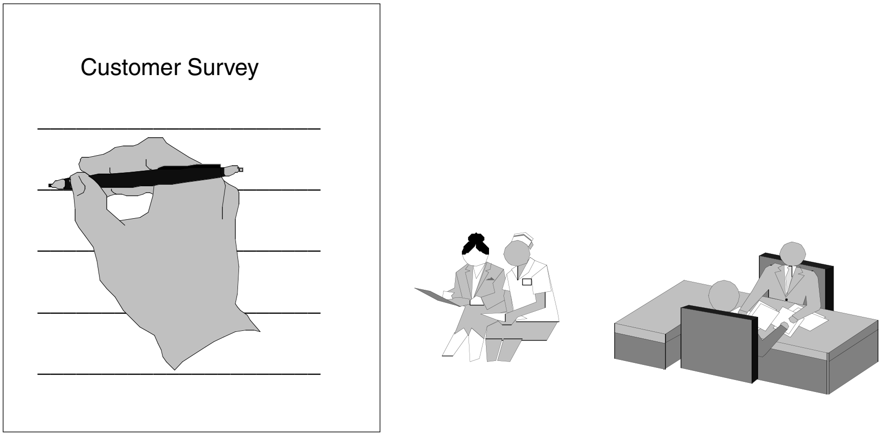


Figure 5 Surveys. (Copyright 1980–1998 Associates in Process Improvement)

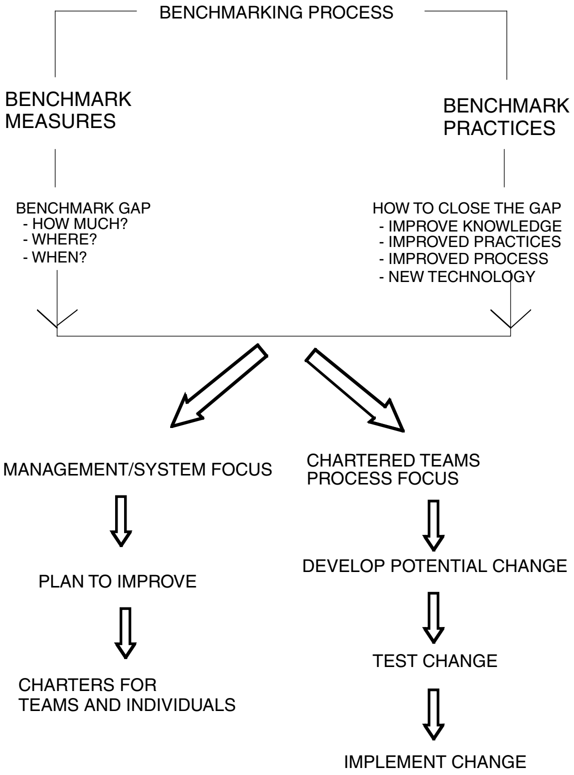

BENCHMARKING PROCESS
BENCHMARK MEASURES
BENCHMARK PRACTICES
BENCHMARK GAP
- HOW MUCH?
- WHERE?
- WHEN?
HOW TO CLOSE THE GAP
- IMPROVE KNOWLEDGE
- IMPROVED PRACTICES
- IMPROVED PROCESS
- NEW TECHNOLOGY
MANAGEMENT/SYSTEM FOCUS
CHARTERED TEAMS PROCESS FOCUS
PLAN TO IMPROVE
DEVELOP POTENTIAL CHANGE
CHARTERS FOR TEAMS AND INDIVIDUALS
TEST CHANGE
IMPLEMENT CHANGE

Figure 6 Benchmarking. (Copyright 1980–1998 Associates in Process Improvement)

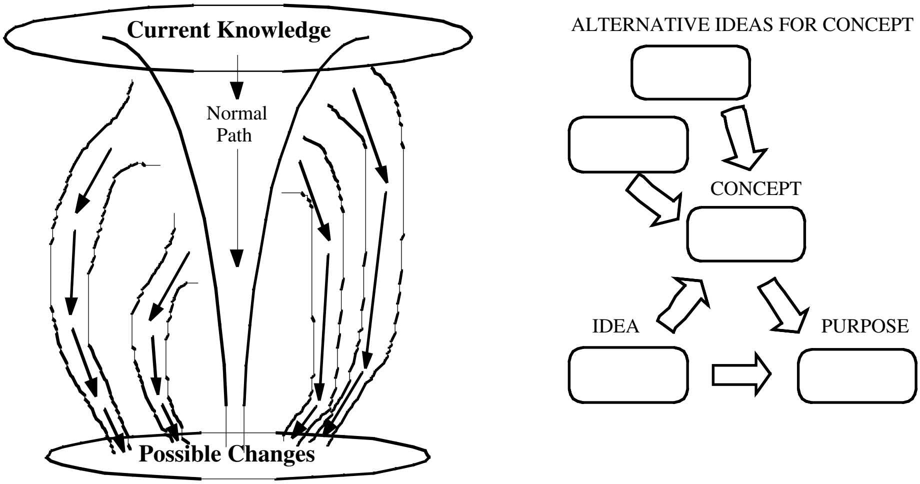

Current Knowledge
Normal Path
Possible Changes
ALTERNATIVE IDEAS FOR CONCEPT
CONCEPT
IDEA
PURPOSE

Figure 7 Creativity Methods. (Copyright 1980–1998 Associates in Process Improvement)

Each manager is doing the next level job	Products routinely miss the market window	Customer complaints about product consistency
Managers are task oriented	Increase in unprofitable products	Customer complaints about product consistency
Employee grievances have increased 14%	Change over from one product campaign to another is difficult and time consuming	Product variability causes downtime of equipment & maintenance PM time increase
Turnover rate is high - 23%	Costs associated with product introduction are typically 30% over estimates	Customer inspection results do not match our documented results
Decision making process is cumbersome		
Globalization of market requires managers that appreciate other cultures		

Figure 8 Affinity Diagram. (Copyright 1980–1998 Associates in Process Improvement)

- *Other tools:* Tools for gathering information that result in qualitative data can lead to an affinity analysis. After constructing an affinity diagram, the results may be summarized in a Pareto chart. Categories developed during an affinity analysis can be used for stratification with the other tools.
- *Mechanics:* Make a clear statement of the issue to be considered; record the qualitative data on cards; sort into related groups; create the main subgroup themes.
- *What?* The force field analysis is a method to analyze situations that support and inhibit a planned change.
- *Why?* The force field analysis is typically used in planning to overcome restraining forces and reinforce driving forces when implementing a change.
- *Planning:* Develop a clear definition of the planned change. The group will need knowledge of the organization as well as its outside environment.
- *Other tools:* The force field analysis is a planning tool. Information to do the analysis can come from any of the methods to gather information.

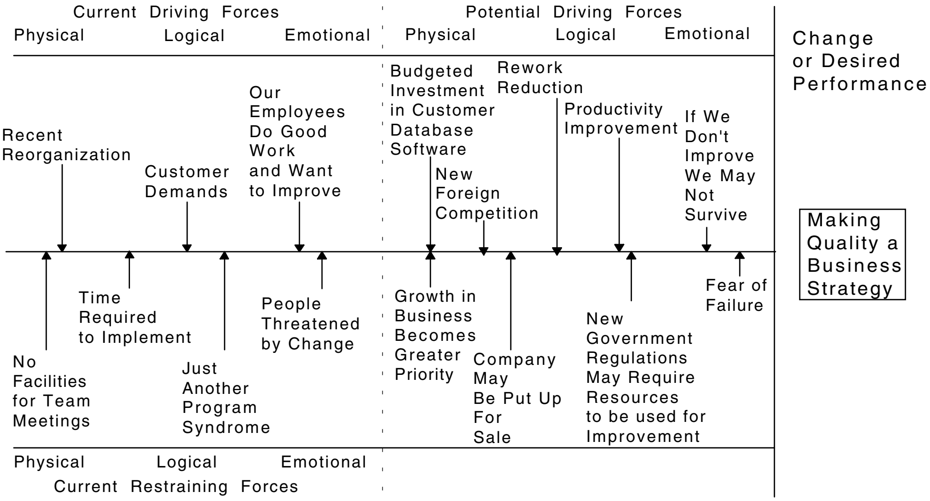

Current Driving Forces
Physical
Logical
Emotional
Potential Driving Forces
Physical
Logical
Emotional
Change or Desired Performance
Recent Reorganization
Customer Demands
Our Employees Do Good Work and Want to Improve
Budgeted Investment in Customer Database Software
Rework Reduction
New Foreign Competition
Productivity Improvement
If We Don't Improve We May Not Survive
Making Quality a Business Strategy
No Facilities for Team Meetings
Time Required to Implement
Just Another Program Syndrome
People Threatened by Change
Growth in Business Becomes Greater Priority
Company May Be Put Up For Sale
New Government Regulations May Require Resources to be used for Improvement
Fear of Failure
Physical
Logical
Emotional
Current Restraining Forces

Figure 9 Force Field Analysis. (Copyright 1980–1998 Associates in Process Improvement)

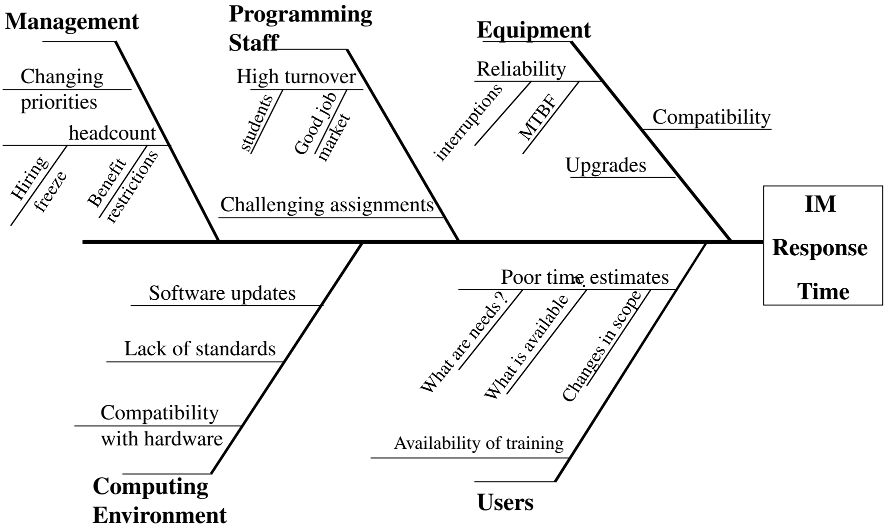

Management
Changing priorities
headcount
Hiring freeze
Benefit restrictions
Programming Staff
High turnover
students
Good job market
Challenging assignments
Equipment
Reliability
interruptions
MTBF
Compatibility
Upgrades
IM Response Time
Software updates
Lack of standards
Compatibility with hardware
Computing Environment
Poor time estimates
What are needs?
What is available
Changes in scope
Availability of training
Users

Figure 10 Cause-and-Effect Diagram. (Copyright 1980–1998 Associates in Process Improvement)

- *Mechanics:* Develop a clear statement of the change to be considered, brainstorm restraining, and reinforcing forces. Place these into categories, if useful.
- *What?* The cause-and-effect diagram (C&E) is a tool for organizing a group's current knowledge about causes of a problem or variation in a quality characteristic. The tool is also called a fishbone diagram or an Ishikawa diagram.
- *Why?* It is useful for a group to share knowledge when each member has a different perspective on a problem, e.g., management, operations, maintenance, and accounting. The C&E diagram can be updated and serve as a method to organize a team's current knowledge.
- *Planning:* It is important to define the problem or issue clearly. Putting the diagram together requires people with all of the knowledge relevant to the situation working together and sharing ideas.
- *Other tools:* Any of the data-analysis tools might be useful after it is decided what should be measured using the C&E diagram. Causes on the C&E are likely to show up on a flowchart as

	Criteria for Selection						
Computer Type	Disk size	Memory size	Processing speed	Features	Standard software	Costs	**Total Score**
Brand A	1	2	4	3.5	1	2.5	14
Brand B	4	4	3	3.5	3	2.5	20
Brand C	2	1	1	2	2	1	9
Brand D	3	3	2	1	4	4	18

Figure 11 Matrix Diagram. (Copyright 1980–1998 Associates in Process Improvement)

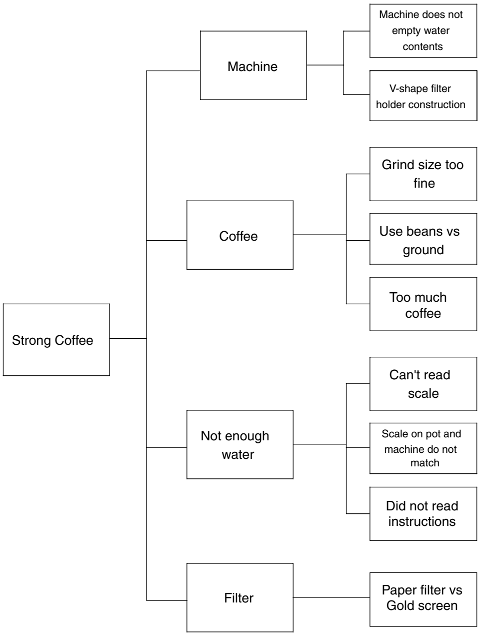

Strong Coffee
Machine
Machine does not empty water contents
V-shape filter holder construction
Coffee
Grind size too fine
Use beans vs ground
Too much coffee
Not enough water
Can't read scale
Scale on pot and machine do not match
Did not read instructions
Filter
Paper filter vs Gold screen

Figure 12 Tree Diagram. (Copyright 1980–1998 Associates in Process Improvement)

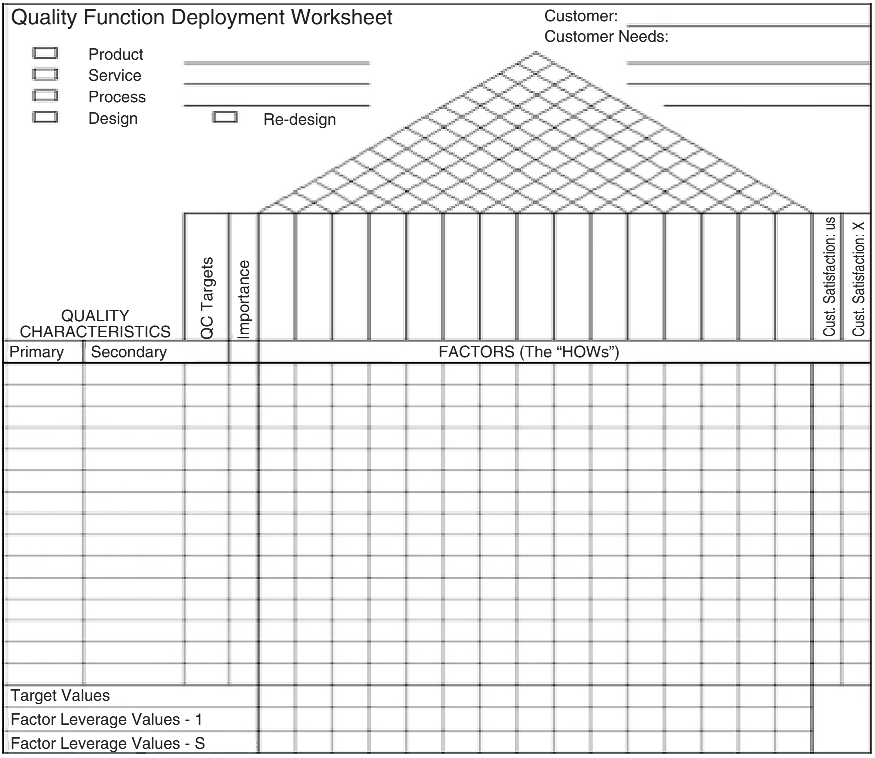

Quality Function Deployment Worksheet
Product
Service
Process
Design
Re-design
Customer:
Customer Needs:
QUALITY CHARACTERISTICS
QC Targets
Importance
Cust. Satisfaction: us
Cust. Satisfaction: X
Primary
Secondary
FACTORS (The "HOWs")
Target Values
Factor Leverage Values - 1
Factor Leverage Values - S

Figure 13 Quality Function Deployment. (Copyright 1980–1998 Associates in Process Improvement)

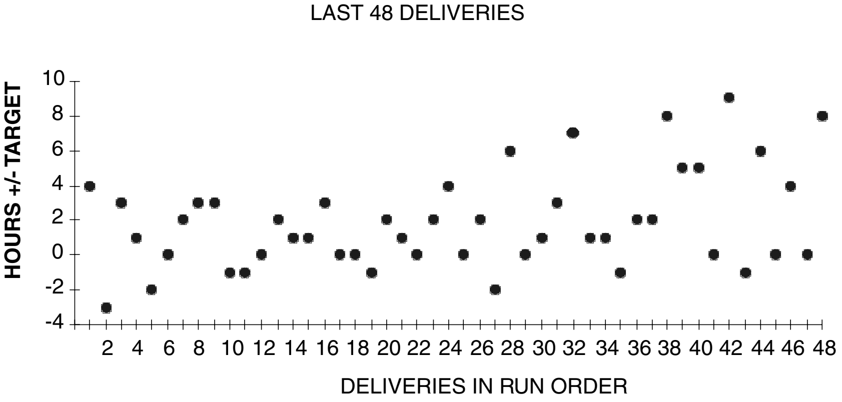


Figure 14 Run Chart (Trend Chart). (Copyright 1980–1998 Associates in Process Improvement)

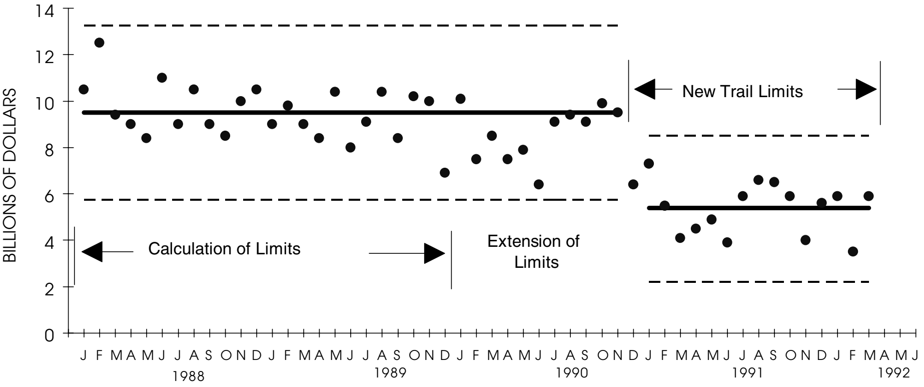


Figure 15 Control Chart (Shewhart Chart). (Copyright 1980–1998 Associates in Process Improvement)

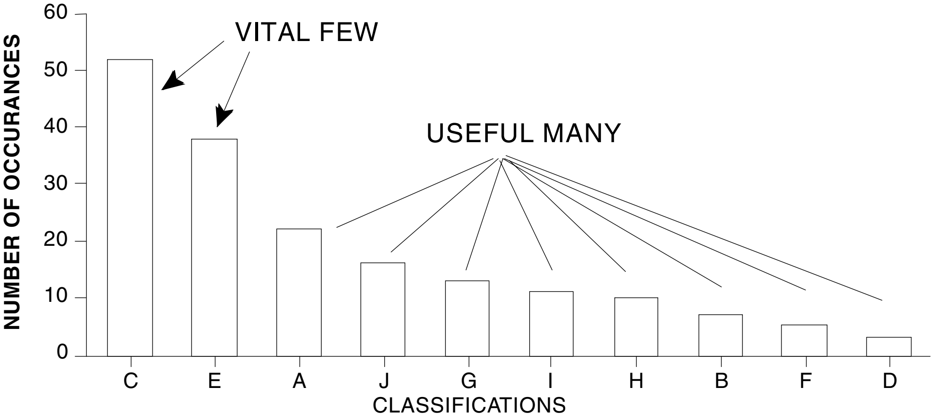


Figure 16 Pareto Chart. (Copyright 1980–1998 Associates in Process Improvement)

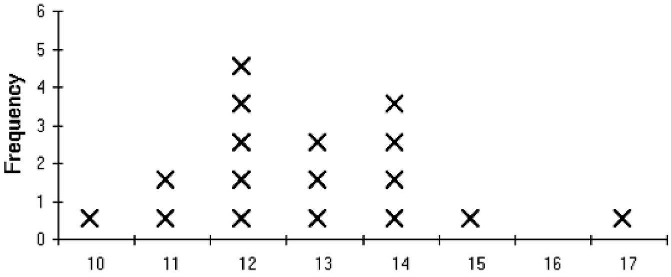


Figure 17 The Frequency Plot. (Copyright 1980–1998 Associates in Process Improvement)

complexity or awkward stages of a process. A C&E diagram is useful when planning an experiment to identify factors and background variables.

- *Mechanics:* Define the issue, brainstorm the causes, and organize the causes into categories. Display the categories on a diagram.
- *What?* A matrix diagram is a method used to arrange data to help the user understand important relationships. The diagram displays the relationship between two groupings (e.g., steps in a process and departments, customer needs and features offered with your service, vendors and selection criteria).
- *Why?* The matrix diagram has many uses in improvement efforts. It is especially useful for developing a change and deciding where to test it. Some of the typical uses include showing the changes planned for a process and where in the organization they will have an impact; summarizing information to help make a decision; and showing the relationship between factors and measure of quality in a planned experiment.
- *Planning:* Understand the information available and the relationships that are important.
- *Other tools:* Information from any of the methods for gathering information can be summarized using a matrix diagram.
- *Mechanics:* Establish lists of items whose relationship is under study. Place one list across the top of the matrix and the other down the side. Draw the matrix. Pick the data or relationship

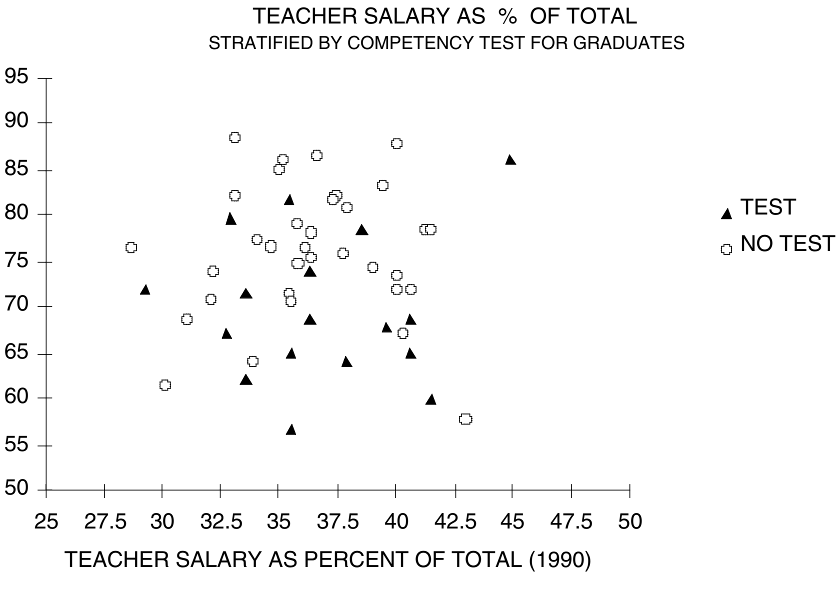


Figure 18 Scatterplot. (Copyright 1980–1998 Associates in Process Improvement)

	New System	
Current System	**on-time**	**late**
on-time	**28**	**2**
late	**16**	**4**

Figure 19 Two-Way Table. (Copyright 1980–1998 Associates in Process Improvement)

symbols to be used. Place the relevant symbols or data on the matrix that corresponds to the proper relationship.

- *What?* A tree diagram is used to visualize the structure of a problem, a plan, or any other opportunity of interest. The diagram helps you think systematically about each aspect of the problem. It has also been called a systematic diagram. The tree diagram allows the graphical view of different level of details about a problem.
- *Why?* The diagram allows a system of strategies for solving a problem or means of achieving an objective to be developed systematically and logically, making it less likely that any essential items will be omitted. The diagram facilitates dialogue and agreement among group members. Tree diagrams are convincing in presentations because they identify and clearly display the details of complex issues.
- *Planning:* Understand the information available and which of the many varieties of tree diagrams is appropriate for the opportunity of interest.
- *Other tools:* Information from any of the methods for gathering information can be summarized using a tree diagram. A tree diagram is similar to a cause-and-effect diagram when causes of an event are being evaluated. Standard symbols are used with tree diagrams for applications like fault tree analysis or failure mode and effects analysis (FMEA).
- *Mechanics:* A tree diagram is simple and natural to construct and thus is a common tool used to organize information. This is done by developing the branches on the tree into different levels of detail. The tree can be developed either horizontally or vertically on the page.
- *What?* Quality function deployment (QFD) is a tool for organizing information when designing a process or product.
- *Why?* QFD is especially useful as an interdisciplinary method that takes customer-centered marketing information and translates this into choices for the organization.
- *Planning:* The group developing the QFD will need marketing information on quality characteristics with customer weights. Also, operations knowledge will be needed to relate quality characteristics to process choices.
- *Other tools:* QFD is typically a jumping-off point for a new product or other major change. Often QFD will establish the need for planned experimentation or customer surveys.
- *Mechanics:* Create a matrix; enter quality characteristics on one side. Enter process choices on the other side; weight the effects. Identify interrelationships. Note where tests and new information are needed.

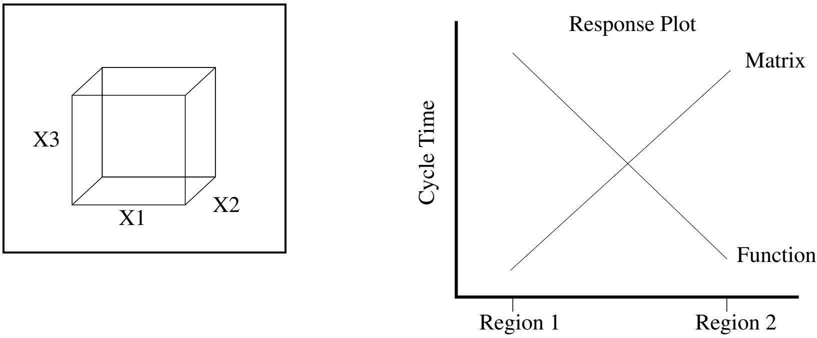


Figure 20 Planned Experimentation. (Copyright 1980–1998 Associates in Process Improvement)

5. TOOLS FOR UNDERSTANDING VARIATION

- *What?* The run chart is a graphical record of a quality characteristic measured over time.
- *Why?* The run chart is useful when it is important to study variation over time. Studying the order of the data enables important information to be learned from measurements.
- *Planning:* Scale the chart appropriately for the variation of interest. Choose a useful interval of time.
- *Other tools:* A run chart can be converted into a control chart after sufficient data have been collected. A run chart can be collapsed into a frequency plot to study the spread and shape of the data.
- *Mechanics:* Scale the chart such that the expected variation will take up about 75% of the chart. (Sometimes this will not be possible on the first try if not much is known about the variation in the process.) Record the measurements directly on the chart in the time order they are generated.
- *What?* A control chart is a tool for studying variation in data, distinguishing between common cause and special cause variation.
- *Why?* To decide on a course of action for improving a process and learn about the extent of variation and its degree of predictability; to evaluate the effectiveness of a change; and study dynamic complexity of a system.
- *Planning:* What subgrouping strategy will help us to learn the most about a process? How often should measurements be made? What is our reaction plan for out-of-control points?
- *Other tools:* A frequency plot would yield useful information about the spread and shape of the data. The flowchart might be helpful in understanding where the measurement was taken, and for what purpose. The cause-and-effect diagram of the process might help in understanding causes of variation. A scatter plot can be used to study measurements from different control charts.
- *Mechanics:* Define the variable and subgrouping; collect and plot the data; calculate control limits; identify special causes and learn from them.
- *What?* The Pareto chart is a tool for helping to focus our efforts by identifying the relative importance of certain categories of events.
- *Why?* This tool should be used when we want to focus an improvement effort on those areas that will have the greatest impact.
- *Planning:* Develop useful definitions for the categories of observations. Select the number of occurrences of each category or another useful measure (time, cost, etc.) of importance.
- *Other tools:* The cause-and-effect diagram may give useful information regarding which categories to measure. A Pareto chart is often used to analyze further a characteristic evaluated using an attribute control chart.
- *Mechanics:* The X-axis categories of the chart need to be well defined and useful. The Y-axis should be a measure of what is most important. Often this is frequency of occurrence, but it may be time, cost, or other meaningful measures.
- *What?* The frequency plot (or histogram) is a tool to display data. It presents to the user basic information about the location, shape, and spread of a set of data. The frequency plot is similar to the Pareto chart. The frequency plot displays quantitative data, while the Pareto chart displays categorical data.
- *Why?* This tool should be used when there is a need to understand the spread, location, and/or shape of the data. It is useful for summarizing a large amount of data.
- *Planning:* Have a clear objective for constructing the plot. What is the likely spread of the data? Is your scale appropriate?
- *Other tools:* A control chart provides important information about the stability of the process and should always be used when interpreting a frequency plot.
- *Mechanics:* The scale for the horizontal measure needs to be developed. Sometimes groupings of the values need to be determined. Individual values can be recorded or bars can be used to represent the frequency.

6. TOOLS FOR UNDERSTANDING RELATIONSHIPS

- *What?* The scatterplot is a tool for analyzing associations or relationships between two quantitative variables.
- *Why?* If a cause-and-effect relationship exists between the variables, the scatter plot will show this relationship.

- *Planning:* Which two variables may be associated? What is the range of data for both variables? Consider whether there might be a third variable causally related to one or more variables on the chart. Are the processes delivering the data stable?
- *Other tools:* Run charts and control charts would be important to determine whether the data is stable or trending. The cause-and-effect diagram could give insights as to which variables should be plotted together.
- *Mechanics:* Select the two variables. Record pairs of measures. Plot the data on a scale such that the range of variation takes up the full range of data. The axes should be of approximately equal length for each variable.
- *What?* The two-way table is a tabular representation of the relationship between pairs of variables or categories. (Similar to scatter plot.)
- *Why?* If a cause-and-effect relationship exists between the variables, the two-way table will show this relationship.
- *Planning:* Which two variables are associated? What are the levels of the two variables? Consider whether there might be a third variable causally related to one or both variables on the chart. Are the processes delivering the data stable?
- *Other tools:* The scatterplot can be used instead of a two-way table when quantitative data are available. Run charts and control charts would be important to determine whether the data are stable or trending. The cause-and-effect diagram could give insights as to which variables should be on a table together.
- *Mechanics:* Select the two variables. Record events. Develop range of the data or categories for the table. Place the data into the table. Evaluate the ratios across the tables.
- *What?* Planned experimentation is a set of tools for understanding the causes of variation in a variable of interest. It is of particular interest when there are several factors that all contributed considerably to the variation under study.
- *Why?* To study the important common causes of variation in a process. Planned experimentation is most effective when we want to learn about the effects of numerous factors in a study.
- *Planning:* What subject matter knowledge is available? How much time and expense can we devote to the study? Are the processes that are part of the study stable? How will the various variables in the process be treated during the study?
- *Other tools:* Control charts, cause-and-effect diagrams, scatter plots, matrices, run charts, and frequency plots are used to analyze data from planned experiments.
- *Mechanics:* Variables are classified as response variables, factors, or background variables. A design matrix is created for the factors. Background variables are held constant or grouped in blocks. Randomization and replication are used to minimize the impact of unknown variables. Graphical methods are used to summarize the response variables.

7. INTEGRATION OF THE TOOLS FOR IMPROVEMENT

When working with the tools used in improvement, it should be kept in mind that they form a set of tools. Sometimes a single tool will answer all of the important questions, but more often several tools may be needed together. It is a common mistake to reach for the ‘‘favorite tool’’ and expect it to be useful in all situations. One of the particularly useful aspects of using the model for improvement to guide the improvement effort is that it will encourage the careful consideration of the questions to be answered. Choosing the correct tool or tools to use will follow naturally from asking the right questions.

It may be helpful, when learning the tools and methods, to consider not only what each tool tells us but also how another tool might give us information that would complement that given by the first tool. For example, a flow diagram can give us useful information about how a process works and may suggest where improvements can be made. It may also show us where complexity in the process might be causing difficulties. We may see more clearly where data could be collected to learn about the process. This is all very useful information, but what other types of information might be helpful? The flowchart gives us no quantitative information from which we might learn. If it were important to understand the variation in a process, we might need to make use of control charts, scatter diagrams, histograms, or Pareto charts. One or several of these might be useful to answer questions about the performance of the process.

Stratifying the data may also enhance the tools of data analysis—that is, separating the data according to some logical grouping, such as by machine, shift, supplier, or method of operation.

There are some particularly strong connections between some of the tools that deserve special mention. The Pareto diagram should only be used with adequate knowledge about the stability of the characteristic being measured. If the process is stable, the Pareto diagram displays the important failure modes or problem classifications produced by the common cause system. If the process is unstable, then stratification of the data should be performed to separate the data obtained when special causes were present from data produced by common causes.

A similar issue arises with histograms. Adequate knowledge of stability is necessary to interpret the histogram. It is a common misuse of a histogram to display data that have a symmetric ‘‘normal’’ shape and imply that the data came from a stable process. An unstable process can also appear as a ‘‘normal’’ distribution on a frequency plot.

As you gain experience using the tools, it will become clear that the tools actually just formalize ways of thinking that we often use in ordinary life. For example, we all have some idea of how we go about a task, such as coming to work in the morning. If we were to write down the steps, using some agreed-upon symbols, we would have a flowchart. A similar statement could be made regarding each of the tools presented, even those requiring calculations. A control chart could be used to chart your arrival time at work. Special causes could be learned from, and the extent of common cause variation could be seen. Yet you probably already have some mental estimate of these, along with some reaction plan.

By using the tools, we will apply some formality and will be more effective in learning and making changes. Also, we will have a ready means to share the information with others. This is one of the great benefits of using the improvement tools, particularly in combination; they put information into a form that can be effectively shared and acted upon.

8. CASE STUDY: USING THE TOOLS TO IMPROVE

The following example illustrates how several of the tools can be used together, along with stratification, to improve a process. The ABC Distribution Company delivered several different types of products mostly within a five-state region. Discussions with the customers showed that on-time delivery and complete orders were their main concern. On-time generally meant that the delivery should arrive on the agreed-upon day. If, for some reason, the delivery could not be made as scheduled, the customers wanted to know ahead of time, if possible. One major customer was a grocery store that expressed a desire for the shipments to arrive within a window of ± two hours. This standard had not been met consistently. At the company’s planning session, management decided that improving arrival time would be a way of differentiating themselves from competitors.

Using the model for improvement as a guide, the team decided to try to improve the processes affecting delivery. To find out how the delivery processes had performed in the past, they gathered some historical data from recent shipments and constructed the frequency plot shown in Figure 21. From this chart it can be seen that although most of the shipments were close to the target time, there are nonetheless a considerable number of deliveries that strayed from target. The frequency plot gives the following information about the historical performance: (1) average, (2) spread, and (3) shape. This chart gives only a summary of what happened in the past.

The next question that the team asked was: How had the delivery performance at ABC varied over time? There was quite a lot of disagreement among the team members over this. Some remembered a day when “everything seemed to work better.” Others thought the problems had always been there but had just floated from one place to another. Fortunately, time order of the

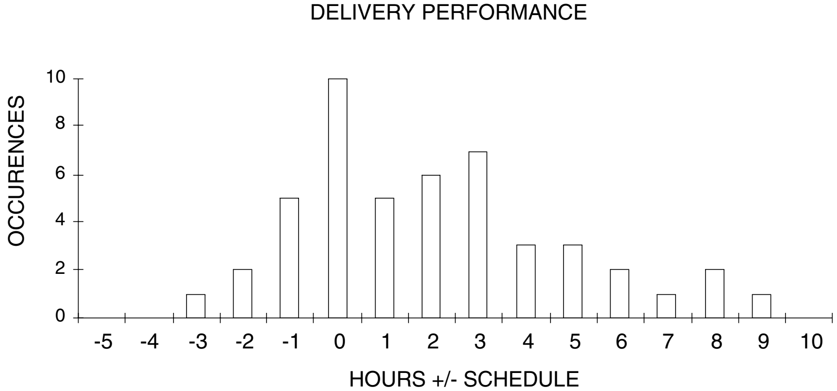


Figure 21 Frequency Plot of Time from Target. (Copyright 1980–1998 Associates in Process Improvement)

recent data used to construct the frequency plot had been retained. A run chart (Figure 22) was constructed using the same data as in the frequency plot from both warehouses over the last 46 deliveries.

There seemed to be a slight trend, and some of the team members thought there might be some increase in variation. In deciding what to do next, the team asked the question: What are some factors that could be contributing to variation in delivery performance?

The ABC business consisted of two warehouses located about 200 miles apart. Each carried about the same line of products, although there were some differences and some seasonal variations. Some of the team members thought the locations might be performing a little differently. It was decided to construct a run chart of the data but to stratify the data by warehouse 1 and 2. This run chart with stratification is shown in Figure 23.

A trend upward is visible from warehouse 2. The team decided to construct control charts to answer the question: Are the deliveries stable? The team was fairly sure about the trend in warehouse 2, but there was considerable uncertainty about warehouse 1. Control charts were constructed for each location. These are shown in Figures 24 and 25.

The trend seen in the run chart is clearly visible with the special cause signal, while warehouse 1 appears to be stable. The center lines for the two warehouses are also different. The team also constructed a scatterplot to see whether the results from the two warehouses on any particular week were related. The scatterplot in Figure 26 showed no relationship in results.

The team then more thoroughly considered the factors that might be affecting the delivery performance. There were several opinions that generally reflected the experience of the individual drivers and shipping personnel. One interesting theory was that the minor sideline-type products were a major source of delay in getting off on a run. When the run started late, more problems were likely to creep into the delivery, such as traffic delays and slow unloading. These sidelines were deliveries that were not part of the main customer deliveries and were things shipped only on some trips. These included some types of machinery, special building materials, and some clothing. The team decided to use a two-way table to investigate whether these deliveries might be tied to the late deliveries in a systematic manner. Figure 27 shows these data.

From this two-way table, it can be seen that these sideline deliveries are strongly associated with late deliveries. This can be seen by looking at the ratios across the table. For the on-time deliveries, 1 out of 16 were associated with special deliveries; for the late deliveries, 19 out of 30 were associated with special deliveries. Furthermore, the team looked into the portion of these deliveries from each warehouse and saw that location 2 had experienced an increase in these kinds of shipments, while location 1 had seen no change recently.

The team was now ready to make, develop, and test some changes. Management agreed that generally these sideline items were less time sensitive than the mainstay deliveries. This was verified in discussions with these customers. It was decided to split these types of deliveries into separate, smaller deliveries to be made in smaller trucks. For the test cycle, these trucks would be rented to try out this system. A month after the change was implemented, an updated control chart of the ABC delivery performance was constructed. Figure 28 shows the impact of these changes.

From this example, it can be appreciated that we need to be very careful about falling into the habit of trying to use only one tool to answer all questions. There should be no great difficulty

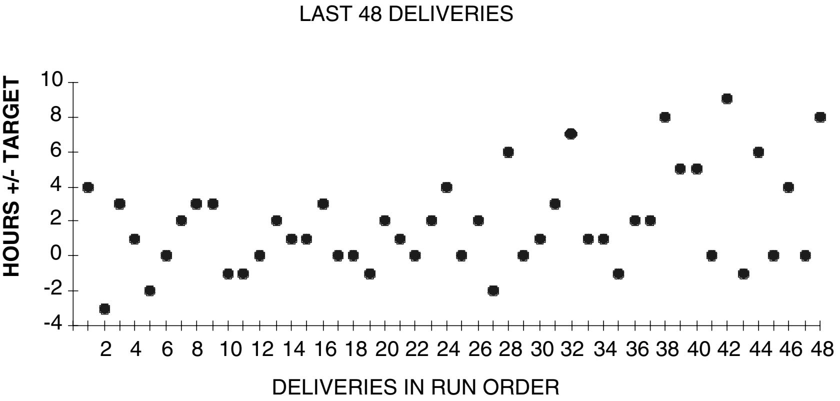


Figure 22 Run Chart of Delivery Data. (Copyright 1980–1998 Associates in Process Improvement)

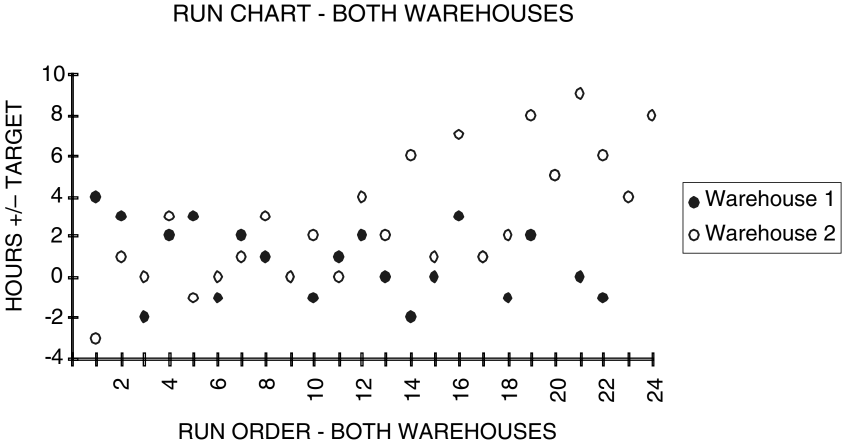


Figure 23 Run Chart Stratified by Location (Warehouse). (Copyright 1980–1998 Associates in Process Improvement)

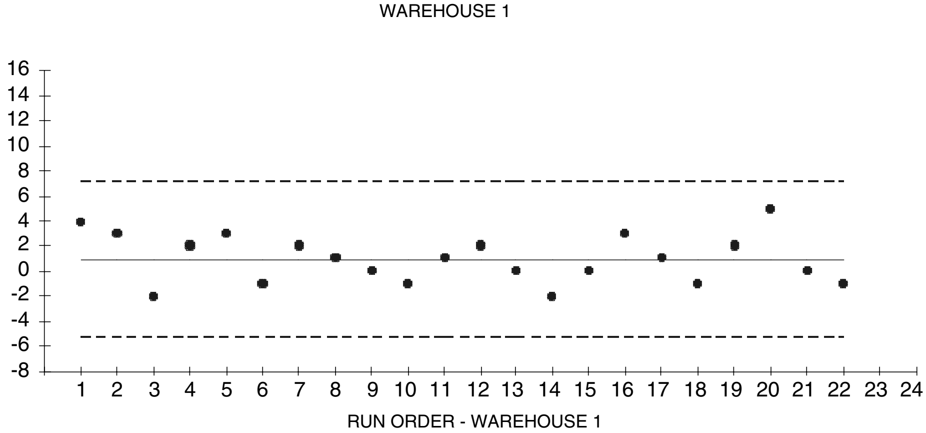


Figure 24 Control Chart of Warehouse 1. (Copyright 1980–1998 Associates in Process Improvement)

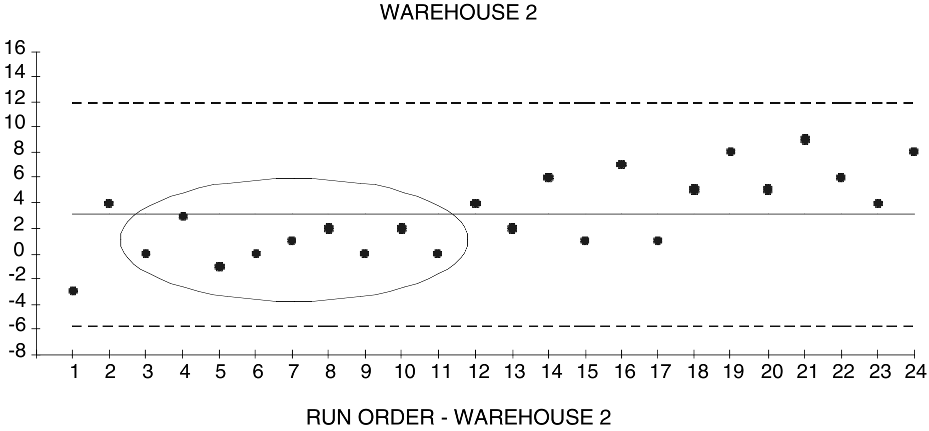


Figure 25 Control Chart of Warehouse 2. (Copyright 1980–1998 Associates in Process Improvement)

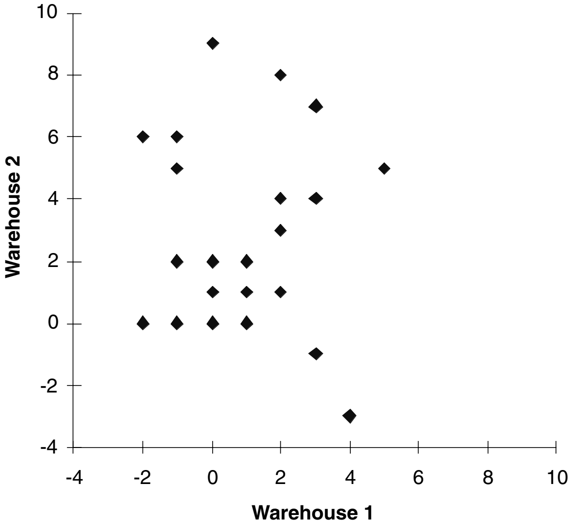


Figure 26 Scatterplot of Delivery Performance for the Two Warehouses. (Copyright 1980–1998 Associates in Process Improvement)

	Special Delivery	No Special Delivery	
On-Time	**1**	**15**	**16**
Late	**19**	**11**	**30**
	20	**26**	**46 = total**

Figure 27 Two-Way Table of Deliveries and Special Products. (Copyright 1980–1998 Associates in Process Improvement)

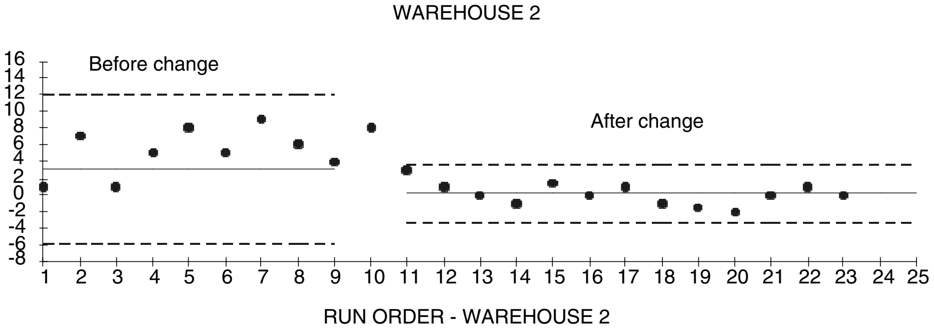


Figure 28 Control Chart of Deliveries after the Changes. (Copyright 1980–1998 Associates in Process Improvement)

imagining situations in which practically any combination of the tools would be used together. It is your subject matter knowledge of your business, along with knowledge of the tools, that will guide you to select the best tool for your situation. Using the model for improvement will also provide a helpful discipline and assist in selecting the best tool when help is needed to answer the three fundamental questions for improvement.

ADDITIONAL READING

Associates in Process Improvement (API), *The Improvement Handbook: Model, Methods, and Tools for Improvement*, API, Austin, TX, 1997.

Ishikawa, K., *Guide to Quality Control*, Asian Productivity Organization, 1982.

Langley, J., Nolan, T., and Nolan, K., ‘‘The Foundation of Improvement,’’ *Quality Progress,* June 1994, pp. 81–86.

Langley, G., Nolan, K., Nolan, T., Norman, C., and Provost, L., *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance,* Jossey-Bass, San Francisco.

Moen, R. D., and Nolan, T. W., ‘‘Process Improvement.’’ *Quality Progress,* September 1987, pp. 62–68.

CHAPTER 68

Understanding Variation

LLOYD PROVOST
Associates in Process Improvement

1. VARIATION IN DATA **1828**
2. APPLICATION OF THE CONCEPTS OF COMMON AND SPECIAL CAUSES **1830**
  2.1. Operation of Processes 1830
  2.2. Management of Processes 1830
  2.3. Improvement of Quality 1831
  2.4. Supervision and Leadership 1832
3. TOOLS FOR LEARNING FROM VARIATION IN DATA **1832**
4. SHEWHART CONTROL CHARTS **1834**
  4.1. Rationale for Shewhart Control Limits 1834
  4.2. Interpretation of a Control Chart 1835
  4.3. Control Charts for Different Data Types 1836
  4.4. Subgrouping and Stratification 1837
  4.5. Planning a Control Chart 1839
  4.6. Control Chart for Individual Measurements 1841
  4.7. Control Charts for Attribute Data 1844
    4.7.1. The P Chart for Classification Data 1844
    4.7.2. Control Charts for Count Data 1847
  4.8. X-bar and R Control Charts 1851
REFERENCES **1855**
ADDITIONAL READING **1855**

1. VARIATION IN DATA

Data from observations or measures will vary over time or location, and analysis of this variation is often used as a basis for action on the process. Sometimes this action is inappropriate or counterproductive because of a lack of understanding of the concept of common and special causes of variation.

One approach to analyze the performance of a process is to compare measures to an established standard, set of specifications, or customer requirements. The outcomes of the process can then be classified as acceptable or unacceptable. The unacceptable product is then scrapped, reworked, repaired, blended, or sold at a lower price. An unacceptable service usually requires rework, as well as management of an unhappy customer. This approach to the analysis of a process is an application of inspection. Inspection is useful to sort the good product or service from the bad, but without further analysis of the data, it provides no help in determining what should be done to improve the performance of the process.

A fundamental concept for the study and improvement of processes, due to Walter Shewhart (1931), is that variation in a measure of quality has its origins in one of two types of causes:

1. *Common causes:* those causes that are inherent in the process over time, affect everyone working in the process, and affect all outcomes of the process

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

2. *Special causes:* those causes that are not part of the process all the time or do not affect everyone but arise because of specific circumstances*

For example, the variation in cycle time in an assembly process is affected by causes common to the process and to all the workers in the process. Some possible examples of common causes of variation in cycle time are line speed, equipment reliability, staffing levels, complexity of orders, and supplier performance. If a high cycle time is due to these common causes, changes in the system by management will be required to reduce the times. If the high cycle times are due to special causes (e.g., broken belt, an absent worker, a fire in a supplier's plant), reduced cycle time will require specific actions by process workers and managers to remedy these issues. This example illustrates the importance of knowing whether the process is dominated by common or special causes before assigning responsibility for improvement. This example is used only to illustrate the concept. In practice, the distinction between common and special causes must be made with the aid of a control chart.

A process that has only common causes affecting the outcomes is called a *stable process,* or one that is in a state of statistical control. A stable process is one in which the cause system for the variation remains essentially constant over time. This does not mean there is no variation in the outcomes of the process, the variation is small, or the outcomes meet the requirements. A stable

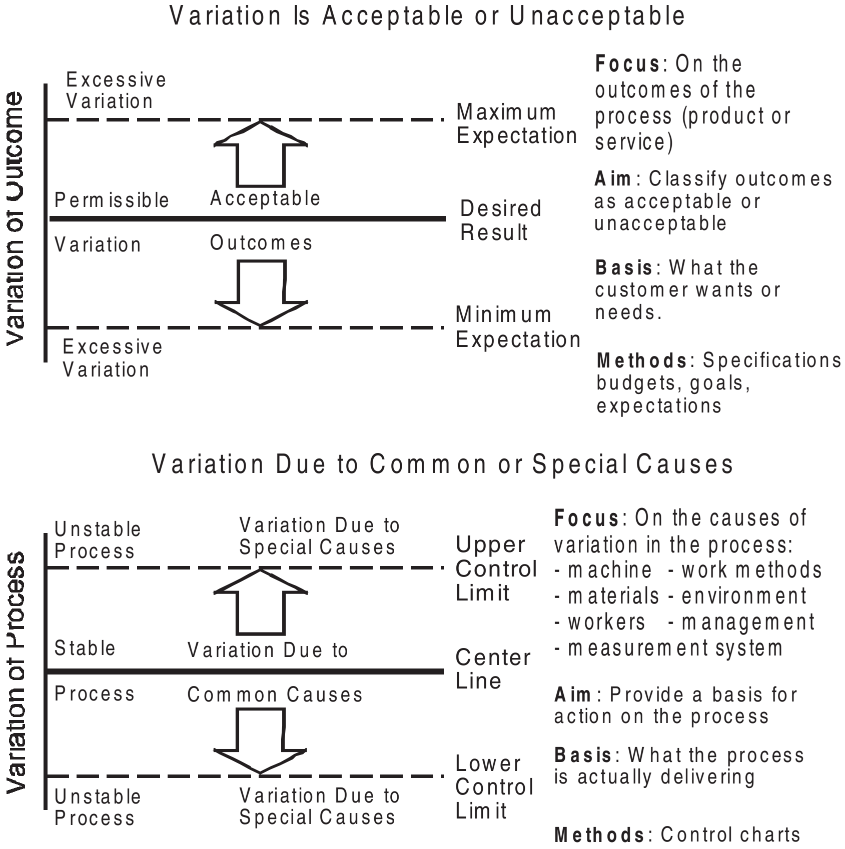

Variation Is Acceptable or Unacceptable
Variation of Outcome
Excessive Variation
Maximum Expectation
Permissible Variation
Acceptable Outcomes
Desired Result
Minimum Expectation
Excessive Variation
Focus: On the outcomes of the process (product or service)
Aim: Classify outcomes as acceptable or unacceptable
Basis: What the customer wants or needs.
Methods: Specifications budgets, goals, expectations
Variation Due to Common or Special Causes
Variation of Process
Unstable Process
Variation Due to Special Causes
Upper Control Limit
Stable Process
Variation Due to Common Causes
Center Line
Lower Control Limit
Unstable Process
Variation Due to Special Causes
Focus: On the causes of variation in the process:
- machine - work methods
- materials - environment
- workers - management
- measurement system
Aim: Provide a basis for action on the process
Basis: What the process is actually delivering
Methods: Control charts

Figure 1 Two Views of Variation. (Copyright 1980–1998 Associates in Process Improvement)

process implies only that the variation is predictable within statistically established limits. A process whose outcomes are affected by both common causes and special causes is called an *unstable process*. An unstable process does not necessarily mean one with large variation. It means that the magnitude of the variation from one time period to the next is unpredictable. The two views of variation are contrasted in Figure 1.

---

* Shewhart (1931) used the terms *assignable* and *chance* rather than *special* and *common* to describe these two types of causes. Deming (1986) popularized the latter two terms.

As special causes are identified and removed, the process becomes stable. Deming (1986, p. 340) gives several benefits of a stable process. Some of them are:

- The process has an identity; its performance is predictable.
- Costs and quality are predictable.
- Productivity is at a maximum and costs at a minimum under the present system.
- It is relatively easy to evaluate the effect of changes in the process.
- Stability provides a solid basis for altering specifications that cannot be met economically.

Besides providing the basic concepts, Shewhart also provided a tool, the Shewhart control chart, for determining whether a process is dominated by common or special causes. The control chart is the means to operationally define the concept of a stable process. There are many different types of control charts. The appropriate chart to use in a particular application depends in part on the type of data obtained from the process or product.

2. APPLICATION OF THE CONCEPTS OF COMMON AND SPECIAL CAUSES

Although Shewhart focused his initial work on manufacturing processes, the concepts of common and special causes and of stable and unstable processes have implications in many areas, including:

- Operation of processes
- Management of processes
- Improvement of quality
- Supervision and leadership

2.1. Operation of Processes

In manufacturing processes, it is often easy to make adjustments to the average of a process. It is a mistake to make these adjustments on the basis of inspection results without the aid of a control chart. For example, if a dimension of a machined part is inspected and is found to exceed the upper specification, an adjustment is made to the machine so that the average dimension of future parts is lowered. If a batch of a particular chemical is outside of specifications, an adjustment is made by changing the amount of catalyst added to the next batch. In both of these cases, there are circumstances in which the adjustments described will improve the performance of the process and circumstances in which the adjustment will result in even worse performance. It is vital that both managers and operators be able to distinguish between these two sets of circumstances. Fortunately, there is a simple way to do this.

Adjustment to reduce the variability of a stable process, that is, one whose output is dominated by common causes, will make the performance worse. Improvement of a stable process is achieved through a fundamental change in the process that results in the removal of some of the common causes. If a special cause is found and will persist for some time, for example a lot of raw material, an adjustment of the process to counteract the special cause may be helpful in the short term. The control chart is an important tool to help the operator know when an adjustment to the process is needed.

2.2. Management of Processes

Tools such as specifications, standards, forecasts, and budgets are useful for planning, pricing of product, and other functions of management. They are used to communicate what the customer or manager expects or wants from the process. It is important to keep in mind that they do not communicate reality, that is, they do not communicate how the process is doing or what it is capable of doing.

A control chart of important measures such as costs, material usage, volume of production, sales and profit, and an analysis of the capability of the process (if the process is stable) communicates a realistic view of the performance of the process. Without the aid of a control chart and an understanding of the concept of common and special causes of variation, the tools for planning are mistaken for reality or the capability of the process. Workers or other managers are often asked to conform to that "reality." If the salesman does not meet the forecast, his performance is unacceptable. When the production worker does not achieve the production standard, his performance is unacceptable.

When a manager compares a measure of performance of the process, such as costs or sales to a planning tool such as a forecast or standard, and uses this comparison as a basis for action, his actions are analogous to the operator adjusting the machine on the basis of specifications. Sometimes his actions will be appropriate, other times they will not. Just as in the case of the operator, there is a simple way to know which set of actions is appropriate.

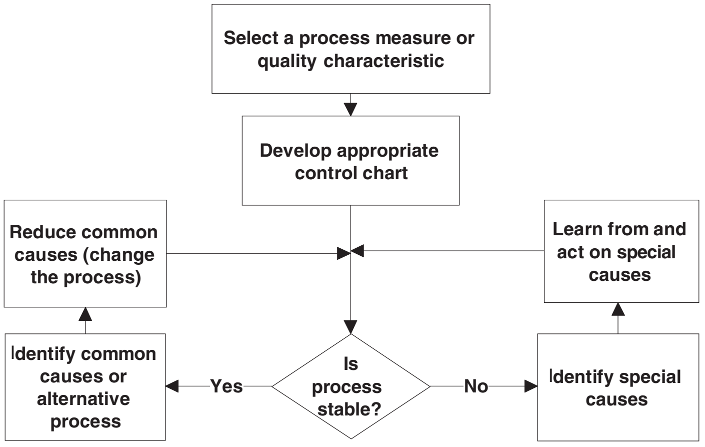
If the process is stable with respect to a particular measure of performance such as costs, then a fundamental change in the process, the responsibility of management, will be needed to reduce cost. Exhortations to lower-level managers or workers in the process to meet the forecast or standard will make things worse. Deming (1986) calls this type of action ‘‘tampering.’’ Webster’s dictionary defines ‘‘tamper’’ as follows:

- To interfere so as to weaken or change for the worse
- To alter for an improper purpose or in an improper way

It is vital in managing processes that planning tools are kept in their proper place and tools such as control charts and capability analysis are used as a basis for action on the process.

2.3. Improvement of Quality

To improve the quality of a process, it is useful to recognize whether the process is dominated by common causes or special causes. This will determine who is responsible for specific steps of improvement, what resources are needed, and what statistical tools will be useful (Figure 2). Since


Select a process measure or quality characteristic
Develop appropriate control chart
Reduce common causes (change the process)
Learn from and act on special causes
Identify common causes or alternative process
Yes
Is process stable?
No
Identify special causes

Stable Process	Unstable Process
Primary Methods for Learning	
1. Planned experimentation	1. Control charts
2. Subgrouping/stratification	2. Subgrouping/stratification
3. Control charts	3. Planned experimentation
Responsibility for Identification	
1. Technical experts	1. Workers in the process
2. Supervisors	2. Supervisors
3. Workers in the process	3. Technical experts
Responsibility for Improvement	
1. Management	1. Supervisors
2. Technical experts	2. Technical experts
3. Supervisors	3. Management
4. Workers in the process	4. Workers in the process

Figure 2 Using the Concepts of Variation to Guide Improvement. (Copyright 1980–1998 Associates in Process Improvement)

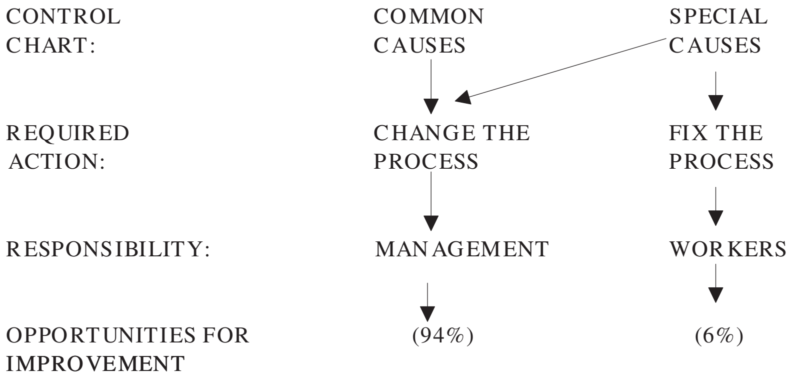


Figure 3 Opportunities for Improvement. (Copyright 1980–1998 Associates in Process Improvement)

unacceptable product or service can result from either common or special causes, the comparison of quality characteristics to requirements (product inspection) is not a basis for action on the process. Product inspection is useful to sort good products or services from bad and to set priorities on which processes to improve.

Activities to improve quality include the assignment of various people in the organization to work on common causes and special causes. The appropriate people to identify special causes are usually different than those needed to identify common causes. The same is true of those needed to remove causes. Removal of common causes is the responsibility of management, often with the aid of experts in the process such as engineers, chemists, and systems analysts. Special causes can frequently be handled at a local level by those working in the process, such as supervisors and operators. Without a knowledge of the concepts of common and special causes, it is difficult to allocate human resources efficiently to improve quality.

Many leaders of quality improvement have emphasized that most of the improvements in quality will take action by management. For example, Deming (1986) states that in almost all cases the removal of common causes will take a fundamental change in the process initiated by management. Some special causes can be removed by operators or supervisors. Others will require action by management in another process, possibly one of management or administration. For example, a special cause of variation in a production process may result when there is a change from one supplier's material to another. To prevent the special cause from occuring in the particular production process or other production processes, a change in the way the organization chooses and works with suppliers is needed. Figure 3 contains a summary of these concepts.

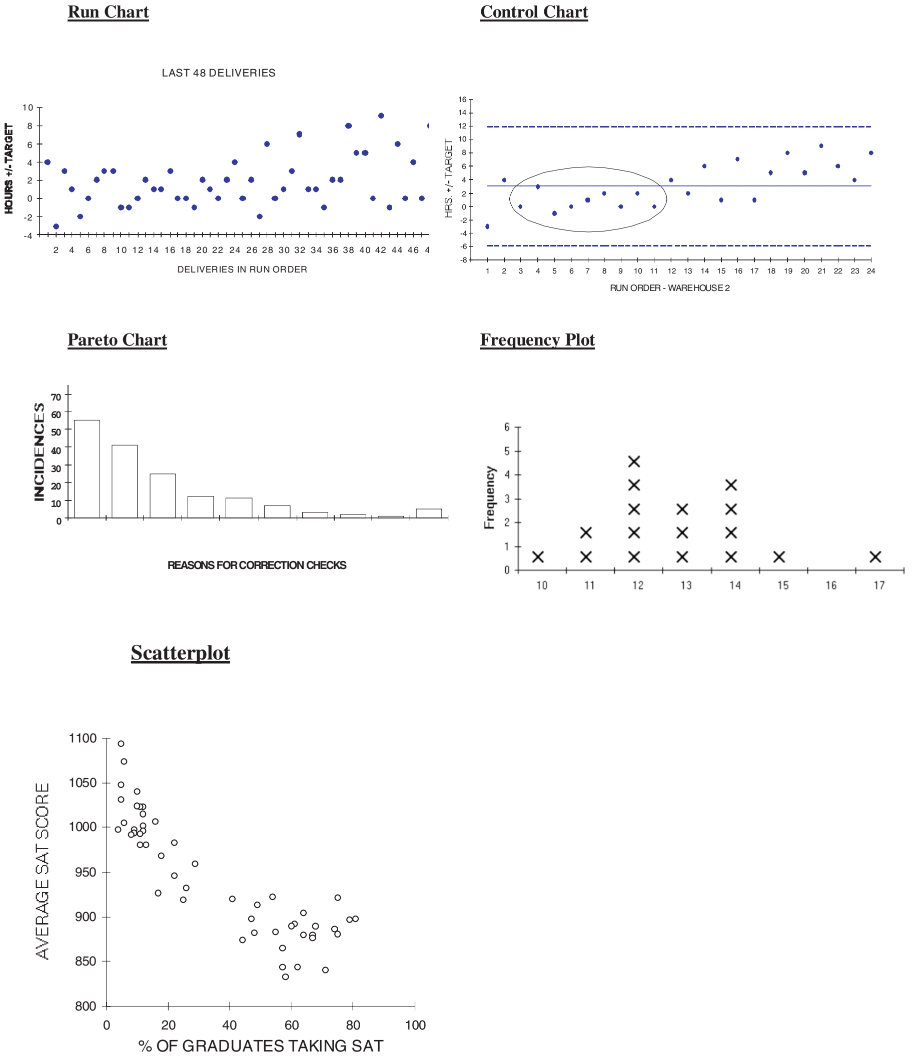
2.4. Supervision and Leadership

Another area in which the knowledge of common and special causes of variation is vital is in the supervision and leadership of people. A frequently made mistake is the assignment of faults of the process (common causes) to those working in the process, such as operators and clerks, rather than to those in charge of the process, management. It is obviously important for a supervisor or manager to know whether problems, mistakes, or rejected material are a result of common causes, special causes related to the system, or special causes related to the people under his or her supervision. Again, the use of a control chart will help the supervisor and manager accomplish this. For a thorough explanation of the role of statistical thinking in supervision, see Deming (1986).

3. TOOLS FOR LEARNING FROM VARIATION IN DATA

Some tools for learning from the variation in data are presented in Chapter 67. The primary tools are shown in Figure 4. Each of these tools looks at a particular aspect of the variation:

1. *Run charts:* View variation in data over time; study the impact of changes on measures.
2. *Control charts:* Distinguish between special and common causes of variation.
3. *Pareto charts:* Focus on areas of improvement with greatest impact.
4. *Frequency plots:* Understand location, spread, shape, and patterns of data.
5. *Scatterplots:* Analyze the associations or relationship between two variables.



Run Chart

Control Chart

Pareto Chart

Frequency Plot

Scatterplot

Figure 4 Tools to Learn from Variation. (Copyright 1980–1998 Associates in Process Improvement)

The run chart is simply a graphical record of a measure or characteristic plotted over time. Some type of run chart should always be a part of the study of variation in a process or system. The run chart focuses on dynamic complexity in a system (complexity over time) as well as the detail complexity of specific measures. The very simplicity of the chart is what makes it so powerful (Deming 1986). Everyone connected with the process can use and understand a run chart. Run charts are commonly used in business and economic documents.

The control chart (discussed extensively below) is an extension of the run chart. The control chart method provides a more formal way to learn from variation and guide the development of changes for improvement. The Shewhart control chart is a fundamental tool to guide improvement of processes.

The Pareto chart is a tool to help focus quality improvement efforts. It is useful whenever general classifications of problems, errors, defects, customer feedback, and so on can be classified for further study and actions. Often a few (the "vital few") classifications dominate the problem of interest while all the rest (the "useful many") contribute only a small proportion. To improve a process, it is important to find out which are the vital few problem areas.

A frequency plot is a tool to display data. It presents to the user basic information about the location, shape, and spread of a set of data. The frequency plot is widely used as a tool to help one understand variability. The frequency plot should only be used with adequate knowledge of the stability of the characteristic being measured. If the process is stable, the frequency plot serves as a prediction of the performance of the process in the future. If the process is unstable, then the frequency plot is simply a summary of what the process has done in the past. A basic type of frequency plot is the histogram, which is constructed by putting the scale for the characteristic of interest on the horizontal axis and the number of occurrences on the vertical axis.

A scatterplot is a tool used to study such relationships between possible causes and effects. It can also be used to study the association (or correlation) between different quality characteristics. A scatterplot is a graphic representation of the association between pairs of data. This pairing of data is the result of associating different measurements of a certain cause (e.g., pressure) with the corresponding measurement of the quality characteristic (e.g., paint thickness). The paired data could also be the measurements of two causes (e.g., pressure and temperature) or two quality characteristics (thickness and glossiness). Each pair becomes one point of the scatterplot.

4. SHEWHART CONTROL CHARTS

The control chart method provides an operational definition of the two types of causes of variation in a measure: common and special causes. Besides providing a new theory of variation, Shewhart also provided the method, the Shewhart control chart, for determining whether a system is dominated by common or special causes. The control chart is a statistical tool used to distinguish between variation in a measure of quality due to common causes and variation due to special causes. The name used to describe the chart (“control”) is misleading because the most common uses of these charts are to learn about variation and to evaluate the impact of changes. A better name might be “learning charts.” But Shewhart’s name has persisted. Figure 5 shows an example of a typical control chart.

The construction of a control chart typically involves:

- Plotting the data or some summary of the data in a run order (time is the most common order)
- Determining some measure of the central tendency of the data (such as the average)
- Determining some measure of the common cause variation of the data
- Calculating a centerline and upper and lower control limits (see Figure 5)

In developing the control chart method, Shewhart emphasized the importance of the economic balance between looking for special causes when they do not exist and overlooking special causes that do exist. It is also necesssary to develop rules that will give an acceptable economic balance for all types of measures in a variety of systems, processes, and products. Figure 6 illustrates the impact of these two mistakes.

4.1. Rationale for Shewhart Control Limits

Shewhart called the control limits three-sigma control limits and gave a general formula to calculate the limits for any statistic. Let T be the statistic to be charted, then:

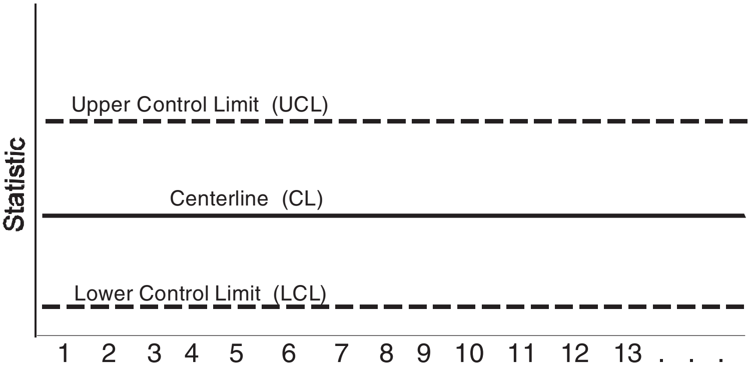


Figure 5 Illustration of a Control Chart. (Copyright 1980–1998 Associates in Process Improvement)

MISTAKE 1: To react to an outcome as if it came from a special cause, when actually it came from common causes of variation.

MISTAKE 2: To treat an outcome as if it came from common causes of variation, when actually it came from a special cause.

	ACTUAL SITUATION	
ACTION TAKEN	**NO CHANGE**	**CHANGE**
Take action on individual outcome (special)	**– $**	**+ $**
Treat outcome as part of system; work on changing the system (common)	**+ $**	**– $**

Figure 6 Mistakes Made in Attempts to Improve. (Copyright 1980–1998 Associates in Process Improvement)

$$\text{The centerline: CL} = \text{U}_t$$

$$\text{The upper control limit: UCL} = \text{U} + 3 * \sigma_t$$

$$\text{The lower control limit: LCL} = \text{U} - 3 * \sigma_t$$

where U is the expected value of the statistic and σ_t is the standard deviation of the statistic. Shewhart emphasized that statistical theory can furnish the expected value and standard deviation of the statistic, but empirical evidence justifies the width of the limits (the use of "3" in the control limit calculation).

The challenge for any particular situation is to develop appropriate estimates of the expected value and standard deviation of the statistic to be plotted. Appropriate statistics have been developed for control charts for a wide variety of applications.

The rationale for the use of Shewhart's three-sigma limits is:

- The limits have a basis in statistical theory.
- The limits have proven in practice to distinguish between special and common causes of variation.
- In most cases, use of the limits will approximately minimize the total cost due to overreaction and underreaction to variation in the process.
- The limits protect the morale of workers in the process by defining the magnitude of the variation that has been built into the process.

4.2. Interpretation of a Control Chart

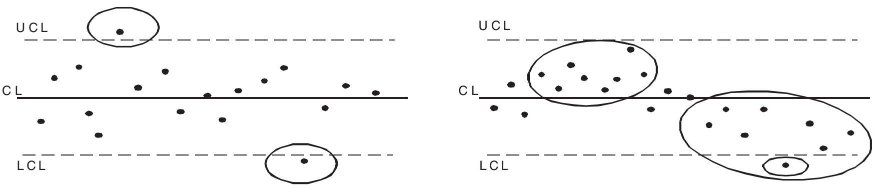
The control chart provides a basis for taking action to improve a process. A process is considered to be stable when there is a random distribution of the plotted points within the control limits. For a stable process, action should be directed at identifying the important causes of variation common to all of the points. If the distribution (or pattern) of points is not random, the process is considered to be unstable and action should be taken to learn about the special causes of variation.

There is general agreement among users of control charts that a single point outside the control limits is an indication of a special cause of variation. However, there have been many suggestions for systems of rules to identify special causes that appear as nonrandom patterns within the control limits. Figure 7 contains five rules that are recommended for general use with control charts. These rules are consistent in the sense that the chance of occurrence of rules #2 through #5 in a stable process is close to the chance of rule #1 occurring in a stable process.

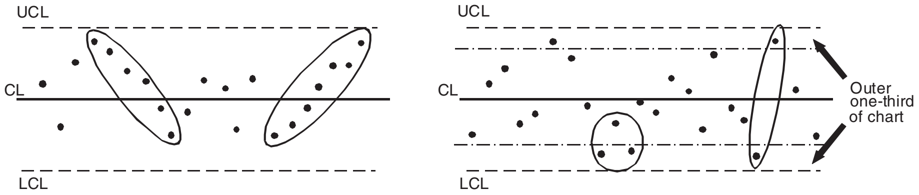
When applying the rules, the following guidelines will help with consistent interpretation of charts:

- Ties between two consecutive points do not cancel or add to a trend (rule 3).
- A point exactly on a control limit is not considered outside the limit (rule 1).

1. A single point outside the control limits.
2. A run of eight or more points in a row above (or below) the centerline



3. Six consecutive points increasing (trend up) or decreasing (trend down).
4. Two out of three consecutive points near (outer one-third) a control limit.



5. Fifteen consecutive points close (inner one-third of the chart) to the centerline

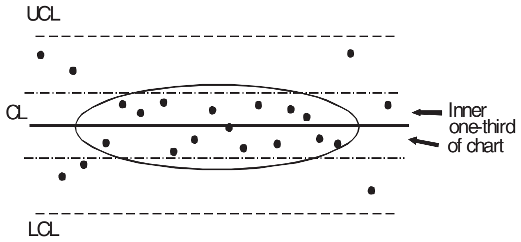


Figure 7 Rules for Determining a Special Cause. (Copyright 1980–1998 Associates in Process Improvement)

- When control charts have varying limits due to varying numbers of measurements within subgroups, rule 3 should not be applied.
- A point exactly on the centerline does not cancel or count towards a run (rule 2).
- When there is not a lower or upper control limit (for example, on a range chart with less than seven measures in a subgroup or on a P chart with 100% as a possible result for the process), rules 1 and 4 do not apply to the missing limit.

Rule 5 is especially useful in detecting a reduction of variation with an individual chart or for detecting improper subgrouping with an X-bar chart. Special circumstances may warrant use of some additional tests given by Nelson (1984). Deming (1986) emphasizes that the most important issue is the necessity to state in advance what rules to apply to a given situation.

4.3. Control Charts for Different Data Types

The different kinds of control charts are based on two groupings of types of data: attribute data and variable data. Attribute data includes classification, count, and rank data. Variable data refers primarily to continuous data, but rank data are often analyzed using a variable-control chart (realizing that the arithmetic functions are not theoretically valid). Otherwise the ranks can be converted to classification data and analyzed using attribute charts. Figure 8 contains examples of each of these categories of data.

Type of Data	Quality Characteristic	Recorded Data
Classification	Delivery Performance	On-time/late delivery
	Rework	OK the first time/rework
	Scratches	OK/excessive scratches
Count	Changes	Number of changes/design
	Accidents	Number of accidents/month
	Scratches	Scratches/surface
Continuous	Time	Minutes early or late
	Weight	Grams using a laboratory scale
	Scratches	Length in cm of each scratch

Figure 8 Examples of Types of Data. (Copyright 1980–1998 Associates in Process Improvement)

As can be seen from Figure 8, data for some characteristics can be recorded as any one of three types. For example, for a part with a large number of dimensional characteristics, the data could be recorded in the following ways:

- Classification: part meets or does not meet specification
- Count: number of dimensions not meeting specification
- Continuous: measured value for selected dimensions

Continuous data can be converted to attribute data by applying an operational definition for the count or classification. A recorded dimension can be classified as meeting or not meeting the specification; however, this conversion does not work in reverse. The measured dimensions are unknown for a part that is recorded as not meeting specifications.

For classification data, quality attributes are recorded in one of two classes. Example of these classes are conforming units/nonconforming units, go/no-go, and good/bad. To obtain count data, the number of incidences of a particular type is recorded: number of mistakes, number of accidents, or number of sales leads. For continuous data, a measured numerical value is recorded: a dimension, physical attribute, or calculated number.

Examples of continuous data include height, weight, density, elapsed time, viscosity, and costs. Continuous data can be converted to attribute data by applying an operational definition for the count or classification. In general, data should be collected as continuous data whenever possible because learning can occur with many fewer measurements compared to attribute classifications or counts. The control charts for continuous data require fewer measurements in each subgroup than the attribute control charts. Typical subgroup sizes for charts for continuous data range from 1 to 10, while subgroup sizes for attribute data range from 30 to 1000.

Figure 9 contains a summary of the frequently used control charts and the type of data to which they apply.

4.4. Subgrouping and Stratification

The concept of subgrouping is one of the most important components of the control chart method. Shewhart said the following about subgrouping (Shewhart 1931, p. 299):

> Obviously, the ultimate object is not only to detect trouble but also to find it, and such discovery naturally involves classification. The engineer who is sucessful in dividing his data into rational subgroups based upon rational hypotheses is therefore inherently better off in the long run than the one who is not thus successful.

Shewhart's concept was to organize data from the process in a way that is likely to give the greatest chance for the data in each subgroup to be alike and the greatest chance for data in other subgroups to be different. The aim of rational subgrouping is to include only common causes of variation within a subgroup, with all special causes of variation occurring between subgroups.

The most common method to obtain rational subgroups is to hold time constant within a subgroup. Only data taken at the same time (or for some selected time period) are included in a subgroup. Data from different time periods will be other subgroups. This use of time as the basis of subgrouping allows the detection of causes of variation that come and go with time.

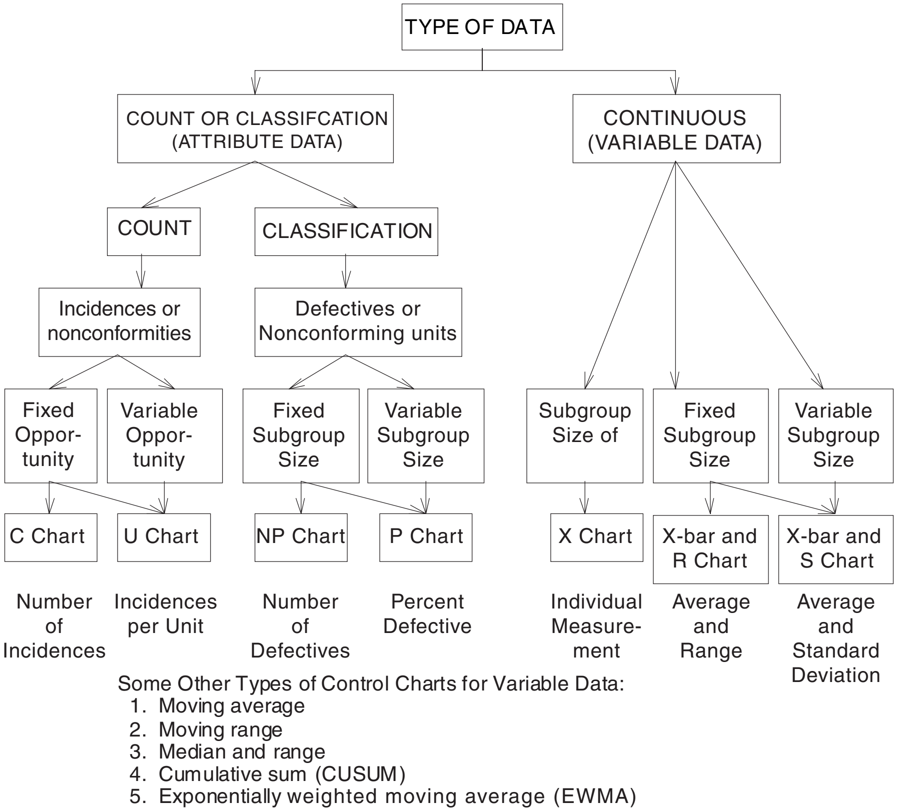

TYPE OF DATA
COUNT OR CLASSIFCATION (ATTRIBUTE DATA)
CONTINUOUS (VARIABLE DATA)
COUNT
CLASSIFICATION
Incidences or nonconformities
Defectives or Nonconforming units
Fixed Opportunity
Variable Opportunity
Fixed Subgroup Size
Variable Subgroup Size
Subgroup Size of
Fixed Subgroup Size
Variable Subgroup Size
C Chart
U Chart
NP Chart
P Chart
X Chart
X-bar and R Chart
X-bar and S Chart
Number of Incidences
Incidences per Unit
Number of Defectives
Percent Defective
Individual Measurement
Average and Range
Average and Standard Deviation
Some Other Types of Control Charts for Variable Data:
1. Moving average
2. Moving range
3. Median and range
4. Cumulative sum (CUSUM)
5. Exponentially weighted moving average (EWMA)

Figure 9 Selection of Particular Type of Control Chart. (Copyright 1980–1998 Associates in Process Improvement)

As an example of subgrouping, consider a study planned to reduce late payments. Historical data from the accounting files will be used to study the variation in late payments. What is a good way to subgroup the historical data on late payments? The data could be grouped by billing month, receiving month, major account, product line, or account manager. Knowledge or theories about the process should be used to develop rational subgroups. Some combination of time (either receiving or billing month) and one or more of the other variables in the process would be a reasonable way to develop the first chart.

After selecting a method of subgrouping, the user of the control chart should be able to state which sources of variation in the process will be present within subgroups and which sources will occur between subgroups. The specific objective of the control chart will often help determine the strategy for subgrouping the data. For example, if the objective is to evaluate differences between raw material suppliers, then only material from a single supplier should be included in data within a subgroup.

Since there is no grouping of the measurements for X charts, the power of rational subgrouping is not available. The use of stratification and rational ordering of the measurements with X charts provides an alternative to rational subgrouping for individual charts. Stratification is the separation and classification of data according to selected variables or factors. Stratification on a control chart is done in two different ways.

1. Plotting a symbol (instead of the usual ● or x) to indicate a classification for the measurement or statistic being plotted. For example, plot the symbol A, B, or C to indicate which of the three offices the measurements came from.
2. Ordering the measurements, or subgroups of measurements, by stratification variables such as laboratory, classroom, material type, supplier, shift, programmer, part position, etc., to investigate the importance of these factors.

4.5. Planning a Control Chart

Constructing a control chart is a relatively simple process. Anyone can get started by selecting a measure of quality and plotting it in order of time. When enough data become available, a centerline and control limits can be calculated (e.g., using the individual control limit formulas). Many useful control charts have been developed with this minimal amount of planning.

In other cases, lack of planning and preparation has made attempts to use control charts unsuccessful. In these more complex situations, the effective use of control charts requires careful planning to develop and maintain the chart. Figure 10 contains a planning form that can be used to guide the planning of a control chart.

Every control chart should be associated with one or more specific objectives. The objective might be to improve the yield of the process, identify and remove special causes from a process, or establish statistical control so that the capability of the process can be determined. The objectives should be summarized on the control chart form. After a period of time, the objective may be met. The control chart should be discontinued at that time, or a new objective developed.

A number of issues related to measurement and sampling must be resolved prior to beginning a control chart. The type of data for each variable to be charted will determine the type of chart to

1. **OBJECTIVE OF THE CHART:**

2. **SAMPLING, MEASUREMENT, AND SUBGROUPING:**

   Measure to be charted:

   Type of data:

   Type of control chart:

   Method of measurement:

   Quality of measurement process:

   Location of sampling:

   Strategy for subgrouping:

   Frequency of subgroups:

3. **MOST LIKELY SPECIAL CAUSES:**

4. **NOTES REQUIRED:**

Note	Responsibility

5. **REACTION PLAN FOR OUT OF CONTROL POINTS:** (attach copy)

6. **ADMINISTRATION:**

Task	Responsibility
Making measurements:	
Recording data on charts:	
Computing statistics:	
Plotting statistics:	
Extending/changing control limits:	
Filing:	

7. **SCHEDULE FOR ANALYSIS:**

Figure 10 Form for Planning a Control Chart. (Copyright 1980–1998 Associates in Process Improvement)

use. Information about the variability of the measurement system to be used should be documented. If the variability is not known, or if the stability of the measurement process is not documented, an effort to develop that information should be planned.

Important issues of sampling for control charts include the location in the process for measuring or sampling, the frequency of sampling, the number of samples, and the strategy for subgrouping mesurements (see section 4.4).

The documentation of information about the process is a most important part of many control charts. This documentation includes changes in the process, identification of special causes, investigations of special causes, and other relevant process data. Flow charts and cause-and-effect diagrams can be used to identify particular notes that should be recorded. Responsibility for recording this critical information should be clearly stated.

A plan for reaction to special causes on the chart should be established. Often a checklist of items to evaluate or a flow chart of the steps to follow is useful. The reaction plan should state the transfer of responsibility for identification of the special cause if it cannot be done at the local level. A plan for reaction to special causes on the chart should be established. Often a checklist of items to evaluate or a flow chart of the steps to follow is useful. The reaction plan should state the transfer of responsibility for identification of the special cause if it cannot be done at the local level. As an example, a reaction plan for a control chart in a laboratory to monitor a measurement system might have the following reaction plan:

1. Run the quality control standard.
2. Notify operations of a potential problem.
3. Review the log book for any recent changes in instrumentation.
4. Prepare a new QC standard and test it.
5. Replace the column in the instrument.
6. Notify the supervisor and call instrument repair.
7. Document the results of these investigations on the control chart.

There are a number of administrative duties required to maintain an effective control chart. Responsibility for measurement, recording data, calculating statistics, and plotting the statistics on the chart must be delineated. Proper revision and extension of control limits is an important consideration.

Control limits for the chart should be established using 20–30 subgroups from a period when the process is stable. If it is desirable to extend the control limits, any points affected by special causes should be removed and the control limits recalculated. The limits should only be extended when they are calculated using data without special causes.

Revision of the control limits should only be done when the existing limits are no longer appropriate. There are four circumstances when the original control limits should be recalculated:

1. When the initial control chart has special causes and there is a desire to use the calculated limits for analysis of data to be collected in the future. In this case, control limits should be recalculated after removing the data associated with the special causes.
2. When "trial" control limits have been calculated with fewer than 20–30 subgroups (note: trial limits should be calculated with fewer than 12 subgroups). In this case, the limits should be recalculated when 20–30 subgroups become available.
3. When improvements have been made to the process and the improvements result in special causes on the control chart. Control limits should than be calculated for the new process.
4. When the control chart remains out of control for an extended period of time (20 or more subgroups) and approaches to identify and remove the special cause(s) have been exhausted. Control limits should be recalculated to determine if the process has stabilized at a different operating level.

The date the control limits were last calculated should be a part of the ongoing record for the control chart. Some notation (such as vertical lines on the chart) should be used to indicate subgroups used to calculate control limits.

The form to record the data and to plot the control chart is another important consideration. The form should allow for a continuing record and not have to be restarted every day or week. The control chart form should include space to document the important decisions and information about the process from the planning form. The recorded data should include the time and place and the person making the measurements as well as the results of the measurements. The scale on the charts should be established to give a clear visual interpretation of the variation in the process. With the control limits centered on the chart, about one-half of the scale should be included inside the control limits.

A schedule for analysis should be established for every active control chart. The frequency of analysis will vary depending on the objective of the chart. For example, the quality improvement team might meet to analyze the chart weekly to assist in their improvement effort, while the department manager might be interested in a monthly review of the chart for planning purposes. The production vice president might review the charts with the department manager at the end of each quarter for planning and evaluation purposes.

Figure 11 shows an example of a completed planning form for a control chart maintained by an accounting group. Taking the time to plan a control chart before data collection is begun will help ensure that the chart leads to learning about the process or system.

4.6. Control Chart for Individual Measurements

One of the most useful control charts is the control for individual measurements, or the X chart. This control chart is a simple extension of the run chart. The control chart for individuals is useful when:

1. OBJECTIVE OF THE CHART: *To learn about the causes of returned invoices in order to reduce the number of returned invoices that have to be billed again.*

2. SAMPLING, MEASUREMENT, AND SUBGROUPING:
Measure of quality to be charted: *Percent of invoices returned that are not paid*
Type of data: *Classification*
Method of measurement: *Accounting supervisor records number of invoices sent each week and number returned unpaid.*
Quality of measurement process: *Complete, accurate counts can be made. The totals can be validated.*
Location of sampling: *Master list and returns that cross the supervisor's desk.*
Strategy for subgrouping: *Subgroup will be all invoices mailed in a given week (historically 35-90 invoices)*
Frequency of subgroups: *One per week—100% of invoices for that week.*
Type of control chart: *P chart*

3. MOST LIKELY SPECIAL CAUSES:
New customers, price changes, computer program updates, new employees in the Accounting Department.

4. NOTES REQUIRED:

Note	Responsibility
Number of new customers each week	*Supervisor*
New employees	*Supervisor*
Changes in computer program	*Systems*

5. REACTION PLAN FOR OUT OF CONTROL POINTS:
Supervisor will call meeting of Department to discuss all special causes.

6. ADMINISTRATION:

Task	Responsibility
Making measurements:	*Supervisor*
Recording data on charts:	*Supervisor*
Computing statistics:	*Supervisor*
Plotting statistics:	*Supervisor*
Extending/changing control limits:	*Dept. QI Team*
Filing:	*Supervisor*

7.SCHEDULE FOR ANALYSIS: QI team review each month

Figure 11 Example of a Completed Control Chart Planning Form. (Copyright 1980–1998 Associates in Process Improvement)

- There is no rational way to organize the data into subgroups (see later section on X-bar and R charts for a detailed discussion of control charts).
- Measures of performance of the process can only be obtained infrequently.
- The variation at any one time (within a subgroup) is insignificant relative to the between subgroup variation.

Examples of situations and data where a control chart for individuals can be useful include batch processes, accounting data, maintenance records, shipment data, yields, efficiencies, sales, costs, and forecast or budget variances. Often the frequency of data collection cannot be controlled for these situations and types of data.

Instrument readings such as temperatures, flows, and pressures often have minimal variation at any one time but will change over time. The study of tool wear is another example of insignificant short-term variation relative to variation over time. Control charts of the individual measurements can often be useful in these cases.

Some advantages of the control chart for individuals (compared to other types of control charts) are:

- The chart is an extension of the familiar run chart.
- No calculations are required when plotting on the chart.
- Plotting is done each time a measurement is made, providing fast feedback. Study of the process does not have to wait for additional measurements.
- Because only one chart is required for each measure of quality, charts for multiple measures of performance can be grouped on one form for presentation purposes to facilitate evaluation of a process.
- The capability of a process can be evaluated directly from the control limits on the chart.

Because of these advantages, the control chart for individuals is sometimes used when another type of control chart is more appropriate. The X chart is somewhat less sensitive than other variable control charts with larger subgroup sizes in its ability to detect the presence of a special cause. Sometimes data analyzed with an X chart will indicate a stable process, but the same data analyzed with a more appropriate chart (P chart, C chart, or X-bar and R chart, discussed in later sections) will clearly indicate the presence of special causes.

Besides this lesser sensitivity, there are some other disadvantages to using an X chart to study variation in data:

- Because each individual measure is plotted on the chart, there is no opportunity to focus on different sources of variation through subgrouping.
- All sources of variation are combined on one chart, sometimes making identification of the important sources of variation difficult.
- The X chart is sensitive to a nonsymmetric distribution of data and may require data transformation to be used effectively.

To develop a control chart for individuals, 20–30 measurements are required. The symbol for the number of measures used to calculate control limits is ‘‘k.’’ The individual measurements are plotted on the X chart and the average of the individual measurements is used for the centerline of the chart. The moving ranges of consecutive measurements are used to estimate the variation of the process and develop control limits for the X chart.

The moving range is calculated by pairing consecutive measurements. The range is calculated for each set of two measurements by subtracting the low value from the high value. Each individual measurement is considered twice in the calculation of the moving ranges. Because a previous measurement is not available for the first measurement in the set, only $k - 1$ moving ranges can be calculated. The average of the moving ranges ($\overline{\text{MR}}$) is used for control limit calculations. Because the X chart of individual measurements contains all the information available in the data, it is not necesssary to plot the moving ranges.

A example of an X chart concerns a chemical product that is shipped in hopper cars with a sample taken from each car during loading. Laboratory tests are made on each sample for product certification, and this test becomes one dot on the chart. The cars are loaded from storage bins that are filled on an intermittent basis from a process unit. Laboratory results for the concentration of an additive for the last 25 cars loaded are used to develop control charts for the product shipped.

The control chart for the additive is shown in Figure 12, and the calculations of control limits are shown in Figure 13. As can be seen in the two figures, the moving ranges for car numbers 14 and 15 are greater than the moving range upper control limit of 12.8. These two values are removed

Chart133 | Chart Name Additive at Hopper Car Loading

Objective Study variation and density special causes | Subgrouped by: Hopper Car

Process Hopper car loading | Product Q100 through Q209 | Target 225 +/- 25 | date 9/93

Chart Responsibility: Lab Post #2 | Characteristic: Additive | Measurement Method: GC | Unit: PPM | zero=0

car	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25					
time																														
measure	215	218	222	217	216	214	219	221	216	220	218	218	221	236	222	221	216	218	223	217	218	221	220	219	215					
MR	-	3	4	5	1	2	5	2	5	4	2	0	3	15	14	1	5	2	5	6	7	3	1	1	4					

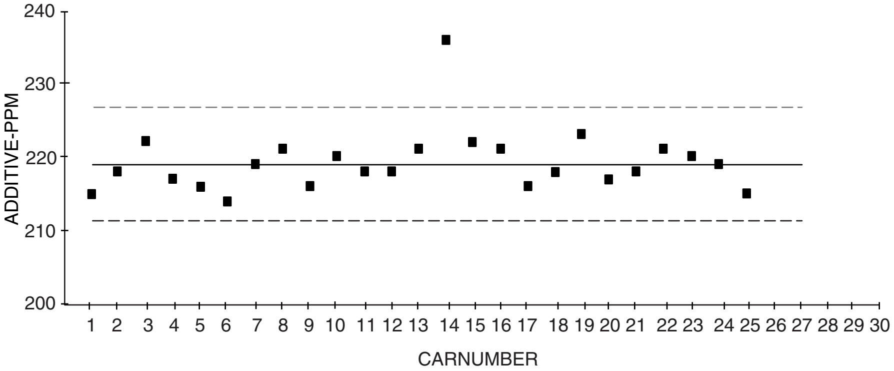


Figure 12 Individual Control Chart for Additive. (Copyright 1980–1998 Associates in Process Improvement)

NAME __Additive______________________________ DATE ____9/93_______________

PROCESS ___Hopper Car Loading _________ SAMPLE DESCRIPTION __Composite_____

NUMBER OF SUBGROUPS (k) ____ 25______ BETWEEN (dates) Car 1 - 25

$\overline{X} = \frac{\Sigma X}{k} = \frac{5481}{25} = 219.2$ $\qquad \overline{MR} = \frac{\Sigma MR}{k-1} = \frac{94}{24} = 3.92$

X CHART **MR Calculation**

UCL = $\overline{X}$ + (2.66 $*$ $\overline{MR}$)
UCL = 219.2 + (2.66 $*$ 2.95)
UCL = 219.2 + 7.8
UCL = 227.0

UCL_{MR} = 3.27 $*$ $\overline{MR}$
UCL_{MR} = 3.27 $*$ 3.92
UCL_{MR} = _12.8____

LCL = $\overline{X}$ – (2.66 $*$ $\overline{MR}$)
LCL = 219.2 – (2.66 $*$ 2.95)
LCL = 219.2 – 7.8
LCL = 211.4

Recalculate $\overline{MR}$ after removing MR's greater than UCL_{MR}
$\overline{MR}$ = 3 MR / k - ?
$\overline{MR}$ = _65__ / _22__
$\overline{MR}$ = 2.95___

Figure 13 Control Chart Calculations for Additive X Chart. (Copyright 1980–1998 Associates in Process Improvement)

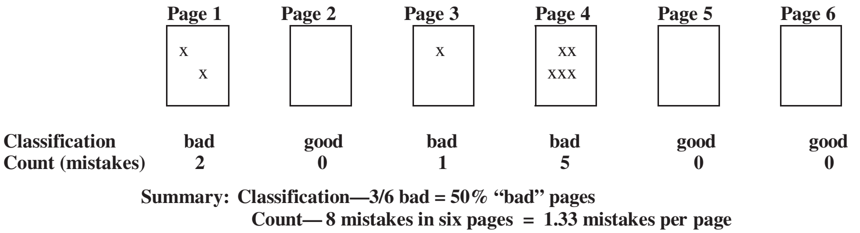


Figure 14 Two Types of Attribute Data. (Copyright 1980–1998 Associates in Process Improvement)

and the average moving range recalculated. The revised $\overline{MR} = 2.95$ was used to calculate the control limits for the X chart. The control chart indicates there is a special cause present for car number 14. Note that the 236 ppm concentration for car 14 is associated with the two moving ranges that were above the upper control limit.

4.7. Control Charts for Attribute Data

The two basic types of attribute data were discussed in Section 4.3:

1. *Classifications of units:* conforming units/nonconforming units, blue/not blue, go/no-go, etc.
2. *Count of incidence:* number of nonconformities, defects, accidents, trips, calls, etc.

Often data can be collected as either type. For example, in evaluating spelling errors in a manuscript (see Figure 14), each page could be classified as (1) having one or more spelling mistakes or (2) having none. This would be classification of units, with each page as a unit. Alternatively, the number of spelling mistakes on each page could be counted.

To develop an attribute control chart, a subgrouping strategy must first be determined. The subgroup size (n) is the number of units tested for classification data, or the area of opportunity for the incidence to occur for count data. There are four commonly used control charts for attribute data, depending on the type of attribute data and the constancy of the subgroup size. Table 1 summarizes these charts.

4.7.1. The P Chart for Classification Data

The P chart is appropriate whenever classifications are made in two categories, such as good parts and scrap parts. The P chart is usually preferred over the NP chart because percentages are more easily interpreted than counts in most applications and the P chart can be used with either a constant or variable subgroup size. The percentage of units in one of the categories (either the positive or the negative one, i.e., percent good product or percent scrap) is then calculated and graphed to develop the chart. Twenty to 30 subgroups are desirable for calculating the control limits, with at least 30 units in each subgroup.

Many times it is desirable to construct and use a P chart when the subgroup size is variable. This is usually the case when a set time period, such as a day or week, rather than a specific number of units, is used to define the subgroup. However, it is not necessary for each subgroup to contain exactly the same number of units to be considered constant. If the maximum and minimum subgroup sizes are within 20% of the average subgroup size, there will be an insignificant effect on the control limits if the average subgroup size is used for all calculations. If this is not the case, the subgroup

TABLE 1 Types of Attribute Control Charts

Chart Name	Type of Attribute Data	Statistic Charted	Subgroup Size
NP Chart	classification	number of nonconforming units (D)	constant
P Chart	classification	percent nonconforming units (P)	may vary
C Chart	count	number of incidents (C)	constant
U Chart	count	incidents per unit (U)	may vary

Copyright 1980–1998 Associates in Process Improvement.

size must be considered variable and different sets of control limits must then be calculated for each subgroup size (or for sets of subgroup sizes with the individual subgroups within 20% of the average for the set). Methods for accommodating variable subgroup sizes are given under Additional Reading.

Once a subgrouping strategy has been determined, the following steps should be followed when constructing a P chart (see Figure 15 for calculation form):

1. Calculate p $\{p$ = (number in a certain category/number in subgroup) $*$ 100$\}$ for each subgroup.
2. Calculate $\overline{\text{P}}$, the average of the p's and the centerline for the chart
3. Determine the control limits for the P chart.
4. Figure and draw a scale on appropriate graph paper so the upper control limit is placed approximately one quarter of the way from the top. If there is a lower control limit, it should be placed 10–25% above the bottom of the chart. (Note: the scale should begin with zero for most situations.)
5. Plot the p's on the chart and draw in the control limits and centerline.

d = Nonconforming sample units per subgroup
n = Number of sample units per subgroup
k = Number of subgroups
p = Percent nonconforming units = 100*d/n

- -

Control limits when subgroup size (n) is constant:

$$\overline{\text{P}} = \frac{\Sigma p}{k} = \frac{\_\_\_\_\_\_}{} = \_\_\_\_\_\_\_\_ \quad \text{(centerline)}$$

$$\hat{\sigma}p = \sqrt{\frac{\overline{p} \; * \; (100 - \overline{p})}{\text{n}}} = \sqrt{\frac{\quad * \; (100 - \qquad)}{}} = \_\_\_\_\_\_$$

UCL = $\overline{\text{p}}$ + (3 * $\hat{\sigma}p$) LCL = $\overline{\text{p}}$ − (3 * $\hat{\sigma}p$)

UCL = ____ + (3 * ____) LCL = ____ − (3 * ____)

UCL = ____ + ______ LCL = ____ − ______

UCL = ______ LCL = ______

- -

Control limits when subgroup size (n) is variable:

$$\overline{\text{p}} = \frac{\Sigma d}{\Sigma n} * 100 = \frac{\_\_\_\_\_\_}{} * 100 = \_\_\_\_\_\_ \quad \text{(centerline)}$$

$$\hat{\sigma}p = \frac{\sqrt{\overline{p} \; * \; (100 - \overline{p})}}{\sqrt{n}} = \frac{\sqrt{\quad * \; (100 - \qquad)}}{\sqrt{n}} = \frac{\_\_\_\_\_\_}{\sqrt{n}}$$

UCL = $\overline{p}$ + (3 * $\hat{\sigma}p$) LCL = $\overline{p}$ − (3 * $\hat{\sigma}p$)

UCL = ___ + (3 * ____ / $\sqrt{\text{n}}$) LCL = ___ − (3 * ____ / $\sqrt{\text{n}}$)

UCL = ___ + (____ / $\sqrt{\text{n}}$) LCL = ___ − (____ / $\sqrt{\text{n}}$)

n:	____	____	____	____	____
$\sqrt{\text{n}}$:	____	____	____	____	____
3* $\hat{\sigma}$p:	____	____	____	____	____
UCL:	____	____	____	____	____
LCL:	____	____	____	____	____

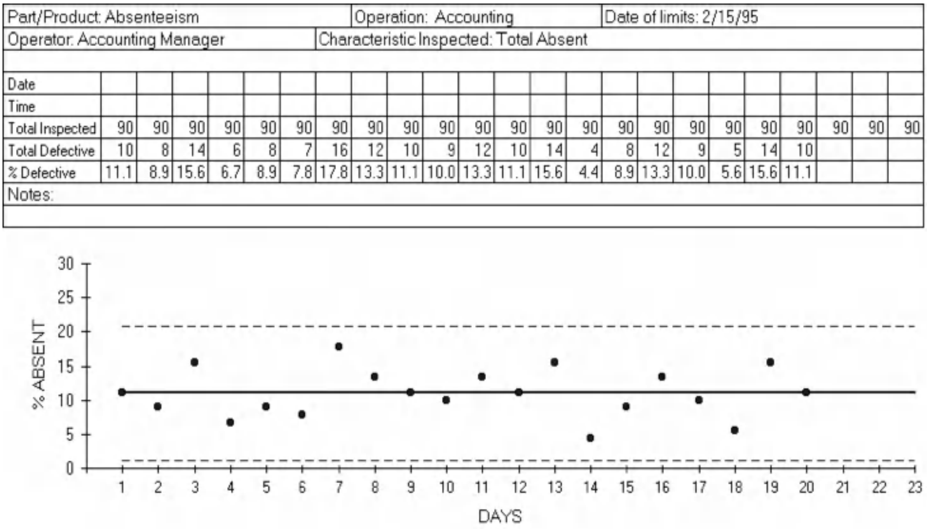
Figure 15 P Chart Calculation Form. (Copyright 1980–1998 Associates in Process Improvement)

TABLE 2 Data on Absenteeism in the Accounting Department

	Absenteeism (90 Employees)			
Day	Total Absences	p	Unexcused Absences	p
1	10	11.1	2	2.2
2	8	8.9	3	3.3
3	14	15.6	1	1.1
4	6	6.7	1	1.1
5	8	8.9	1	1.1
6	7	7.8	2	2.2
7	16	17.8	0	0.0
8	12	13.3	3	3.3
9	10	11.1	1	1.0
10	9	10.0	8	8.8
11	12	13.3	1	1.1
12	10	11.1	2	2.2
13	14	15.6	0	0.0
14	4	4.4	4	4.4
15	8	8.9	3	3.3
16	12	13.3	1	1.1
17	9	10.0	0	0.0
18	5	5.6	2	2.2
19	14	15.6	1	1.1
20	10	11.1	0	0.0
	$\Sigma d = 198$	$\Sigma p = 220.0$	$\Sigma d = 36$	$\Sigma p = 40.0$

Copyright 1980–1998 Associates in Process Improvement.

A P chart example deals with a situation where a constant subgroup size is appropriate. The manager of the accounting department of a company decided to gather information on the absenteeism of her 90 employees. Each day for one month, the number of employees who were absent and whether their absence was unexcused was noted. Table 2 contains the data collected during that month.



Part/Product: Absenteeism
Operation: Accounting
Date of limits: 2/15/95
Operator: Accounting Manager
Characteristic Inspected: Total Absent

Date																							
Time																							
Total Inspected	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
Total Defective	10	8	14	6	8	7	16	12	10	9	12	10	14	4	8	12	9	5	14	10			
% Defective	11.1	8.9	15.6	6.7	8.9	7.8	17.8	13.3	11.1	10.0	13.3	11.1	15.6	4.4	8.9	13.3	10.0	5.6	15.6	11.1			

Notes:

Figure 16 P Chart for Total Absences. (Copyright 1980–1998 Associates in Process Improvement)

d = nonconforming sample units per subgroup
n = number of sample units per subgroup
k = number of subgroups
p = percent nonconforming units = 100**d*/*n*

- -

Control limits when subgroup size (*n*) is constant:

Total Absent $\bar{p} = \frac{\Sigma p}{k} = \frac{220.0}{20} = \underline{\quad 11.0 \quad}$ (centerline)

$$\hat{\sigma}p = \sqrt{\frac{\bar{p} * (100 - \bar{p})}{n}} = \sqrt{\frac{11.0 * (100 - 11.0)}{90}} = \underline{\quad 3.3 \quad}$$

$$\text{UCL} = \bar{p} + (3 * \hat{\sigma}p) \qquad \text{LCL} = \bar{p} - (3 * \hat{\sigma}p)$$

$$\text{UCL} = \underline{11.0} + (3 * \underline{3.3}) \qquad \text{LCL} = \underline{11.0} - (3 * \underline{3.3})$$

$$\text{UCL} = \underline{11.0} + \underline{\quad 9.9 \quad} \qquad \text{LCL} = \underline{11.0} - \underline{\quad 9.9 \quad}$$

$$\text{UCL} = \underline{\ 20.9 \quad} \qquad \text{LCL} = \underline{\ 1.1 \quad}$$

Figure 17 Calculations for Total Absence. (Copyright 1980–1998 Associates in Process Improvement)

Figure 16 shows the control chart for total absences. Figure 17 shows the calculations for total absences. Since there were 90 employees, calculations for a constant subgroup size were utilized. Figure 18 shows the chart and calculations for unexcused absences. The calculations resulted in no lower control limit for the chart on unexcused absences. The P chart for total absences is stable. A fundamental change to the system is required in order to reduce the average daily absenteeism of 11%. The P chart for unexcused absences indicates a special cause on day 10. Reasons for this special cause should be investigated and used to help develop a strategy to reduce unexcused absenteeism.

4.7.2. Control Charts for Count Data

When actual counts of incidence (often nonconformities) rather than classification of units are made, either a C chart or a U chart is usually the appropriate control chart. Figure 14 illustrated the difference between counts and classifications. Since the subgrouping method for counts is not always based on the selection of a certain number of units, a subgroup is defined as an *area of opportunity,* when working with count data.

An area of opportunity is simply the region selected for the count and could be of the following forms:

- Number of units (e.g., five television sets, requisitions per day)
- Space (e.g., 200 feet of yarn, 15 square yards of coated paper, one-quart sample of a product)
- Time (e.g., three months, one shift)

The decision whether to use a C chart or a U chart is made by determining whether the area of opportunity will be constant or will vary for each group of counts. For example, an area of opportunity could be the number of bills received in an office each week. If the number of errors on these bills is counted, the count will be distorted if the number of bills received from week to week is different. How to deal with this situation will be included in the discussion of when and how to use the C chart or U chart in the remaining part of this section. Table 3 lists examples of applications of C and U charts.

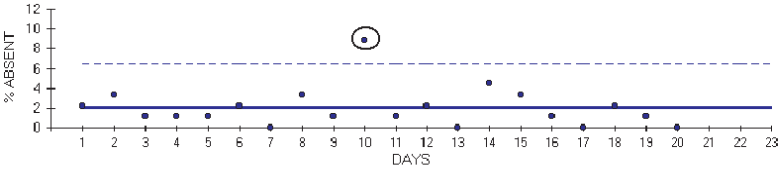
A C chart is used when the area of opportunity is constant for each subgroup. This would be the case in the example given above if 50 bills were received in the office each week. The statistic plotted for a C chart is simply the number of incidents (errors) in each area of opportunity (a week or 50 bills). It is not necessary for the area of opportunity to be exactly the same for each subgroup in order to use a C chart. The area of opportunity in any analysis can be considered constant if each region (number of units, time, or space) on which the counts are taken is within 20% of the overall average.

Part/Product: Absenteeism | Operation: Accounting | Date of limits: 2/15/95

Operator: Accounting Manager | Characteristic Inspected: Unexcused Absence

Date																							
Time																							
Total Inspected	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
Total Defective	2	3	1	1	1	2	0	3	1	8	1	2	0	4	3	1	0	2	1	0			
% Defective	2.2	3.3	1.1	1.1	1.1	2.2	0.0	3.3	1.1	8.9	1.1	2.2	0.0	4.4	3.3	1.1	0.0	2.2	1.1	0.0			

Notes:



Notes:

d = nonconforming sample units per subgroup
n = number of sample units per subgroup
k = number of subgroups
p = percent nonconforming units = 100*d/n

Control limits when subgroup size (n) is constant:

Total Absent $$\bar{p} = \frac{\Sigma p}{k} = \frac{220.0}{20} = \underline{\quad 11.0 \quad} \quad \text{(centerline)}$$

$$\hat{\sigma}p = \sqrt{\frac{\bar{p} \ * \ (100 - \bar{p})}{n}} = \sqrt{\frac{11.0 \ * \ (100 - 11.0)}{90}} = \underline{\quad 3.3 \quad}$$

$$\text{UCL} = \bar{p} + (3 \ * \ \hat{\sigma}p) \qquad \text{LCL} = \bar{p} - (3 \ * \ \hat{\sigma}p)$$

$$\text{UCL} = \underline{11.0} + (3 \ * \ \underline{3.3}) \qquad \text{LCL} = \underline{11.0} - (3 \ * \ \underline{3.3})$$

$$\text{UCL} = \underline{11.0} + \underline{\quad 9.9 \quad} \qquad \text{LCL} = \underline{11.0} - \quad 9.9$$

$$\text{UCL} = \underline{\ 20.9 \quad} \qquad \text{LCL} = \underline{\ 1.1 \quad}$$

Figure 18 C and U Chart Calculation Form. (Copyright 1980–1998 Associates in Process Improvement)

Once it has been determined that the area of opportunity will be constant for each subgroup, the following steps should be followed to construct a C chart:

1. Record the count c for 20 to 30 subgroups.
2. Compute $\bar{c}$, the centerline for the C chart.
3. Compute the control limits for the C chart.
4. Calculate and draw a scale on the charting form such that the upper control limit is 25% below the top of the chart. Plot the individual c's, the centerline, and the control limits.

The example that follows illustrates some of the important points concerning construction of a C chart. In an effort to improve safety in their factory, a company decided to chart the number of injuries that required first aid each month. Since approximately the same number of hours were worked each month, the area of opportunity (total man-hours worked in one month) was constant and a C chart was utilized. Table 4 contains the data collected over a two-year period.

Figure 19 shows the control chart and Figure 20 shows the calculations of the control limits. In July 1998, the reporting of 23 injuries resulted in a point above the upper control limit. This special cause was the result of a large amount of vacation leave taken during July. Untrained people and excessive overtime were needed to achieve the normal number of hours worked for a month. There

TABLE 3 C Chart and U Chart Situations

Area of Opportunity	Use a C Chart if	*c* Statistic	Use a U Chart if	*u* Statistic
Three documents (number of units)	Number of total pages is the same	Number of errors in the three documents	Number of total pages is different	Number of errors is per 50 pages (or other number)
Roll of carpet (space)	Square yards on each roll are the same	Number of visual defects per roll	Square yards on each roll are different	Number of visual defects per 25 square yards
One month (time)	Number of hours worked each month is the same	Number of accidents per month	Number of hours worked each month is different	Number of accidents per 100,000 hours

Copyright 1980–1998 Associates in Process Improvement.

TABLE 4 Injury Data for C Chart

Month/Year	Number of Injuries [c]
Jan. 1998	6
Feb.	2
March	3
April	8
May	5
June	4
July	23
Aug.	7
Sept.	3
Oct.	5
Nov.	12
Dec.	7
Jan. 1999	10
Feb.	5
March	9
April	4
May	3
June	2
July	2
Aug.	1
Sept.	3
Oct.	4
Nov.	3
Dec.	1
	$\Sigma c = 133$

Copyright 1980–1998 Associates in Process Improvement.

was also a run of nine points in a row below the centerline, starting in April 1999. This indicated that the average number of reported first aid cases per month had been reduced. This reduction was attributed to a switch from wire to plastic baskets for the carrying and storing of parts and tools, which greatly reduced the number of injuries due to cuts. If this trend continues, the control limits should be recalculated when sufficient data is available.

It should be noted that there is no lower control limit in the control chart of the previous example. Therefore, a run of eight or more points is required to demonstrate improvement. Combining data by quarter (three months) is one way to increase c and thus obtain a LCL.

Part/Product: Safety | Operation: Entire plant — 300K hours/month | Date of limits: 1/95
Operator: Entire plant | Characteristic Inspected: Injuries requiring first aid

	1993												1994											
Date	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
Time																								
Total Injuries	6	2	4	8	5	4	23	7	3	5	12	7	10	5	9	4	3	2	2	1	3	4	3	1
Notes:																								

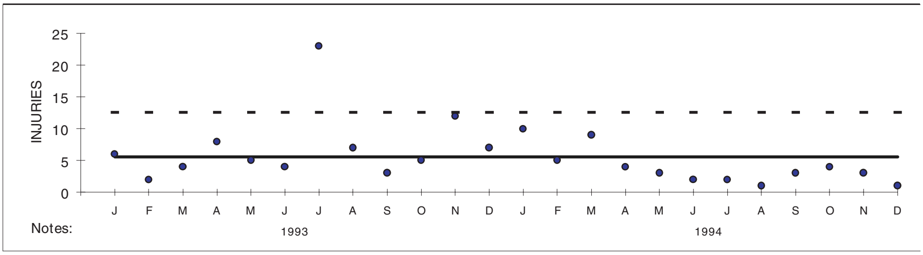


Figure 19 C Chart—Injury Data. (Copyright 1980–1998 Associates in Process Improvement)

C CHART CONTROL LIMITS (area of opportunity constant)

c = number of incidences per subgroup
k = number of subgroups
note: The subgroup size is defined by the “area of opportunity” for incidences and must be constant.

$\bar{c} = \frac{\Sigma c}{k} = \frac{133}{24} = \underline{\ 5.5\ }$ (centerline)

$UCL = \bar{c} + (3 * \sqrt{\bar{c}})$ $LCL = \bar{c} - (3 * \sqrt{\bar{c}})$

$UCL = \underline{5.5} + (3 * \underline{2.3})$ $LCL = \underline{5.5} - (3 * \underline{2.3})$

$UCL = \underline{5.5} + \underline{6.9}$ $LCL = \underline{5.5} - \underline{6.9}$

$UCL = \underline{12.4}$ $LCL = \underline{\ —\ }$

U-CHART CONTROL LIMITS (area of opportunity may vary)

c - number of incidences per subgroup
n - number of “standard areas of opportunity” in a subgroup (n may vary)
u - incidences per standard area of opportunity = c/n
k - number of subgroups
note: The standard area of opportunity will be defined by the people planning the control chart in units such as man-hours, miles driven, per ten invoices, etc.

$\bar{u} = \frac{\Sigma c}{k} = \_\_\_\_\_\_ = \_\_\_\_\_\_$ (Center Line)

$UCL = \bar{u} + (3 * \sqrt{\bar{u}}) / \sqrt{n}$ $LCL = \bar{u} - (3 * \sqrt{\bar{u}})$

$UCL = \_\_\_ + (3 * \_\_\_) / \sqrt{n}$ $LCL = \_\_\_ - (3 * \_\_\_) / \sqrt{n}$

$UCL = \_\_\_ + \_\_\_ / \sqrt{n}$ $LCL = \_\_\_ - \_\_\_ / \sqrt{n}$

$UCL = \_\_\_$ $LCL = \_\_\_$

n:					
$\sqrt{n}$					
UCL:					
LCL:					

Figure 20 C Chart Calculations—Injury Data. (Copyright 1980–1998 Associates in Process Improvement)

4.8. X-Bar and R Control Charts

When continuous data are obtained from a process, it is sometimes of interest to learn about both the average level of the process and the variation about the average level. In these cases, two control charts are often used to study the process: the X-bar chart and the R chart.

An important aspect of the collection of data for the construction of X-bar and R control charts is that the collection is done in subgroups. A subgroup for continuous data is a set (usually three to six) of measurements of some characteristic in a process, which were obtained under similar conditions or at about the same time. The X-bar chart contains the averages and the R chart the ranges calculated from the measurements in each subgroup. These averages and ranges are usually plotted over time.

Figure 21 contains a form used to calculate the appropriate control limits. There are a number of symbols associated with X-bar and R charts:

X = individual measurement of quality characteristic
n = subgroup size (number of measurements per subgroup)
k = number of subgroups used to develop control limits

NAME ____ DATE ________________

PROCESS _______________________ SAMPLE DESCRIPTION ____________________

NUMBER OF SUBGROUPS (k) ________ BETWEEN (DATES) _______-

NUMBER OF SAMPLES OR MEASUREMENTS PER SUBGROUP (n) _____________

$\bar{\bar{X}} = \frac{\Sigma \bar{X}}{k}$ = ________ = ________

$\bar{R} = \frac{\Sigma R}{k}$ = _______ = _____

$\bar{X}$ CHART

UCL = $\bar{\bar{X}}$ + (A_2 * $\bar{R}$)
UCL = + (*)
UCL = +
UCL = ___________

LCL = $\bar{\bar{X}}$ - (A_2 * $\bar{R}$)
LCL = - (*)
LCL = -
LCL = ___________

R CHART

UCL = D_4 * $\bar{R}$
UCL = *
UCL = __________

LCL = D_3 * $\bar{R}$
LCL = *
LCL = __________
—

FACTORS FOR CONTROL LIMITS

n	A_2	D_3	D_4	d_2
2	1.88	-	3.27	1.128
3	1.02	-	2.57	1.693
4	0.73	-	2.28	2.059
5	0.58	-	2.11	2.326
6	0.48	-	2.00	2.534
7	0.42	0.08	1.92	2.704
8	0.37	0.14	1.86	2.847
9	0.34	0.18	1.82	2.970
10	0.31	0.22	1.78	3.087

PROCESS CAPABILITY

If the process is in statistical control, the standard deviation is:

$\hat{\sigma}$ = $\bar{R}$ / d_2
$\hat{\sigma}$ = /
$\hat{\sigma}$ = _____________

the process capability is:

$\bar{\bar{X}}$ - 3 * $\hat{\sigma}$ to $\bar{\bar{X}}$ + 3 * $\hat{\sigma}$
- to +
_________ to _______

Figure 21 X-bar and R Control Chart Calculation Form. (Copyright 1980–1998 Associates in Process Improvement)

Σ =summation symbol
$\overline{X}$ =(X-bar) subgroup average
$\overline{\overline{X}}$ =(X-double bar) average of the averages of all the subgroups
R =subgroup range (largest–smallest)
$\overline{R}$ =(R-bar) average of the ranges of all the subgroups
$A2$, $D3$, $D4$, d_2 =factors for computing control limits and process capability
$*$ =multiplication symbol

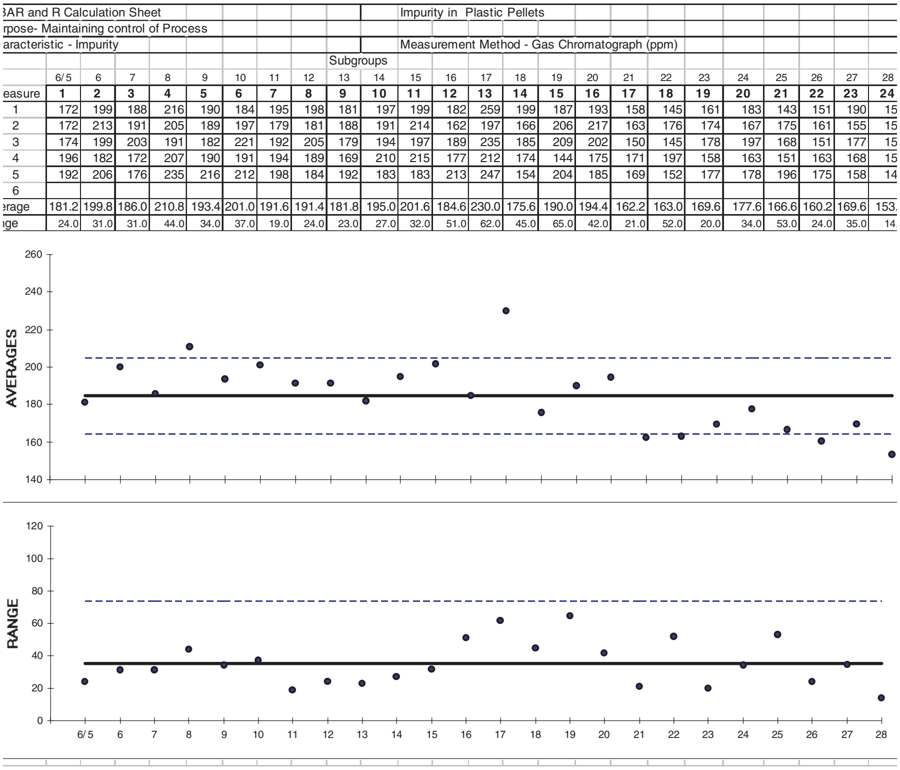
The steps for developing X-bar and R control charts follow. All averages that are calculated should be rounded to one more decimal place (significant figure) than the values being averaged.

1. Calculate $\overline{X}$ ($\overline{X} = \Sigma\ \overline{X}/n$) for each subgroup.
2. Calculate R (largest − smallest value) for each subgroup.
3. Calculate $\overline{\overline{X}}$ ($\overline{\overline{X}} = \overline{\Sigma X}/k$), the centerline of the X-bar chart.
4. Calculate $\overline{R}$ ($\overline{R} = \Sigma R/k$), the centerline of the R chart.
5. Calculate the control limits for the X-bar chart using:

$$\text{UCL} = \overline{\overline{X}} + (A2 * \overline{R})$$

$$\text{LCL} = \overline{\overline{X}} - (A2 * \overline{R})$$

Note: $A2$ is a constant based on n obtained from Figure 21.



AR and R Calculation Sheet — Impurity in Plastic Pellets
rpose- Maintaining control of Process
aracteristic - Impurity — Measurement Method - Gas Chromatograph (ppm)

Subgroups

	6/5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
easure	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
1	172	199	188	216	190	184	195	198	181	197	199	182	259	199	187	193	158	145	161	183	143	151	190	15
2	172	213	191	205	189	197	179	181	188	191	214	162	197	166	206	217	163	176	174	167	175	161	155	15
3	174	199	203	191	182	221	192	205	179	194	197	189	235	185	209	202	150	145	178	197	168	151	177	15
4	196	182	172	207	190	191	194	189	169	210	215	177	212	174	144	175	171	197	158	163	151	163	168	15
5	192	206	176	235	216	212	198	184	192	183	183	213	247	154	204	185	169	152	177	178	196	175	158	14
6																								
erage	181.2	199.8	186.0	210.8	193.4	201.0	191.6	191.4	181.8	195.0	201.6	184.6	230.0	175.6	190.0	194.4	162.2	163.0	169.6	177.6	166.6	160.2	169.6	153.
ge	24.0	31.0	31.0	44.0	34.0	37.0	19.0	24.0	23.0	27.0	32.0	51.0	62.0	45.0	65.0	42.0	21.0	52.0	20.0	34.0	53.0	24.0	35.0	14.

Figure 22 X-bar and R chart for Chemical Process. (Copyright 1980–1998 Associates in Process Improvement)

NAME Impurity in Pellets DATE 6/28

PROCESS Chemical SAMPLE DESCRIPTION Grab sample about every 5 Hours

NUMBER OF SUBGROUPS (k) 24 BETWEEN (DATES) 6/25 and 6/28

NUMBER OF SAMPLES OR MEASUREMENTS PER SUBGROUP (n) 5

$\bar{\bar{X}} = \frac{\Sigma \bar{X}}{k} = \frac{4430.3}{24} = 184.59$	$\bar{R} = \frac{\Sigma R}{k} = \frac{844}{24} = 35.4$
XCHART	**R CHART**
UCL = $\bar{\bar{X}}$ + (A_2 * R)	UCL = D_4 * R
UCL = 184.59 + (.58 * 35.4)	UCL = 2.11 * 35.4
UCL = 184.59 + 20.53	UCL = 74.7
UCL = 205.12	
LCL = $\bar{\bar{X}}$ - (A_2 * R)	LCL = D_3 * R
LCL = 184.59 - (.58 * 35.4)	LCL = *
LCL = 184.59 - 20.53	LCL = -
LCL = 164.06	

Figure 23 Calculations for X-bar and R chart for Chemical Process. (Copyright 1980–1998 Associates in Process Improvement)

6. Calculate the control limits for the R chart using:

$$\text{UCL} = D4 * \bar{R}$$
$$\text{LCL} = D3 * \bar{R}$$

Note: $D3$ and $D4$ are factors that depend on the size of the subgroup and can be obtained from Figure 21. Note that there is no lower control limit for R when n is less than 7.

7. Calculate a scale for the X-bar chart such that the control limits enclose the inner 50% of the charting area. Calculate the scale for the R chart such that the upper control limit is placed 25–35% below the top of the chart.
8. Plot the $\bar{X}$'s on the X-bar chart and the R's on the R chart.
9. Draw the control limits and centerline on the X-bar chart.
10. Draw the control limits and centerline on the R chart.

The following example illustrates some of the important aspects concerning X-bar and R control charts. In a chemical process, a control chart was to be constructed to monitor the concentration of an impurity in finished pellets. Customers wanted the impurity to be stable below 200 ppm. Five grab samples were selected from the continuous process each day (approximately one every five hours). Data were collected for 24 days before control limits were calculated. Therefore, 24 subgroups were used in the calculations. Figure 22 contains the control chart and Figure 23 shows the calculations of the control limits. Since each subgroup contains five measurements, there is no lower control limit for the R chart.

After review of the control chart, the process was determined to be unstable. On June 8 and 17, points were above the upper control limit on the X-bar chart. Beginning on June 21, four points were below the lower control limit on the X-bar chart and there was a run of eight points in a row below the average. Since the process was unstable, action was taken to eliminate the special causes of variation. The special cause detected on June 8 was associated with poor color of the feedstock supplied by the Quality Chemical Company. Discussions with this supplier were initiated immediately and the problem was corrected. Material with better color was introduced into the process on June

21. The lower values of impurity that resulted were detected as a special cause on the X-bar chart. The special cause detected on June 17 was the result of a temporary 10% drop in the production rate. The production planning department was notified to make them aware of the effect of rates on impurity levels.

REFERENCES

Deming, W. E. (1986), *Out of the Crisis,* MIT Center for Advanced Engineering Study, Cambridge, MA.

Shewhart, W. A. (1931), *Economic Control of Quality of Manufactured Product,* Van Nostrand, reprint, by the American Society for Quality Control, Milwaukee, 1980.

ADDITIONAL READING

American Society for Quality Control (ASQC), Z1.1-1985, *Guide for Quality Control Charts;* Z1.2-1985, *Control Chart Method of Analyzing Data;* Z1.3-1985, *Control Chart Method of Controlling Quality during Production,* ASQC, Milwaukee, 1985.

American Society for Quality Control (ASQC). *Industrial Quality Control,* Special Memorial Issue, 1976.

Associates in Process Improvement, *The Improvement Handbook: Model, Methods, and Tools for Improvement,* API, Austin, TX, 1997.

Associates in Process Improvement, *Statistical Process Control,* API, Austin, TX, 1995.

Berwick, D. M., ‘‘Controlling Variation in Healthcare,’’ *Medical Care,* Vol. 29, No. 12, 1991, pp. 1212–1225.

Grant, E. L., and Leavenworth, R. S., *Statistical Quality Control,* 5th Ed., McGraw-Hill, New York, 1980.

Nelson, L. S., ‘‘Control Charts for Individual Measurements,’’ *Journal of Quality Technology,* Vol. 14, No. 3. 1982, pp. 172–173.

Nelson, L. S., ‘‘The Shewhart Control Chart—Test for Special Causes,’’ *Journal of Quality Technology,* Vol. 16, No. 4, 1984, pp. 237–239.

Nolan, T. W., and Provost, L. P., ‘‘Understanding Variation,’’ *Quality Progress,* May 1990, pp. 22–31,

Norman, C., and Provost, L., ‘‘Variation through the Ages,’’ *Quality Progress,* Special Variation Issue, December 1990, pp. 39–44.

Shewhart, W. A., *Statistical Method from the Viewpoint of Quality Control,* W. E. Deming, Ed., Department of Agriculture, Washington, DC, 1939.

CHAPTER 69

Statistical Process Control

John R. English
Terry R. Collins
University of Arkansas

1. INTRODUCTION 1856
2. TOOLSET FOR STATISTICAL PROCESS CONTROL 1857
- 2.1. Histogram 1857
- 2.2. Check Sheet 1858
- 2.3. Pareto Chart 1859
- 2.4. Cause-and-Effect Diagram 1859
- 2.5. Defect Concentration Diagram 1860
- 2.6. Scatter Diagram 1860

3. OVERVIEW OF CONTROL CHARTS 1861
- 3.1. Data Patterns on Control Charts 1863
- 3.2. AT&T Runs Rules 1863

4. CONTROL CHARTS FOR VARIABLES 1864
- 4.1. $\overline{X}$ and R Control Charts 1864
  - 4.1.1. Control Limits: Standards Known 1864
  - 4.1.2. Control Limits: Standards Not Known 1866
- 4.2. Other Charts 1868
- 4.3. Process Capability Analysis 1869
  - 4.3.1. Capability Ratios 1869
  - 4.3.2. C_{pk} 1869
  - 4.3.3. Inferential Approaches 1869

5. CONTROL CHARTS FOR ATTRIBUTE DATA 1871
- 5.1. Attribute Data 1871
- 5.2. Control Chart for Percent Nonconforming: *p* Chart 1872
- 5.3. Control Chart for Number of Defectives: *c* Chart 1874

REFERENCES 1876
APPENDIX 1876

1. INTRODUCTION

Product-to-product and service-to-service variation is potentially observable in all organizations. For example, a group of size 13 running shoes vary from shoe to shoe in form, fit, and finish. Even though the variation may not be observable by the human eye, it is observable if a more accurate measurement system is employed. Consumers of the product declare the product or service to be of inferior quality. Product variation may become increasingly observable to the consumer in complex systems. For example, consumers readily agree that there exists variation of cars of the exact same model and installed accessories. Otherwise, consumers would not insist upon driving the vehicle before purchase and the producer would not have to provide a warranty period to satisfy the concern of purchasing a failure-prone vehicle (commonly called a ‘‘lemon’’).

Examples such as these provide motivation for professionals to consider the implications of product-to-product and service-to-service variation. Clearly, product variation is key to customer satisfaction and loyalty; therefore, successful producers realize this and take appropriate actions to control the variation or quality of the products or services provided.

To monitor the variation of a process or service, data are collected and analyzed for product critical performance metrics. Two types of data are common. Data may be measured on a continuous scale, for example, length, weight, and so on. Such measurements are called variable data. Alternatively, if process observations are of the classification type, they are called attribute data. Examples

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

of such data include the number of defects in an inspection unit, number of defective units in a sample, and so on.

Well-known and documented techniques exist to monitor product variation while it is within the producer's environment. Most of the techniques require the observations or data to be statistically independent. That is, the data for a specific performance measurement are assumed to have no relationship to prior or successive observations. It is assumed that no correlation exists between data collected prior to or following a specific observation. The techniques used to monitor such data are collectively called statistical process control (SPC). These techniques are utilized in consumer-oriented industries. Some of the more prominent or useful techniques are presented in this chapter. Specifically, seven tools for SPC are reviewed, and their applicability is examined. Furthermore, common and improved approaches for process capability analysis are presented.

2. TOOLSET FOR STATISTICAL PROCESS CONTROL

Ishikawa (1985) presents seven common tools for SPC. He claims that 95% of quality-related problems can be resolved with these tools. The tools have been called the "Magnificent Seven" and are as follows:

1. Histogram
2. Check sheet
3. Pareto chart
4. Cause-and-effect diagram
5. Defect concentration diagram
6. Scatter diagram
7. Control chart

In the following subsections, brief overviews and practical examples are presented with the exception of the control charts, which are the topics of detailed discussions in Sections 3, 4, and 5.

2.1. Histogram

The histogram is a graphical tool that presents the relative frequency (in number or percentage) of observations within predefined cells. In other words, the histogram provides a graphical representation of a population or process under examination. The natural spread or distribution, the central tendency, and the variation of the process are readily observable in the histogram. Since the data supporting the histogram are defined by individual observations, the comparison of the process variation to the allowable spread, as defined by specifications, is easily demonstrated by plotting the specifications on the histogram. The histogram is a simple tool that provides a wealth of information regarding the natural tendencies of a product or process under study. Furthermore, from a probability perspective, the histogram provides an empirical estimate of the probability density function defining the product or process.

Example 1 presents an example of the use of the histogram.

Example 1. The following data have been collected from your process. In particular, you produce a child's toy. You have collected 90 values of a critical performance measurement that reflects surface quality for the manufactured product. The upper specification limit (USL) and the lower specification limit (LSL) are known to be 54 and 47, respectively. It is desired to demonstrate the natural tendency, variability, and relationship to the specification limits of the process by constructing a histogram of the data. With the collected data, the histogram is calculated by accumulating the number of observations within prescribed bands. The number of observations can be used to determine the relative frequency of each cell as demonstrated below.

Data for Example 1

The resulting histogram may be plotted with most spreadsheet software systems.

For this example, the resulting histogram is as follows:

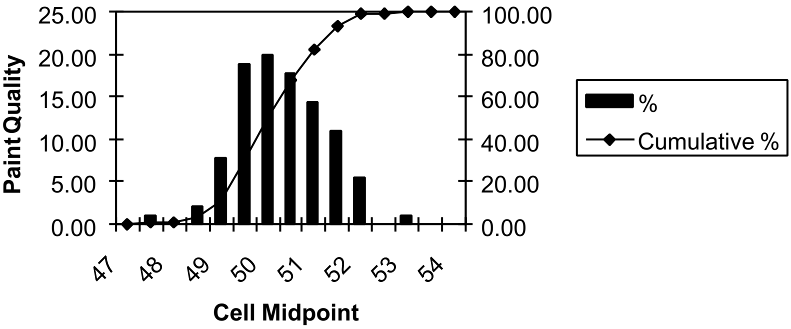


Figure 1 Histogram.

Cell Mid Point	# Observations	%	Cumulative %
47	0	0.00	0.00
47.5	1	1.11	1.11
48	0	0.00	1.11
48.5	2	2.22	3.33
49	7	7.78	11.11
49.5	17	18.89	30.00
50	18	20.00	50.00
50.5	16	17.78	67.78
51	13	14.44	82.22
51.5	10	11.11	93.33
52	5	5.56	98.89
52.5	0	0.00	98.89
53	1	1.11	100.00
53.5	0	0.00	100.00
54	0	0.00	100.00
Total =	90		

As observed, the paint quality measurements are contained within the specifications and are relatively symmetric.

2.2. Check Sheet

The check sheet is a simple and useful tool that is often used in the early stages of a quality control program. It provides a uniform and consistent means for data collection and analysis. Like many of the traditional tools, the check sheet is defect oriented and is used to classify the types of defects frequently found for a product or service. It is used to spot problems areas by identifying defect types that frequently occur. As a result, a check sheet is often used as the input to the next tool, called the Pareto chart. The check sheet is most often tabular in format as presented in Example 2.

Example 2. For the same toy of Example 1, defect data have been collected on a weekly basis to attempt to classify the underlying sources of product defects. In particular, several defect types have been established, and the operators are given a matrix simply to track the number of defects of each category for a multiday study period. The resulting check sheet is presented as follows:

Toy 1 Check Sheet: Defect Count—Finished Surface

Type	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Total	%	Cumulative %
Paint drips	11	9	10	11	10	12	17	80	42.3	42.3
Scratches	8	7	7	6	6	9	8	51	27.0	69.3
Handling marks	3	5	4	7	4	5	5	33	17.5	86.8
Chips	2	1	0	0	3	2	2	10	5.3	92.1
Poor color match	1	2	3	0	2	0	0	8	4.2	96.3
Poor assembly	2	1	0	0	1	0	0	4	2.1	98.4
Failure of functional test	0	1	0	0	1	0	0	2	1.1	99.5
Other	0	0	0	0	1	0	0	1	0.5	100.0
						Total	189			

As demonstrated above, the percent breakdown and accumulation of percentages are readily determined for each defect type. If the percent values are ranked from largest to smallest, a Pareto chart (discussed in the next section) is effectively computed, which makes the Pareto chart an obvious companion to the check sheet.

2.3. Pareto Chart

The Pareto chart was named after the Italian economist Vilfredo Pareto (1848–1923). Pareto observed that in certain cultures or economies a majority of the wealth was held by a small segment of the population.

In the context of quality or the observance of defects, time has proved over and over that there are usually a disproportionately small number of defect types that cause a majority of the problems. As a general rule of thumb, it is commonly observed that around 20% of the defect types can cause around 80% of the problems. Pareto analysis and the resulting Pareto chart are used to identify what is commonly known as the "vital few and the trivial many." Most quality professionals agree that the Pareto chart is one of the best tools, if not the best tool, within the Magnificent Seven. Montgomery (1996) states that "In general, the Pareto chart is one of the most useful of the 'magnificent seven'. Its applications to quality improvement are limited only by the ingenuity of the analyst."

In the context of the check sheet, the resulting Pareto chart is simply the relative frequency of the defect types as observed during a study period. Pareto charts are effectively the plot of the relative frequency or cumulative frequency, of defect types.

Example 3 provides an enlightening example in the use of the Pareto chart.

Example 3. From the toy data of Example 2, plot the associated Pareto chart. From Example 2, you see that the relative frequencies and cumulative frequencies have been calculated. This is, in fact, the Pareto analysis. The resulting Pareto chart is as follows:

Over 70% of the defect types are caused by two of the eight defect types. The resulting Pareto chart can be used to identify priorities for defect-reduction programs. Assuming that paint drops and scratches are of critical concern to the customer and are yielding excessive costs, paint drops and scratches should be the focus of the defect-reduction program.

2.4. Cause-and-Effect Diagram

Once a defect-reduction program has been defined and the focus of the study has been identified, it is necessary to determine the cause of each of the defect types. The cause-and-effect diagram has proven to be an invaluable team-oriented approach to brainstorm the cause of a specific product defect, or more generally a consumer concern. The cause-and-effect diagram is a graphical tool that is most often constructed in a team setting to identify the contributing factors and their relationship to a defect type. The cause-and-effect diagram is used to identify areas of future data collection and analysis to identify process-improvement strategies. The chart has also been called the fishbone diagram or the Ishikawa diagram (Ishikawa 1985).

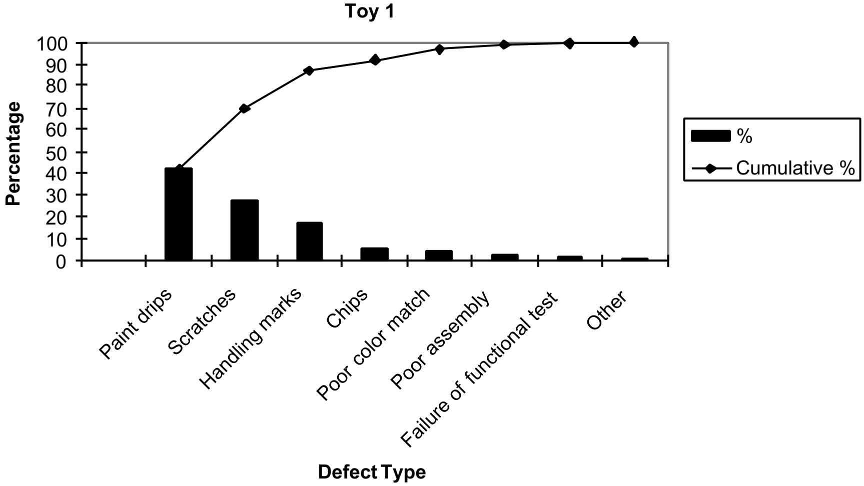


Figure 2 Pareto Chart.

The cause-and-effect diagram basically dissects a defect type into the associated causes. The causes can be further subdivided into additional groupings if the team believes such structure is necessary. Japanese-based tools such as the cause-and-effect diagram are useful for creative problem solving, and teams using such tools should be allowed flexibility in customizing tools for their specific needs. Example 4 presents the use of a cause-and-effect diagram.

Example 4. Within your facility, part accuracy is a consistent problem. You have established a cross-functional team to investigate the causes of this problem in an effort to identify opportunities of process improvement. The resulting cause-and-effect diagram is as follows:

2.5. Defect Concentration Diagram

Of the seven tools of the Magnificent Seven, the defect concentration diagram is most likely the simplest, and one of the most effective, for process-improvement activities. The defect concentration diagram is a schematic diagram of the part being produced. It provides all relevant views, and all defect types are recorded in accordance with a preset key. The operator simply records each observed defect on the defect concentration diagram. This diagram is often used in conjunction with the cause and effect diagram in an effort to identify process opportunities for improvement. The diagram is used to determine the specific location of defect types, and as a result, provides insightful information on process problems and the source of product defects.

Example 5 presents an example of a defect diagram.

Example 5. You produce a metal painted part that serves as the shell of the toy of Example 1. It is effectively a triangular column. The resulting defect concentration diagram for the operators is presented in Figure 4:

The defect concentration diagram is placed at a strategic point in the process, and the operator is instructed to document all observed defects according to location and type. For example, if five paint drips are observed on the top right corner of side 1, five A's would be placed on the top right corner of side 1 of the defect concentration diagram. The collective view of the completed defect concentration diagram provides conclusive evidence supporting the investigation of process-improvement activities.

2.6. Scatter Diagram

The sixth tool of the Magnificent Seven is the scatter diagram. For readers familiar with simple linear regression, the scatter diagram is considered the first step in determining the functional relationship between two variables. The scatter diagram is effectively the plotting of a process or product performance metric against some controllable variable. Most spreadsheet software systems provide the graphical tools to minimize the technical details of constructing such diagrams. Example 6 provides an illustration of the use of the scatter diagram.

Example 6. At the Department of Industrial Engineering at the University of Arkansas, miniature catapults are used to demonstrate the usefulness of statistical modeling. The teaching tool is effec-

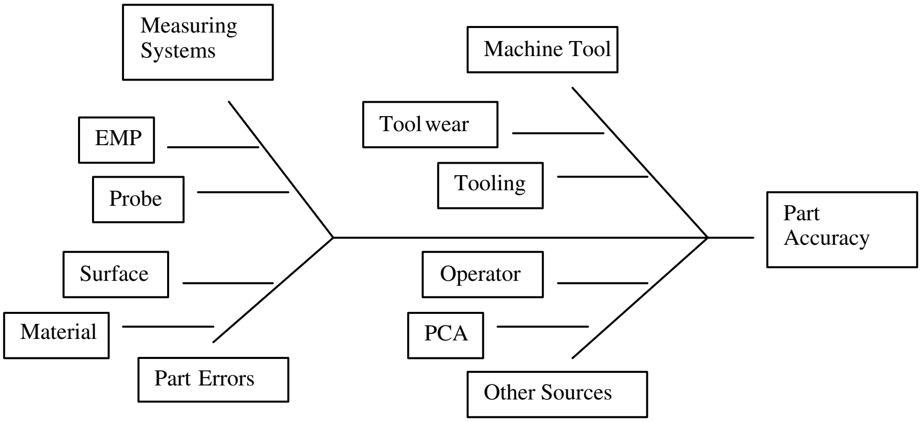


Figure 3 Cause-and-Effect Diagram.

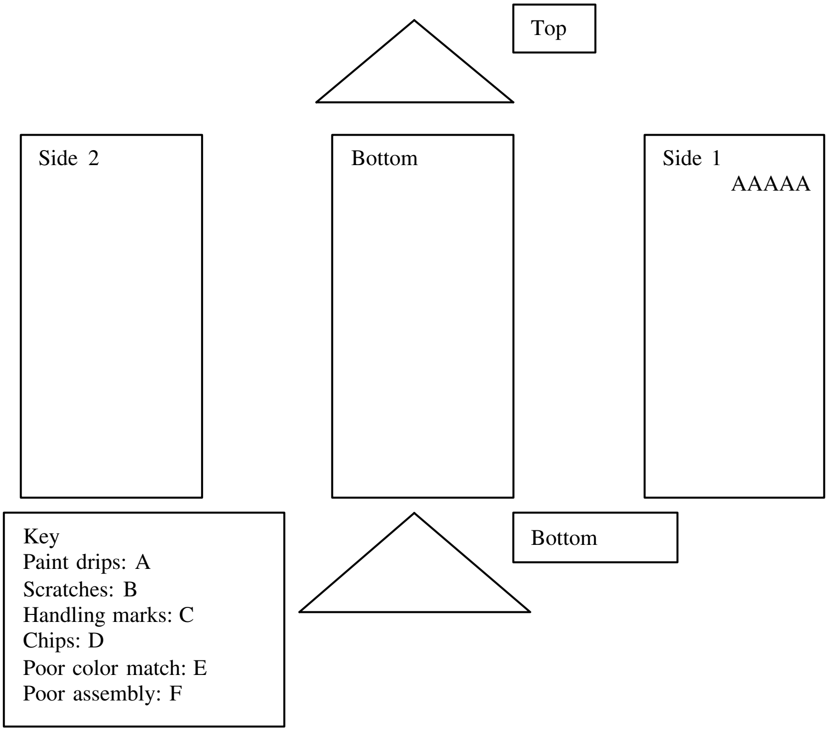


Figure 4 Defect Concentration Diagram.

tively a miniaturization of the classic weapon of ages past. As the arm of the catapult is pulled back further—the angle is called the throwing angle—a given projectile is launched to realize increasing distances. The first step in modeling this process is to demonstrate graphically the relationship between the throwing angle and the resulting throwing distances. In particular, the projectile is launched five times from six different throwing angles. The resulting scatter diagram is given in Figure 5.

There appears to be a fairly linear relationship between the throwing angle and the distance thrown.

3. OVERVIEW OF CONTROL CHARTS

As previously described, almost every product, process, or service varies. The control chart is used to monitor natural process variation in an effort to expose times that the process has shifted such that nonnatural variation is realized. Within the SPC domain, terms have been coined to define various types of variation. The natural variation in the process that cannot be avoided is commonly called common cause variation. It is assumed that such variation is not controllable. Alternatively, undesirable changes occur, causing a process to change in its natural behavior. The variation of the process may become excessively large in view of its common causes, or the central tendency of the data may shift in a positive or negative direction. Such unnatural variations in the data are known as special causes.

In view of these types of variation, it is desired to design SPC approaches such that the user maximizes opportunities for detecting special causes in the midst of the uncontrollable chance causes. In the application of SPC, samples or subgroups are collected and analyzed and are used for this judgment. The logical collection of the subgroups or samples such that opportunities for detecting special causes are maximized is commonly known as rational subgrouping. Most introductory texts in quality control have detailed discussion of this topic (i.e., Grant and Leavenworth 1980).

The tool that is used to monitor process variation over time is known as the control chart. Control charts originate from the work of Walter Shewhart (1927) and are often referred to as Shewhart control charts. Effectively, process observations based upon collected samples or subgroups, at fixed points in time, are plotted in accordance to time. As long as the current observation is within fixed

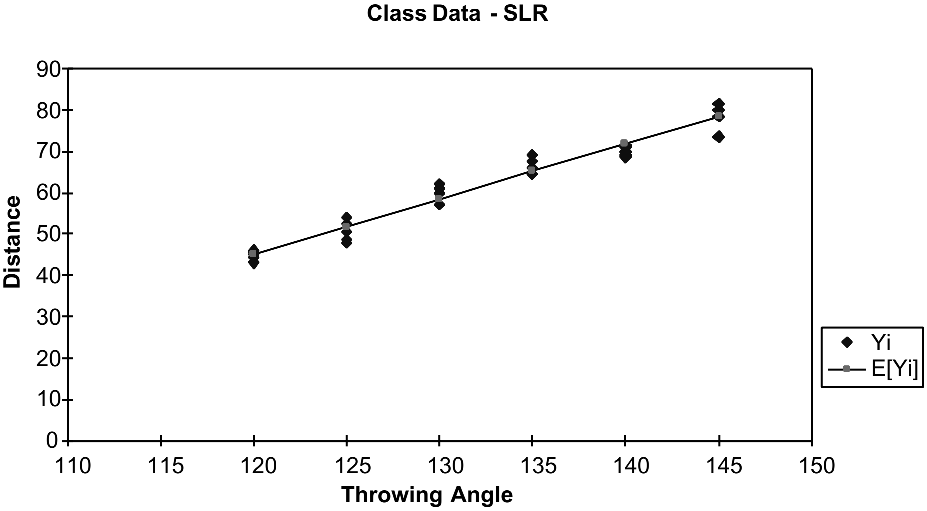


Figure 5 Scatter Diagram.

intervals, called control limits, or no unnatural trends in the data are observed (defined in Section 3.2), it is concluded that the process is operating in the presence of common causes only. Such a process is declared to be operating in a state of statistical control. Alternatively, if the process yields observations that exceed the control limits, or unnatural variation exists, it is concluded that special causes exist in addition to the common causes. Commonly, this process state is known as being out of control.

Figure 6 presents this graphical tool—the control chart. As discussed, there exists a line defining the central tendency as well as the control limits. This process is obviously operating under the presence of both special causes and common causes and should be judged out of control.

The basis of the centerline and the control limits are essentially the same for all charts with the exception of the type of data. Assume that the quality characteristic being monitored is Y_t, for which the mean is known to be $E[Y_t]$, and the variance is known to be $V[Y_t]$. The general centerline (CL) and resulting control limits are:

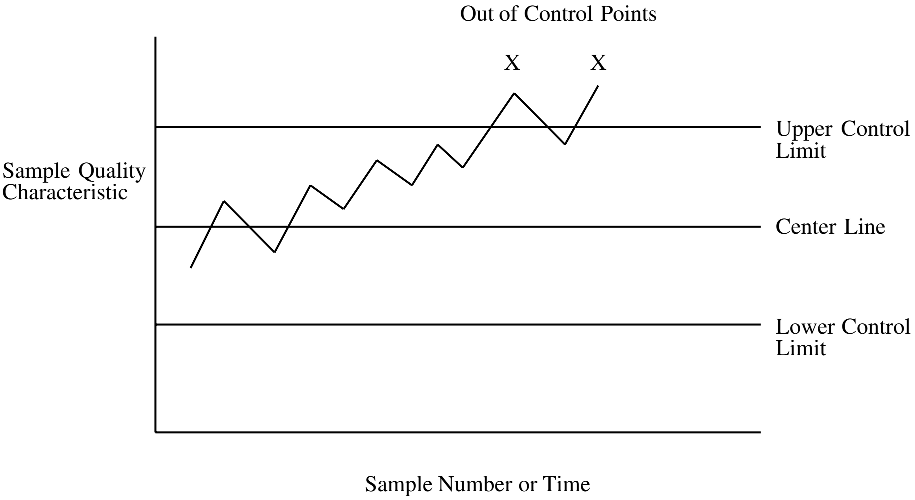


Figure 6 Example Control Chart.

$$\text{Upper control limit (LCL)} = E[Y_t] + k\sqrt{V[Y_t]}$$

$$\text{Centerline (CL)} = E[Y_t]$$

$$\text{Lower control limit (UCL)} = E[Y_t] - k\sqrt{V[Y_t]}$$

where k is defined by the user. In most applications within the United States, k is assumed to be 3 regardless of the type of data.

In Section 4, two control charts for variable data are presented. Both the $\overline{X}$ control chart for monitoring the process mean and the R control chart for monitoring the process variation are presented. In Section 5, two control charts for attribute data are discussed: the p control chart for monitoring percent nonconforming and the c chart for monitoring the number of defects in a sample. Furthermore, brief discussions of data patterns on control charts and recommended supplemental rules for judging nonrandom trends on a control chart are presented.

3.1. Data Patterns on Control Charts

Control charts are powerful tools for monitoring the variation of a process. Furthermore, the nonnatural trends on a control chart can provide significant diagnostic information regarding the cause of a process disturbance. In this section, prevalent trends from out-of-control processes are presented and the suspect causes are briefly discussed. References such as AT&T (1985) and Montgomery (1996) provide more in-depth presentations of this discussion.

First, it is the assumption of most control chart applications that the process from which observations are collected is stable. That is, the statistical behavior of the process is time invariant in that the underlying distribution is fixed, yielding a fixed mean and variance. In some applications, process observations are collected such that multiple processes may feed the data-collection station. As a result, the observations may have a tendency to cluster around the control limits and be sparsely observed around the CL or the mean. This observation is commonly referred to as a mixture.

Secondly, processes may be overly sensitive to factors within the day; for example, temperature cycles. Such trends in the data are known as cycles and are not well suited for traditional applications of control charts, even though the cycling behavior is a natural part of the process variation. When faced with cycling or autocorrelation in the data, it is suggested the advanced techniques, such as the moving centerline exponentially weighted moving average chart, be considered, as presented in Montgomery and Mastrangelo (1991).

Third, a trend (see Figure 6) in the process is typically realized on the control chart as a generally increasing or decreasing trend in the data. Such trends in the data can be caused by process degradation, such as tool or machine wear. In some cases, such trends cannot be tolerated and can be adequately detected with standard control charts. In situations where such trends are a natural part of the process (i.e., tool wear), alternative approaches should be considered. The simplest approach is to fit a reasonable regression model to the process and monitor the process accordingly. Alternatively, Quesenberry (1988) develops an alternative SPC approach for a tool-wear process.

If a shift has occurred in the central tendency of the process, the process data will cluster around the new mean. Such shifts are often caused by changes in raw materials or process settings. In most all cases, such process behavior is not desirable and can be detected quickly with traditional control charts.

3.2. AT&T Runs Rules

Clearly, the term *nonrandom* trend must be consistently defined for the shop floor operators. Such definitions are made by the consistent adoption of rules for judging the current status of a process. Probably the most common set of rules for examining the state of the process is known as the AT&T runs rules (AT&T 1985). In AT&T (1985), these rules are called “Tests for Instability.” In applying these tests, each half of the control chart (the area between the CL and UCL and the CL and LCL) is divided into three equally spaced zones, as shown in Figure 7.

In accordance with the AT&T handbook, a process is judged out of control or unstable if any one of the following rules apply:

Test 1: A single point falls outside of the control limits (beyond zone A).

Test 2: Two out of three successive points fall in zone A or beyond for a given side.

Test 3: Four out of five successive points fall in zone B or beyond for a given side.

Test 4: Eight successive points fall in zone C or beyond for a given side.

If these rules are consistently employed throughout a plant, type I errors should be adequately controlled while ensuring reasonable detection of undesirable process shifts. A type I error, in the context of control charts, is the event that the operator falsely concludes that the process is out of control.

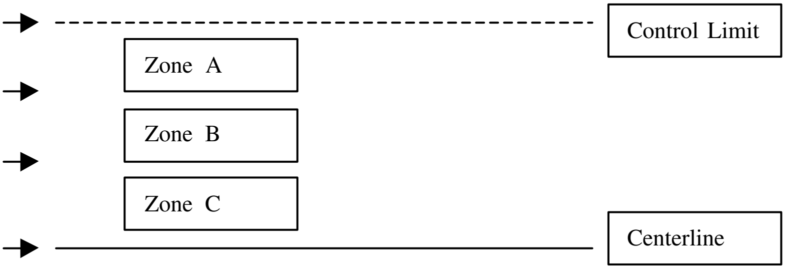


Figure 7 AT&T Runs Rules Zones.

4. CONTROL CHARTS FOR VARIABLES

4.1. $\overline{X}$ and R Control Charts

The most common means of monitoring variable data is with $\overline{X}$ the R and control chart combination. It is assumed that the observations collected from the process are independent and normally distributed. The $\overline{X}$ chart is utilized to monitor the process mean, and the R chart is used to monitor the process variation. Without exception, the $\overline{X}$ and R charts should always be used together. The normality assumption embeds the fact that two parameters, the mean and variance, completely characterize the process; therefore, both control charts are necessary to monitor the process completely.

From an applications perspective, n observations, X_{i1}, X_{i2}, . . . X_{in} are collected from the process at fixed points in time, i. These n observations from sampling interval i are commonly referred to as subgroup i. As a rule of thumb, the number of observations in each subgroup, n, should be between 4 and 6. Furthermore, for slight departures from normality, such sample averages, $\overline{X}$, can be assumed to be approximately normal, due to the central limit theorem (in as much as the underlying distribution is not excessively skewed). Estimates for the process mean and variance at a point in time i are based upon the sample average and range. That is,

$$\text{Sample average}_i = \overline{X}_i = \sum_{i=1}^{n} x_{ij}/n$$

$$\text{Range}_i = R_i = \text{maximum}(X_{i1}, X_{i2}, \ldots, X_{in}) - \text{minimum}(X_{i1}, X_{i2}, \ldots X_{in})$$

4.1.1. Control Limits: Standards Known

If the mean, μ, and the variance, σ^2, are known for the underlying process, the upper and lower control limits for these charts ($\overline{X}$ and R) can be determined as follows:

For the $\overline{X}$ chart:

$$\text{UCL}_{\overline{X}} = \frac{\mu + 3\sigma}{\sqrt{n}}$$

$$\text{CL}_{\overline{X}} = \mu$$

$$\text{LCL}_{\overline{X}} = \frac{\mu - 3\sigma}{\sqrt{n}}$$

For the R chart:

$$\text{UCL}_R = D_2\sigma$$

$$\text{CL}_R = d_2\sigma$$

$$\text{LCL}_R = D_1\sigma$$

where μ = the process mean
σ = the process standard deviation
d_2, D_2, and D_1 = tabled values for a given n (subgroup size) and are included in the Appendix.

The AT&T runs rules should always be employed for the $\overline{X}$ chart and should be used for the R chart when the sample size is five or more (AT&T 1985). The first rule of the AT&T runs rules should always be used for both charts.

To illustrate the use of the $\overline{X}$ and R charts, the Example 7 is presented:

Example 7. Five observations have been collected from a process where the mean, μ, and the standard deviation, σ, are known to be 50 and 1, respectively. Construct and plot the necessary control limits and calculated observations for both the $\overline{X}$ and R charts. The data are collected and the resulting $\overline{X}_i$'s and R_i's are computed as follows:

i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	$\overline{X}_i$	R_i
1	51.69	49.59	52.51	51.07	50.01	50.97	2.92
2	49.37	49.13	49.77	49.61	49.89	49.55	0.77
3	50.59	50.08	50.4	47.82	48.67	49.51	2.77
4	51.24	49.79	50	49.89	51.98	50.58	2.19
5	50.78	48.82	50.59	51.48	49.12	50.16	2.66
6	49.86	48.43	49.1	48.76	49.45	49.12	1.43
7	50.86	50.78	50.43	52.12	49.7	50.78	2.42
8	50.94	51.61	50.2	49.08	51.63	50.70	2.55
9	49.25	49.46	51.71	49.08	51.55	50.21	2.63
10	50.52	50.92	49.06	50.43	51.46	50.48	2.40
11	49.45	49.02	51.23	49.96	49.9	49.91	2.20
12	50.12	51.06	50.99	50.98	50.09	50.65	0.97
13	49.32	49.93	51	50.68	48.14	49.81	2.86
14	50.04	50.1	49.08	49.14	50.3	49.73	1.22
15	50.25	49.78	50.51	49.37	48.21	49.62	2.30
16	50.07	49.69	48.81	50.72	50.25	49.91	1.91
17	49	50.25	49.93	48.87	49.32	49.47	1.38
18	49.64	51.22	49.84	50.51	50.68	50.38	1.58
19	51.5	49.49	48.24	48.02	50.86	49.62	3.48
20	51	49.58	49.57	50.08	51.75	50.39	2.19
21	49.66	49.41	47.43	49.68	49.39	49.11	2.25
22	48.64	48.53	50.03	49.85	48.77	49.16	1.50
23	50.35	49.34	51.2	50.82	49.72	50.29	1.87
24	51.49	51.44	50.78	48.91	49.86	50.50	2.57
25	49.81	49.45	49.43	51.46	49.76	49.98	2.03
26	49.82	49.99	51.3	49.4	50.28	50.15	1.90
27	48.59	49.31	49.95	50.63	48.78	49.45	2.04
28	51.67	51.13	49.33	50.93	48.61	50.33	3.05
29	50	50.85	49.53	50.7	49.3	50.08	1.55
30	50.2	48.74	50.42	51.27	48.57	49.84	2.71

The resulting control limits are as follows:
For the $\overline{X}$ chart:

$$\mathrm{UCL} = 50 + \frac{3(1)}{\sqrt{5}} = 51.342$$

$$\mathrm{CL} = 50$$

$$\mathrm{LCL} = 50 - \frac{3(1)}{\sqrt{5}} = 48.658$$

For the R chart:

$$\mathrm{UCL}_R = 4.918(1) = 4.918$$

$$\mathrm{CL}_R = 2.326(1) = 2.326$$

$$\mathrm{LCL}_R = 0(1) = 0$$

The resulting control charts are given in Figures 8 and 9.

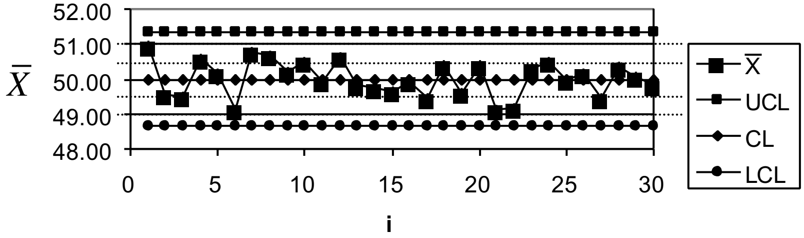


Figure 8 $\overline{X}$ Chart: Standards Known.

From the control charts above, the process exhibits stability using the AT&T runs rules on the $\overline{X}$ chart (see zones marked on Figure 8) and for rule number 1 for the R chart; therefore, the process is judged to be in a state of statistical control.

In most all cases, the process mean, μ, and the process standard deviation, σ, must be estimated from process data and the associated calculation of control limits must be adjusted. In the next section, the calculation of the control limits when the standards are not known is presented.

4.1.2. Control Limits: Standards Not Known

As stated previously, the true values of μ and σ are typically estimated with sampled data from a process. For the quality of estimation, it is typically assumed that at a minimum, 25 to 30 subgroups should be available before estimates of μ and σ are made ($\hat{\mu}$ and $\hat{\sigma}$). The following relationships are used to determine the necessary estimates:

$$\hat{\mu} = \overline{\overline{X}} = \sum_{i=1}^{m} \overline{X}_i / m$$

$$\hat{\sigma} = \overline{R} = \left[\sum_{i=1}^{m} R_i / m \right] \bigg/ d_2$$

where m is the number of subgroups
d_2 is a table look-up value from the Appendix

The control limits when standards are unknown are calculated as follows:
For the $\overline{X}$ chart:

$$\mathrm{UCL}_{\overline{X}} = \frac{\hat{\mu} + 3\hat{\sigma}}{\sqrt{n}}$$

$$\mathrm{CL}_{\overline{X}} = \hat{\mu}$$

$$\mathrm{LCL}_{\overline{X}} = \frac{\hat{\mu} - 3\hat{\sigma}}{\sqrt{n}}$$

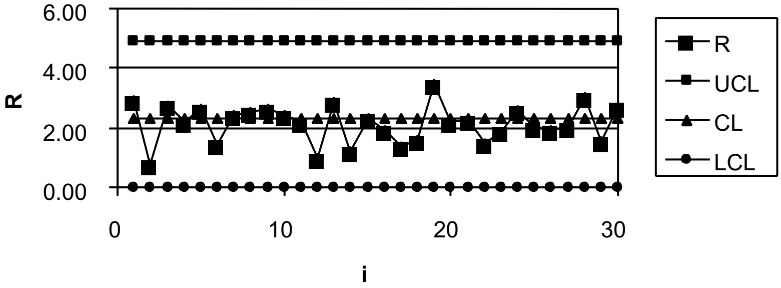


Figure 9 R Chart: Standards Known.

If A_2 is set equal to $3/d_2\sqrt{n}$ and $\overline{\overline{X}}$ and $\overline{R}/d_2$ are substituted for $\hat{\mu}$ and $\hat{\sigma}$, respectively, the resulting control limits can be determined as follows:

$$\mathrm{UCL}_{\overline{X}} = \overline{\overline{X}} + A_2\overline{R}$$

$$\mathrm{CL}_{\overline{X}} = \overline{\overline{X}}$$

$$\mathrm{LCL}_{\overline{X}} = \overline{\overline{X}} - A_2\overline{R}$$

For the R chart:

$$\mathrm{UCL}_{\overline{R}} = D_3\overline{R}$$

$$\mathrm{CL}_{\overline{R}} = \overline{R}$$

$$\mathrm{LCL}_{\overline{R}} = D_4\overline{R}$$

where A_2, D_3, and D_4 are tabled values for subgroup size n (see Appendix).

The AT&T runs rules should always be employed for the $\overline{X}$ chart and should be used for the R chart when the sample size is five or more (AT&T 1985). The first rule of the AT&T runs rules should always be used for both charts.

In practice, the initial estimation of the control limits may require an iterative process in securing control limits that are statistically valid. In particular, it may be necessary to estimate the limits initially. If a given $\overline{X}$ or R plots beyond the initial control limits, it is recommended that the cause of the extraordinary point be resolved and then removed from the calculation. That is, the control limits should be recalculated with the out-of-control point(s) removed. In some extreme cases, this iterative process may be repeated several times. Once the observed data are contained within the trial limits, continued production should be used to collect more data to further validate the limits. Once process stability is obtained, that is control is maintained for an extended period, the new control limits should become the operational standard. Furthermore, once the trial limits are established, the AT&T Runs Rules should be deployed in totality to insure protection against process disturbances.

Example 8 demonstrates the construction of $\overline{X}$ and R charts when standards are not known.

Example 8. Consider the process data of Example 7, except do not assume that the standards are known. Determine the appropriate trial control limits for this process.

i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	$\overline{X}_i$	R_i
1	51.69	49.59	52.51	51.07	50.01	50.97	2.92
2	49.37	49.13	49.77	49.61	49.89	49.55	0.77
3	50.59	50.08	50.4	47.82	48.67	49.51	2.77
4	51.24	49.79	50	49.89	51.98	50.58	2.19
5	50.78	48.82	50.59	51.48	49.12	50.16	2.66
6	49.86	48.43	49.1	48.76	49.45	49.12	1.43
7	50.86	50.78	50.43	52.12	49.7	50.78	2.42
8	50.94	51.61	50.2	49.08	51.63	50.70	2.55
9	49.25	49.46	51.71	49.08	51.55	50.21	2.63
10	50.52	50.92	49.06	50.43	51.46	50.48	2.40
11	49.45	49.02	51.23	49.96	49.9	49.91	2.20
12	50.12	51.06	50.99	50.98	50.09	50.65	0.97
13	49.32	49.93	51	50.68	48.14	49.81	2.86
14	50.04	50.1	49.08	49.14	50.3	49.73	1.22
15	50.25	49.78	50.51	49.37	48.21	49.62	2.30
16	50.07	49.69	48.81	50.72	50.25	49.91	1.91
17	49	50.25	49.93	48.87	49.32	49.47	1.38
18	49.64	51.22	49.84	50.51	50.68	50.38	1.58
19	51.5	49.49	48.24	48.02	50.86	49.62	3.48
20	51	49.58	49.57	50.08	51.75	50.39	2.19
21	49.66	49.41	47.43	49.68	49.39	49.11	2.25
22	48.64	48.53	50.03	49.85	48.77	49.16	1.50
23	50.35	49.34	51.2	50.82	49.72	50.29	1.87
24	51.49	51.44	50.78	48.91	49.86	50.50	2.57
25	49.81	49.45	49.43	51.46	49.76	49.98	2.03
26	49.82	49.99	51.3	49.4	50.28	50.15	1.90
27	48.59	49.31	49.95	50.63	48.78	49.45	2.04
28	51.67	51.13	49.33	50.93	48.61	50.33	3.05
29	50	50.85	49.53	50.7	49.3	50.08	1.55
30	50.2	48.74	50.42	51.27	48.57	49.84	2.71
						50.02	2.14
						$\overline{\overline{X}}$	$\overline{R}$

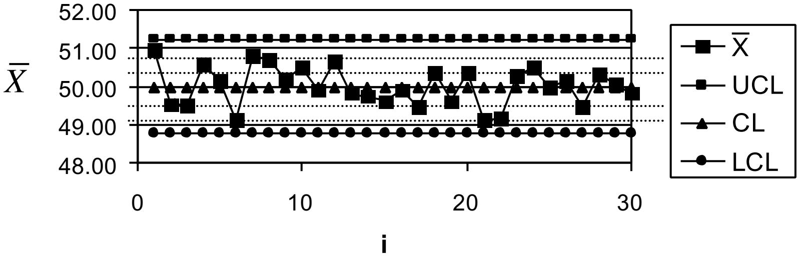


Figure 10 $\overline{X}$ Chart.

The resulting control limits:
For the $\overline{X}$ chart:

$$\mathrm{UCL}_{\overline{X}} = 50.02 + 0.577(2.14) = 51.255$$

$$\mathrm{CL}_{\overline{X}} = 50.02$$

$$\mathrm{LCL}_{\overline{X}} = 50.02 - 0.577(2.14) = 48.785$$

For the R chart:

$$\mathrm{UCL}_{\overline{R}} = 2.115(2.14) = 4.526$$

$$\mathrm{CL}_{\overline{R}} = 2.14$$

$$\mathrm{LCL}_{\overline{R}} = 0(2.14) = 0$$

The resulting control charts are given in Figures 10 and 11.

As demonstrated, all points plot within the control limits; therefore, these trial limits should be employed for production. Furthermore, it is highly recommended that the production also utilize the AT&T runs rules on the $\overline{X}$ chart (see zones on Figure 10) for subsequent production to monitor the state of the process.

4.2. Other Charts

There are many other control charts to monitor variable data. In each case, there should be a control chart for monitoring the mean and a control chart for monitoring the variance. These additional charts include the cumulative sum charts, exponentially weighted moving average charts, modified control charts, trend charts, control charts for short production times, individuals charts, moving average charts, moving range charts, control charts for variable sample sizes, S chart, variance charts, and acceptance control charts. The reader is referred to comprehensive textbooks such as Montgomery (1996) for a complete treatise on these more advanced techniques.

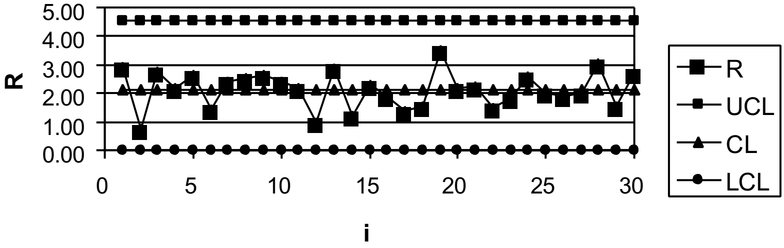


Figure 11 R Chart.

4.3. Process Capability Analysis

Once the process has generated sufficient data to support the conclusion that it is stable or in a state of statistical process control, it is desirable to judge the process's ability to meet customer specifications, that is, the upper specification limit (USL) and lower specification limit (LSL). Within most industries, process capability ratios are used to measure a given process's ability to meet specifications. The effort employed to make this judgment is known as process capability analysis (PCA).

The PCA ratios common to industry are C_p and C_{pk}. The ratios provide different perspectives to PCA. Generally, it is desirable that the ratios exceed 1 to ensure that the process is capable of meeting specifications.

4.3.1. Capability Ratios

The first ratio considered is C_p. This capability ratio estimates the potential that a process has to meet specifications. As demonstrated in the formulation below, the mean of the process is not accounted for.

$$C_p = \frac{\text{USL} - \text{LSL}}{6\hat{\sigma}}$$

4.3.2. C_{pk}

With the C_p calculation above, it is clear that a very good C_p could be determined while the process is generating practically all-bad product. As a result, C_{pk} has been formulated to embed the concern of process centering. C_{pk} is calculated as follows:

$$C_{pk} = \text{minimum}\left(\frac{\text{USL} - \hat{\mu}}{3\hat{\sigma}}, \frac{\hat{\mu} - \text{LSL}}{3\hat{\sigma}}\right)$$

In the formulas above for C_p and C_{pk}, $\hat{\sigma}$ may be determined with almost any known technique. Since the process must be assumed to be in a state of statistical control, it is usually found that $\overline{R}/d_2$ is the most convenient estimate of σ.

It is important to note that the now-famous Motorola 6σ program is based upon PCA. This highly effective program simply requires that C_p and C_{pk} be greater than 2.

Example 9 presents the use of the PCA ratios.

Example 9. For the data presented in Example 8, compute both C_p and C_{pk} and make the appropriate PCA conclusions. The USL is 54 and the LSL is 47.

For the data collected,

$$\hat{\mu} = 50.02$$

$$\hat{\sigma} = \frac{\overline{R}}{d_2} = \frac{2.14}{2.326} = 0.920$$

$$C_p = \frac{54 - 47}{6(0.920)} = 1.268$$

$$C_{pk} = \text{minimum}\left(\frac{54 - 50.02}{3(0.920)}, \frac{50.02 - 47}{3(0.920)}\right) = \text{minimum}(1.442, 1.094) = 1.094$$

As a result, it appears that the process is capable of meeting specifications, but the performance is not at the 6σ level, since C_p and C_{pk} are both less than 2.

4.3.3. Inferential Approaches

In the previous section, the statistical inference of PCA using the capability ratios, C_p and C_{pk}, is questionable. This stems from the fact that only a single point estimate is made and compared to a standard of 1. Clearly, due to the random nature of any process, sample estimates or observed values of C_p and C_{pk} can be greater than 1 when, in fact, the true process values are less than 1.

Chou et al. (1990) provide an inferential approach to PCA using C_p and C_{pk}. Their work necessarily assumes the observations collected from the process in question to be independent and to follow the normal distribution. They establish minimum critical values for both of the parameters that provide a statistical basis to concluding the capability of the process. Essentially, they establish a critical value for both C_p and C_{pk} for the appropriate test of hypothesis. The test of hypotheses for the ratios is as follows:

For C_p,

$$H_o : C_p < C_0$$
$$H_A : C_p \geq C_0$$

where C_0 is the desired minimum value of C_p such that the process is deemed capable.

For C_{pk},

$$H_o : C_{pk} < C_k$$
$$H_A : C_{pk} \geq C_k$$

where C_k is the desired minimum value of C_{pk} such that the process is deemed capable.

For the estimates of C_p and C_{pk}, the unbiased estimate of σ^2, S^2, should be utilized for the estimates of the capability ratios. Furthermore, the estimate of the mean, μ, should be the sample average, commonly known as $\overline{X}$. In Chou et al. (1990), as long as the calculated or estimated value of the specific capability ratio is greater than the critical value, as defined in Table 1 for various sample sizes (n = 10, 20, 30, and 50), it is concluded that the "process is considered capable 95% of the time." Table 1 presents a small yet practical set of the values presented in Chou et al. (1990).

It is interesting to note that if it is desired that C_p and C_{pk} be both greater than 1, the calculated values of C_p and C_{pk} should be considerably larger than 1 for both cases for the practical sample sizes shown in Table 1. The Motorola 6σ performance criteria ($C_p = C_{pk} = 2.0$) are even greater. Example 10 presents applications of the inferential approach to PCA.

Example 10. Thirty observations have been collected from the in-control process of Examples 8 and 9.

1. If minimum values of one are required for C_p and C_{pk}, is the process capable.
2. Is the process of 6σ quality?

Data:

TABLE 1 Critical Values for Estimated C_p and C_{pk}

C_p:				
C_o	n = 10	20	30	50
1.0	1.65	1.37	1.28	1.20
1.2	1.97	1.64	1.54	1.44
1.4	2.30	1.92	1.79	1.68
1.6	2.63	2.19	2.05	1.92
1.8	2.96	2.47	2.30	2.16
2.0	3.29	2.74	2.56	2.40
C_{pk}:				
C_k	n = 10	20	30	50
1.0	1.80	1.46	1.35	1.25
1.2	2.12	1.73	1.61	1.49
1.4	2.45	2.01	1.86	1.73
1.6	2.78	2.28	2.12	1.97
1.8	3.11	2.55	2.37	2.21
2.0	3.44	2.83	2.63	2.45

i	Observation (X_i)
1	50.13
2	49.65
3	52.41
4	49.19
5	51.16
6	50.01
7	49.99
8	48.76
9	49.59
10	49.73
11	50.36
12	49.19
13	50.02
14	49.46
15	49.47
16	49.75
17	50.42
18	51.84
19	51.14
20	51.74
21	48.07
22	48.9
23	49.95
24	48.85
25	50.24
26	48.87
27	50.5
28	53.13
29	48.82
30	48.39

$S^2 = 1.401$
$\bar{X} = 49.99$

1. The resulting calculated values of C_p and C_{pk} are as follows (assuming the USL of 54 and LSL of 47 from Example 9):

$$C_p = \frac{54 - 47}{6\sqrt{1.401}} = 0.986$$

$$C_{pk} = \text{minimum}\left(\frac{54 - 49.99}{3\sqrt{1.401}}, \frac{49.99 - 47}{3\sqrt{1.401}}\right) = \text{minimum}(1{,}129, 0.842) = 0.842$$

   In both cases, the calculated values are less than the critical values. That is, the critical value for C_p when C_0 is 1 is 1.28, and the critical value for C_{pk} when C_k is 1 is 1.35. As a result, a strong, statistically based conclusion can be made that the process is not capable.

2. Obviously, if the process is not capable at values of 1 for the PCA ratios, it is not capable to meet the 6σ performance criteria. For this sample of 30, minimum critical values for C_p and C_{pk} would have been 2.56 and 2.63, respectively, if C_0 and C_k values of 2 were required (which is implied by the 6σ program).

5. CONTROL CHARTS FOR ATTRIBUTE DATA

5.1. Attribute Data

At this point, we have reviewed $\bar{X}$ and R control charts. Both of these charts are used for variable data. Many more control charts exist for varying conditions for variable data, as described above. In this section, two control charts are presented that are most useful in monitoring attribute data. In particular, a control chart known as the p chart is presented to provide a tool for monitoring the percent nonconforming from a process. Secondly, the *c* chart is used to monitor the number of defects being produced in a process.

5.2. Control Chart for Percent Nonconforming: *p* Chart

The percentage of nonconforming units being produced is of utmost concern to all organizations. A nonconforming unit represents the reality of a process's failure to produce product that meets customers' expectations or specifications. As such, the percent nonconforming control chart can also be thought of as an ongoing estimate of the process's ability to meet customer expectations.

For the *p* chart, the percent nonconforming is estimated with sampled data observations. Typically, 25–100 observations are collected for a given subgroup. The current percent nonconforming, *p*, is estimated as follows:

$$\hat{p} = \frac{D_i}{n}$$

where D_i = the number of nonconforming observations in the current subgroup i
n = the number of observations or samples in each subgroup

It is assumed that the number of observations that are defective in a subgroup of n samples follows the binomial distribution with parameters p (the true population percent nonconforming) and n. For the formulation presented here, n is assumed to be fixed. This assumption can be relaxed, but a modified approach for the varying sample size must be employed as demonstrated in most standard quality control textbooks (i.e., Montgomery 1996). The p chart is essentially the plotting of the observed values of $\hat{p}$. These values are individually compared to the control limits (shown below) to determine the current status of the process (i.e., state of statistical control or out of control). Furthermore, the AT&T runs rules should be deployed when possible.

Obviously, p, the true population percent nonconforming, is typically estimated by averaging the observed values of p. As a rule of thumb, it is suggested that 25–30 subgroups be collected before the estimate of p is made. This parameter, known as $\overline{p}$, is estimated as follows:

$$\overline{p} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m}$$

where m is the number of subgroups collected (recommended: 25–30).

Both $\hat{p}$ and $\overline{p}$ are known to be unbiased estimators of p. Once $\overline{p}$ is determined, the control limits can be determined as follows:

$$\begin{aligned}
\text{UCL}_p &= \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \\
\text{CL}_p &= \overline{p} \\
\text{LCL}_p &= \text{maximum}\left(0, \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right)
\end{aligned}$$

The LCL is truncated to 0 when the calculated value falls below 0. The AT&T runs rules can be employed in situations where the control limits are reasonably symmetric (AT&T 1985).

The establishment of the trial control limits for the p chart follows the same practice as described for the $\overline{X}$ and R charts. That is, if a given point plots beyond the initial control limits, it is recommended that the cause of the extraordinary point be resolved and then removed from the calculation. That is, the control limits should be recalculated with the out-of-control point(s) removed. In some extreme cases, this process may have to be repeated. Once the observed data are contained within the trial limits, continued production should be used to collect more data in an effort to validate the limits. Once process stability is obtained, that is, control is maintained for an extended period, the new control limits should become the operational standard. Furthermore, once the trial limits are established, the AT&T runs rules should be deployed in totality to ensure protection against process disturbances.

Example 11 demonstrates the use of the p chart.

Example 11. Thirty subgroups of 100 observations each have been collected from your process of producing the toy presented earlier in this chapter. The data collected and calculated $\hat{p}_i s$ are as follows:

i	D_i	p_i
1	1	0.01
2	0	0
3	2	0.02
4	1	0.01
5	2	0.02
6	1	0.01
7	0	0
8	0	0
9	1	0.01
10	1	0.01
11	1	0.01
12	1	0.01
13	2	0.02
14	2	0.02
15	1	0.01
16	1	0.01
17	1	0.01
18	1	0.01
19	0	0
20	4	0.04
21	0	0
22	0	0
23	1	0.01
24	3	0.03
25	2	0.02
26	0	0
27	0	0
28	1	0.01
29	1	0.01
30	3	0.03
	$\overline{p}$	0.011333

Computing the trial limits:

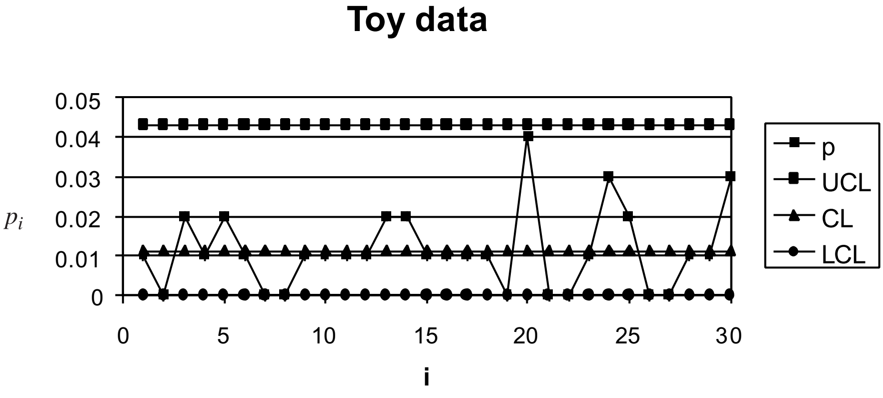


Figure 12 *p* Chart.

$$\text{UCL}_p = 0.011333 + 3\sqrt{\frac{0.011333(1 - 0.011333)}{100}} = 0.0431$$

$$\text{CL}_p = \bar{p} = 0.011333$$

$$\text{LCL}_p = \text{maximum}\left(0,\ 0.011333 + 3\sqrt{\frac{0.011333(1 - 0.11333)}{100}}\right) = 0$$

As demonstrated, all points plot within the control limits; therefore, these trial limits should be employed for production. Since the control limits are not symmetric, only rule number one of the AT&T runs rules should be used. AT&T (1985) describes an approach that may be used for customizing runs rules to improve the detection strength of a p chart.

5.3. Control Chart for Number of Defective Units: *c* Chart

An equally important concern in production is the number of defects that are produced. Examples of defects include the type of data included on the check sheet described in Example 2 (i.e., paint drips, scratches). This type of data collected and monitored on a consistent basis in an organization gives rise to the organizational understanding of the defect types and clearly document sustained performance. For this type of data, the sample should be a fixed amount of product, that is, linear feet, number of units, pounds of product. Once the sample is collected, the number of defects, c_i, is determined through the careful analysis of the sampled product. Typically, although not necessarily, all defects are grouped together to construct one control chart.

It is assumed that the number of defects in a sample follows the Poisson distribution with parameter λ (the average defect rate for this application). The c chart is essentially the plotting of the observed values of c_i. These values are individually compared to the control limits (shown below) to determine the current status of the process (i.e., state of statistical control or out of control). Furthermore, the AT&T runs rules should be deployed when the resulting control limits are symmetric (AT&T 1985).

Obviously, c (or λ for the classic Poisson distribution), the true population average defect rate, is never known with certainty. As a result, it is estimated by averaging m c_i's. As a rule of thumb, it is suggested that 25–30 subgroups be collected before the estimate of c, called $\bar{c}$, is made. This parameter is estimated as follows:

$$\bar{c} = \frac{\sum_{i=1}^{m} c_i}{m}$$

where m is the number of samples collected

With an accurate estimate of the average defect rate in hand, the control limits for the c chart are computed as follows:

$$\text{UCL}_c = \bar{c} + 3\sqrt{\bar{c}}$$

$$\text{CL}_c = \bar{c}$$

$$\text{LCL}_c = \text{maximum}(0, \bar{c} - 3\sqrt{\bar{c}})$$

The LCL is truncated to 0 when the calculated value falls below 0. The AT&T runs rules should be employed when the control limits are reasonably symmetric (AT&T 1985). Rule number one should always be used. AT&T (1985) describes an approach that may be used for customizing runs rules to improve the detection strength of a p chart.

The establishment of the trial control limits for the c chart follows the same practice as described for the $\bar{\bar{X}}$ and R and the p charts. Once the observed data are contained within the trial limits, continued production should be used to collect more data in an effort to validate the limits. Once process stability is obtained, that is, control is maintained for an extended period, the new control limits should become the operational standard. Furthermore, once the trial limits are established, the AT&T runs rules should be deployed in totality to ensure protection against process disturbances when appropriate.

The following example demonstrates the use of the c chart.

Example 12. c Chart: 30 samples of five toys have been collected from your process of producing the toy presented earlier in this chapter. The data collected and calculated $c_i s$ are as follows:

i	c_i
1	2
2	4
3	9
4	2
5	1
6	4
7	5
8	1
9	3
10	6
11	1
12	8
13	7
14	2
15	4
16	3
17	5
18	4
19	3
20	5
21	7
22	6
23	2
24	2
25	6
26	7
27	3
28	4
29	5
30	7
$\bar{c}$	4.266667

The resulting trial control limits are as follows:

$$\text{UCL}_c = 4.266667 + 3\sqrt{4.266667} = 10.463$$

$$\text{CL}_c = 4.266667$$

$$\text{LCL}_c = \text{maximum}(0,\ 4.266667 - 3\sqrt{4.266667}) = 0$$

The resulting control chart is as follows:
The process clearly exhibits stability.

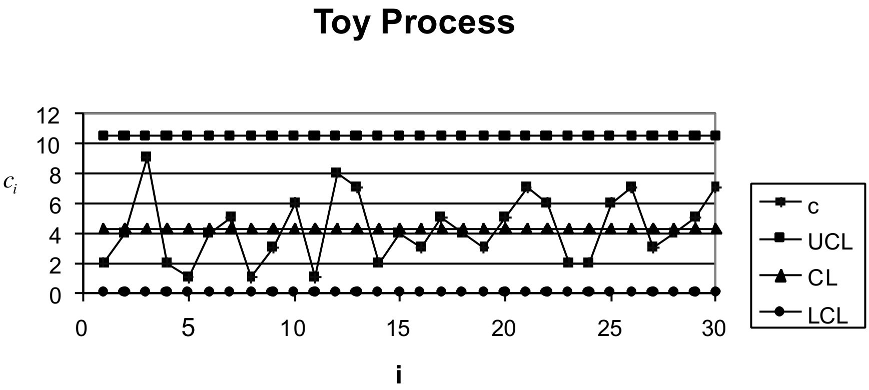


Figure 13 *c* Chart.

REFERENCES

AT&T (1985), *Statistical Quality Control Handbook.*

Chou, Y. M., Owen, D. B., and Borrego, S. A. (1990), ‘‘Lower Confidence Limits on Process Capability Indices,’’ *Journal of Quality Technology*, Vol. 22, No. 3, pp. 223–229.

Grant, E. L., and Leavenworth, R. S. (1980), *Statistical Quality Control*, 5th Ed., McGraw-Hill, New York.

Ishikawa, K. (1985), *What Is Total Quality Control: The Japanese Way*, D. J. Lu, Trans., American Society for Quality Control Press, Milwaukee.

Montgomery, D. C. (1996). *Introduction to Statistical Quality Control*, 3rd Ed., John Wiley & Sons, New York.

Montgomery, D. C., and Mastrangelo, C. M. (1991), ‘‘Some Statistical Process Control Methods for Autocorrelated Data,’’ *Journal of Quality Technology*, Vol. 23, No. 3, pp. 179–193.

Quesenberry, C. P. (1988), ‘‘An SPC Approach to Compensating a Tool-Wear Process,’’ Journal of Quality Technology, Vol. 3, No. 4, pp. 220–229.

Shewhart, W. A. (1927). ‘‘Quality Control Charts,’’ *Bell System Technical Journal*, Vol. 6, pp. 722–735.

APPENDIX
Table Values for the $\overline{X}$ and R Charts

The table values of statistical constants in this appendix are a portion of the values that can be found in any basic quality control text (i.e., Grant and Leavenworth 1980; Montgomery 1996).

Subgroup size, n	A_2	d_2	D_1	D_2	D_3	D_4
2	1.880	1.128	0	3.686	0	3.267
3	1.023	1.693	0	4.358	0	2.575
4	0.729	2.059	0	4.698	0	2.282
5	0.577	2.326	0	4.918	0	2.115
6	0.483	2.534	0	5.078	0	2.004
7	0.419	2.704	0.204	5.204	0.076	1.924
8	0.373	2.847	0.388	5.306	0.136	1.864
9	0.337	2.970	0.547	5.393	0.184	1.816
10	0.308	3.078	0.687	5.469	0.223	1.777

CHAPTER 70

Measurement Assurance

S. CHANDRASEKAR
Purdue University

1. ELEMENTS OF A MEASUREMENT SYSTEM 1877
2. CHARACTERIZATION OF MEASUREMENT SYSTEM ELEMENTS 1878
  2.1. Characterization of Measurement System Elements 1879
    2.1.1. Range 1879
    2.1.2. Linearity 1879
    2.1.3. Sensitivity(s) 1879
    2.1.4. Environmental Effects, Wear and Hysteresis 1879
  2.2. Statistical Characteristics 1880
    2.2.1. Repeatability 1880
3. IDENTIFICATION OF STATIC CHARACTERISTICS—CALIBRATION 1881
4. ACCURACY OF MEASUREMENT SYSTEMS IN STEADY STATE 1882
  4.1. Error-Reduction Methods 1883
5. DYNAMIC CHARACTERISTICS 1884
  5.1. Transfer Function of a Measurement System 1884
6. NOISE IN MEASUREMENT SYSTEM 1885
7. SUMMARY 1885
ADDITIONAL READING 1885
APPENDIX: MEASUREMENT-RELATED TECHNIQUES 1886

Measurements play a major role in modern technology. Their purpose is typically threefold: to validate or invalidate a theory or model by comparing the measured value of a variable with a prediction; to determine whether a process or product meets specifications or quality requirements; and to enable closed-loop control of systems and processes. More generally, the purpose of measurement is to present an observer with a numerical value for the variable being measured. The input to a measurement system is the true value of the variable being measured, while the output is the measured value of this variable. In general, this measured value does not equal the true value of the variable. For example, a force sensor may read a value of 10.2 N when the true value of the force is 10.25 N; the speed of an engine as measured by a tachometer may be 3000 rpm when the true value is 3010 rpm; and so on. The problems involved in establishing the true value of the variable, an assessment of measurement error, a determination of the static and dynamic characteristics of measurement systems, and some remarks about noise in measurement systems constitute this chapter.

1. ELEMENTS OF A MEASUREMENT SYSTEM

A modern measurement system typically consists of the four distinct elements, as shown in Figure 1.

1. The *sensing element* probes the process, either in a contact or noncontact mode, and gives an output that depends on the variable being measured. Examples are:

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

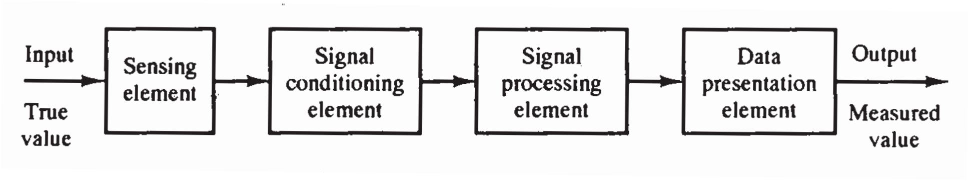

Input
True value
Sensing element
Signal conditioning element
Signal processing element
Data presentation element
Output
Measured value

Figure 1 General Structure of Measurement System.

- Thermocouple, where an output voltage depends on temperature
- Piezoelectric force sensor, where an output charge depends on force
- Capacitive displacement transducer, where the capacitance output is a function of position or displacement

2. The *signal conditioning* element converts the output of the sensing element into a form more suitable for processing, usually a current, voltage, or frequency signal. Examples are:
   - Charge amplifier for a piezo force sensor, which converts a charge to a voltage.
   - Amplifier, which magnifies, say, a millivolt signal to a signal of several volts.
3. The *signal processing* element, which converts the output of the conditioning element into a form suitable for presentation. Typically, this element is an analog-to-digital (A/D) converter, which digitizes the analog signal output of the conditioning element.
4. The *data-presentation* element, which presents the measured value in an easily recognizable form. Examples are a chart recorder, graphical or numerical output on a computer terminal, and an oscilloscope.

Figure 2 shows a force-measurement system that incorporates each of these elements. Some measurement systems may have several of these elements.

2. CHARACTERIZATION OF MEASUREMENT SYSTEM ELEMENTS

In the previous section, we described the various elements that constitute a typical measurement system. The characteristics of the elements, as far as their input–output relationship impact the overall performance of the system, and as such it is important to describe these characteristics. In this section, we describe the static and dynamic characteristics of the system or system elements. We begin with a discussion of static or steady-state characteristics; these are the relationships that exist between the output (O) and input (I) of an element when I is either constant or changing slowly.

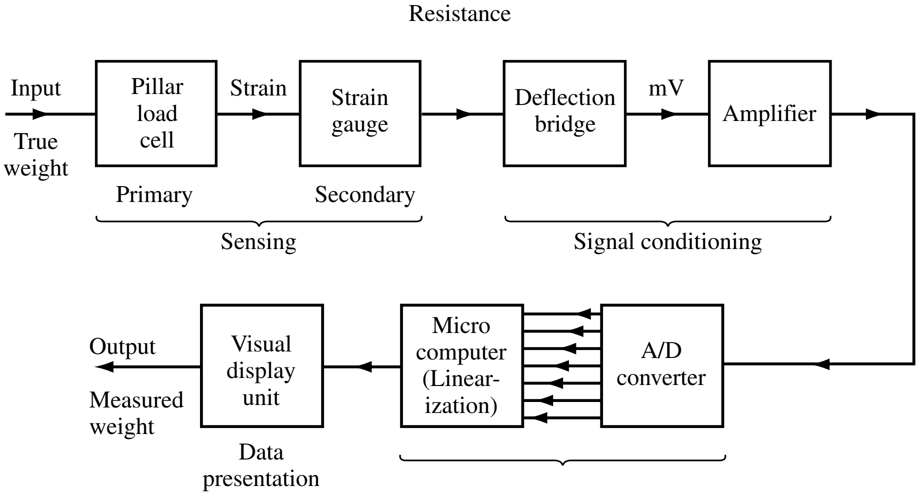

Resistance
Input
True weight
Pillar load cell
Strain
Strain gauge
Deflection bridge
mV
Amplifier
Primary
Secondary
Sensing
Signal conditioning
Output
Measured weight
Visual display unit
Micro computer (Linear-ization)
A/D converter
Data presentation

Figure 2 Weight-Measurement System.

2.1 Systematic Characteristics

Systematic characteristics are those that can be accurately quantified by analytical or empirical methods. These are to be distinguished from statistical characteristics, which will be considered later.

2.1.1. Range

This gives the range of values allowable for the input and output. For example, a piezo force sensor may have an input range of 0–10,000 N and a corresponding output range of 0–10 V.

2.1.2. Linearity

The operating ranges of most measurement systems are characterized by a linear relation between input (I) and output (O). In such instances:

$$O(I) = KI + a \tag{1}$$

where

$$K = \text{slope} = \frac{O_{\text{max}} - O_{\text{min}}}{I_{\text{max}} - I_{\text{min}}}$$

In practice, measurement systems show small deviations from linearity even within their operating range. These nonlinearities, which can be obtained from calibration experiments, may be ignored except in specific instances of measurement. In the presence of nonlinearity, equation (1) becomes:

$$O(I) = KI + a + N(I) \tag{2}$$

where $N(I)$ is the nonlinear term.

This nonlinearity is typically expressed as a percentage of the full-scale output range; for most measurement systems this value is less than 1% of the full-scale range.

If we look at the voltage output (E) of a chromel-alumel thermocouple in temperature (T) range of 0–200°C, where $E\ (T)$ is given by $E\ (T) = 40.1\ T + 2.09 \times 10^{-2}\ T^2 + 4.31 \times 10^{-4}\ T^4 +$ higher-order terms

Then, with reference to Eq. (2):

$$K = 40.1,\ a = 0, \text{ and}$$
$$N(I) = 2.09 \times 10^{-2}\ T^2 + \ldots$$

2.1.3. Sensitivity(s)

Sensitivity(s) is the rate of change of output with input, that is:

$$S = \frac{dO}{dI}$$

A high value of S is often a common requirement for measurement systems in their operating range.

2.1.4. Environmental Effects, Wear, and Hysteresis

The output O is often influenced by fluctuations in the environment, such as changes in humidity, ambient temperature, and supply voltage. These fluctuations add additional terms to Eq. (2) or cause changes in the sensitivity-related parameter K. Aging of system elements is also a common source of changes in system sensitivity.

Hysteresis refers to the differing dependence of O on I, depending on whether the measurement is being made while I is increasing or decreasing. An example of hysteresis is backlash in gears and rack–pinion arrangements. The input–output relationship for a measurement system with hysteresis and non-linearities is shown in Figure 3.

Interfering and modifying inputs are environmental inputs that cause the linear sensitivity and intercept of the system, respectively, to change. An example of an interfering input is fluctuations in the reference junction temperature of a thermocouple.

We can now collect together the various characteristics of a measurement system, except for hysteresis, and express the general model for the system in the form of a block diagram. Such a block diagram representation is shown in Figure 4. The dynamic and static characteristics are shown decoupled in the diagram. The dynamic characteristics will be discussed shortly.

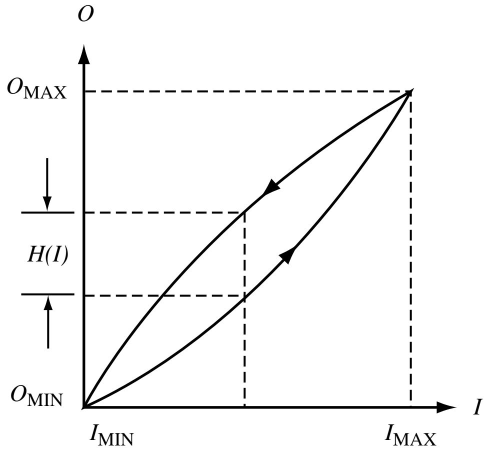


Figure 3 Hysteresis and Nonlinearity.

2.2. Statistical Characteristics

2.2.1. Repeatability

Suppose we apply a constant dead-weight load to a force-measurement system and keep the load applied for several hours. If the output O is monitored over this period of time, it is often likely that the output value will fluctuate about an expected value for the output. For example, if the load is 100 N, for which the expected output value of the sensing element should be 1 V, it is likely that over time the output will assume values such as 1.01, 0.98, 0.99, 0.98, 1.02, etc. This effect is termed as a lack of repeatability. Repeatability is the ability of a system to give the same output for the same input, when this input is repeatedly applied to it. The most common causes for lack of repeatability are random variations in the measurement system elements and their environment. By making reasonable assumptions about the fluctuations of the various inputs, including environment-related ones, it is possible to analytically characterize the fluctuations expected in the output.

Consider an output O that is a function of two types of inputs I and I_E. Further, let us assume, as is typically done, that the fluctuations in I and I_E can be described by normal distributions with standard deviations σ_1 and σ_2, respectively.

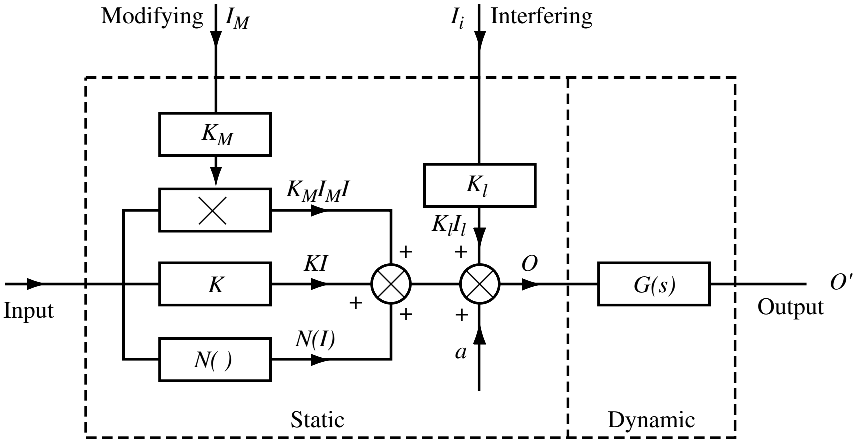


Figure 4 General Model of Measurement Element.

Then for

$$O = f(I, I_E)$$

$$\Delta O = \frac{\partial f}{\partial I}\Delta I + \frac{\partial f}{\partial I_E}\Delta I_E \tag{3}$$

and the standard deviation of O about its mean is given by

$$\sigma_o = \left(\frac{\partial f}{\partial I}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial I_E}\right)^2 \sigma_2^2 \tag{4}$$

Thus, σ_o can be calculated from a knowledge of σ_1 and σ_2. Alternatively, a calibration experiment can be done on the system, which allows for σ_o to be estimated directly from the experimental results. The mean value $\overline{O}$ for the output can be estimated from the experimental data. The corresponding probability density function for O can then be estimated as

$$P(O) = \frac{1}{\sigma_o\sqrt{2\pi}} \exp\left(\frac{-(O - \overline{O})^2}{2\sigma_o^2}\right) \tag{5}$$

3. IDENTIFICATION OF STATIC CHARACTERISTICS—CALIBRATION

The static characteristics of an element can be found experimentally by measuring corresponding values of the input I, the output O and the environmental inputs I_M, I_I, when I is either at a constant value or changing slowly. This type of experiment is referred to as calibration. The measurement of the variables I, O, I_M, I_I must be accurate if meaningful results are to be obtained. The instruments and techniques used to quantify these variables are referred to as standards.

The accuracy of measurement of a variable is the closeness of the measurement to the true value of the variable. It is quantified in terms of measurement error, that is, the difference between the measured value and the true value. Thus, the accuracy of a laboratory standard pressure gauge is the closeness of the reading to the true value of pressure. This brings us to the problem of how to establish the true value of a variable. We define the true value of a variable as the measured value obtained with a standard of ultimate accuracy. Thus, the accuracy of the above pressure gauge is quantified by the difference between the gauge reading for a given pressure and the reading given by the ultimate pressure standard. However, the manufacturer of the pressure gauge may not have access to the ultimate standard to measure the accuracy of his products. They can, however, measure the accuracy of the gauges relative to a portable intermediate or transfer standard, such as a dead-weight pressure tester. The accuracy of the transfer standard must be established by calibration against the ultimate pressure standard. This introduces the concept of a traceability ladder, which is shown in simplified form in Figure 5.

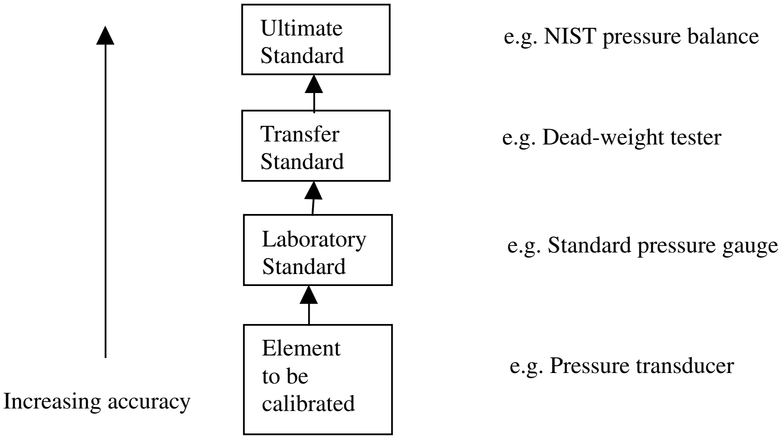


Figure 5 Schematic of Traceability Ladder.

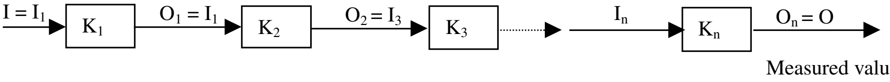


Figure 6 Measurement System Consisting of n Measuring Elements in Series. Input to the System is the True Value of the Variable to be Measured, and Output is the Measured Value of this Variable. The K_i's are the Gains of Each Measuring Element.

The measuring element is calibrated using the laboratory standard, which should itself be calibrated using the transfer standard, and this in turn should be calibrated using the ultimate standard. Each element in a traceability ladder should be significantly more accurate than the one below it.

4. ACCURACY OF MEASUREMENT SYSTEMS IN STEADY STATE

The input to a measurement system is the true value of the variable being measured, while the output corresponds to the measured value of the variable. Also, assuming the measurement system is complete, the system output is the measured value of the variable. Accuracy is defined in terms of measurement error, that is, the difference between the measured and true values of the variable. It follows, therefore, that accuracy is a property of a complete measurement system rather than a single element. Accuracy is quantified using measurement error E where:

$$\begin{aligned} E &= \text{measured value} - \text{true value} \\ &= \text{system output} - \text{system input} \end{aligned}$$

We can use the static model of a single element, developed previously, to calculate the output and thus the measurement error for a complete measurement system that is made up of several measuring elements. Consider the system shown in Figure 6 consisting of n elements in series. Suppose each element is ideal, that is, perfectly linear and not subject to environmental inputs. If it is also assumed that the intercept or bias is zero, that is, $a = 0$, then:

$$O_I = K_i I_I$$

for $i = 1, \ldots, n$, where K_i is the linear sensitivity or slope. It follows that $O_2 = K_2 I_2 = K_2 K_1 I$, $O_3 = K_3 I_3 = K_3 K_2 K_1 I$, and for the whole system:

$$O = O_n = K_1 K_2 K_3 \ldots K_i \ldots K_n I$$

If the measurement system is complete, then $E = O - I$, giving:

$$E = (K_1 K_2 K_3 \ldots K_n - 1) I$$

Thus, if

$$K_1 K_2 K_3 \ldots K_n = 1$$

we have $E = O$ and the system is perfectly accurate. The temperature measurement system shown in Figure 7 appears to satisfy the above condition. The indicator is simply a moving coil voltmeter with a scale marked in degrees Celsius so that an input change of 1 V to the indicator causes a change in deflection of 25°C. This system has $K_1 K_2 K_3 = 40 \times 10^{-6} \times 10^3 \times 25 = 1$ and thus appears to be perfectly accurate. The system is not accurate, however, because none of the three elements present is ideal. For instance, the thermocouple is nonlinear, so that as the input temperature

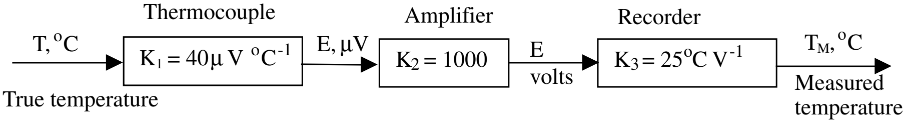


Figure 7 Schematic of a Temperature-Measurement System.

changes the sensitivity is no longer 40 $\mu V°C^{-1}$. Also, changes in reference junction temperature cause the thermocouple emf. to change. The output voltage of the amplifier may also be affected by changes in ambient temperature. The sensitivity K_3 of the indicator usually depends on the stiffness of the restoring spring in the moving coil assembly. This is affected by changes in environmental temperature and wear, causing K_3 to deviate from 25°C V^{-1}. Thus, the condition $K_1K_2K_3 = 1$ cannot be always satisfied and the system is in error.

In general, the error of any measurement system depends on the nonideal characteristics—for example. nonlinearity, environmental, and statistical effects—of every element in the system. Thus, in order to quantify this error as precisely as possible, it is necessary to use the general model for a single element.

The error probability density function for a measurement system comprising of n elements in series as in Figure 6 can be estimated from a knowledge of the corresponding density functions for each of the elements. This can be done analytically for a variety of density functions (see, for e.g., Bentley 1995). Alternatively, the density function for the entire system can be estimated from a calibration experiment.

4.1. Error-Reduction Methods

The error of a measurement system depends on the nonideal characteristics of each of the elements in the system. It is possible to identify which of the elements show significant nonideal behavior using calibration experiments. Compensation strategies may then be devised for each of these elements so that significant reductions can be accomplished for the overall system error. Some of these compensation methods are briefly highlighted, especially for nonlinear and environmental effects.

A common method for correcting a nonlinear element is to introduce a *compensating non-linear element* into the system such that the overall characteristic of the system is as linear as possible within the operating range of the system. This is illustrated in Figure 8 for a resistance thermometer. Depending on the degree of linearity desired, the bridge in Figure 8 can be suitably designed.

The most common method for reducing the effects of environmental inputs is *isolation*. Here, each of the measuring elements is effectively isolated from environmental changes. Examples are the placement of reference junction of a thermocouple in a temperature-controlled enclosure and the use of active vibration-isolation tables to isolate a measuring system (e.g., atomic force microscope) from external mechanical vibrations. Of course, it is possible to reduce environmental influences by selecting a transducer material that is completely insensitive to a specific environmental parameter. An example is the use of a metal alloy in strain gauges that has a zero coefficient of thermal expansion. But such an ideal material is often difficult to find and quite expensive.

A successful method for neutralizing environmental inputs is by providing an *opposing environmental input*. This is done by incorporating a second element(s) into the measuring system that is exposed to the same environmental input as the element of interest, such that the two effects tend to cancel out. An example is the use of two identically matched strain gauges in adjacent arms of a Wheatstone bridge to compensate for changes in the ambient temperature. In such an arrangement,

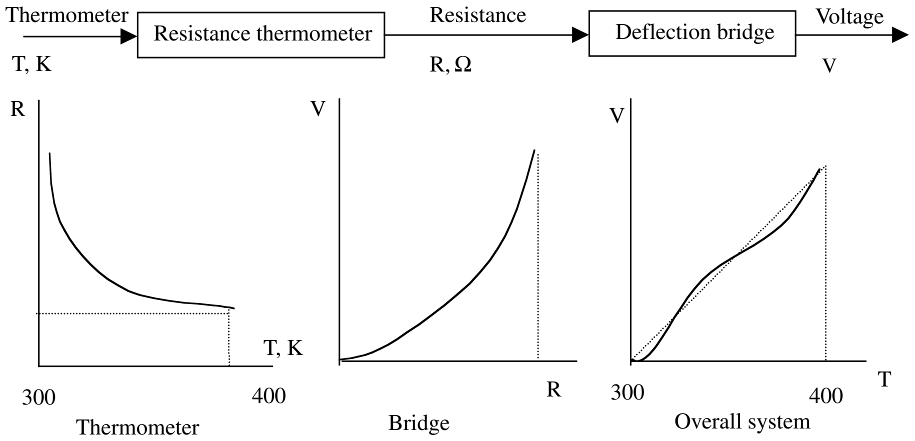


Figure 8 Nonlinearity Compensation in a Temperature-Measuring Element Using a Nonlinear Compensation Element.

one gauge is measuring a mechanical tensile strain and the other an equal compressive strain. However, because temperature-induced strains will have the same sign and magnitude in both the gauges, the bridge design effectively subtracts the resistances corresponding to the strains in the two gauges, thereby canceling out the thermal strain while doubling the mechanical strain being measured.

Other methods of compensation include the use of high-gain negative feedback and computer estimation of measured values. A discussion of these methods can be found in Doebbelin (1990).

5. DYNAMIC CHARACTERISTICS

The discussion thus far has centered on measurement system characteristics that are relevant to measurements of parameters that are either constant or changing very slowly with time. That is, the input (I) to the measurement system or a measuring element is varying rather slowly. By slow, we mean here that the variation in I with time is small over a duration equal to the response time of the measurement system or measuring element. However, in practice there are many instances where the input to a measurement system may change quite rapidly or even, in some cases, almost instantaneously. An example in this regard is the sudden change in the temperature input to a thermocouple from an ambient temperature of 30°C to 90°C when it is brought into contact with, say, an object at 90°C. In this case, the output of the thermocouple will change with time and reach a steady state output value that is equal or close to 90°C only after a finite amount of time. This finite amount may be a few microseconds or a few seconds depending on the type of thermocouple. The characteristics of the measurement system that describe its response to rapid changes in input are known as *dynamic characteristics*, and these are most conveniently described using the concept of a transfer function from linear systems theory.

5.1. Transfer Function of a Measurement System

Consider the measurement system comprising of n measuring elements in series (Figure 6) The input–output relationship for each element in this figure has been specified by an amplifier gain K_I, which means that the output can change instantaneously with time in the same manner as the input. In practice, this is not the case. The variation of $O(t)$ with $I(t)$ for a measuring element is often best described for most elements by a first-order or second-order linear differential equation. This description can be translated into an equivalent description in the form of a block diagram using the concept of a transfer function.

Consider a thermocouple that is brought into contact with a temperature source T. Then the output of the thermocouple, T_o, translated into a temperature reading, can often be described by the first-order differential equation

$$\tau \frac{dT_o(t)}{dt} + T_o(t) = T_i(t)$$

assuming that $T_o(0) = 0$ and $T_I(0) = 0$. The measured value of the variable T_i is T_o. The transfer function of this thermocouple element is given by

$$G(s) = \frac{T_o(s)}{T_i(s)} = \frac{1}{\tau s + 1}$$

The parameter τ is called the time constant of the element, and its value completely describes the response of the element. The smaller the value of τ, the faster the response of the element, which is an extremely desirable characteristic in measurement elements and systems.

Another common type of measurement system is a second-order, linear system described by a linear differential equation of the form

$$\frac{1}{\omega_n^2} \frac{d^2O}{dt^2} + \frac{2\xi}{\omega_n} \frac{dO}{dt} + O(t) = I(t)$$

where $O(t)$ and $I(t)$ are the output and input, respectively. The transfer function of this system is

$$G(s) = \frac{O(s)}{I(s)} = \frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2}$$

and the system behavior is completely determined by the two system parameters, ξ and ω_n, which are the damping ratio and natural frequency, respectively. Together, the two parameters determine the system response. A small value of ξ results in a fast response, but the overshoot in the response is also high. A high value of ω_n also results in a fast response. Furthermore, a high value for ω_n in a measurement element increases the range of signal frequencies over which measurements of the signal can be made accurately. Roughly speaking, signal frequencies up to a third of ω_n can be measured accurately without much attenuation. For further discussion of first- and second-order systems and their responses, see Franklin and Powell (1994).

We shall illustrate the computation of output and measurement error in a piezoelectric force sensor that is subjected to a time varying force. Figure 9 shows the block diagram of a piezoelectric force sensor or measurement system consisting of three elements. The charge amplifier is a first-order system with time constant of 0.1 sec, while the recording element is a second-order system with ξ and ω_n given in the figure. The measurement system is complete in the sense that the product of the elemental gains, $20 \times 10^{-3} \times 50$, is 1.

Let the system be subjected to a time varying input force,

$$F_i(t) = 100 \left[\sin 10t + \frac{1}{3} \sin 30t + \frac{1}{5} \sin 50t \right]$$

This is the true value of the force to be measured. The output ($F_o(t)$), which is the measured force, can be estimated in the usual way either analytically, or more easily using MATLAB. The error $E(t)$ is then obtained as $F_i(t) - F_o(t)$. An increase in the value of τ or ω_n can be used to reduce the magnitude of the error, as is easily verified by a calculation.

We will not discuss the calibration of dynamic characteristics of measurement systems here. Any one of the books under Additional Reading dealing with measurement systems can be referred to for this purpose.

6. NOISE IN MEASUREMENT SYSTEMS

Unwanted signals that interfere with the measurement process are often present in measurement systems. These signals can be deterministic or random; collectively, they are referred to as noise.

Random signals can be quantified using the following statistical functions: mean, standard deviation, probability density, power spectral density, and autocorrelation. Deterministic noise, such as in rotating machinery, can be quantified by their power spectrum. A variety of methods are available to reduce the effects of different types of noise signals. These include electromagnetic shielding, electrostatic screening, filtering, modulation, averaging, and modulation. See Additional Reading for detailed discussions of these issues.

7. SUMMARY

A brief tutorial on measurement systems has been presented. Description of measurement system elements and their characteristics, assessment of measurement error, and a short discussion of noise in measurement systems have constituted this tutorial. Additional Reading lists books related to these aspects where more detailed discussions of the topics reside. Included are references in which elements of measurement systems for sensing of specific parameters, such as force, temperature, and displacement, are discussed at length.

Acknowledgment

The author would like to thank Linda Weybright and Sridhar Kompella for their help in the preparation of the manuscript.

ADDITIONAL READING

Doebelin, E. O., *Measurement Systems,* McGraw-Hill, New York, 1990.

Bentley, J. P., *Principles of Measurement Systems,* Longman Scientific & Technical, Harlow, UK, and John Wiley & Sons, New York, 1995.

These two books provide a good overview and considerable detail about all aspects of measurement systems, including specific systems for force, temperature, displacement, pressure measurement, etc.

Lyons, L., *A Practical Guide to Data Analysis for Science Students,* Cambridge University Press, Cambridge, 1991.

Taylor, J. R., *An Introduction to Error Analysis,* University Science Books, Sausalito, CA, 1997.

These two books are useful guides for analysis of experimental data.

Franklin, G. F., Powell, J. D., and Emami-Naei, A., *Feedback Control of Dynamic Systems,* 3rd Ed., Addison-Wesley, Reading, MA, 1994.

Discusses linear systems and transfer functions.

Neubert, K. H. P., *Instrument Transducers: An Introduction to their Performance and Design,* 3rd Ed., Clarendon Press, Oxford, 1983.

Instrument Society of America (ISA), *ISA Transducer Compendium,* ISA, Pittsburgh.

Sydenham, P. H., *Mechanical Design of Instruments,* Instrument Society of America, Research Triangle Park, NC, 1986.

Compendium of transducers for various applications.

Whitehouse, D. J., *Handbook of Surface Metrology,* Institute of Physics Publishing, Bristol, UK, 1994.

A one-step solution for manufacturing metrology. Discusses surface and dimensional metrology.

Jones, R. V., *Instruments and Experiences,* John Wiley & Sons, New York, 1988.

Some perspective on the evolution of measuring elements written by a person who headed radar development in the United Kingdom during World War II.

APPENDIX
Measurement-Related Journals

1. *The Review of Scientific Instruments*
2. *Journal of Physics E: Scientific Instruments* (UK)
3. *Transactions of Instrument Society of America*
4. *Experimental Mechanics*
5. *Industrial Laboratory* (Russian; English translation)
6. *Instruments and Control Systems*
7. *Control Engineering*
8. *Journal of Instrument Society of America*
9. *Archiv für Technisches Messen* (Germany)
10. *Journal of Research of the National Bureau of Standards* (National Institute of Standards and Technology after 1988)
11. *Transactions of the Society of Instrument Technology* (Great Britain)
12. *ASME Journal of Dynamic Systems, Measurement and Control*
13. *Biomedical Engineering* (London)
14. *Medical Electronics and Data*
15. *Experimental Techniques*
16. *Optical Engineering*
17. *Sensors and Actuators*
18. *Transactions of Institute of Measurement and Control* (London)
19. *IEEE Transactions in Instruments and Measurements*
20. *Precision Engineering*
21. *Journal of Optical Sensors*
22. *Applied Optics*
23. *ASME Journal of Manufacturing Science and Engineering*
24. *Precision Engineering*
25. *Journal of Japan Society of Precision Engineering*

CHAPTER 71

Human Factors and Automation in Test and Inspection

COLIN G. DRURY
State University of New York at Buffalo

1. DECISION FUNCTIONS IN A GLOBAL BUSINESS ENVIRONMENT **1887**
  1.1. Increasing Decision Options through IT 1889
  1.2. Quality Control/Assurance/Management 1889
  1.3. Decisions on Quality Made at Source 1889
  1.4. Distribution of Test and Inspection Effort 1889
2. TEST AND INSPECTION REQUIREMENTS **1890**
  2.1. Inspection Measures and Scales 1890
3. LOGICAL STRUCTURE OF TEST AND INSPECTION **1892**
  3.1. Human and Automated Test and Inspection 1892
  3.2. Mission and Function in Test and Inspection 1892
4. THE HUMAN ROLE IN TEST AND INPECTION **1894**
  4.1. Human Inspection Function by Function 1894
    4.1.1. Setup 1894
    4.1.2. Present 1895
    4.1.3. Search 1895
    4.1.4. Decision 1896
    4.1.5. Respond 1899
  4.2. Overarching Considerations: Job Design 1899
5. AUTOMATION IN TEST AND INSPECTION **1900**
  5.1. Automated Inspection Function by Function 1901
    5.1.1. Setup 1901
    5.1.2. Materials Handling 1902
    5.1.3. Sensing 1902
    5.1.4. Signal Processing 1904
  5.2. Image Processing in Automated Visual Inspection Systems (AVIS) 1904
    5.2.1. Examples 1906
6. NONPRODUCTION TEST AND INSPECTION **1907**
  6.1. Maintenance Inspection 1908
7. LOGICAL FUNCTION ALLOCATION IN TEST AND INSPECTION **1912**
  7.1. A Methodology for Test and Inspection Systems Design 1914
8. CONCLUSIONS ON TEST AND INSPECTION **1916**
REFERENCES **1916**

1. DECISION FUNCTIONS IN A GLOBAL BUSINESS ENVIRONMENT

Throughout the world, the business environment is changing, primarily because of globalization. This chapter examines the human factors aspects of one function in industry, test, and inspection to summarize the state of knowledge and practice. Test and inspection are the sense organs of industry; many decisions are based upon their data. They may even be the decision functions themselves when inspection is closely coupled to the control of a business process. Changes in the business environment

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

directly impact the functioning of industry, including in the areas of test and inspection, so these impacts must be considered throughout the chapter. While much of this chapter concentrates on test and inspection in a manufacturing context, there are many other applications, such as in medicine, maintenance, security, and design review, some of which will be considered in Section 7.

Globalization is the combination of market and political forces that has reshaped business and politics since the end of the Cold War in the late 1980s. Some of these forces, such as technological change, are obvious to those who study work, but other forces are changing the nature of jobs, perhaps even more profoundly. Globalization of customers, finance, and production of goods and services has been driven by forces of deregulation, inexpensive transportation, and rapid diffusion of distributed computing (Friedman 1999). Industry is becoming spread across more regions of the world and is shifting away from manufacturing and agriculture towards communications and service. Global capital markets force "creative destruction," the often brutal flow of capital away from enterprises with low shareholder value to enterprises where the capital will generate the greatest return. Investment moves rapidly, forcing industries to respond equally quickly to changing customer demands. We have moved from the managerial capitalism of the first part of this century to investor capitalism with more demanding shareholders (e.g., large pension funds) and more information available instantly (Whitman 1999).

While these changes may seem remote from the lives of human factors professionals, they in fact have very direct effects as they drive industry into new modes of operation beyond technological changes. Perhaps the most important change from the viewpoint of test and inspection is that most workers will be more closely involved with customers because consumers are increasingly demanding a combination of high quality, customization, and low price (Whitman 1999). Exposing the workforce to customers will change the skill requirements of jobs to include communications skills, as has always been the case in service industries such as travel and banking. It also means that the effective batch size is driven towards unity, the individual customer. Quality control in such environments poses a number of challenges because the concepts of continuous control or large-batch sampling no longer apply. Customer involvement can increase job satisfaction but can also increase performance pressure.

Secondly, many operations will be outsourced, leading to reduced levels of job security and more temporary jobs (Whitman 1999). In fact, layoffs in U.S. industry peaked in 1992–1995 at a time of maximum job growth. The fastest-growing category was in temporary jobs, which rose six-fold between 1972 and 1995 (Whitman 1999). The total number of temporary jobs is still small but is a major concern of workers (Kanter 1995, chap. 6). Global competition has forced many companies to downsize their workforces to remain competitive and increase shareholder value. Budros (1999) examines the reasons for downsizing as separate from reorganization. He sees the downsizing trend as being caused by technological innovation and by the existence of highly paid long-term employees, concluding that downsizing is not always effective. From a work perspective, downsizing can be expected to increase workloads for those remaining and to remove some of the company expertise. Work hours overall may be increasing for Americans, according to Schor (1991). She analyzes national data on long-term employment hours of work, vacation time, and work in the home, concluding that total hours of work have increased by about one month per year over the past 40 years or so. While her data and analyses have been questioned, we seem to find very few people who are not working harder than they used to.

Global changes are also driving job demands in ways beyond employment security. Increasingly, work at the world-class levels demanded by global competition generates greater worker skill requirements and a greater rate of worker knowledge obsolescence. Kanter (1999) shows that even in manufacturing, physical assets represented 63% of company capitalization in 1982 but only 38% in 1991. The remainder of the assets are largely composed of company knowledge and competence. Indeed, Siemieniuch and Sinclair (1999) show that even if useful industrial knowledge has a half-life as long as 10 years, only 6% of the knowledge at the start of a working life will be useful at the end 40 years later. Before we retire, we will be producing unknowable offerings (goods and services) with unborn people and uninvented techniques. In turn, this creates a demand for life-long training. Whitman (1999) notes that companies with a heavy emphasis on training show a 19% greater productivity gain over a three-year interval than other companies.

There are other industrial changes taking place at the same time that are not specifically part of the globalization of work. Information technology is becoming a part of ALL jobs as computing power becomes less expensive and more distributed. We not only use information technology to replace workers, but to change the nature of their jobs. While only a few years ago the National Research Council was investigating the gap between IT investment and productivity improvement (NRC 1997), it now appears that the gap has closed and that computing power is having a significant effect on both productivity growth and the nature of work.

As far as the sensing and decision functions of test and inspection are concerned, the factors considered above have large impacts. The pressures for competitiveness have increased the demands on the test and inspection systems for simultaneous improvement of both effectiveness (quality) and efficiency (low cost). In turn, some of these demands have been met by new forms of production. In the past decades, production changes have been driven by influential, although not always effective, movements in industry such as the quality movement and business process re-engineering. Some of the major impacts are described below.

1.1. Increasing Decision Options through IT

Information technology increases the options open to designers while creating a climate in which automation at any price is seen as a virtue per se. Major breakthroughs in vision systems, visual-scene analysis, and decision systems have allowed systems designers to consider replacing human labor in test and inspection with automation. In addition, the increased speeds of production now being experienced and the legal pressure for a return to 100% inspection have meant that in some industries automated systems are the only feasible answer. Throughout this chapter, we will consider the roles of humans and automation within test and inspection as parallel options for allocation of function so that the full impact of information technology can be brought to bear in the most effective manner rather than be a design goal in itself.

1.2. Quality Control/Assurance/Management

A major influence on industry since the 1980s has been the quality movement. The quality revolution has grown over two decades, beginning in quality technology and proceeding through the quality circles movement. It has stabilized over the past 10 years or so under the general titles of total quality control (Hancock et al. 1992) or total quality management (Evans and Lindsay 1993). Quality itself is defined as *meeting or exceeding customer expectations* (fitness for use) or *conformance to specifications* (manufacturing quality). TQM is seen as having both technical and managerial components, in that quality requires both technical knowledge and organizational knowledge.

The most basic quality requirement is freedom from design error, or ‘‘off-line quality,’’ as Taguchi and Wu (1979) characterize it. Agricultural products, manufactured goods, and delivered services must meet the needs of their customer over a wide range of customer environments. Products must be functional and must be free from user/product mismatches because any mismatch is by definition a design error. The implications are that designers know their customers’ needs and have the techniques to turn these needs into product design. For test and inspection the implications are largely in the design stage, where human factors test and evaluation is an important component of customer service (O’Brien and Charlton 1997).

Secondly, globalized production cannot work unless all elements of the company’s global operations fit together without error. Thus, new production systems based on pervasive computer-mediated design and manufacturing can assemble parts from up and down the supply chain without first-time errors of fit. Modern civil and military aircraft production give prime examples of this freedom from physical error. In a customer-oriented system, any over-cost or delayed delivery is an error with the same effects on company performance as a defective product. The implication for test and inspection is that we must be able to measure these customer needs and convert them into measurable precursors of error states.

1.3. Decisions on Quality Made at Source

From TQM and sociotechnical systems design (Taylor and Felten 1993) comes the concept of controlling key variances at their source. One major effect of the quality movement has been to renounce post-production inspection to a large extent and replace it by in-process inspection. With the introduction of more tightly coupled production systems, error cannot be tolerated without widespread consequences, so that control of quality at source is a key component of advanced manufacturing systems such as flexible manufacturing systems (FMS) cells. The corresponding push towards just-in-time (JIT) manufacturing has also made early control of errors a priority.

Control at source means not producing defective products in the first place rather than trying to sort defective products from good products later in the process. With the ultralow defect rates now being demanded and achieved, even highly effective sorting leaves the error rate too high. If we require defective rates in parts per million ($\sim 10^{-6}$) or at a six-sigma standard, then even a hit rate on inspection of 99% will not give the necessary result unless the production defective rate is $\sim 10^{-4}$. Even this level may be difficult with human inspectors because they tend to move their criterion for reporting a defect to more and more stringent values as the overall quality improves. Thus, we look instead for necessary precursors to error rather than error. If we know the mean of a process has shifted by a given amount, then we can deduce that defects will be more common*, even if we find no defects*. This is the principle of in-process quality control, often using control charts (e.g., Vardeman and Jobe 1998).

1.4. Distribution of Test and Inspection Effort

Given the effects of globalization on all enterprises, it should not be surprising that quality is demanded with minimum expenditure of resources. Test and inspection resources are inherently difficult to justify because they can be easily dismissed as not contributing value added to the goods or services produced. They do contribute by providing process control, but not necessarily directly. A product is physically unchanged after test or inspection, except perhaps for the addition of a sticker with the inspector's name on it.

In a broader perspective, production systems design had considered for many years the optimal distribution of inspection effort in control of complex processes. For example, Nurani and Akella (1996) studied the best strategies for sampling semiconductor wafers. A later paper (Shindo et al. 1999) showed that the classification of defects by family allowed the accurate in-line yield prediction required in modern production systems. Such work implies that in designing inspection and test systems, we need to look beyond the individual inspection work point and examine the whole range of defects possible, where they are generated, and how they can best be detected using the technology and models available. This theme will be returned to in Section 7.

2. TEST AND INSPECTION REQUIREMENTS

Test and inspection are decision functions in manufacturing and service industry. They provide decisions concerning the fitness for use of either an item of product or a production process. As such, their prime responsibility is in decision quality, and these functions of test and inspection are often placed in a department with ''quality'' in its title. Decisions made by the test and inspection subsystem should be:

- **Precise:** Enough depth of information should be incorporated into the decision process so that its conclusions are unbiased by underspecification or rounding errors.
- **Valid:** Decisions should be the same as would be reached after the product or process has been allowed to proceed to actual use.
- **Reliable:** Valid decisions must be possible repeatedly, with different items of product or different states of the process having consistently correct decisions. An inspection or test system should not need frequent recalibration for reliable operation.
- **Robust:** The inspection system must be capable of detecting and classifying a range of defect types large enough to cover customer concerns. This aspect has also been termed ''flexibility'' (Drury 1992a)
- **Rapid:** Information on a decision is needed rapidly enough for the system to react before many, or even any, defective items have been produced.

Both ''test'' and ''inspection'' are from Latin roots, with the former defined by Websters as ''to view closely in critical appraisal: look over,'' and the latter implying ''a critical examination, observation, or evaluation'' or ''a procedure, reaction, or reagent used to identify or characterize a substance or constituent.'' Precision, depth, and validity are key elements in these definitions, while reliability appears of necessity in a manufacturing context where a sequence of decisions is required. To differentiate between test and inspection, test typically requires a determination of functional suitability for use, while inspection is more usually confined to indications of fitness for purpose short of actual use testing. For example, each new aircraft produced will be flight tested before delivery to determine whether it fulfils its functional requirements. Prior to this flight test, considerable inspection of components and subassemblies will have taken place to ensure that they are free from manufacturing or assembly defects.

2.1. Inspection Measures and Scales

Note that the rather loose term *accuracy* of inspection and test is related to precision, validity, and reliability. To measure the performance of an inspection system, we need to consider how well its decisions match the decisions that should have been made given complete knowledge of the system and the items inspected. First, note that decisions can be of three types, following quality control terminology:

1. Decisions about the fitness of the design of a product. This is off-line quality control, as discussed in Section 1.2. above.
2. Decisions about the fitness of a single item (or group of items) of product. This is the traditional customer protection aspect of any test and inspection system. It is the area traditionally classified as (statistical) quality control, although it can include physical sorting of product items.
3. Decisions about the fitness of a process to continue manufacturing. This is the *jidoka* concept (Monden 1992), aimed at preventing a production system from ever producing a single defective product. A decision not to continue the process does not always imply that individual items produced are unfit (see Section 1.3). Decisions about process fitness are traditionally termed (statistical) process control.

TABLE 1 Definition of Outcome Probabilities

Decision of Test and Inspection System	True State of Conforming	Item (or Process) Nonconforming	Total
Conforming	$p_1\,(1 - p')$	$(1 - p_2)\,p'$	$p_1 + p'\,(1 - p_1 - p_2)$
Nonconforming	$(1 - p_1)\,(1 - p')$	$p_2\,p'$	$(1 - p_1) - p'(1 - p_1 - p_2)$
Total	$1 - p'$	p'	1.0

Each of these types of decision can only have two final outcomes: fitness and unfitness (of item or process). There are many names for these outcomes:

Conforming/nonconforming
In control/out of control
Good/faulty
Effective/defective

A Note on Measurement Scales. Each decision is a binary one, although subdivisions are possible. For example, an individual item may be conforming or nonconforming (scrap) or nonconforming (rework). Similarly, a process may be fit to continue, or fit to continue with increasing monitoring, or unfit to continue. This brings up the question of level of measurement. Final outcome (above) is always measured on a nominal scale with a discrete number of categories, although some implied ordering may exist. Nominal scales are used directly in test and inspection for the recording of discrete defects on an item—for example, scratches on roller bearings (Drury and Sinclair 1983). Ordered, or ordinal, scales are rarely used as such in test and inspection, although they have occasionally been proposed (e.g., Kelly, 1955). Interval scales, where a continuous measurement is taken, or even ratio scales (which are interval scales with a true zero), are often a part of the measurement process leading to a decision. Thus, the outside diameter of a roller bearing may be measured with high precision and plotted upon a statistical process control chart, although the application of standard rules to this measurement will give a decision outcome about whether the process should continue.

The outcomes of test and inspection can thus be classified by the level of measurement upon which they are based and the type of decision required. These outcomes in turn define the typical statistical methods used, such as prototype testing for decision 1 above. For in-process quality control (decision 3 above), nominal scale decisions lead to attributes control charts, while interval or ratio scales lead to variables control charts (Vardeman and Jobe 1998). The second type of decision above would lead to either attributes of variables sampling plans, although the whole concept of sampling a batch to determine its quality has largely been abandoned in favor of in-process quality control.

Because the final decisions are binary ones, we can define performance in terms of the outcomes for product items (or processes) that were in fact either fit or not fit (conforming or non-conforming). If these probabilities are defined as:

p_1 = probability of deciding that a conforming item or process is conforming
p_2 = probability of deciding that a nonconforming item or process is nonconforming

then the probabilities of the various outcomes are as shown in Table 1, where p' is the true fraction of items (or processes) nonconforming.

In the special case of inspection of items for attributes, these cells have common names, as shown in Table 2.

TABLE 2 Outcomes of Inspection of Items by Attributes

Decision of Test and Inspection System	True State of Conforming	Item (or Process) Nonconforming
Accept	Correct Accept	Miss
Reject	False Alarm	Hit

However, as noted above, in cellular and JIT manufacturing, final outcome should be predicted rather than measured. Thus, statistical process control becomes the technique most used for deciding upon process fitness, with attributes inspection of individual product items playing a progressively smaller role. To move from output control to process control, however, requires detailed, predictive models of all processes. Until these are achieved, the dream of eliminating all final functional test and inspection will perhaps remain a rarely attained goal. In particular, large process changes arising through continuous improvement will mean that predictive models still lag behind the need to deliver product. Thus, the practical strategy to achieve fitness with high reliability would appear to be cyclic changes from item to process control as each process cycles through maturity and periodic replacement. The rapidity with which predictive control is achieved can be expected to increase on successive cycles as parts of predictive models can be expected to be usable across innovation cycles.

3. LOGICAL STRUCTURE OF TEST AND INSPECTION

If we are to proceed towards predictive control of any process, detailed models of the process are needed. This applies equally to the test and inspection system itself. In this section, we develop a classification system for test and inspection systems and generic structure for test and inspection activities. These structures will help guide our more detailed examination of human and automated systems in later sections while showing explicitly the similarities between human and automation. Such similarities are not emphasized, and indeed rarely addressed, in the test and inspection literature.

3.1. Human and Automated Test and Inspection

Logically, no inspection or test system can be either fully automated or fully manual. These extremes are approached, for example with the automated functional tests for integrated circuit chips or the visual examination in proofreading. But there is always some human operator involved in the system, even if only for setting up the system and maintaining it. Equally, an apparently unaided human inspector still uses technology in the workplace, from desks and chairs to flashlights and hand tools.

We can thus define the degree of automation of inspection from one extreme of fully manual, through various levels of hybrid inspection, to the other extreme of fully automated. In this chapter, we will use *manual inspection* as a shorthand for inspection primarily using the human operator's minimally aided senses and minimally aided decision mechanisms. At the other extreme, *automated inspection* will be used to denote an inspection system where the human operator is not involved in on-line sensing or decision functions. Most real systems will involve hybrid inspection, as some functions involve humans and some automation.

The design of systems, such as test and inspection, that can logically incorporate alternative designs (manual or automated) of different system functions has a long history in human factors under the label of allocation of function. This is a primary concern to system designers who must choose which parts of the inspection and test system to automate, although in practice few papers on automated inspection mention function allocation explicitly. The typical automation paper (e.g., Komatsu et al. 1999) list the shortcomings of humans as inspectors and uses this as the justification for developing an automated system to replace the humans in the system. Often the automated system is characterized as being able to detect smaller defects, or automated-system performance is assessed by comparing how well the system classifies a test set of items previously classified by human inspectors. Allocation of function has never had much influence as a formal discipline outside human factors circles (e.g., McCarthy et al. 2000), but it still provides a logical framework for considering a range of automation alternatives. In fact, the allocation of function at the design stage is becoming recognized as perhaps too rigid a design tool, with flexible allocation during operations seen as a more natural and useful approach. For inspection systems this means designing alternate means of fulfilling each objective and having these parallel means available during operations. This requires that we still consider automation alternatives function by function.

3.2. Mission and Function in Test and Inspection

In the classical systems design process (e.g., Singleton 1972), a top-down design starts by defining the system mission, then splits the mission into logical functions. Each function is a unit of system behavior that can be assigned to either human or machine without consideration (at least at first) of other function allocations. Functions can be specified by their goals (or outcomes) in a way analogous to the mission specification of the overall system. In a test or inspection system, we have implicitly defined the mission in Section 2 by defining characteristics of the final system (precision, validity, etc.). Now we must be more explicit about such a definition, both to guide our later evaluation of test and inspection systems (Section 7.1) and to help define the system's logical functions.

The implicit mission in Section 2 concerns determination of functional suitability, so this can become the basis for a more formal definition:

> *The test and inspection system determines the suitability of a product or process to fulfil its intended function, within given parameters of accuracy, cost, and timeliness.*

This definition incorporates both the idea of evaluation of suitability and the notions of predefined system performance. It also forces us to concentrate on the ultimate suitability of the product from the point of view of the customer, so that we do not perform test and inspection activities unrelated to customer needs. Finally, the definition covers both products and processes, so that it can apply equally to output control and process control.

Before we can define the mission for any particular test or inspection system we must be able to specify customer needs. While a detailed framework for designing inspection systems is given in Section 7, we must consider now how to define such needs. One way is to apply a failure modes and effects analysis (FMEA) to the product and design a test and evaluation system to cover each of the potential failure modes. But this technique does not make the customer an explicit part of the design process, whereas we have seen earlier (Section 1) that direct customer input is increasingly needed in more customized products. A preferable technique is to begin with customer function and quality requirements as the basis for a list of product attributes that form the basis of test and inspection. In attributes inspection (Section 2.1), this list is often a ‘‘defect list’’ or ‘‘fault list’’ defining the discrete defects that the inspection system must ensure the customer never experiences.

An explicit approach has been used in a series of studies of paint-inspection systems for automotive painted body panels culminating in Lloyd et al. (2000). Instead of starting with the list of defects traditionally used by process engineers, they sampled many body panels with and without defects and had customers directly rate their dissatisfaction with each defect. This technique produced a list of the most important defects from the customer’s point of view and associated criteria for defects to be below the customer’s threshold of dissatisfaction. They could then design improved inspection systems, for example, using better lighting for human inspection, that reduced warranty data significantly and had a payback time of only six months.

Function lists for inspection have been defined generically by Drury (1978), Sinclair (1984), and Wang and Drury (1989), and for the specific case of inspecting aircraft structures by Drury et al. (1990). Table 3 defines five functions of inspection, with the goal (or outcome) and representative errors for each. Note that the functions are allocation independent in the systems-design sense, in that the functions can be assigned to either human or machine (mechanical, optical, electrical, computer) components. The outcomes and errors remain the same, although the detailed error-causation mechanisms will obviously differ. Thus, error 3.1, ‘‘Indication missed,’’ in the Search function could be caused by the human searcher not detecting a visual indication through inattention or by an automated visual inspection system (AVIS) failing to reach the threshold required for a response to a pixel configuration. The result is identical, although the error mechanism is different.

Because each function can (in principle) be allocated to human or machine, parallel models need to be presented for both types of component. This is done in the next two sections, which provide detail and models for the human and automated components available for building integrated, usually hybrid, systems.

TABLE 3 Generic Function, Outcome, and Error Analysis of Test and Inspection

Function	Outcome	Errors
Setup	Inspection system functional, correctly calibrated, and capable	1.1. Incorrect equipment 1.2. Nonworking equipment 1.3. Incorrect calibration 1.4. Incorrect or inadequate system knowledge
Present	Item (or process) presented to inspection system	2.1. Wrong item presented 2.2. Item mispresented 2.3. Item damaged by presentation
Search	Indications of all possible nonconformities detected, located	3.1. Indication missed 3.2. False indication detected 3.3. Indication mislocated 3.4. Indication forgotten before decision
Decision	All indications located by Search correctly measured and classified and correct outcome decision reached	4.1. Indication incorrectly measured 4.2. Indication incorrectly classified 4.3. Wrong outcome decision 4.4. Indication not processed
Respond	Action specified by outcome taken correctly	5.1. Nonconforming action taken on conforming item 5.2. Conforming action taken on nonconforming item

4. THE HUMAN ROLE IN TEST AND INSPECTION

The framework presented in Table 3 suggests a linear sequence of functions, but in practice there are some branches and reentries in the sequence. For example, an indication of one type may be located, a decision made, and action taken without consideration of any other indications. This occurs when a discrete fault with major consequences is located, such as a missing chip on a circuit board. For simplicity, this treatment will consider only sequential steps.

However, the concept of a one-dimensional sequence is inappropriate for human inspection functions for another reason. Humans can operate at several different levels in each function, depending upon the requirements. Thus, in search, the operator functions as a low-level detector of indications but also as a high-level cognitive component when choosing and modifying a search pattern. It is this ability which makes humans uniquely useful as self-reprogramming devices, but, equally, it leads to more error possibilities. As a framework for examining inspection functions at different levels the skills/rules/knowledge classification of Rasmussen (1983) will be used. Within this system, decisions are made at the lowest possible level, with progression to higher levels being invoked only when no decision is possible at the lower level.

In inspection, two of the functions (2. Present and 5. Respond) have no logical higher-level functions. Note, however, that if test and inspection is regarded as part of a total process-control system, function 5 (Respond) involves considerable higher-level cognitive functioning for diagnosis and prediction of the optimum changes to be made (e.g., Moray et al. 1986; Bainbridge 1990; Umbers 1981). Each function will be considered in turn, but first the terms used in classifying human errors need to be reviewed. System-specific errors have been classified by Hollnagel (1989) as error phenotypes. They are instances of errors generated by an error-generating model or mechanism, which provides the prototypical errors, or error genotypes. Thus, the errors in Table 3 are all phenotypes, but for humans each is an instance of a genotype. Each level of the Rasmussen S/R/K classification has its associated genotypes. Reason (1990) identifies these as skill-based slips, rule-based mistakes, and knowledge-based mistakes. Following Reason (1990), mistakes are choosing the wrong action (or decision), whereas slips are failures to implement the chosen action correctly. Thus, mistakes represent wrong intentions, while slips represent the failure of correct intentions.

With these classifications in mind, the functions will be considered.

4.1. Human Inspection Function by Function

4.1.1. Setup

In setting up a test and inspection system, the measurement devices, decision aids, and recording mechanisms must be procured, checked for functionality, and calibrated. For optical inspection, for example, the correct lighting and magnification levels must be implemented. Calibration consists of checking that the measurement devices respond in the correct manner to known inputs. At the lowest level, these are a sequence of psychomotor activities, equivalent to the items on a pilot's preflight checklist. Job aids consist of checklists, while the tasks themselves are typically closing switches and observing outputs or, more usually now, interacting with computer-based equipment. All of the factors important in good control panel design (Chapter 39) or in human–computer interaction (Chapter 44) become important if these steps are to have a high reliability. Possible slips are those common to any sequential activity, listed by Hollnagel (1989) as:

Wrong place	repetition reversal omission
Wrong time	omission delay premature action
Wrong type	replacement
Not in current plan	insertion intrusion

Note that the errors possible can be mapped onto the set in Table 3 in many different ways, that is, the outcome phenotypes do not necessarily specify the error genotypes.

At the rule-based level, there are typically changes in the setup to accommodate different customers, different products, or different process conditions. Thus, in Drury and Kleiner (1990), a job aid was produced to help inspectors of roller bearings reason more effectively from customer specifications to choice of inspection standards. In aircraft-structure inspection, the calibration of an eddy-current meter must be changed by known rules to accommodate different thicknesses of fuselage skin or detect cracks of different depths. Errors at this level consist of misapplication of rules, particularly reasoning within multiple conditions. Again, principles for designing job aids to assist in rule application are well documented (Johnson 1990).

Knowledge-based reasoning should rarely be required of a setup inspector under normal conditions. It will be invoked when new products are introduced; after process changes due to continuous improvement; occasionally to substitute an alternate inspection device for one that has failed; to troubleshoot and diagnose errors in the inspection device; or even as an innovation on the inspector's part to improve the inspection function. Examples are specifying an increased magnification for detection of smaller severities of visual defects in circuit chip inspection, or increasing the threshold of an eddy current meter to avoid frequent false alarms in lap joint inspection of an aircraft fuselage. Errors can arise from incomplete models of process, product, and/or inspection device or from faulty application of these models. Examples of such knowledge-based reasoning, and of its error genotypes, are more commonly found in the process-control literature (e.g., Moray et al. 1986). System improvement consists of knowledge-based training of the set-up inspector as well as job aids such as schematics and computer programs (e.g., Johnson 1990) to aid understanding of the various components.

4.1.2. Present

Each selected item of product, or test point in the process, must be presented to, and interfaced with, the inspector. Although typically a machine function, it can be given to the human, for example, picking items from a mass-production process and placing each in turn into a gaging device. In more extended inspection tasks such as inspecting aircraft for structural defects (Wenner et al. 1997), the inspector must move to the inspected area rather than the area being moved to the inspector. In such cases, inspectors must move their whole bodies or their limbs to ensure that each area of the item is available to the (visual) sense. The tasks are again psychomotor, typically picking up, orienting, placing, attaching sensors, and disposing. The reliability of such processes is high, with errors due to either misperception of orientation, or to slips in musculoskeletal execution. Models of such activities are widely available (e.g., Knight and Salvendy 1992; Holden 1981). Standard design techniques for preventing misassembly (e.g., the *poka-yoke* concept of Japanese JIT manufacture, Monden 1992) can be used to reduce error rates. Training in manual skills (e.g., Salvendy and Seymour 1973) should not be ignored to improve reliability.

4.1.3. Search

Search is an active process in any inspection context. The item or process inspected must be searched in stages. At times these stages can involve moving an area of limited extent across the item, as in using a flashlight to inspect the inside of an aircraft fuel tank or a microscope's limited field of view to cover the whole area of a microchip. Within a field of view, a target is visible to the inspector only within a limited area, called the visual lobe. This visual lobe must be actively moved in successive fixations across the field of view with saccadic eye movements. Within a single visual lobe, information can only be extracted at a limited rate (Eriksen 1990), so that an attention area can be said to move successively within the visual lobe.

Note that search is a serial sequential process, terminating upon successful detection of a discrete indication (such as a surface scratch) called a target in visual search. Because of its sequential nature, it is a resource-limited rather than data-limited process (Wickens 1991). As such, the probability of detection increases with time spent searching in a manner dependent upon the search plan.

Visual search is so well practiced as to be automatic at the visual lobe and field of view levels, and experienced inspectors' movements of the field of view across the item can also require little cognitive effort. Thus, to a large extent, the search process occurs at the skill level. Aspects of this skill, such as the attention conditions (Eriksen 1990; Engel 1971) and the optical factors affecting the visual lobe (e.g., Overington 1973) have received considerable study so that reasonable models of the probability of detection are possible (e.g., Greening 1975). Errors at the skill level are failure to detect/locate a target and detection/location of a nontarget. However, there are still strategic issues in visual search as evidenced by error phenotypes. Targets that are fixated do go unreported: parts of the item are not fixated at all, while others receive multiple fixations: search terminates too quickly, or continues unproductively.

At the rule-based level, it is the search plan that is altered. Thus, the choice of the next fixation point for the visual lobe is made. This is a decision partly based upon a fixed top-down search plan, but also partly based upon bottom-up information gained during the search. For example, Kundel and LaFollette (1972) show that searchers of medical X rays return a number of times to fixate places where a possible indication was seen. Search is usually modeled as either a fixed systematic process or as a purely random process (e.g., Morawski et al. 1980), although these are simplifications for mathematical convenience. As noted earlier, the probability of detection increases with search time. Typical cumulative search time distributions are shown in Figure 1 for the extreme cases of random search, where each fixation is placed randomly on the field of view, and repeated systematic search,

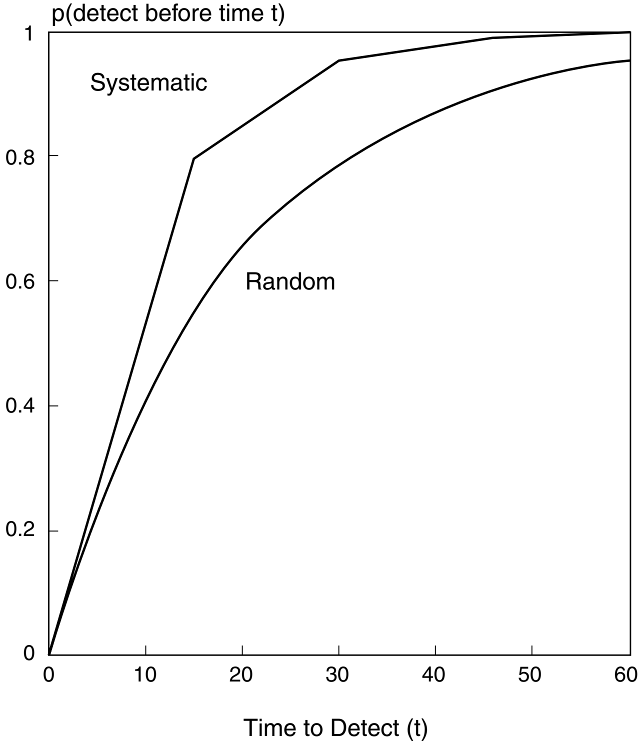


Figure 1 Speed–Accuracy Trade-off Curves for Random and Systematic Visual Search.

where fixations are only placed in unsearched parts of the field of view within each complete scan. Errors at the rule-based level are failing to choose a plan that will lead to the target being fixated.

Knowledge-based functioning in search consists of optimizing choice of parameters for the search. For example, prior information concerning the likelihood of a target being in a particular location will lead to that location being searched first. Choice of how long to spend searching before deciding that no target exists is another optimization aspect of search. There is evidence (Baveja et al. 1996; Chi and Drury 1995) that humans do choose a close-to-optimal stopping policy for search. Errors at this level are stopping too soon, neglecting an area entirely, or the lesser error of failing to detect the target in an optimally short time.

To improve search, the conspicuity of the target in the background must be increased (e.g., by overlays, Chaney and Teel 1967), or the search time increased (e.g., Schoonhard et al. 1973), or the search plan improved by training (e.g., Kundel and LaFollette 1972; Gramopadhye et al. 1997a). If the task requires simultaneous search for more than one type of indication, performance degrades rapidly (e.g., Gallwey and Drury 1986). Recent research (Wenner 1999) on large-scale search tasks, such as inspection of aircraft or searching offices for files, has shown that people need good information support if they are to devise effective and efficient search strategies.

4.1.4. Decision

The output from the search function is either zero, one, or many indications that must be evaluated by the decision process. Logically, if no indications have been found, the item or process must be classified as fit for use. Note that a false alarm is not possible unless at least one indication reaches the decision stage; thus, false alarms are directly diagnostic of decision error rather than search error. Given that decisions will be nontrivial only with an indication, the nature of the decision or indication must be examined.

Decision conditions must always be provided in terms of rules to the inspector, although whether these rules are well formulated depends upon the organization. At times they can be passed on by on-the-job experience from more senior inspectors, while in other instances complex, written, decision procedures exist. Typically, the rules are of one of three types:

Rule 1: IF the magnitude M_i of an indication of type (i) exceeds a severity S_i, THEN item is not fit.
Rule 2: IF the magnitude M_i of an indication of type (i) under circumstances (j) exceeds a severity $S_{i,j}$, THEN item is not fit.
Rule 3: IF the number of indications of type (i) with magnitude M_i exceeding severity S_i exceeds N_i, THEN item is not fit.

The existence of such defining rules would naturally suggest rule-based behavior, but other factors enter into the decision. For extreme indications, such as total absence of an indication or very high severity of indication, the decision will be trivial and essentially skill based. Thus, a missing component in an assembly will ''automatically'' trigger a rejection response, as will a major crack in a magnesium casting. In the former case, the implication is that

$S_i = 0$
for i = missing component

while the latter case has

$M_i >> S_i$
for i = crack

Most decisions, however, are nontrivial. In inspection, the decision component has been identified with the theory of signal detection (TSD) (e.g., Drury and Addision 1973), although not without necessary warnings about validity (e.g., Megaw 1992). Here, the perceived magnitude of an evidence variable is compared with a perceived criterion, and a decision to reject as unfit made if the evidence variable exceeds the criterion (McNichol 1972). If we modify the notation above with primes indicating perceived variables, we have:

Perceived magnitude for type (i) = M'_i
Perceived criterion (standard) for type (i) = S'_i

TSD states that the decision is based upon $(M'_i - S'_i)$ so that

IF $(M'_i - S'_i) < 0$ THEN accept item
IF $(M'_i - S'_i) > 0$ THEN reject item

The distribution of $(M'_i - S'_i)$ depends upon both the true signal, and the noise in the decision, thus:

$$\begin{aligned}\text{Signal} &= \text{mean } (M'_i - S'_i)\\ &= \text{mean } (M'_i) - \text{mean } (S'_i)\\ \text{Noise} &= \text{var } (M'_i - S'_i)\\ &= \text{var } (M'_i) + \text{var } (S'_i)\end{aligned}$$

To maximize decision reliability, that is, the fraction of total decisions made correctly, then the signal must be as large as possible and the noise as small as possible. This gives a maximum signal-to-noise ratio. In inspection terms, this means magnifying the difference between indication magnitude and standard, for example, using optical devices or computer-enhanced images (e.g., Komorowski et al. 1991) to increase the signal. To reduce the noise, the variability in both the indication and the standard should be minimized. Standard variability can be reduced by having comparison standards available at the workplace (Harris and Chaney 1969). Perceived magnitude variability can be reduced by training (e.g., Gramopadhe et al. 1997b; Drury and Kleiner 1990).

For simple rules, such as rule 1, used in the above exposition, performance is skill based and the only possible errors are:

1. Failing to invoke the decision process
2. Deciding that a conforming item is nonconforming (false alarm)
3. Decision that a nonconforming item is conforming (miss)

All are slip errors, but the first is an example of Hollnagel's sequence errors. Misses and false alarms are classic examples of correct intentions failing.

Rule-based decision making is well supported by training and job aids for complex rules (rules 2 and 3) or for lengthy lists of rules of type 1. Errors can be due to invoking the wrong rule, misapplying the correct rule, or failing to invoke the decision stage. Examples of the first of these errors were given by Drury and Sinclair (1983) in aircraft roller bearing inspection, where there was confusion about defect names among the inspectors, leading to choice of rule appropriate to the wrong type of indication.

In a manufacturing setting, it is rare that the false alarm and miss errors will have equal weights. Thus, in inspection of fuel flow valves for a space shuttle, the consequences of a miss are potentially catastrophic, whereas a false alarm leads only to the delay of component replacement. The weighting of these consequences is an essential component of TSD and represents a form of knowledge-based behavior when consciously applied. An inspector can define an optimum strategy as one that maximizes the expected value of outcomes rather than maximizing the fraction of correct decisions. Expected value depends upon:

1. The costs and values associated with the four decision outcomes, correct accepts, false alarms, misses, hits (V1, V2, V3, V4)
2. The prior probabilities of an item being conforming $(1 - p')$ or non-conforming (p').
3. The criterion (S_i') chosen by the inspector for reaching an accept or reject decision

It is the third of these factors that is under the inspector's control, based upon the perceived values of the other two factors. It is possible to calculate an optimum placement of criteria so as to maximize the expected value across the decision outcomes (McNichol 1972). Experimental evidence from both laboratory studies (Chi and Drury 1995) and field studies (Drury and Addison 1973) shows that inspectors do indeed modify their decision criterion in the direction indicated by the optimum criterion. Changes in chosen criterion are usually less than optimal predictions, an effect dubbed the sluggish beta phenomenon (Wickens 1991).

Clearly, knowledge-based functioning in decision depends upon accurate knowledge of costs, probabilities, and the process of optimization. Misperception of costs and values and misperception of prior defect probabilities can lead to incorrect decisions on criterion placement. Decision support for cost and value perceptions is simple in principle, although it is almost never formalized in an inspection task. Support for estimation of the true defective rate for each type of defect is much more difficult because it is based upon the output of the inspector, or of another inspector who is likely to be equally error prone. It is possible to provide such data to the inspector using another inspector who is allowed more time and/or resources for the decision. When these data are provided as feedforward of likely defect types and probabilities (Sheehan and Drury 1971) or as performance feedback (Drury and Addison 1973), dramatic reductions in error are found.

An alternative way of approaching the decision function in inspection is through the notion of fuzzy sets. Indications of defects such as wear, corrosion, scratches, stains, and quality of cloth have imprecisely defined criteria, and judgments cannot often be represented by a single precise number (Karwowski et al. 1990; Watson et al. 1979). The framework of fuzzy sets (Zadeh 1965) provides us a way of dealing with the category of problems where an absence of sharply defined criteria of class membership is the source of imprecision. Thus, inexact knowledge can be represented, and so can situations where membership in sets cannot be defined purely on an yes/no basis.

Wang et al. (1986) postulate that if the relative contribution of the different cognitive skills towards the performance of a human–machine system can be either known or inferred, then the corresponding subjective assessment of these skills can help in the designing cognitive aids for the human. They developed a fuzzy set approach to formulate a multicriteria decision-making problem to determine whether individuals can prioritize cognitive skills considered important for inspection performance.

A fuzzy set approach was more appropriate than a probabilistic approach because the question was not whether a cognitive factor belongs to a set representing cognitive skills important for inspection, but how strongly it belongs to this set. Thus, the decision alternatives (cognitive factors) for the individuals were imprecise.

The problem was formulated as follows: $X = (X_1, X_2, X_3, X_4)$ constitutes the set of decision-making alternatives corresponding to four cognitive factors. These factors are X_1 = memory, X_2 = attention, X_3 = perception, and X_4 = judgment. The objective was prioritizing the weights associated with these factors and determining the correspondence between these weights and their representation in a multiple-regression model as predictors of inspection performance. The weights are given by

$$FD\,(X_i) = \text{MIN}\,[F_{s,1}\,(X_i), \ldots, F_{s,12}\,(X_i)] \text{ for each } X_i \text{ in } X|S_j$$

where S_j = the fuzzy set of cognitive skills important for inspection performance corresponding to subject j.

$F_{sj}\,(X_i)$ = the membership function associating with each of the four cognitive factors, a number in interval (0, 1) that indites the grade of membership of factor i in S_j

D = the fuzzy subset that results from selecting, for each X_i, the smallest membership value from any of the fuzzy subsets S_1 through S_{12}, under the assumption that the judgments of all subjects are equally reliable.

The membership function was derived using the method proposed by Saaty (1974, 1977), which determined the grade of membership through the process of magnitude estimation derived from pairwise comparisons. The normalized membership for each factor in D was found to be:

$$D = \begin{array}{cccc} [\ 0.130, & 0.354, & 0.183, & 0.333\ ] \\ \text{memory} & \text{attention} & \text{perception} & \text{judgment} \end{array}$$

The priority of importance is seen to be attention, judgment, perception, and memory.

The above study demonstrates an application of fuzzy sets in modeling inspection performance, namely the human's perception of relative importance of various cognitive processes; this could serve as a selection device for inspection tasks. It has to be noted, however, that fuzzy set approaches are applicable in situations where humans must cope with the inexact or imprecise knowledge of the process or system being observed or controlled. It is a descriptive theory that attempts to model the way humans actually cope with a complex problem. Thus, if the space of actions is precise (e.g., inspection of variables such as length, diameter, etc.), then conventional approaches should be followed. Some of the issues that need to be addressed are methodologies for determining membership functions, clarifying the relationship between probability theory and fuzzy set theory, and interpreting of fuzzy utilities (guidelines to interpret the fuzzy advice).

4.1.5. Respond

Actions taken in response to a decision involve both the item itself and a data-capture system. Thus, at the simplest level, defective items can be removed from a production system, or a process can be stopped because it is no longer in control. At the same time, the inspector may need to capture the data in a form usable by the manufacturing system. Again, in a simple form, counters for different defect types are often observed at inspection workstations, with the counter readings recorded at specific times. Another example of a manual system is for aircraft structural inspection (Drury et al. 1990), where an inspector provides details of each defect on a nonroutine repair card.

With modern manufacturing system, evidence of a process or item not being fit to function is expected to be an increasingly rare event, so that action needs to be immediate and data dissemination widespread. Errors in response are, like errors of presentation, likely to be slips rather than mistakes. Actions taken on the basis of the test and inspection response may be complex and require cognitive processing, but the response itself is confined to the skill level. This is not to diminish its importance, as it is the final function of a sequence of processes that can have involved considerable thought and effort. Even if an inspector uses information intelligently to optimize search and decision, the outcome error is just as bad if an item is wrongly tagged or if part of the defect report is omitted. To improve the reliability of the response on the item, principles of workplace design should be followed. Space should be left for rejected items to be stored, and care should be taken to ensure that the response required when an item is rejected does not become so onerous as to discourage a rejection response. Data capture should be efficient and can be simply automated as an alternative to relying on error-prone handwritten forms. One airline company, for example, uses bar codes of all possible faults so that nonroutine repairs (NRRs) can be documented with little written or even keyboard input from the inspector. Portable computer-based systems for aircraft inspection (e.g. Patel et al. 2000) include functionality to allow the inspector to complete NRR forms directly in the computer, ensuring correct identification of the airframe number and the inspector.

4.2. Overarching Considerations: Job Design

Because we are unlikely to reach total automation of the whole manufacturing organization, there will be humans involved somewhere in test and inspection for the foreseeable future, even if only in a supervisory role (Sheridan 1987). Hence, organizational design will continue to impact upon the test and inspection function, requiring at least some familiarity with organizational variables even in highly automated systems.

Many prescriptive models of organizational design exist (e.g., Hackman 1990), but relatively few studies applied specifically to test and inspection. Early work by Jamieson (1966), Thomas and Seaborne (1961), and McKenzie (1958) established that humans change their inspection behavior, and hence performance, in predictable ways when social and organizational variables are changed. Inspection, McKenzie notes, is always of people. The inspector is always judging the work of others, or even his or her own work. Thus, pressures on the inspector are to be expected. More recent work (e.g., Taylor 1991) has examined the sociotechnical systems context of inspection in aviation maintenance and shows that similar pressures still exist. As industry moves from an inspection-based philosophy towards control at the point of manufacture, inspection activities are increasingly a part of all manufacturing jobs. Many years ago (Stok 1965), the visual presentation of information using X-bar and R charts was shown to be highly effective in reducing the production of defects by operators who measured their own production quality.

While the change from item to process control has been taking place, management in manufacturing has also been interested in organizational design for quality, also known as total quality management (TQM) (Evans and Lindsay 1993), as noted in Section 1.2 above. The interaction between such quality initiatives and human factors has been considered elsewhere in some detail (Drury 1996).

Job demands for the operator as inspector and controller logically include an enriched range of tasks. Job enrichment has been shown to be highly effective in improving inspection performance. Thornton and Matthews (1982) found missed defects reduced from 35% to 11%, while false alarms also fell slightly when an enriched inspection job was implemented, with the improvement in performance continuing well beyond the initial change. In a more broad-reaching job-enrichment program, Maher et al. (1970) found a halving of both human errors and time for inspection. To reach the level of knowledge and skill required in those new tasks, operators will need expanded training in process control and inspection procedures (Drury and Kleiner 1990).

The group-oriented aspects of organizational changes have been aimed at empowering a small, self-contained group of operators in the same way that job design empowers the individual (Hackman 1990). Good results have been reported in group inspection work within a maintenance environment (Rogers 1991; Diehl 1990). However, when groups such as manufacturing cells are implemented, the training needs again increase. Drury (1991) notes that operators in such environments require communication and interaction skills as well as group decision-making training to function effectively.

5. AUTOMATION IN TEST AND INSPECTION

Automated inspection and test systems are now finding a regular place in many industrial applications, alongside or replacing human inspection. Their main advantage is that they can perform a relatively limited number of inspection services rapidly and reliably, suiting them to high-volume repetitive test and inspection tasks. In this section, we provide a classification of automated systems based on their functions and give some detailed examples of automated systems in the literature. A typical system (Steinmetz and Delwiche 1993) is first introduced to illustrate many of the points raised.

The Steinmetz and Delwiche (1993) example is a Franco-American design of a system for automatic grading of cut roses. It was chosen because it uses a product of considerable geometric and chromatic complexity and variability but that is instantly recognizible and is likely to be well known to readers from different industries. In fact, a human grading was used to develop the test batches required for performance evaluation of the system. A second consideration was that the paper contains some numerical evaluation of system performance.

The paper begins, as is typical, with making the economic case for automation, that is, that the industry has high throughput and the grading and packaging costs (currently human activities) account for about half of total production costs. Direct cost savings should be possible through automation of the grading process. The cut rose has a set of quality characteristics that form the basis for the current grading system:

1. Stem length
2. Stem diameter
3. Stem straightness
4. Bud maturity (degree of opening)
5. Bud color

First, a holding and lighting system was designed, supporting the rose below the bud on a simple hook and surrounding it with diffuse fluorescent light of appropriate spectral characteristics. Note the similarity to the Setup and Present functions of Section 4. A standard color video camera was used to give a 480 × 512 pixel image (a low figure by today's standards), giving about 1.4 mm/pixel.

Next, a set of algorithms was devised for the computer to "understand" the image captured by the camera. Because the rose was in a constant position and orientation, defined by the support hook, the stem and bud could be unambiguously differentiated if the coordinates of the hook were known. Stem length was derived from the distance from the hook to the first cross-section of the image containing a nonwhite pixel, starting from the side farthest from the hook. At this point, other stem features could be extracted. The leaves were differentiated from the stem itself by searching column by column along the length of the stem for wide groups of nonwhite pixels. At each end of the stem, an authorized region (AR) was defined as a rectangle, while where leaves were encountered, triangular ARs were defined. These regions were large enough to include thorns, which had also to be differentiated from the stem proper. Logical detection of stem and leaf boundaries was performed using an edge-detection algorithm. Feature recognition made use of the geometric fact that a stem must be a set of rectangular segments of approximately equal widths located in the space defined by the two ends of the stem already found and oriented approximately parallel to the line between the stem end points. Note the similarity to the early stages of the human function of Search, where overall salient features of the item are used to guide the detailed search process in a top-down manner. Stem straightness was defined by the maximum angle between any of the rectangular stem segments, or by the maximum distance between any segment and the line joining the stem end points. Bud maturity measurement started by searching above the hook for an area of nonwhite pixels. This area was recognized by its edges, transformed to a convex shape and the centroid calculated. Maturity was defined as the maximum distance of any bud edge to either side of the centroid. Bud color was determined by averaging separately the red, green and blue chromaticity coordinates of the convex region of the bud.

From this measurement phase, a set of nine measures was available as a vector for each rose. In fact, the measurements were repeated for two orthogonal views of each rose to ensure that straightness was not an artifact of the direction of view. To test the system, two sets of roses were first graded by experienced graders, and the grades were refined in the laboratory using the raw camera images to improve grading consistency. The two sets were a training set and a test set, for use with the later classification evaluation. For the two direct measurements (length and diameter of the stem), standard errors were used to show that the automated system was accurate enough for use. For each of the other characteristics, the training set was used to train a multilevel neural net (or alternatively a Bayes classifier) to reach the correct classification of each rose. After training, the neural net then classified the test set to measure performance. Compare this technique with the Decision function for human inspectors, where training was again found to be useful. For straightness, the neural net classified 21 of 27 straight roses correctly and 28 of 33 crooked roses correctly, for an overall error rate of 18%. Bud maturity was classified in the same way, but onto a three-point scale rather than into two classes. The misclassification rate was between 15% and 21%, depending on rose color sample. Finally, comparison to a simple threshold of the average red chromaticity was used to classify red and white roses with 0% errors. The final human function of Respond was not used in this prototype system.

Recommendations were finally developed for improvement of the system, including a low-pass filter to prevent thorns interfering with diameter measures, removal of lower leaves to make length detection more reliable, and extension of the system to the more difficult task of detecting defects on the leaves.

From this description, the logic of automated systems is clear, as are many parallels with human inspection. The functions can be separated and implemented, performance measured, and recommendations made for future improvement. The functions of automated inspection will now be considered in turn because they involve quite different mechanisms than the equivalent human functions.

5.1. Automated Inspection Function by Function

Although automated inspection must logically fulfill the five functions presented in Section 4, these functions do not represent the most natural breakdown for automated inspection. Some functions are combined in a single piece of hardware, while some automated functions (e.g., sensing) have a major classificatory role. One example of combined roles is that the Present and Search functions can overlap in a CCD readout. Another is that Search and Decision can be part of one computer program rather than separated as they were in the rose-inspection example. A more natural breakdown for automated inspection would be as follows:

1. Setup
2. Materials handling
3. Sensing
4. Signal processing

Figure 2 shows how these two classification schemes map onto each other. The Materials-handling function of automation covers both the Access and Respond functions of our more general scheme, while, as noted above, sensing and signal processing may be difficult to differentiate physically. All are, however, separable logically.

5.1.1. Setup

The setup function in automated inspection corresponds to configuring the system for the task. Several environmental factors affect the setup function. In a machine vision system, lighting factors such as illumination levels, type of illumination, reflectivity, and contrast affect performance. Most vision systems are quite sensitive to these variables. In many cases, the automated system has failed because

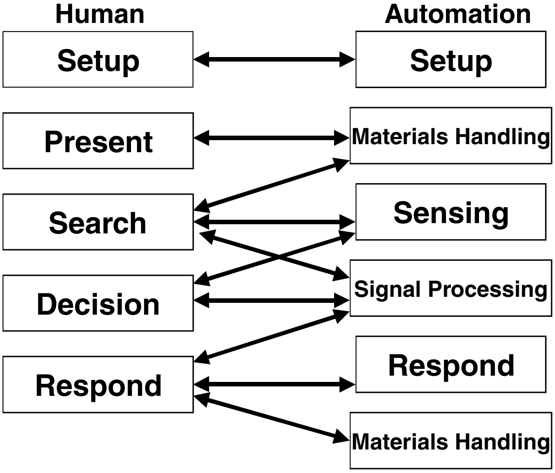


Figure 2 Mapping of Classification Schemes for Manual and Automated Inspection.

the stability of the environment and the effect that environmental factors may have on the measurements have not been correctly understood. In fabric inspection, cloth brightness can change due to environmental conditions that can lead to incorrect detection of dirt on cloth. Thus, we need to have either vision algorithms that can detect defects independent of illumination changes or a tightly constrained environment (e.g., Steinmetz and Delwiche 1993). Lighting intensity is a critical factor and must be sufficient to swamp interference from ambient sources to improve environmental reliability. Similarly, the contrast of the object against its background must be greater than the local lighting variation around the feature that is of interest.

5.1.2. Materials Handling

The materials-handling function for manufacturing test and inspection consists of devices to select items for test, mount them into the test fixture, move them within the fixture (if needed) and transport the item to the next stage. Thus, in the inspection of aircraft engine blades (Rosemau et al. 1999), a robot is used to position the blade in different orientations so as to cover all of the cooling holes with a sequence of video images. Often the fixture for inspection will be that used for transport throughout the system, for example in a flexible manufacturing system. Parts are held in the fixtures by gravity, quick-release clamps, or magnetic couplings, depending upon stability and human compatibility requirements. Pick-and-place robots can be used if an item needs moving from its fixture to the test/inspection device. In nonmanufacturing applications, the sensor must be moved physically around the item inspected, as in the robotic inspection of aircraft skins by Siegel et al. (1998) or the inspection of water pipes by Moraleda et al. (1999).

Very small items produced and tested at very high rates (e.g., fasteners, microcircuit chips) can be handled by vibratory bowl feeders, slides, and small conveyors to ensure that they are correctly presented to the inspection device. These systems often have automated disposal mechanisms built in to gate any defective items along a different track from the good items, an example of a materials-handling component equivalent to the human Response function.

The materials-handling system can be automated or manual. In a system with manual material handling, an operator manually loads the part and then initiates the inspection cycle. In some cases, automated materials handling is suitable and eliminates lifting and lowering of heavy components. The type of materials handling system (and the degree of its automation) depends on the inspection system and its place in the production system (i.e., preprocess, online, or postprocess inspection).

5.1.3. Sensing

In both inspection and process control, we can sense a wide range of properties of the product or item produced, or of the process itself. Thus, in test and inspection, a classification of these sensors aids in choosing the most suitable one from the many available. Here we use a standard listing from a recent computer automation text (Boucher 1996). All use the transducer principle, converting energy of one form (e.g., light, water flow) into a common electrical form with well-defined properties. The electrical signal can then be processed in many ways, independently of the type of transducer used.

A necessary part of sensing is signal generation. Typically some form of energy must be input into the sensor to elicit a response. Thus, a temperature sensor works because the electrical resistance changes, but detecting such a change requires that a voltage be applied across the resistance element and an output voltage measured. Signal generation involves the application of a suitable constant voltage or the illumination of an object to be sensed by appropriate lighting etc. We will consider this explicitly only when we deal with parts of nondestructive inspection in the next section.

5.1.3.1. Discrete-Event Sensors Note that these can be combined into arrays, for example to detect complex logical events, or even items of different sizes when some elements are activated and some not.

- *Mechanical limit switches:* These sense when an object displaces an actuator, either on a lever or by pushing/pulling. Their logic can be normally open or normally closed. Limit switches are used to sense when a machine (e.g., robot) has reached a particular position or when an item arrives on a conveyor.
- *Proximity switches:* Two types are inductive and capacitive, both of which measure when an object enters a defined sensing field. Inductive switches sense only conducting materials such as metals, while capacitive switches indicate the proximity of any object. They are particularly useful because they do not require contact with the item sensed and hence cannot damage it.
- *Photoelectric sensors:* An emitter unit gives out a beam of light or infrared radiation, which is sensed by a receiver unit. The emitter and receiver can be used in an opposed sensing mode, where an electrical event is registered if the beam is blocked. They can also be used in the reflective mode to sense objects of high reflectivity, and even used to sense only objects at a given range of distances using a convergent-beam system.
- *Fluid-flow switches:* An analog of proximity switches for fluid detection, fluid-flow sensors use a sprung or magnetic valve to detect when flow takes place in a pipe.

5.1.3.2. Continuous Sensors Continuous transducers convert energy directly from one form to electrical voltages or currents. A measurement circuit, such as a bridge circuit, is used to make the conversion from the direct sensor signal. The voltage or current is then digitized using an analog-to-digital converter, which provides input to the signal processor. Each type of transducer has a fixed relationship between the input energy and the output digital signal, a form known as the calibration curve.

- *Linear position transducers:* The simplest form is a linear potentiometer, where the position of the slider is proportional to the output voltage. Linear variable differential transformers (LVDIs) move a metal core between primary and secondary coils to produce a voltage proportional to core position.
- *Rotary (angluar) transducers:* Rotary potentiometers sense angle in a way analogous to linear potentiometers. Resolvers use rotary transformers similar to linear variable differential transformers. Optical encoders go directly from angular position to a digital signal by passing light beams through a disc attached to the rotating shaft. The disc has sectors printed in a pattern of black and white so that each angle corresponds to a unique pattern of transmitted and blocked light beams.
- *Float transducers:* As in an automobile gas tank, a float transducer uses a rotary potentiometer to produce a voltage proportional to the angle of an arm with a float attached.
- *Ultrasonic sensor:* The reflection time of a high-frequency sound pulse can be measured to give an accurate estimate of distance from an object or from a liquid level. In the latter case, an ultrasonic sensor is a noncontact alternative to a float transducer.
- *Velocity encoders:* Both velocity (speed) of solid objects and liquid flow rates can be measured. One way is to use an optical encoder and measure time between position changes to give the differential of position, i.e., velocity. Tachometers use an AC or DC generator to give a voltage proportional to angular velocity. Flow rate transducers can use the differential pressure in a pipe to measure a flow rate if pipe diameters are known. Alternatively, a small turbine blade can be inserted into the flow and its rotation velocity used to measure flow rate.
- *Force/pressure transducers:* The simplest force measures use load cells based on an LVDT, which displaces by a small amount as a force is applied. Strain gauges can measure the (small) deformation of thin film attached to a surface by measuring the resistance of an attached conducting element.
- *Temperature transducers:* The most common temperature sensor is a thermocouple made from two different metal wires. As the temperature changes, so does the relative resistance of the wires, giving a signal proportional to the temperature. Thermistors are semiconductors with the property of changing resistance with temperature.

5.1.3.3. Scene Sensors: While the above sensing devices convert discrete or continuous information into electrical signals, they are all essentially one-dimensional. Their output can be characterized by the changing of a single quantity over time, such as position or temperature. However, many automated inspection and test systems need to utilize two-dimensional or even three-dimensional data, for example in inspection of sheet materials or circuit boards. Such systems demand two-dimensional sensing.

The typical sensor for these applications is a charge-coupled device (CCD), which consists of an array of picture elements (pixels) that are light sensitive. Electrons emitted from each pixel are integrated for a field time (typically 1/60 sec) and a voltage proportional to this value is sent out. This voltage is thus proportional to the light intensity at the pixel, although the relationship is not linear. CCD pixels have a number of limitations, such as saturation and bloom, that can degrade image quality.

- **One-dimensional arrays:** Two-dimensional information can be obtained by sweeping a one-dimensional row of sensors across the two-dimensional surface of interest. With a one-dimensional array, such as a row of charge-coupled devices, the necessary condition is that the two-dimensional surface remain constant while the scan is in progress. Thus, any fixed object, such as a circuit board, can be scanned in this way to produce a two-dimensional image for later processing. The essential condition is that the sensor be moved relative to the material to provide an undistorted two-dimensional image. Alternatively, the inspected item itself may be moving at a known velocity, so that data collected from a one-dimensional array will give a complete picture of the material. In such a way, continuous sheet production, such as float glass or sheet steel, can be scanned continuously.
- **Two-dimensional arrays:** If the array is composed of rows and columns of sensors, as in a television picture sensor such as a two-dimensional CCD, then two-dimensional information is available simultaneously for all parts of the inspected material. A lens may be used to form a sharp image on the sensor array. For three-dimensional information, height maps may be constructed with appropriate two-dimensional sensors. The common methods for 3D image acquisition include laser radar, confocal microscopes, and high-speed triangulation-based 3D systems. It is currently believed that triangulation based systems offer the greatest potential, although this is somewhat application dependent (Svetkoff et al. 1989; Gieles et al. 1989).

5.1.4. Signal Processing

While the output from a discrete-event sensor is essentially digital, digitization is first required for all of the sensors of continuous one-dimensional or two-dimensional information,. Digitization is typically performed by an analog-to-digital converter (A to D converter). Such a device quantitizes the continuous information by comparing the signal with known levels or thresholds. If a signal exceeds threshold (i) but not ($i + 1$), then it is digitized at the level corresponding to threshold (i). The reverse process of digital-to-analog (D to A) conversion will not be considered here because it pertains to system effectors rather than receptors. D to A conversion is required when moving from the test and inspection task to a process control task.

For scene sensing, signal processing is more complex and will thus be covered in a separate section.

5.2. Image Processing in Automated Visual Inspection Systems (AVIS)

Because so much of test and inspection is now concerned with vision systems, whether human or automated, this section will use automated vision as an extended example through which to discuss issues of automation in general. A typical example has already been presented, that of rose grading, and will be used to illustrate concepts as they arise.

Lumai (1994) reviews the applications of vision systems and image processing in manufacturing, not all of which are concerned with test and inspection. For example, automated vision systems are applied to control of guided vehicles, control of robot positioning, and determination of orientation of parts for assembly. They may use the same sensors and signal-processing algorithms as an automated visual inspection system (AVIS), but they will not be considered further here.

Figure 3 shows the basic functions of image processing, adapted from Lumai (1994). Note as in Figure 2 that these partially correspond to the generic functions given earlier, with Classification very similar to Decision, Respond the same in both lists, and the other functions partially overlapping with Search. Ventura and Chen (1994) use a similar but more specific breakdown of functions in their system for classifying component shapes based on a two-dimensional image and the original CAD specification.

Signal generation here is the lighting system, as for example in the careful design of even illumination in the rose-grading study. The rest of the image capture system, after the lighting, is composed of a lens and a CCD. The lens is chosen to give the correct ratio between angular coverage

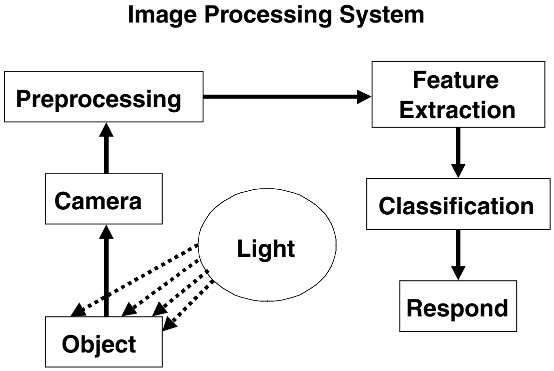


Figure 3 Functions in Image Processing.

and resolution, defined as the solid angle subtended by individual pixels. Note that lenses have their own aberrations, but modern lenses are amazingly free from distortions and also inexpensive. CCD devices are also increasing rapidly in the total number of pixels and the grayscale dynamic range. As of 2000, inexpensive CCDs contain well over 2 million pixels and have at least an 8-bit grayscale range. Note that for color imaging, very important in such inspection applications as food processing, three pixels with red, green and blue filters are used to capture the RGB image, giving $3 \times 8 = 24$ bits per data point. The signal is typically stored in a frame grabber so as to be available as an ensemble to the computer system for further processing. The signal at this stage consists of a grayscale level or RGB value for each pixel for each sampling interval.

Signal preprocessing now has the task of preparing the signal for feature extraction. In the rose-grading example, very little preprocessing was required because the lighting ensured that the object was captured as a dark stem and leaves against a uniform white background. In most industrial applications, however, dirt, visual noise, and stray reflections can degrade both the object inspected and the background. A typical visual noise-reduction technique is smoothing, performed by a low-pass filtering technique. This operation consists of convoluting a small template, say 3×3 pixels, with the original image. The value of the pixel at the center of the template is replaced by the value calculated by summing the product of the 3×3 array of pixel grayscale values and the template values around the center pixel. Thus (Lumia 1994), if the template is composed of nine identical values of 1/9, then the resulting smoothed value is the average of the 3×3 array of pixels. This means that any deviant (noise) value has its effect reduced by a factor proportional to the square root of the template area. The penalty for this smoothing is that signal information may be reduced along with the noise. If the signal were, to use an extreme example, a single deviant pixel, its discriminability would be reduced as if it were noise. The preprocessing may also consist of changing the overall brightness and contrast of the image, as is done, for example, in image-processing software for photographic purposes. This allows the subsequent feature extraction to utilize the full dynamic range of the image information despite changes in incident illumination.

Feature extraction uses a variety of techniques for converting a pixel value array into a set of discrete features useful for classification. One set of operations involves convolution to detect specific features. For example the template (Sobel operator):

$$\begin{array}{ccc} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{array}$$

when convoluted with a pixel array will yield a high value if the image contains a horizontal edge. If the operator is rotated about its leading diagonal, it will detect vertical edges. If edges are neither horizontal nor vertical, the two Sobel operators can be combined by finding the effective gradient in each direction. Typically, the edge is detected by comparing the value of the convolution at each point to a threshold. This threshold can be a fixed value or chosen algorithmically (e.g., Hou et al. 1993).

In addition to edge detection, there are also operators for thinning edges and for converting a set of pixels forming an edge to a straight line or a circular arc. Additional algorithms can be used to detect circles (e.g., holes drilled in components) or irregular features such as the leaves or thorns of roses. Some, as in Steinmetz and Delwiche (1993), use the concept of an allowed area or search for the maximum extent of a contiguous area such as a stem or bud. In fact, as with human perception, the more closely a feature can be specified, the easier it is to detect.

Classification: The set of features from feature extraction must be classified to reach conclusions about the overall acceptability or otherwise of the item inspected. Features such as end points, edges, and blobs need to be combined and classified in order to classify the object as a whole. While for a single measured parameter (e.g., force or temperature) a single threshold will provide this classification, for complex two- and three-dimensional objects the process is less obvious. For example, with the rose grading, one algorithm finds the size of the area representing the bud, but eventually a three-level scale is needed to judge bud development. This requires thresholds to be set in the simplest case, or more usually an algorithm must be developed to mimic human judgement of these factors. Fuzzy set theory can provide such an integration (see Section 4), but a preferred solution is to use neural nets as classifying algorithms. One such neural net architecture was used in the rose-grading example.

Neural network models have the ability to explore several competing hypotheses simultaneously. This is done using computational elements connected by links having variable weights that form massive parallel nets. Neural networks can be classified based on the nature of input data, that is, whether they process binary or continuous valued inputs. The pixels of a digitized image can be expressed either in binary (black or white pixels only) or grayscale format. Another classification is based on the type of learning. Learning can be either supervised or unsupervised. In supervised learning, the network is told what the correct answer should be during training. It can then determine whether or not its output was correct and use its learning law to adjust its weights. In unsupervised learning, on the other hand, the network has no knowledge of the correct answer (Caudill 1988a, b). For most inspection applications, neural networks with continuous valued inputs that are trained under supervision seem appropriate.

Current application of neural networks has included speech recognition, handwritten character recognition, target recognition, industrial parts inspection, and signature recognition (Buffa 1985; Cormier 1991). Cormier (1991) developed a three-layer, back-propagation neural network for circuit board component inspection. The system verifies presence of correct surface-mount components on board. Graphical images of various components were used. The scale of the images was chosen to accomplish an effective resolution of 1 mm per pixel, which is the most common resolution for surface-mount inspection systems. Noise was added to the images to simulate variations due to inconsistent lighting, surface irregularities, imaging equipment, and so on. The network was found to have false alarm rate, miss rate, and nonclassification rates nearing zero at a speed equivalent to covering a 100 mm × 100 mm area per second.

In a study that will be considered in detail in the section on function allocation, a neural net was developed as one of the automated alternatives to human visual inspection of circuit boards (Hou et al. 1993). It used the output from edge detection and thinning algorithms, followed by a Hough transformation as the inputs to a three-layer back-propagation network, with either a fixed threshold (for fully automated inspection) or a multi-level threshold where a human inspector could help the system make the final decision. The network was given a number of boards as the training set and then given another sample as the test set to measure its performance.

Neural network research and its application to the decision-making component in inspection offer interesting possibilities towards more realistic automated inspection systems. However, a number of issues have to be addressed before neural networks can make a significant impact in the area of test and inspection. The training time for neural networks can be extremely long, especially if there is an overlap of categories in the decision space. The speed of the network also has to meet line-speed requirements. With respect to neural networks, the speed is largely a function of the size of the middle layer. Adding middle layer units can significantly increase computation time while only increasing accuracy up to a point. Accuracy of neural networks in terms of both false alarms and misses has to be established and is always application specific. As with human training, the amount of fine tuning that is needed to get the desired performance may be quite large.

The final output from the overall system is a unit of product classified into one of a small number of categories corresponding to different system responses. Thus, a circuit board could be classified as “Conforming” or “Nonconforming” (Hou et al. 1993) or into three levels of acceptability, such as “Tight,” “Open 1,” or “Open 2,” for rose grades.

5.2.1. Examples

Any recent conference on automated vision from an automation or applied optics viewpoint will provide copious current examples of AVIS applications in a variety of industries. Table 4 lists a number of recent applications of automated vision to show the variety of products and industries covered by the pace of this form of automation. Unfortunately, few technical papers contain the depth of performance evaluation required to assess system reliability or its components. The Steinmetz and

TABLE 4 Examples of Automated Test and Inspection Systems

Industry and Reference	Task	Parameters/ Defects Detected	Technology
Civil Aviation Siegel et al. 1998	In-service structural inspection of aircraft	Cracks, corrosion	3D stereo vision and eddy current
Semiconductor manufacture Komatsu et al. 1999	Detection of defects in lithographic process after development	Contamination, scratches, reduced conductor thickness	Image processing
Aircraft engine manufacture Rosemau et al. 1999	Classification of cooling holes in jet engine fan blades	Cooling hole wrong diameter, blocked cooling holes	Robot movement of fan blade for different views, image processing to classify defects
Food products Legeard et al. 1999	Evaluation of quality of pork hams	Detection and classification of muscle color and fat thickness	Color vision system using hue, saturation, and intensity measures, using thresholding
Ceramics Kälviäinen et al. 1998	Classification of color matches in tile production	Classification of brown tiles by color and feature	Color machine vision system measures RGB pixel values, classify color features using neural nets
Chemical Industry Liu et al., 1998	Detection of welding defects in pipelines	Weld defects of all types	X-radiography followed by image processing and pattern recognition
Automobile Manufacturing Hung and Park 1996	Detection of dents in large automobile steel panels	Dents	3D computer vision measures slope, curvature and depth.

Delwiche (1993) paper was chosen as a worked example partly because such performance data was available and published.

6. NONPRODUCTION TEST AND INSPECTION

Production and manufacturing are not the only realms of inspection and test. Service industries need to inspect goods and services before they are released to the customer. A number of these nonmanufacturing applications are shown below:

- *Regulatory inspection:* To ensure that regulated industries meet meet or exceed regulatory norms. Examples are review of restaurants against local service codes, fire safety inspection of buildings, and safety inspection of workplaces.
- *Maintenance:* To detect failure arising during the service life of a product. This failure detection function can be seen in inspection of road and rail bridges for structural determination or of civil airlines for stress cracks or corrosion (see below).
- *Security:* To detect items deliberately concealed. These may be firearms or bombs carried onto aircraft, drugs smuggled across borders, or camouflaged targets in aerial photographs. They can also be suspicious happenings on a real-time video monitor at a security station. Law enforcement has many examples of searching crime sites for evidence.

- *Design review:* To detect discrepancies or problems with new designs. Examples are the checking of building drawings for building code violations, of chemical plant blueprints for possible safety problems, or new restaurant designs for health code violations.
- *Functionality testing:* To detect lack of functionality in a completed system. This functional inspection can often include problem diagnosis, as with checks of avionics equipment in aircraft. Often functional inspection is particularly dangerous and costly, as in test flying aircraft or checking out procedures for a chemical process.

6.1. Maintenance Inspection

The example chosen here is that of inspection of civil aircraft as part of the system for assuring the public that airworthiness is maintained throughout the service life of airframes, avionics, and aircraft structures. It is part of a maintenance process and is typical of many transportation applications, such as for maritime transport, heavy goods vehicles, or even the space shuttle.

Airworthiness of civil aircraft depends upon a process by which a team composed of aircraft manufacturers, regulators, and one or more airlines predicts possible system failures. This process, Maintenance Steering Group 3 (MSG-3), considers possible failure pathways (e.g., in structures, controls, avionics) and for each pathway determines a recovery strategy. For structural failure, this may be replacement after a fixed service life, regular inspection to ensure detection, or an indication to crew of the malfunction. The concern here is with the reliability of the primary failure recovery system for aircraft structural inspection: regular inspection to ensure detection.

Failure modes of aircraft structures can be cracks, corrosion, fastener/bonding failure, or deformation beyond the plastic limit. Inspection systems are designed to detect all of these in a timely manner, that is, before the failure has a catastrophic effect on structural integrity. For example, crack growth rates can be predicted probabilitistically from material properties and applied stresses, so that the MSG-3 process can schedule inspections before a potential crack becomes dangerous. However, the detection system has certain limits on the size of crack that can be detected, so MSG-3 typically schedules several inspections between the time the crack becomes detectable and the time it becomes dangerous. If too many inspections are scheduled, the costs are driven up in a highly competitive industry and the risk of collateral damage is increased due to the handling activities involved in the inspection process itself. Conversely, if too few inspections are scheduled, the probabilistic rate of the crack growth-prediction process may combine with the probabilistic nature of the detection process to cause dangerous cracks to remain undetected. Spectacular failures of this inspection process have occurred both for aircraft structures (Aloha incident, Hawaii, 1988) and engine components (Pensacola incident, Florida, 1997).

The MSG-3 process thus requires quantitative data on inspection reliability to function correctly. In addition, no rule-based prediction system can foresee all possible malfunctions, so that once an aircraft is in service, regular detailed inspections are made of the whole structure to discover any unexpected cracks. When such ‘‘new’’ cracks are found, the information is typically shared among manufacturers, operators, and regulators in the form of supplementary inspections. Similar considerations apply to other failure modes, such as corrosion.

This whole reliability assurance process thus rests upon an inspection system that checks both points where malfunctions are expected and points where they are not expected, for a variety of malfunctions. For good reasons, human inspectors are part of this inspection system, and thus human inspection reliability is an essential element in ensuring structural integrity and hence airworthiness.

The inspection task combines two goals: detection of expected malfunctions and detection of unexpected malfunctions. Neither detection is particularly easy or particularly rapid, so inspection can be a difficult and time-consuming task. In some ways inspection can be classified as an ill-structured task (Wenner 1999) because there is no simple step-by-step procedure that will ensure success and usually no knowledge of task success available during the task. Finding (n) malfunctions in a structure still leaves an unknown number (hopefully zero) potentially undetected.

In addition, inspection is typically scheduled at the beginning of an aircraft’s maintenance visit so that malfunctions can be detected early and their repair scheduled to overlap in time with other maintenance activities. As airlines streamline their parts inventory to reduce holding costs, the lead time for replacement components can increase, again pressuring the inspection system to ensure early detection. Aircraft typically arrive following scheduled service, that is, after the last flight of the day. Following opening up and cleaning processes, maximum inspection resources are committed to the initial inspection. In practice, this means inspectors working overtime, even double shifts, starting with a night shift, under some implied pressure for early detection. Human inspection reliability may not be optimal under these conditions.

The inspection task itself is classified in aviation as either visual inspection or nondestructive inspection. Regulatory bodies have issued formal descriptions of both of these tasks (e.g., Bobo 1989, for the FAA), and both have somewhat different characteristics in aviation

Nondestructive inspection (NDI) includes a set of techniques to enhance the ability to detect small and/or hidden malfunctions. One set of NDI techniques is those that enhance what is essentially still a visual inspection task, such as X ray, fluorescent particle, magnetic particle, or D Sight. They show cracks that are very small (fluorescent particle) or hidden within other structures (X ray). Apart from the steps necessary to ensure a good image, they have many of the human interface characteristics of visual inspection. The other set of NDI techniques is focused on specific malfunctions in specific locations, such as eddy current and ultrasound. For this reason, they are useful only for detection of malfunctions already predicted to exist. In practice, such NDI techniques are much more proceduralized than visual inspection or NDI techniques, which contain a human visual inspection component.

Visual inspection is much more common, making up 80% of all inspection (Goranson and Rogers 1983). It consists of using the inspector's eyes, often aided by magnifying lenses and supplementary lighting, as the detection device. Inspectors must visually scan the whole structure of interest, typically using portable mirrors to examine areas not directly visible. Whether the task is categorized as visual inspection or NDI, its aim is to detect flaws (indications) before they become hazardous. Next we consider the bodies of knowledge potentially applicable to aircraft inspection reliability (e.g., Drury and Spencer 1997).

Over the past two decades there have been several studies of human reliability in aircraft structural inspection (Rummel et al. 1989; Spencer and Schurman 1995; Murgatroyd et al. 1994). All of these to date have examined the reliability of nondestructive inspection (NDI) techniques, such as eddy current and ultrasonic technologies.

From NDI reliability studies have come human–machine system detection performance data, typically expressed as a probability of detection (PoD) curve (e.g., Spencer and Schurman 1995). This curve expresses reliability of the detection process (PoD) as a function of a variable of structural interest, such as crack length, providing in effect a psychophysical curve as a function of a single parameter. Sophisticated statistical methods (e.g., Hovey and Berens 1988) have been developed to derive usable PoD curves from relatively sparse data. Because NDI techniques are designed specifically for a single fault type (e.g., cracks) and much of the variance in PoD can be described by just crack length, the PoD is a realistic reliability measure. It also provides the planning process with exactly the data required because remaining structural integrity is largely a function of crack length.

Both the FAA (1993, pp. 26, 35) and the Air Transport Association (ATA) have recognized the need for equivalent studies of the reliability of visual inspection as a research priority.

Aircraft inspection has already benefited from models of human inspection such as those in Section 2. Thus, the ECRIRE program (Spencer and Schurman 1995) examined one NDI technique, eddy current inspection, incorporating human factors variables. They were able to test one-person vs. two-person teams (no consistent effects) and gross body posture (a small decrease in detection performance when the inspector had to work at about knee height). A FAA program on human factors in aviation maintenance and inspection (e.g., Drury et al. 1997) has had some success in improving documentation design, lighting, and communications. This program expanded the search-plus-decision model following industrial inspection findings to include the five generic inspection functions presented earlier.

Such a task description invites task analysis, which would lead naturally to human reliability analysis (HRA). Indeed, perhaps the earliest work in this field applied HRA techniques to construct fault trees for aircraft structural inspection (Lock and Strutt 1985). The HRA tradition lists task steps, such as expanded versions of the generic functions above, lists possible errors for each step, then compiles performance shaping factors for each error. Such an approach was tried early in the FAA's human factors initiative (Drury et al. 1990) but was ultimately seen as difficult to use because of the sheer number of possible errors and PSFs. It is occasionally revised, such as in the current FRANCIE project (Haney 1999), using a much expanded framework that incorporates inspection as one of a number of possible maintenance tasks. Other attempts have been made to apply some of the richer human error models (e.g., Reason 1990; Hollnagel 1997; Rouse 1985) to inspection activities (Latorella and Drury 1992; Prabhu and Drury 1992; Latorella and Prabhu 2000) to inspection tasks. These have given a broader understanding of the possible errors but have not helped better define the PoD curve needed to ensure continuing airworthiness of the civil air fleet.

One outcome of this work has been a detailed analysis of a single process to develop recommendations of human factors good practices for use by industry managers and inspectors (Drury 1999). The project was performed for the regulatory body, the FAA, as part of its response to recommendations arising from investigation of the engine hub failure at Pensacola in 1997. That failure was due in part to failure of the fluorescent penetrant inspection (FPI) system for titanium hubs. Hence a human factors analysis of engine FPI systems was performed.

The human inspection models of visual search and human decision making were shown to be particularly applicable leading to a task analytic framework using hierarchical task analysis (HTA). In this way, human factors knowledge could be applied systematically to observed FPI processes. Visits were made to several engine repair facilities owned by major air carriers and engine manufacturers. The site visit team consisted of a human factors engineer specializing in inspection reliability and two senior Federal Aviation Administration (FAA) personnel specializing in NDI.

At each visit the HTA model of FPI was further developed and both good and poor human factors practices were noted. The HTA had seven major tasks, most of which would be classified under Setup in our five-function model. The tasks are shown in a HTA format in Figure 4. Note that Load/transport part in FPI would be equivalent to the Present function, and Read part for defects to Search, Decision, and Respond. For each of these seven tasks, two levels of subtasks were developed and potential errors listed. A fault tree for the primary failure, defect not reported, was developed and used to help generate a set of human factors good practices. Table 5 shows an example of the most detailed level for function 3, Apply penetrant, with the listing of human factors issues and process variances.

Using this process, a set of 90 human factors good practices was generated, with each good practice keyed to one of the seven major tasks listed above. Each was also keyed to the potential errors that that good practice can prevent. In this way, users are given the reasons for the recom-

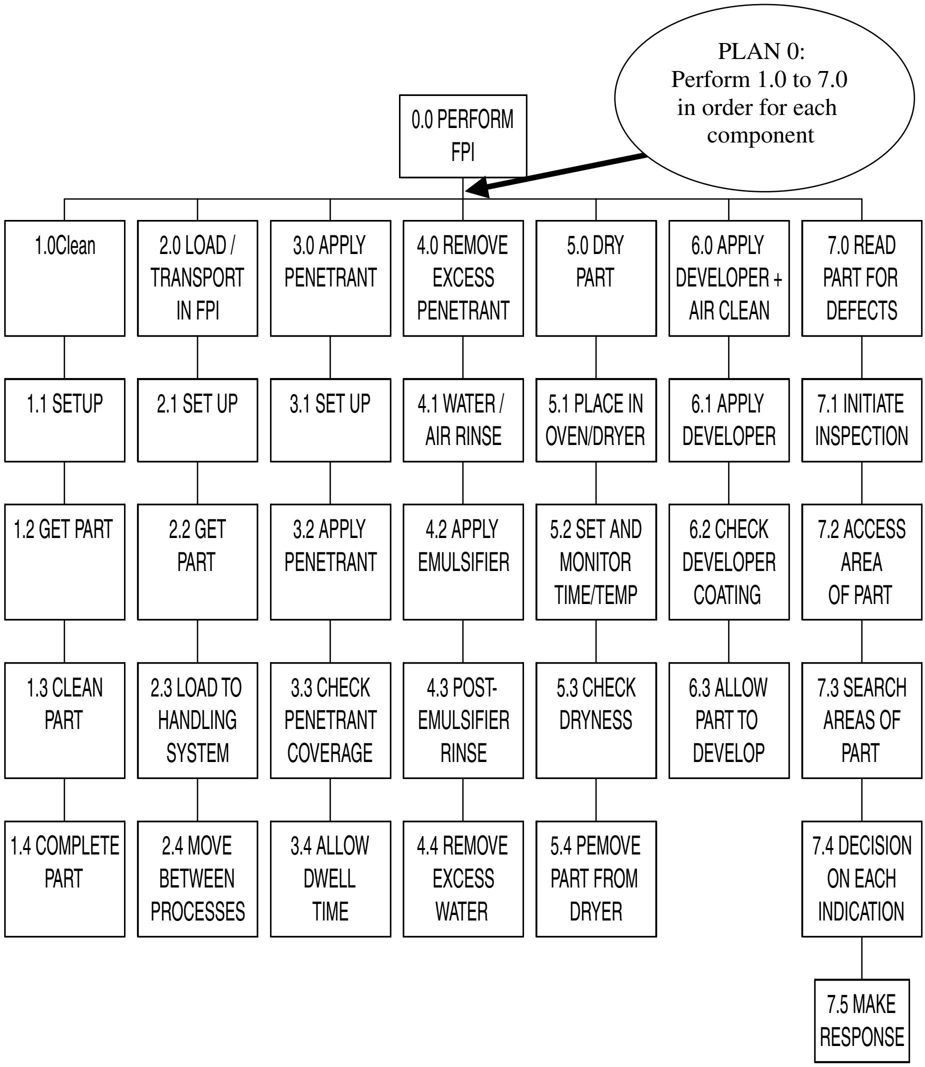

PLAN 0:
Perform 1.0 to 7.0
in order for each
component
0.0 PERFORM FPI
1.0Clean
2.0 LOAD / TRANSPORT IN FPI
3.0 APPLY PENETRANT
4.0 REMOVE EXCESS PENETRANT
5.0 DRY PART
6.0 APPLY DEVELOPER + AIR CLEAN
7.0 READ PART FOR DEFECTS
1.1 SETUP
1.2 GET PART
1.3 CLEAN PART
1.4 COMPLETE PART
2.1 SET UP
2.2 GET PART
2.3 LOAD TO HANDLING SYSTEM
2.4 MOVE BETWEEN PROCESSES
3.1 SET UP
3.2 APPLY PENETRANT
3.3 CHECK PENETRANT COVERAGE
3.4 ALLOW DWELL TIME
4.1 WATER / AIR RINSE
4.2 APPLY EMULSIFIER
4.3 POST-EMULSIFIER RINSE
4.4 REMOVE EXCESS WATER
5.1 PLACE IN OVEN/DRYER
5.2 SET AND MONITOR TIME/TEMP
5.3 CHECK DRYNESS
5.4 PEMOVE PART FROM DRYER
6.1 APPLY DEVELOPER
6.2 CHECK DEVELOPER COATING
6.3 ALLOW PART TO DEVELOP
7.1 INITIATE INSPECTION
7.2 ACCESS AREA OF PART
7.3 SEARCH AREAS OF PART
7.4 DECISION ON EACH INDICATION
7.5 MAKE RESPONSE

Figure 4 Hierarchical Task Analysis for the Top Level of Fluorescent Penetrant Inspection.

TABLE 5 Detailed Level of HTA for Function 3: Apply Penetrant

Task Step	Task Description	Task Analysis
3.1. Set-up	3.1.1. Monitor penetrant type, consistency (for electrostatic spray) or concentration, chemistry, temperature, level (for tank)	Are measurements conveniently available? Are measurement instruments well human-engineered? Do recording systems require quantitative reading or pass/fail?
3.2. Apply		
3.2.1. Electrostatic spray	3.2.1.1. Choose correct spray gun, water-washable or postemulsifiable penetrants available.	Are spray guns clearly differentiable? Can feeds be cross-connected?
	3.2.1.2. Apply penetrant to all surfaces.	Can sprayer reach all surfaces?
3.2.2. Tank	3.2.2.1. Choose correct tank, water washable or postemulsifiable penetrants available.	Are tanks clearly labeled? Is handling system well-designed to use for part placement?
	3.2.2.2. Place in tank for correct time, agitating/turning as needed.	Does operator know when to agitate/turn? Does carrier interfere with application?
	3.2.2.3. Remove from tank to allow to drain for specified time.	Is drain area available?
3.2.3. Spot	3.2.3.1. Choose correct penetrant, water-washable or postemulsifiable penetrants available.	Are spot containers clearly differentiable? Does operator know which areas to apply penetrant to?
	3.2.3.2. Apply to specified areas with brush or spray can.	Can operator reach all areas with spray can/brush? Is handling systems well human-engineered at all transfer stages?
3.3. Check coverage	3.3.1. Visually check that penetrant covers all surfaces, including holes.	Can operator see penetrant coverage? Is UV light/white light ratio appropriate?
	3.3.2. Return to 3.2 if not complete coverage.	Can operator see all of part? Can handling system back up to reapplication?
3.4. Dwell time	3.4.1. Determine dwell time for part.	Does operator know correct dwell time?
	3.4.2. Allow penetrant to remain on part for specified time.	How is it displayed? Are production pressures interfering with dwell time? Is timer conveniently available, or error-proof computer control?
	Errors/Variances for 3.0 Apply Penetrant	
	Process measurements not taken	
	Process measurements wrong	
	Wrong penetrant applied	
	Wrong time in penetrant	
	Insufficient penetrant coverage	
	Penetrant applied to wrong spots	
	No check on penetrant coverage	
	Dwell time limits not met	

mendations so that they can develop a knowledge base in addition to the rule-based good practices. In addition, there were five general control mechanisms where applications of human factors principles should lead to improved FPI reliability: Operator selection, training and turnover, Hardware design, Software and job aids, Interpersonal systems design, and Environmental control. Each was considered to show how human factors could be applied at a higher level than the 90 specific recommendations. Finally, the analysis established a set of research and development needs to provide better support for improved FPI reliability: Improved solvent and developer, Better magnifying loupe, Better process test panel validity, Job aids for search strategy, and Realistic expectation control.

It was concluded that the FPI system for critical rotating engine components can reach significantly higher reliability through application of specific and general human factors good practices. For the current purposes, the analysis provides a worked example of how human factors knowledge can be applied to find practical changes in an existing inspection system.

7. LOGICAL FUNCTION ALLOCATION IN TEST AND INSPECTION

Given that test and inspection systems must be designed and redesigned on an accelerated cycle, systematic design procedures would be advantageous to modern manufacturing. With the array of devices and human factors interventions available to the designer, a procedure that highlights functional and organizational differences between alternative solutions would assist in the evaluation or convergence phase of design (Jones 1981, p. 68). In Section 3, the functions of inspection were defined and the principles of allocation of function between human and machine proposed (Section 3.1).

Function allocation is a basic technique of human factors engineering, although it has periodically been called into question. Most recently, McCarthy et al. (2000) edited a special issue of a journal devoted to a modern consideration of function allocation in manufacturing. Although test and inspection are not mentioned, the issues raised clearly apply to these manufacturing functions. They argue that function allocation at the design stage is against the principles of localized design and is often rational and atomistic, both currently disfavored concepts in the field. They see function allocation as being at times arrogant instead of growing more naturally out of each particular operating environment. However, in any design there is always allocation of function, even if it is only implicit or a default to automation. Human operators still have too many jobs composed of leftover functions in a system designed for full automation but failing to achieve that in practice.

In test and evaluation, much of the automation literature is frankly antihuman operator. For example, Lumai (1994) states that "the subjective nature of human inspectors prevents achieving a 'six-sigma' manufacturing capability" going on to give the argument that if inspectors only achieve a 90% success rate, this will not meet the six-sigma requirement. In fact, if the defective rate is low enough, even a 90% hit rate can achieve arbitrarily low final defect rates. But 100% inspection by humans is rarely the issue in industry. Most research papers, as noted earlier, start by justifying automation based on human error and/or human variability. The implication is that automation will prevent errors and remove unwanted variability. Much of modern manufacturing is based on continual variability reduction (e.g., total quality management), so this aim is reasonable. There is no justification for variability for its own sake. But variability of *response* may well be required in a practical system to allow the system to cope with unexpected variance in inputs or process conditions. Human inspectors provide the required ability to respond to external change, for example by finding defects not initially defined for the process. In fact, one characteristic of automated systems is that they function best when inspecting for a limited set of defect types, often a single type. Humans share this characteristic but can expand their effective set of detect types much more easily than automated devices.

Three examples of direct tests of function allocation have been reported. The first, Drury and Sinclair (1983), compared two alternative designs for inspection of precision roller bearings for discrete surface imperfections (scratches, toolmarks, nicks, and dents). The current manual method using trained inspectors was compared to a computer-based automated visual inspection system (AVIS). In fact, the systems differed on the functions of Setup, Search, and Decision. The latter two functions could not be separated for the AVIS, indicating one difficulty of performing allocation at the function level. Results, in terms of misses and false alarms, were poor for both systems, with the automated system slightly worse than the human inspectors. In fact, the company embarked upon a very effective retraining program for its inspectors (Kleiner 1986), while redesigning the AVIS for better discrimination between conforming and nonconforming indications.

The second example, Drury and Goonetilleke (1992), used an inspection device for populated printed circuit boards (called a CVC) that could be reconfigured to allocate functions progressively between human and device. Direct comparisons of performance (misses, false alarms, performance speed) were possible between six levels of automation from level 1 (fully manual) to level 6 (automated Setup, Present, and Search), with only the final decision left to the inspector. Response via a control panel was kept constant throughout. Both laboratory and in-plant studies were undertaken, but only the former are presented here. Figure 5 shows performance for a number of measures of defect-detection performance that gave significant effects. Note that performance generally falls with increasing automation until an automated search procedure that detects all indications is reached. Performance differences between the fully manual and fully automated systems were very small.

The final example is from Hou et al. (1993), who examined five function-allocation alternatives for the task of inspecting circuit board height maps for a set of three defects: missing component, misaligned component, and wrong-sized component. The five function allocations were based on the fact that humans are expected to perform worse than computer-based automation on the search function but may be better for decision making. Thus, the extreme allocations of all tasks to the human and all tasks to the machine were complemented by an allocation of machine search/human decision and another of machine search/shared decision. For automated inspection both a template matching system and a neural net system were used. Each of the five systems was optimized, for example by finding the optimum human search time (Morawski et al. 1992) for human search or by choosing parameters for the automated systems that maximized performance. All five systems were tested using four human inspectors for a range of three circuit board sizes and three different image contrast levels to simulate environmental degradation. Results were measured by the set of performance measures [*p*(miss), *p*(false alarm), time for inspection], with the two probabilities being combined using signat-detection theory into the single nonparametric measure of discriminability, *p*(*A*). There were significant differences between the five systems on *p*(*A*), with the two fully automated systems giving the worst discriminability. When inspection time was factored into the evaluation, the graph in Figure 6 could be drawn. This shows the speed–accuracy trade-off (SATO) across the five systems. Optimum performance is in the upper left corner of the figure with maximum discriminability in minimum time. The contours represent potential cost trade-offs for speed and accuracy, shown as straight lines for simplicity. From this figure, it can be seen that the ultimate choice of system depends upon the exact SATO cost weighting, but some systems dominate others no matter what the weightings. For example, template automation dominates neural net automation. Also, the human decision system dominates both the shared decision and the totally human systems. The conclusion from this study was that some form of human/automation hybrid system is preferable to either total allocation to human or total allocation to automation. It was also found that the hybrid systems were more consistent across different levels of environmental degradation.

Note that the set of criteria for deciding on function allocation was different for the three studies. The first used only the error probabilities, the second expanded this set to include performance measures such as misclassification rate, while the third also included performance speed. The issue of flexibility was addressed in the third study by comparing the competing systems across different contrast levels and board sizes. In none of these examples was there any measure of the cost of

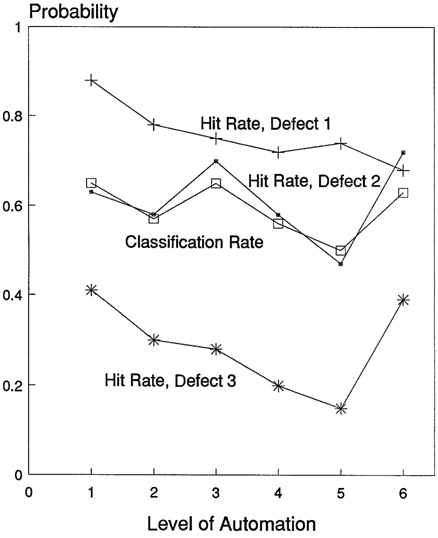


Figure 5 Performance with Increasing Automation.

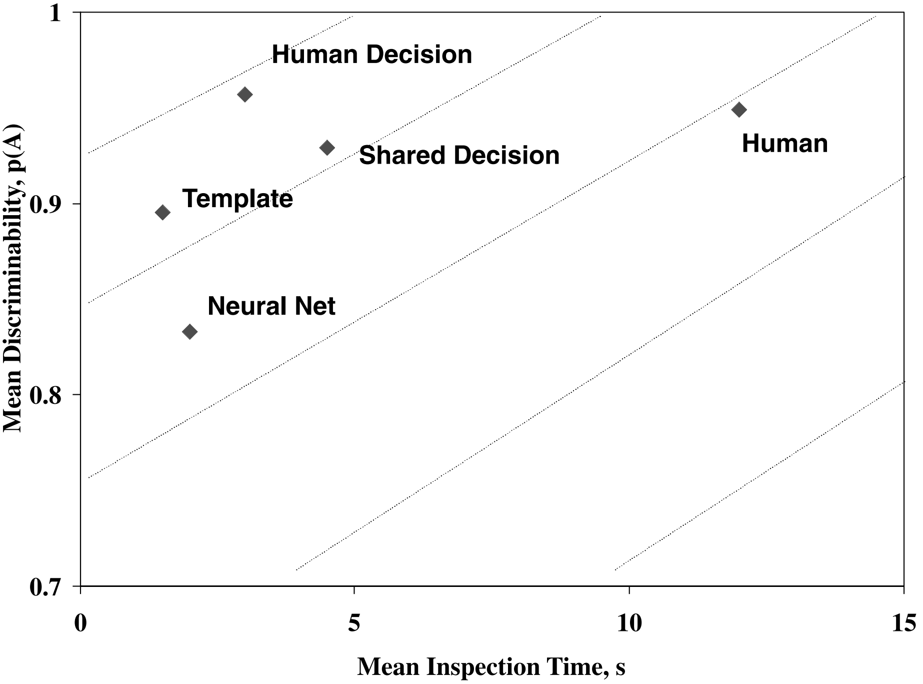


Figure 6 Accuracy and Speed Measures for the Five Alternative Function Allocations of Hou et al. (1993). The Dotted Lines Represent Typical Isocost Contours.

performance to the human inspector, although another part of the second study did use measures of workload and stress. The final choice of allocation is dependent on the set of criteria used to compare the systems.

7.1. A Methodology for Test and Inspection Systems Design

The implication that there are five functions in Table 3 that can each be allocated to either component is perhaps too simplistic. Not only are there practical interactions between functions, but there are multiple potential solutions to the design problem represented by each function. The best function allocation may well change over time, meaning that a dynamic rather than a static allocation procedure is needed. One approach to the design problem is the design for inspectability procedure proposed by Drury (1992b), although this was concerned primarily with the design of a product to enhance its inspectability. Here the procedure is further developed to examine different design alternatives.

Note that the outcome errors for test and inspection were defined as misses, false alarms, and (to a lesser extent) delays. Any system must attempt to minimize some function of these errors, although it has been shown that they can have trade-offs. For example, in the Search function, allowing greater time per item (increased delay) reduces misses. Also, in Decision, variation of the criterion for reporting (S_i) has the effect of causing misses and false alarms to covary inversely. Thus, compromises among all three measures will usually be required. Because the functions are sequential, failure to detect a defect (miss) is usually a much more common error than a false alarm (e.g. Megaw 1992).

If the probability of search success is p_s and the probability of decision success is p_d (here for simplicity both are assumed constant for good and faulty items):

$$p\ (\text{miss}) = p_s\,(1 - p_d) + (1 - p_s)$$

$$p\ (\text{false alarm}) = p_s\,(1 - p_d)$$

Thus,

$$p \text{ (false alarm)} < p \text{ (miss) for}$$
$$0 < p_s < 1$$
$$0 < p_d < 1$$

In addition to being numerically the most important error, misses are usually the most expensive in consequences because they imply failure of the system in use. False alarms, while still costly, can often be reevaluated by slower, off-line testing and usually corrected.

Thus, while the complete set of performance measures [p(miss), p(false alarm), time for inspection] is required to evaluate competing systems and systems components, for design purposes the detection of nonconforming items must be the priority. This is particularly so in modern manufacturing environments, where nonconforming items are increasingly rare and increasingly important to detect early. Thus, rather than optimizing the set of all three measures, in practice it is often imperative to optimize (minimize) p(miss) subject to fixed limits or constraints on p(false alarm) and time for inspection.

The design for inspectability procedure noted above starts from the assumption that a list of defects to be detected can be developed at the product design stage, perhaps from a failure modes and effect analysis (FMEA) of the product's components. These defects (D_1, D_2, . . . , D_i, . . . , D_n) represents the challenges to the test and inspection system because all must be detected, and detected at the appropriate stage of manufacture. Only three components of the inspection system can be changed, or changes can affect:

Product: the item produced and its subcomponents

Process: each test or inspection process used on the item and its subcomponents

Person: the human or humans interfacing with the product or process at each test or inspection

A design procedure must find an optimum mix of product, process, and person design features for each function to ensure that the probability of detection is maximized, subject to false alarm and time constraints. Note that the function-allocation procedure has been transformed into a function design feature, with design suggestions for each function (or even group of functions) being evaluated for how it impacts product, process, and person design and how it affects the probabilities of detection of each defect. Table 6 summarizes this design procedure, with each row representing a design alternative for one function and containing design impacts on both system components (product, process, people) and detection probabilities for the set of all defects. Entries in the System Design Impacts columns are references to notes concerning, for example, impact upon training or product weight. Entries in System Performance Impacts are detection probabilities for each defect or, where these are not available, in indication of the direction of change from a reference test and inspection system, perhaps the current system. Symbols of +, −, or 0 would be suitable in such a procedure, with actual magnitudes included if available.

TABLE 6 Design Procedure for Test and Inspection

Function	Design Alternative	System Design Impacts						
		Product	Process	Person	D_1,	D_2, . . .	D_i, . . .	D_n,
Setup	SU Alternative 1 SU Alternative 2 —							
Present	*P* Alternative 1 *P* Alternative 2 —							
Search	*S* Alternative 1 *S* Alternative 2 —							
Decision	*D* Alternative 1 *D* Alternative 2 —							
Response	*R* Alternative 1 *R* Alternative 2 —							

In many automated inspection systems reported in the literature, there is a concentration on a single performance measure, which makes it difficult to handle allocation decisions. For example, a vision system may be quoted as being able to detect surface defects "in the 20 μm range" without saying what the false alarm rate would be in those conditions. As Drury and Sinclair (1983) have shown, it is possible to change the ratio of misses to false alarms over the complete range by varying the sensitivity or threshold of an automated system. Most papers on human inspection quote at least the miss and false alarm rates and often the time per item. For automated systems, this is not typically true, making direct comparisons currently difficult.

8. CONCLUSIONS ON TEST AND INSPECTION

While much ground has necessarily been covered in this review, certain rather general and simple conclusions are possible. First, there is indeed a role for test and inspection in modern business. Global pressures are likely to increase the quality requirements over time while simultaneously requiring efficient performance with minimal resources. Give that test and inspection will still be needed, their role within the organization is moving from checking output for customer protection to collecting process data for point of manufacture control. This contextual change should in fact benefit human operators by including an explicit test and inspection function within most jobs. In the context of automation, we now have a plethora of sensors and signal processors to choose from, with capability increasing while cost decreases over time. Our choice of human and automation roles is likely to require hybrid systems, at least for image data where pattern recognition is required. Having said this, there appears to be a general conclusion that automation is preferable for the search function, at least where defects can be specified in advance. For routine measurement, such as of forces or temperatures, human roles are much reduced. Fortunately, for such measurements the ultimate goal is process control, where humans have a definite role in interpretation of patterns of results and choice of remedial action.

REFERENCES

Bainbridge, L. (1990), "Verbal protocal analysis," in *Evaluation of Human Work,* J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London, pp. 161–179.

Baveja, A., Drury, C. G., Karwan, M., and Malone, D. M. (1996), "Derivation and Test of an Optimum Overlapping-Lobe Model of Visual Search," *IEEE Tranactions on Systems, Man, and Cybernetics,* Vol, 28, pp. 161–168.

Bobo, S. (1989), "Communication and Transfer of Nondestructive Inspection Information," in *Human Factors Issues in Aircraft Maintenance and Inspection—Information Exchange and Communications, Second Federal Aviation Administration Meeting,* Washington, DC.

Boucher, T. O. (1996), *Computer Automation in Manufacturing: An Introduction,* Chapman & Hall, London, pp. 92–132.

Budros, A. (1999). "A Conceptual Framework for Analyzing Why Organizations Downsize, *Organizational Science,* Vol. 10, No. 1, pp. 69–82

Buffa, M. (1985), "Process Control for Automatic Component Assembly of Surface Mounted Devices Utilizing a Machine Vision Controller," *SME Vision '85 Conference Proceedings,* Society of Manufacturing Engineers, Dearborn, MI, pp. 5.180–5.200.

Caudill, M. (1988a), "Neural Networks Primer—Part I," *Al Expert,* Vol. 31, No. 4, pp. 53–59.

Caudill, M. (1988b), "Neural Networks Primer—Part II," *Al Expert,* Vol. 31, No. 5, pp, 53–59.

Chaney, F. B., and Teel, K. S. (1967), Improving Inspector Performance through Training and Visual Aids," *Journal of Applied Psychology,* Vol. 51, pp. 311–315.

Chi, C.-F., and Drury, C. G. (1995), "A Test of Economic Models of Stopping Policy in Visual Search," *IIE Transactions,* Vol. 27, pp, 392–393.

Cormier, D. R. (1991), "Neural Network Based Surface Mount Component Verification," Master's Thesis, State University of New York at Buffalo, Buffalo, NY.

Diehl, A. (1990), "The Effectiveness of Aeronautical Decision-Making Training," in *Proceedings of the Human Factors Society 34th Annual Meeting,* pp. 1367–1371.

Drury, C. G. (l978), "Integration of Human Factors Models into Statistical Quality Control," *Human Factors,*, Vol. 20, No. 5, pp. 56l–572.

Drury, C. G. (1991), "Ergonomics Practice in Manufacturing," *Ergonomics,*, Vol. 34, pp. 825–839.

Drury, C. G. (1992a), "Inspection Performance," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., New York, John Wiley & Sons, pp. 2282–2314.

Drury, C. G. (1992b), “Design for Inspectability,” in *Design for Manufacturability: A Systems Approach to Concurrent Engineering and Ergonomics,* M. H. Helander and M. Nagamachi, Eds., Taylor & Francis, London, pp. 204–216.

Drury, C. G. (1996), “Ergonomics Society Lecture 1996: Ergonomics and the Quality Movement,” *Ergonomics,* Vol. 40, No. 3, pp. 249–264.

Drury, C. G. (1999), “Human Factors Good Practices in Fluorescent Penetrant Inspection,” in *Human Factors in Aviation Maintenance—Phase Nine, Progress Report, DOT/FAA/AM-99/xx,* National Technical Information Service, Springfield, VA.

Drury, C. G., and Addison, J. L. (1973), “An Industrial Study of the Effects of Feedback and Fault Density on Inspection Performance,” *Ergonomics,* Vol. l6, pp. l59–l69.

Drury, C. G., and Goonetilleke, R. S. (1992), “Stress, Performance and Automation: A Field Study,” in *Proceedings of the 28th Annual Conference of the Ergonomics Society of Australia,* pp. 48–55.

Drury, C. G., and Kleiner, B. M. (1990), “Training in Industrial Environments,” in *Proceedings of the Human Factors Association of Canada Meeting* (Ottawa), pp. 99–108.

Drury, C. G., and Sinclair, M. A. (1983), “Human and Machine Performance in an Inspection Task,” *Human Factors,* Vol. 25, No. 4, pp. 391–400.

Drury, C. G., and Spencer, F. W. (1997), “Human Factors and the Reliability of Airframe Visual Inspection,” in *Proceedings of the 1997 SAE Airframe/Engine Maintenance and Repair Conference* (AEMR ’97) (August).

Drury, C. G., Prabhu, P., and Gramopadhye, A. (1990), “Task Analysis of Aircraft Inspection Activities: Methods and Findings,” in *Proceedings of the Human Factors Society 34th Annual Conference,* pp. 1181–1185.

Drury, C. G., Patel, S. C., and Prabhu, P. V. (2000), “Relative Advantage of Portable Computer-Based Workcards for Aircraft Inspection,” *International Journal of Industrial Ergonomics,* Vol. 26, No. 2, pp. 163–176.

Engel, F. L. (1971), “Visual Conspicuity, Directed Attention, and Retinal Locus,” *Visual Research,* Vol. 11, pp. 563–576.

Eriksen, C. W. (1990), “Attentional Search of the Visual Field,” in *Visual Search,* D. Brogan, Ed., Taylor & Francis, pp. 3–19.

Evans, J. R., and Lindsay, W. (1993), *The Management and Control of Quality,* West, Minneapolis-St. Paul.

Federal Aviation Administration (FAA) (1993), *National Aging Aircraft Research Program Plan,* FAA Technical Center, Atlantic City, NJ.

Friedman, T. L. (1999), *The Lexus and the Olive Tree: Understanding Globalization,* Farrar, Straus & Giroux, New York.

Gallwey, T. J., and Drury, C. G. (1986), “Task Complexity in Visual Inspection, *Human Factors,* Vol. 28, No. 5, pp. 595–606.

Gieles, I., and Venema, W. J. (1989), “Inspection of SMD’s with 3-D Laser Scanning,” *SME Vision ’89 Conference Proceedings* (Chicago), pp. 5.59–5.71.

Gramopadhye, A. K., Drury, C. G., and Prabhu, P. (1997a), “Training for Aircraft Visual Inspection,” *Human Factors and Ergonomics in Manufacturing,* Vol. 3, pp. 171–196.

Gramopadhye, A. K., Drury, C. G., and Sharit, J. (1997b), “Feedback Strategies for Visual Search in Airframe Structural Inspection,” *International Journal of Industrial Ergonomics,* Vol. 19, No. 5, pp. 333–344.

Greening, C. P. (1975), “Mathematical modeling of Air-to-Ground Target Acquisition,” *Human Factors,* Vol. 18, pp. 111–147.

Hackman, J. R. (1990), *Groups That Work,* Jossey-Bass, San Francisco.

Hancock, W. M., Sathe, P., and Edosomwan, J. A. (1992), “Quality Assurance,” in *Handbook of Industrial Engineering,* 2nd Ed., G. Savendy, Ed., John Wiley & Sons, New York, pp. 2221–2234.

Haney, L. (1999), “Framework Assessing Notorious Contributing Influences for Error (FRANCIE)” in *Proceedings of International Workshop on Human Factors in Space* (Tokyo).

Harris, D. H., and Chaney, F. B. (1969), *Human Factors in Quality Assurance,* John Wiley & Sons, New York.

Holding, D. H. (1981), *Human Skills,* John Wiley & Sons, New York.

Hollnagel, E. (1989), “The Phenotype of Erroneous Actions: Implications for HCI Design,” in *Human-Computer Interaction and Complex Systems,* G. R. S. Weir and J. L. Alty, Eds., Academic Press, London.

Hollnagel, E., Ed. (1997), *CREAM—Cognitive Reliability and Error Analysis Method,* Elsevier Science, New York.

Hou, T.-S., Lin, L., and Drury, C. G. (1993), "An Empirical Study of Hybrid Inspection Systems and Allocation of Inspection Function," *International Journal of Human Factors in Manufacturing,* Vol. 3, pp. 351–367.

Hovey, P. W., and Berens, A. P. (1988), "Statistical Evaluation of NDE Reliability in the Aerospace Industry," in *Review of Progress in Quantitative Nondestructive Evaulation,* Vol. 7B, D. D. Thompson and D. E. Chimenti, Eds., Plenum Press, New York, pp. 1761–1768.

Hung, Y. Y., and Park, B. G. (1996), "Automated Inspection and Quantification of Dents in Automotive Bodies Using 3-D Computer Vision Technique," SME Technical Paper IQ, Society of Manufacturing Engineers, Dearborn, MI.

Jamieson, G. H. (1966), "Inspection in the Telecommunications Industry: A Field Study of Age and Other Performance Variables," *Ergonomics,* Vol. 9, pp. 297–303.

Johnson, W. B. (1990), "Application of New Technology for Aviation Maintenance Training: An Industry Status Report," in *Proceedings of the Human Factors Society 34th Annual Meeting,* pp. 1171–1175.

Jones, J. C. (1981), *Design Methods: Seeds of Human Factors,* John Wiley & Sons, New York.

Kälviäinen, H., Kukkonen, S., Parkinnen, J., Hyvärinen, T. (1998), "Quality Control in Tile Production," *Proceedings of SPIE,* Vol. 3522, pp. 355–360.

Kanter, E. M. (1995), *World Class: Thriving Locally in the Global Economy,* Simon & Schuster, New York.

Kanter, E. M. (1999), "Challenges in the Global Economy," *The Washington Quarterly,* Vol. 22. No. 2, pp. 39–58.

Karwowski, W., Kosiba, E., Benabdallah, S., and Salvendy, G. (1990), "A Framework for Development of Fuzzy GOMS Model for Human Computer Interaction," *International Journal of Human–Computer Interaction,* Vol. 2, pp. 287–307.

Kelly, M. L. (1955), "A Study of Industrial Inspection by the Method of Paired Comparisons," *Psychological Monographs,* Vol. 69(9) (No. 394), p. 16.

Knight, J. L., and Salvendy, G. (1992), "Psychomotor Work Capabilities," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 978–1004.

Komatsu, K., Omori, T., Kitamura, T., Nakajima, Y., Aiyer, A., and Suwa, K. (1999), "Automatic Macro Inspection System," in *Proceedings of SPIE,* Vol. 3677, pp. 764–771

Komorowski, J. P., Simpson, D. L., and Gould, R. W. (1991), "Enhanced Visual Technique for Rapid Inspection of Aircraft Structures (D. Sight)," *Materials Evaluation,* Vol. 49, pp. 1486–1490.

Kundel, H. S., and Lafollette, B. S. (1972), "Visual Search Patterns and Experience with Radiological Images," *Radiology,* Vol. 103, pp. 523–528.

Latorella, K., and Drury, C. G. (1992), "A Framework for Human Reliability in Aircraft Inspection," in *Proceedings of the 7th FAA Meeting on Human Factors Issues in Aircraft Maintenance and Inspection,* Federal Aviation Administration, Washington, DC.

Latorella, K., and Prabhu, P. (1998), "Review of Human Error in Aviation Maintenance and Inspection," *International Journal of Industrial Engineering,* Vol. 26, No. 2, pp. 133–161.

Legeard, D., Marty-Mahe, P., Camillerapp, J., Marchal, P., and Leredde, C. (1999), "Real-Time Quality Evaluation of Pork Hams by Color Machine Vision," in *Proceedings of SPIE,* Vol. 3652, pp. 138–149.

Liu, Y., Li, X., Ren, D., Ye, S., Wang, B., and Sun, J. (1998), "Computer Vision Application for Weld Defect Detection and Evaluation," in *Proceedings of SPIE,* Vol. 3558, pp. 354–357.

Lloyd, C. J., Boyce, P., Ferzacca, N., Eklund, N., and He, Y. (2000), "Paint Inspection Lighting: Optimization of Lamp Width and Spacing," *Journal of the Illuminating Engineering Society,* Vol. 28, pp. 99–102

Lock, M. W. B., and Strutt, J. E. (1985), "Reliability of In-Service Inspection of Transport Aircraft Structures," CAA Paper 85013, Civil Aviation Authority, London.

Lumai, R. (1994), "Image Processing in Manufacturing," in *Handbook of Design, Manufacturing and Automation,* R. C. Dorf and A. Kusiak, Eds., John Wiley & Sons, New York.

Maher, J., Overbach, W., Palmer, G., and Piersol, D. (1970), "Enriched Jobs Improve Inspection Performance," *Work Study and Management Services,* October, pp. 821–824.

McCarthy, J. C., Fallon, E., and Bannon, L. (2000), "Dialogues on Function Allocation," *International Journal of Human Computer Studies,* Vol. 52, No. 2, pp. 191–202.

McKenzie, R. M. (1958), "On the Accuracy of Inspectors," *Ergonomics,* Vol. 1, No. 3, pp. 258–272.

McNichol, D. (1972), *A Primer of Signal Detection Theory,* Allen & Unwin, London.

Megaw, E. D. (1992), *Contemporary Ergonomics,* Taylor & Francis, London.

Monden, Y. (1992), "Just-in-Time Production System," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, pp. 2116–2130.

Moraleda, J., Ollero, A., and Orte, M. (1999), "Robotic System for Internal Inspection of Water Pipes," *IEEE Robotics and Automation Magazine,* Vol. 6, pp. 30–41.

Morawski, T., Drury, C. G., and Karwan, M. H. (l980), "Predicting Search Performance for Multiple Targets," *Human Factors,* Vol. 22, No. 6, pp. 707–7l8.

Morawski, T. B., Drury, C. G., and Karwan, M. H. (1992), "The Optimum Speed of Visual Inspection Using a Random Search Strategy," *IIE Transactions,* Vol. 24, pp. 122–133.

Moray, N., Lootseen, P., and Pajak, J. (1986), "The Acquisition of Process Control Skills," *IEEE Transactions,* Vol. SMC-16, pp. 497–505.

Murgatroyd, R. A., Worrall, G. M., and Waites, C. (1994), "A Study of the Human Factors Influencing the Reliability of Aircraft Inspection," AEA/TSD/0173, AEA Technology, Culham, UK.

National Research Council (1997), *Enhancing Organizational Performance,* National Academy Press, Washington, DC.

Nurani, R. K., and Akella, R. (1996), "In-line Defect Sampling Methodology in Yield Management: An Integrated Framework," *IEEE Transactions on Semiconductor Manufacturing,* Vol. 9, No. 4, pp. 506–518.

O'Brien, T. G., and Charlton, S. G. (1997), *Handbook of Human Factors Testing and Evaluation,* Erlbaum, Mahwah, NJ.

Overington, I. (1973), *Vision and Acquisition,* Pentech Press, London.

Prabhu, P., and Drury, C. G. (1992), "A Framework for the Design of the Aircraft Inspection Information Environment," in *Proceedings of the 7th FAA Meeting on Human Factors Issues in Aircraft Maintenance and Inspection,* Federal Aviation Administration, Washington, DC.

Rasmussen, J. (1983), "Skills, Rules and Knowledge: Signals, Signs and Symbols, and Other Distinctions in Human Performance Models," *IEEE Transactions on Systems Man and Cybernetics,* Vol. SME-13, No. 3, pp. 257–266.

Reason, J. (1990), *Human Error,* Cambridge University Press, Cambridge.

Rifkin, J. (1995), *The End of Work,* Putnam, New York.

Rogers, A. (1991), "Organizational Factors in the Enhancement of Military Aviation Maintenance," in *Proceedings of the Fourth International Symposium on Aircraft Maintenance and Inspection,* Federal Aviation Administration, Washington, DC, pp. 43–63.

Rosemau, R. D., Nawaz, S., Niu, A., and Wee, W. G. (1999), "Aircraft Engine Blade Colling Holes Detection and Classification from Infrared Images," *Proceedings of SPIE,* Vol. 3586, pp. 85–93.

Rouse, W. B. (1985), "Optimal Allocation of System Development Resources to Reduce and/or Tolerate Human Error," *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-15, No. 5, pp. 620–630.

Rummel, W. D., Hardy, G. L., and Cooper, T. D. (1989), "Applications of NDE Reliability to Systems," *Metals Handbook,* Vol. 17, pp. 674–688.

Saaty, T. L. (1974), "Measuring the Fuzziness of Sets," *Journal of Cybernetics,* Vol. 4, pp. 53–61.

Saaty, T. L. (1977), "A Scaling Method for Priorities in Hierarchical Structures," *Journal of Mathematical Psychology,* Vol. 15, pp. 234–280.

Salvendy, G., and Seymour, J. W. D. (1973), *Prediction and Development of Industrial Work Performance,* John Wiley & Sons, New York.

Schor, J. B. (1991), *The Overworked American,* Basic Books, New York.

Schoonhard, J. W., Gould, J. D., and Miller, L. A. (1973), "Studies of Visual Inspection," *Ergonomics,* Vol. 16, pp. 365–379.

Sheehan, J. J., and Drury, C. G. (l97l), "The Analysis of Industrial Inspection," *Applied Ergonomics,* Vol. 2, No. 2, pp. 74–78.

Sheridan, J. J. (1987), "Supervisory Control," in *Handbook of Human Factors,* G. Salvendy, Ed., John Wiley & Sons, New York, pp. 1243–1268.

Shindo, W., Wang, E. H., Akella, R., Strojwas, A. J., Tomlinson, W., and Bartholomew, R. (1999), "Effective Excursion Detection by Defect Type Grouping in In-line Inspection and Classification," *IEEE Transactions on Semiconductor Manufacturing,* Vol. 12, No. 1, pp. 3–10.

Siegel, M. Gunatilake, P., and Podnar, G. (1998), "Robotic Assistants for Aircraft Inspectors," *Industrial Robot,* Vol. 25, pp. 389–400.

Siemieniuch, C. E., and Sinclair, M. A. (1999), ''Knowledge Lifecycle Management in Manufacturing Organizations,'' in *Contemporary Ergonomics,* M. A. Hanson, E. J. Lovesey and S. A. Robertson, Eds., Taylor & Francis, London, pp. 322–332.

Sinclair, M. A. (1984), ''Ergonomics of Quality Control.'' Workshop document, *International Conference on Occupational Ergonomics,* Toronto.

Singleton, W. T. (1972), *Man–Machine Systems,* Penguin, Harmondsworth.

Spencer, F., and Schurman, D. (1995), ''Reliability Assessment at Airline Inspection Facilities; Volume III: Results of an Eddy Current Inspection Reliability Experiment,'' *DOT/FAA/CT-92/12,* FAA Technical Center, Atlantic City, NJ.

Steinmetz, V., and Delwiche, M. (1993), ''Machine Vision Techniques for Rose Grading,'' *Proceedings of SPIE Conference on Vision, Sensors, and Control for Automated Manufacturing Systems,* SPIE, Bellingham, WA, pp. 32–43

Stok, T. (1965), *The Worker and Quality Control,* University of Michigan Press, Ann Arbor.

Svetkoff, D., Rohrer, D., Doss, B., Kelley, R., and Jakimcius, A. (1989), ''A High Speed 3-D Imager for Inspection and Measurement of Miniature Industrial Parts,'' in *SME Vision '89 Conference Proceedings,* (Chicago), pp. 10.1–10.9.

Taguchi, G., and Wu, Y. (1979), *Introduction of Off-Line Quality Control,* Meieki Nakamura-Ku Magaya, Japan.

Taylor, J. C. (1991), ''Maintenance Organization,'' in *Human Factors in Aviation Maintenance Phase 1: Progress Report,* DOT/FAA/AM-91/16, Office of Aviation Medicine, National Technical Information Service, Springfield, VA, pp. 15–43.

Taylor, J. C., and Felten, D. F. (1993), ''Designing for Purpose: Understanding the Business We're In,'' in *Performance by Design,* J. C. Taylor, and D. F. Felten, Eds., Prentice Hall, Upper Saddle River, NJ, pp. 35–50.

Thomas, L. F., and Seaborne, A. E. M. (1961), ''The Sociotechnical Context of Industrial Inspection,'' *Occupational Psychology,* Vol. 35, pp. 36–43.

Thornton, D. C., and Matthews, M. (1982), ''An examination of the Effects of a Simple Task Alternation in Quality Control of Cookware,'' in *Proceedings of Human Factors Association of Canada/Association Canadien d'ergonomie* (*HFAC/ACE*) (Toronto), pp. 134–137.

Umbers, I. G. (1981), ''A Study of Control Skills in an Industrial Task, and in a Simulation, Using the Verbal Protocol Technique,'' *Ergonomics,* Vol. 24, pp. 275–293.

Vardeman, S. V., and Jobe, J. M. (1998), *Statistical Quality Assurance Methods for Engineers,* John Wiley & Sons, New York.

Ventura, J. A., and Chen, J-M. (1994), ''Automated CAD-based Vision Inspection,'' in *Handbook of Design, Manufacturing and Automation,* R. C. Dorf and A. Kusiak, Eds., John Wiley & Sons, New York.

Wang, M. J., and Drury, C. G. (1989), A Method of Evaluating Inspector's Performance Differences and Job Requirements,'' *Applied Ergonomics,* Vol. 20, No. 3, pp. 181–190.

Wang, M-J., Drury, C. G., and Sharit, J. (1986), ''An Application of Fuzzy Set Theory for Evaluation of Human Performance on an Inspection Task,'' in *Applications of Fuzzy Set Theory in Human Factors,* W. Karwowski and A. Mital, Eds., Elsevier, Amsterdam, pp. 257–268.

Watson, S. R., Weiss, J. J., and Donnell, M. L. (1979), ''Fuzzy Decision Analysis,'' *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 19, No. 1, pp. 1–9.

Wenner, C. (1999), ''The Role of Instructions in the Aircraft Maintenance Domain,'' Master's Thesis, State University of New York at Buffalo.

Wenner, C. L, Wenner, F., Drury, C. G., and Spencer, F. (1997), ''Beyond 'Hits' and 'Misses': Evaluating Inspection Performance of Regional Airline Inspectors,'' *Proceedings of the 41st Annual Human Factors and Ergonomics Society Meeting* (Albuquerque, NM), pp. 579–583.

Whitman, M. V. (1999), ''Global Competition and the Changing Role of the American Corporation,'' *The Washington Quarterly, Vol. 22, No. 2, pp. 59–82.*

Wickens, C. D. (1991), *Engineering Psychology and Human Performance,* 2nd Ed., Harper Collins, New York.

Zadeh, L. A. (1965), ''Fuzzy Sets,'' *Information and Control,* Vol. 8, pp. 338–353.

CHAPTER 72

Reliability and Maintainability

Kailash C. Kapur
The University of Washington

1. INTRODUCTION 1922
- 1.1. Reliability and Maintainability Activities during System Life Cycle 1923
  - 1.1.1. Step 1: The Need 1923
  - 1.1.2. Step 2: Goals and Definitions 1923
  - 1.1.3. Step 3: Concept and Program Planning 1924
  - 1.1.4. Step 4: Product Assurance Activities 1925
  - 1.1.5. Step 5: Design 1925
  - 1.1.6. Step 6: Prototype and Development 1925
  - 1.1.7. Step 7: Production 1925
  - 1.1.8. Step 8: Field and Customer Use 1925
  - 1.1.9. Step 9: System Evaluation 1925
  - 1.1.10. Step 10: Continuous Feedback 1925
- 1.2. Reliability and Life Characteristic Curve 1925
  - 1.2.1. Infant Mortality Period 1925
  - 1.2.2. Useful Life Period 1927
  - 1.2.3. Wear-Out Period 1927

2. RELIABILITY MEASURES 1927
- 2.1. Mathematics of Reliability Measures 1928
- 2.2. Various Life Distributions 1930
  - 2.2.1. Exponential Distribution 1930
  - 2.2.2. Normal Distribution 1931
  - 2.2.3. Lognormal Distribution 1931
  - 2.2.4. Weibull Distribution (Three Parameters $\theta > \delta$) 1931
  - 2.2.5. Gamma Distribution 1932
- 2.3. Summary 1932

3. SYSTEM RELIABILITY MODELS 1932
- 3.1. Reliability Block Diagram 1933
  - 3.1.1. Series Configuration 1933
  - 3.1.2. Parallel Configuration 1935
  - 3.1.3. *k*-out-of-*n* Configuration 1935
  - 3.1.4. Coherent System 1935

4. FAULT TREE ANALYSIS [FTA] 1936

5. ALLOCATION OF RELIABILITY REQUIREMENTS 1937

6. DESIGN FOR RELIABILITY 1937
- 6.1. Design Review 1939
- 6.2. Failure Mode and Effects Analysis 1940
- 6.3. Probabilistic Approach to Design 1940

7. HUMAN FACTORS IN RELIABILITY 1941

8. RELIABILITY MEASUREMENT 1941
- 8.1. Test Programs 1942
  - 8.1.1. Design Support Tests 1942
  - 8.1.2. Design-Verification Tests 1942
  - 8.1.3. Design-Evaluation Tests 1943
  - 8.1.4. Design-Acceptance Tests 1943
  - 8.1.5. Technical Evaluation Tests 1943
  - 8.1.6. Operational Evaluation Tests 1943
  - 8.1.7. Production Acceptance Tests 1943
  - 8.1.8. Fleet/Field Surveillance Tests 1943
  - 8.1.9. Test Procedures 1943

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

8.2. Reliability Estimation 1944
8.2.1. Reliability Estimation: Exponential Distribution 1944
8.2.2. Reliability Estimation: Weibull Distribution 1945
9. MAINTAINABILITY 1946
9.1. Maintainability Measures 1946
10. AVAILABILITY 1949
10.1. Availability Measures 1949
10.2. Reliability—Maintainability—Availability Trade-Off 1950
11. RELIABILITY GROWTH 1951
11.1. Reliability Growth Models 1952
12. DESIGN AND MANAGEMENT OF RELIABILITY PROGRAMS 1953
12.1. Elements of a Reliability Program 1953
REFERENCES 1954
ADDITIONAL READING 1955

1. INTRODUCTION

The ultimate objective of any system is the performance of some intended function. This function is frequently called the *mission*. The term often used to describe the overall capability of a system to accomplish its mission is *system effectiveness* For consumer products, system effectiveness is related to customer satisfaction, which is related to the overall concept of quality. Quality of products and services is defined and evaluated by the customers and users. Similarly, system effectiveness is defined and evaluated by the customers and users of the product. System effectiveness is defined as the probability that the system can successfully meet an operational demand within a given time when operating under specified conditions. Effectiveness is influenced by the way the system is designed, manufactured, used, and maintained. Thus, the effectiveness of a system is a function of several attributes, such as design adequacy, performance measures, safety, reliability, quality, manufacturability, and maintainability. The disciplines of assurance sciences help to increase the overall effectiveness of any system. The assurance sciences are engineering disciplines that have the common objective of attaining product integrity (product does what it says it is supposed to do) (Carrubba and Gordon 1978). The term *product assurance* is also popular in many companies. This chapter is concerned with the reliability, maintainability, serviceability, and availability aspects of the product assurance system.

Reliability is one of the major attributes determining system effectiveness. It is generally defined as the probability that a given system will perform its intended function satisfactorily, for its intended life, under specified operating conditions. With this definition the obvious problems are (1) the acceptance of the probabilistic notion of reliability; (2) the problems associated with defining adequate performance, particularly for system parameters that deteriorate slowly with time; and (3) the judgment required to determine the proper statement of operating conditions. Thus, reliability is a relative measure, or in terms of probability, it is conditional probability. It is relative to (1) definition of function from the viewpoint of the customer; (2) definition of failure from the viewpoint of the customer; (3) definition of intended life; and (4) customer's operating and environmental conditions.

Reliability is an inherent attribute of a system resulting from design, just as is the system's capacity, performance, or power rating. The reliability level is established at the design phase, and subsequent testing and production will not raise the reliability without a basic design change. Because reliability is an abstract concept that is difficult to grasp and to measure, many organizations find themselves unable to implement a comprehensive reliability program, primarily because of the lack of understanding on the part of both management and technical system design personnel. This is not to say that the system designers or managers in the organization are not interested in a reliable product; rather, the pressures on the design engineer, and very often on the organizational structure, impede the development of an effective reliability program (Kapur and Lamberson 1996).

With increasing system complexity, reliability becomes an elusive and difficult design parameter. It becomes more difficult not only to define and achieve as a design parameter, but to control and demonstrate in production and thus to ensure as an operational characteristic under the projected environmental conditions of use. However, past history has demonstrated that where reliability was recognized as a necessary program development component, with the practice of various reliability engineering methods throughout the evolutionary life cycle of the system, reliability can be quantified during the specification of design requirements, can be predicted by testing, can be controlled during production, and can be sustained in the field (Kapur 1996a). The purpose of this chapter is to present some of the reliability and maintainability methodologies and philosophies that are applicable throughout the life cycle of a system.

1.1. Reliability and Maintainability Activities during the System Life Cycle

Reliability and maintainability activities should span the entire life cycle of the system. Figure 1 shows the major points of reliability practice in a typical system life cycle (AMC 1968). The activities given in the exhibit are briefly explained here. There is a continuous feedback, and designs go through several cycles of the reliability program activities.

1.1.1. Step 1: The Need

The need for a reliability and maintainability program must be anticipated right from the beginning. The need for such programs cannot be overemphasized. These programs are justified based on specific system requirements in terms of life-cycle cost and other operational requirements. As already mentioned, the effectiveness of a system is determined by its reliability and maintainability characteristics.

1.1.2. Step 2: Goals and Definitions

All the requirements must be specified in terms of well-defined and quantitative goals. The goals and requirements are defined by some of the following measures:

1. *Reliability measures.* Mission reliability, a reliability function based on specified failure distribution, mean time to failure (MTTF), and failure rate.
2. *Maintainability measures.* Maintainability function based on time to repair distribution, mean time to repair (MTTR), percentile of time to repair, and maintenance ratio.

There are other measures as well, and the relationship among them is given in Figure 2 (Von Alven 1964).

The term *reliability* has already been defined; some of the other terms are defined as follows:

1. *Mission reliability* is the probability that the product and/or system will successfully complete a given mission with specified operating requirements and time duration.
2. *Maintainability* is defined as the probability that a failed system can be made operable in a specified interval of downtime. As shown in Figure 2, the downtime includes the failure detection time, the active repair time, the logistics time connected with the repairs of the product, and all the administrative time. The maintainability function describes probabilistically how long a system remains in the failed state.

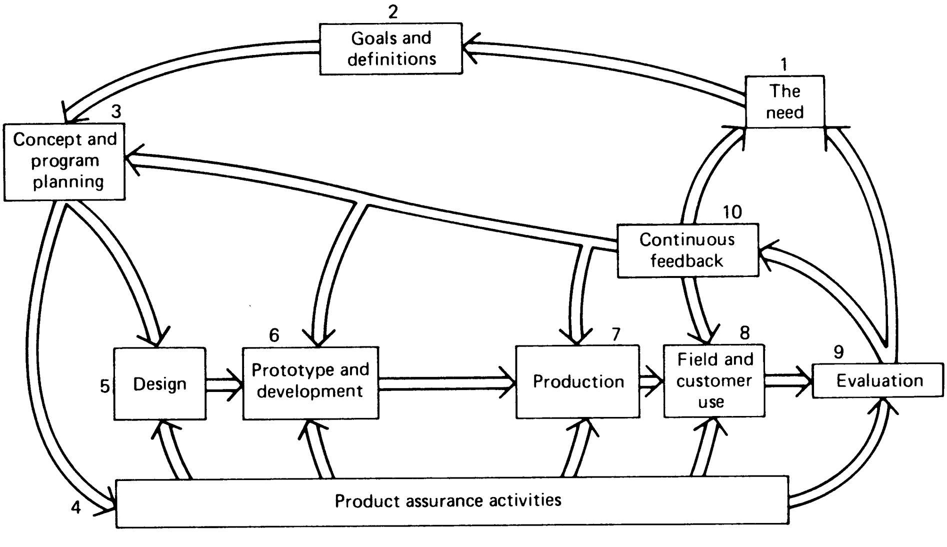


Figure 1 Reliability and Maintainability Activities During System Life Cycle.

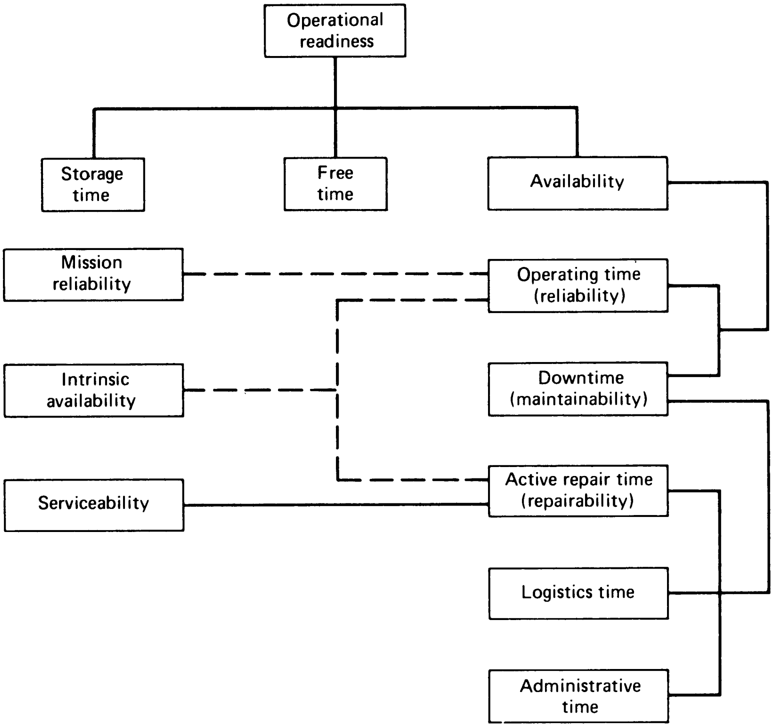


Figure 2 Relationship Among Various Product Assurance Measures.

3. *Repairability* deals only with the active repair time and can be defined by the time to actively repair random variable and the associated distribution. Repairability is defined as the probability that a failed system be restored to a satisfactory operating condition in a specified interval of active repair time. This measure is more valuable to the administration of the repair facility since it helps to quantify the workload for the facility and its workers.
4. *Serviceability* is defined as the ease with which a system can be repaired. Serviceability, like reliability, is a characteristic of the system design and must be planned at the design stage. Serviceability is difficult to measure on a ratio scale; however, it can easily be measured on an ordinal scale by a specifically developed rating and/or ranking procedure, which requires that systems be compared and ranked according to the ease of serviceability.
5. *Operational readiness* is defined as the probability that a system either is operating or can operate satisfactorily when the system is used under stated conditions. Operational readiness deals with all the time elements, including storage time, free time, operating time, and downtime.
6. *Availability* is defined as the probability that a system is operating satisfactorily at any point in time. Availability considers only the operating time and downtime, thus excluding the idle time. Therefore, it is a measure of the ratio of operating time of the system to the operating time plus the downtime. Availability is a function of both the reliability and maintainability of the system.
7. *Intrinsic availability* is more restrictive than availability because it is limited to operating and active repair time only.

These reliability, maintainability, and availability measures should be defined and system requirements specified using these quantitative measures. The effectiveness of the total product assurance program depends on these definitions.

1.1.3. Step 3: Concept and Program Planning

Based on the reliability and other operational requirements, various concepts are developed that can potentially meet these requirements. Also, at this stage, total product assurance program plans must be formulated and responsibilities assigned to different groups. The conceptual stage is an important part of the system life cycle because it has a major impact on the future system. Studies done by the U.S. Department of Defense indicate that 70% of the system life-cycle cost is determined by the decisions made at the concept stage. The detailed nature of the reliability programs will also determine the effectiveness of the total program.

1.1.4. Step 4: Product Assurance Activities

The plans developed in step 3 are implemented and the total program is continuously monitored, as indicated in Figure 1. An organization for the implementation of these plans must exist, with well-defined responsibilities

1.1.5. Step 5: Design

The conceptual system selected in step 3 is designed. Reliability and maintainability of this design are assessed. Various methodologies, such as design review, failure mode and effect analysis, fault tree analysis, and probabilistic design approach, can be applied at this step. Reliability is a design parameter and must be incorporated in the system at the design step.

1.1.6. Step 6: Prototype and Development

Prototypes are developed based on the design specifications. Reliability of the design is verified by testing. If the design has certain deficiencies, they are corrected by redesign. Reliability growth-management plans must be developed for this step in order to monitor continuously the growth and progress of the program. After the system has achieved the required level of reliability, the design is released for production.

1.1.7. Step 7: Production

The system is manufactured based on design specifications. Quality control methodologies are essential during this step. All the parts, materials, and processes are controlled based on methodologies discussed in previous chapters in the area of quality assurance. One of the objectives of the quality control program is to make sure that the inherent reliability of the design is not degraded.

1.1.8. Step 8: Field and Customer Use

Before the system is actually used in the field by its customers, it is very important to develop all the service and maintenance instructions, which are well documented. Just like reliability, maintainability is considered throughout the life cycle, and its purpose is to sustain required levels of reliability and availability in the field. Maintainability program plans are developed at the planning step.

1.1.9. Step 9: System Evaluation

The system in the field is continuously evaluated to determine whether the original reliability and maintainability goals are met by the system. For this purpose a reliability monitoring program and field data-collection program must be established.

1.1.10. Step 10: Continuous Feedback

There must be continuous feedback among all the steps in the system's life cycle. A proper communication system should be developed among all the groups responsible for the various steps. All the field deficiencies must be reported to the appropriate groups. This will help guide the system improvements.

The methodology related to some of the activities during the system life cycle is given in this chapter.

1.2. Reliability and Life Characteristic Curve

Reliability has sometimes been described as ''quality in the time dimension'' (RDG-376 1964) and a ''time oriented quality characteristic'' (Kapur 1986). The reliability characteristics of a product change with time. One of the characteristics is the concept of failure rate, which is defined mathematically later in this chapter. The failure rate, or the hazard rate, changes with the age or life of a product and has three distinct periods, as shown in Figures 3(*a*) and 3(*b*). These three periods are described here (Kececioglu 1991).

1.2.1. Infant Mortality Period

The total item population or a system generally exhibits a relatively high failure rate in the beginning, which decreases rapidly and stabilizes at some approximate time t_1. This initial period is generally called the 'burn-in,'' ''infant mortality,'' or ''debugging'' period. The item population has ''weak'' items, and these fail in the beginning. To understand the nature of these early failures, some of their causes are listed:

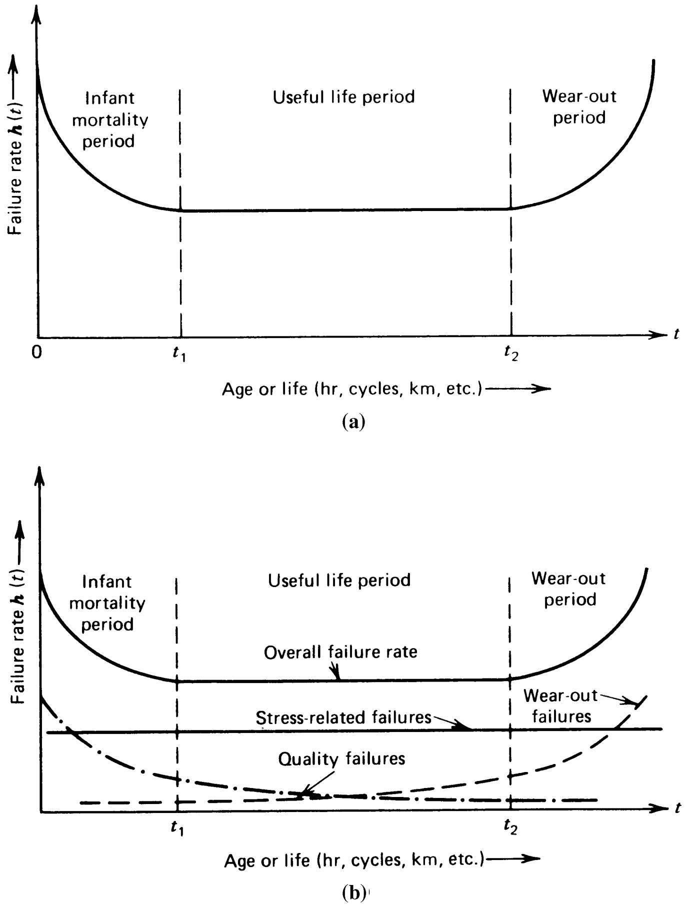


Figure 3 (*a*) Failure Rate-life Characteristic Curve. (*b*) Failure Rate Based on Components of Failure.

Substandard workmanship
Poor quality control
Substandard materials
Insufficient debugging
Poor manufacturing techniques
Poor processes and handling techniques
Problems due to assembly
Contamination
Improper installation
Improper start-up
Human error
Parts failure in storage and transit
Improper packaging and transportation practices

Fundamentally, these failures reflect the "manufacturability" of the product, and many are due to the lack of an effective quality control system in manufacturing. Thus, these early failures would show up during process audits, in-process or final tests, life tests, environmental tests, and so on.

Most manufacturers provide a burn-in period for their products so that the early or infant mortality failures occur in the plant and are not experienced by the customer in the field, where it is much more costly to fix these failures. The duration of the burn-in period determines what portion of the early failures is eliminated. Burn-in is a very expensive way to improve quality. A better way is to improve the process capability in terms of reducing variance. This will decrease the weak items in the population.

1.2.2. Useful Life Period

After having burned in, the item population reaches its lowest failure rate level, remaining relatively constant during this period. This failure rate is related to the inherent design reliability of the product and hence is given the most weight during the design reliability. This is also the most significant period for reliability prediction and assessment activities. Some of the causes of these failures are:

Low safety factors
Stress-related failures—higher-than-expected random loads
Lower random strength than expected
Defects that cannot be detected by the ''best'' available inspection techniques
Abuse
Human errors
Failures that cannot be observed during debugging
Failures that cannot be prevented by the ''best'' preventive maintenance practices
Unexplainable causes
''Act of God'' failures

1.2.3. Wear-Out Period

Most of the products are designed to last for a specified period of useful life. The time t_2 in Figures 3(*a*) and 3(*b*) indicates the end of useful life or the start of the wear-out period. After this point, the failure rate increases rapidly. The wear-out, or deterioration, results from a number of familiar chemical, physical, or other causes, some of which are as follows:

Corrosion or oxidation
Frictional wear or fatigue
Aging and degradation
Creep
Poor maintenance or service practices
Improper overhaul practices
Short designed-in life
Shrinkage or cracking in plastics

A reliability program must consider all three of these distinct periods. It must also be pointed out that not all products have these three periods. The importance of the periods depends on the magnitudes of time t_1 and t_2, where $0 \leq t_1 \leq t_2 < \infty$. Thus, we can develop various types of life characteristic curves, depending on the values t_1 and t_2. The best way to reduce infant mortality failures is to reduce the variance of the manufacturing process and reduce variation from the target. Early failures can also be eliminated by systematic procedures of controlled screening, quality control, and burn-in tests. Stress-related failures during the useful life can be minimized by providing adequate design or safety margins. Wear-out failures can be minimized by preventive maintenance and replacement policies.

2. RELIABILITY MEASURES

Reliability has been defined as the probability that a given system will perform satisfactorily its intended function for its intended life under specified operating conditions. Thus, reliability is related to the probability of the successful performance of any system. It is clear that we must define what the successful performance of any system is or what we mean by the failure of the system; otherwise it is not possible to predict when any system will fail to perform its intended function. The time to failure or ''life'' of a system cannot be deterministically defined, and hence it is a random variable. Thus, we must quantify reliability by assigning a probability function to the time to failure random variable.

2.1. Mathematics of Reliability Measures

Let T denote the time to failure random variable. Then reliability at any time t, denoted by $R(t)$, is the probability that the system will not fail by time t, or mathematically:

$$R(t) = P[\mathrm{T} > t] \tag{1}$$

Let $f(t)$ be the probability density function for the failure random variable T. Then the cumulative distribution function $F(t)$ is given by

$$P[T \leq t] = F(t) = \int_0^t f(\tau)d\tau \tag{2}$$

Hence from Eqs. (1) and (2) we have the following fundamental relationships (Eq. 3) between

$$R(t) = 1 - P[T \leq t] = 1 - F(t) = 1 - \int_0^t f(\tau)d\tau \tag{3}$$

the reliability function, cumulative distribution function, and probability density function.

Figures 4(*a*), 4(*b*), and 4(*c*) show, respectively, the failure probability density function, the cumulative distribution function, and the reliability function for the well-known case when the time to failure is exponentially distributed. Here we have

$$f(t) = \lambda e^{-\lambda t} \qquad t \geq 0, \lambda > 0 \tag{4}$$

and

$$F(t) = \int_0^t \lambda e^{-\lambda\tau}\, d\tau = 1 \geqslant q\, e^{-\lambda t} \qquad t \geq 0 \tag{5}$$

$$R(t) = e^{-\lambda t} \qquad t \geq 0 \tag{6}$$

These functions are all related, and selection of any one determines the others. This is obvious from Eqs. 1, 2, and 3 or 4, 5, and 6.

Obviously, the reliability function inherent in a system, by virtue of its design, dictates the probability of successful system operation during the system's fife. A natural question is then, "How does one know the shape of a reliability function for a particular systems?" There are basically three ways in which it can be determined:

1. Test many systems to failure using a mission profile that is identical to use conditions. This would allow one to develop empirically a curve such as that shown in Figure 4(*c*).
2. Test many subsystems and components to failure where use conditions are recreated in the test environment This allows one to develop empirically the component reliability functions and then to derive analytically the system reliability function.
3. Based on past experience with similar systems, the underlying failure distribution may be hypothesized. Then one can test fewer systems to determine the parameters needed to adapt the failure distribution to a particular situation. For example, the lifetime of many different kinds of electronic components follows the exponential distribution as previously given in Eq. 4. To apply this distribution, one must know the value of the parameter λ for a particular situation. Elaborate studies have been done, so that for a given environment and mission, parameter λ can be determined for most electronic components.
4. In some cases, the failure physics involved in a particular situation may lead one to hypothesize a particular distribution. For example, fatigue of certain metals tends to follow either the lognormal or Weibull distributions. Here again, once a distribution is selected, the parameters for a particular application must be ascertained (MIL-HDBK 1974, 1979; Klion 1992).

Another measure that is frequently used as an indirect indicator of system reliability is the MTTF, which is the expected or mean value of the time to failure random variable. Thus, the MTTF is theoretically defined as

$$\mathrm{MTTF} = E[T] = \int_0^\infty tf(t)dt = \int_0^\infty R(t)dt \tag{7}$$

Sometimes the term *mean time between failures* (MTBF) when the product can be repaired or renewed is also used to denote $E[T]$. The problem with using only the MTTF as an indicator of

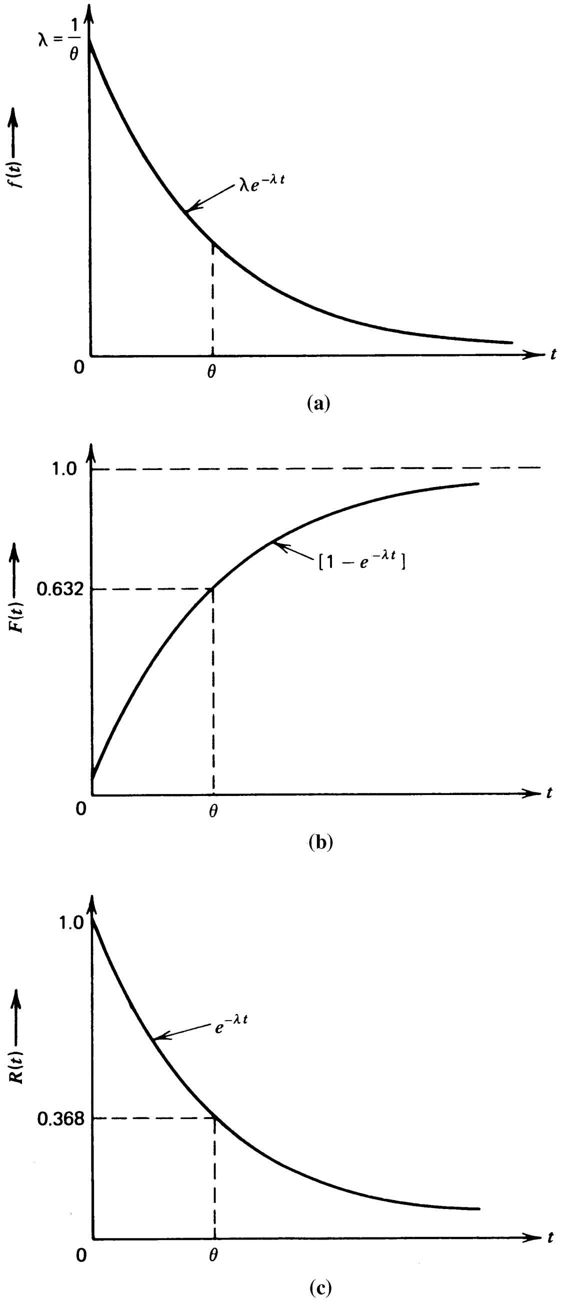


Figure 4 (*a*) Exponential Density Function. (*b*) Exponential Distribution Function. (*c*) Reliability Function for Exponential Distribution.

system reliability is that it uniquely determines reliability only if the underlying time to failure distribution is exponential. If the failure distribution is other than exponential, the MTTF can produce erroneous comparisons and we must develop other moments of the life distribution.

If we have a large population of the items whose reliability we are interested in studying, then for replacement and maintenance purposes we are interested in the rate at which the items in the population that have survived at any specific time will fail. This is the failure rate, or hazard rate, and is given by the following relationship:

$$h(t) = \frac{f(t)}{R(t)} \tag{8}$$

The failure rate for most components follows the curve shown in Figure 3(*a*), which is also called the life characteristic curve.

To help understand the notion of failure rate or hazard rate, basic mathematical relations are given here.

The hazard rate is defined as the limit of the instantaneous failure rate given no failure up to time *t* and is given by

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P[t < T \le t + \Delta t | T > t]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{R(t) - R(t + \Delta t)}{\Delta t R(t)} \\ &= \frac{1}{R(t)} \left[-\frac{d}{dt} R(t) \right] \\ &= \frac{f(t)}{R(t)} \end{aligned} \tag{9}$$

Also

$$f(t) = h(t) \exp \left[-\int_0^t h(\tau) d\tau \right] \tag{10}$$

and thus

$$R(t) = \exp \left[-\int_0^t h(\tau) d\tau \right] \tag{11}$$

2.2. Various Life Distributions

The properties of some life distributions that are used in reliability and maintainability discipline are given below.

2.2.1. Exponential Distribution

$$f(t) = \lambda e^{-\lambda t} \qquad t \ge 0 \tag{12}$$

$$F(t) = 1 - e^{-\lambda t} \qquad t \ge 0 \tag{13}$$

$$R(t) = e^{-\lambda t} \qquad t \ge 0 \tag{14}$$

$$h(t) = \lambda \tag{15}$$

$$\text{MTBF} = \theta = \frac{1}{\lambda} \tag{16}$$

Thus, the failure rate for the exponential distribution is always constant.

2.2.2. Normal Distribution

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \quad -\infty < t < \infty \tag{17}$$

$$F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right) \tag{18}$$

$$R(t) = 1 - \Phi\left(\frac{t-\mu}{\sigma}\right) \tag{19}$$

$$h(t) = \frac{\phi[(t-\mu)/\sigma]}{\sigma R(t)} \tag{20}$$

$$MTBF = \mu \tag{21}$$

Thus, $\Phi(z)$ is the cumulative distribution function and $\phi(z)$ is the probability density function, respectively, for the standard normal variable Z. The failure rate for the normal distribution is a monotonically increasing function. Normal distribution should be used as a life distribution when $\mu > 6\sigma$ because then the probability that T will be negative is exceedingly small. Otherwise, truncated normal distribution should be used.

2.2.3. Lognormal Distribution

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln t-\mu}{\sigma}\right)\right]^2 \quad t \geq 0 \tag{22}$$

$$F(t) = \Phi\left(\frac{\ln t-\mu}{\sigma}\right) \tag{23}$$

$$R(t) = 1 - \Phi\left(\frac{\ln t-\mu}{\sigma}\right) \tag{24}$$

$$h(t) = \phi\left(\frac{\ln t-\mu}{\sigma}\right)\Big/ t\sigma R(t) \tag{25}$$

$$\text{MTBF} = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{26}$$

The failure rate for the lognormal distribution is neither always increasing nor always decreasing. It takes different shapes, depending on the parameters μ and σ.

2.2.4. Weibull Distribution (Three Parameters $\theta > \delta$)

$$f(t) = \frac{\beta(t-\delta)^{\beta-1}}{(\theta-\delta)^{\beta}} \exp\left[-\left(\frac{t-\delta}{\theta-\delta}\right)^{\beta}\right] \quad t \geq \delta \geq 0 \tag{27}$$

$$F(t) = 1 - \exp\left[-\left(\frac{t-\delta}{\theta-\delta}\right)^{\beta}\right] \tag{28}$$

$$R(t) = \exp\left[-\left(\frac{t-\delta}{\theta-\delta}\right)^{\beta}\right] \tag{29}$$

$$h(t) = \frac{\beta(t-\delta)^{\beta-1}}{(\theta-\delta)^{\beta}} \tag{30}$$

$$\text{MTBF} = \delta + (\theta-\delta)\Gamma\left(1 + \frac{1}{\beta}\right) \tag{31}$$

The failure rate for the Weibull distribution is decreasing when $\beta < 1$, constant when $\beta = 1$ (same as the exponential distribution), and increasing when $\beta > 1$.

2.2.5. Gamma Distribution

$$f(t) = \frac{\lambda^{\eta}}{\Gamma(\eta)} t^{\eta-1} e - \lambda t \qquad t \geq 0 \tag{32}$$

$$F(t) = \sum_{k=\eta}^{\infty} \frac{(\lambda t)^k \exp[-\lambda t]}{k!} \qquad \text{when } \eta \text{ is integer} \tag{33}$$

$$R(t) = \sum_{k=0}^{\eta-1} \frac{(\lambda t)^k \exp[-\lambda t]}{k!} \qquad \text{when } \eta \text{ is integer} \tag{34}$$

$$h(t) = \frac{f(t)}{R(t)} \qquad \text{[Using Eqs. (32) and (34)]} \tag{35}$$

$$MTBF = \frac{\eta}{\lambda} \tag{36}$$

The failure rate for the gamma distribution is decreasing when $\eta < 1$, constant when $\eta = 1$, and is increasing when $\eta > 1$.

2.3. Summary

Properties of extreme value distributions are given by Gumbel (Gumbel 1958). Other mathematical properties of the preceding distributions and their use in the reliability field have been discussed extensively by several authors, and results are available in the literature (Barlow and Proschan 1975; Kapur and Lamberson; 1977; Kececioglu 1991, 1993). To use these distributions in the product assurance field, one must understand their nature and properties and the conditions under which they are applicable to describe various physical phenomena (Bane and Engelhardt 1991; Elsayed 1996). The following discussion on the application of life distributions is given in Kapur (1996b).

Exponential distribution is a good model for the life of a complex system that has a large number of components. Because the exponential distribution has a constant failure rate, it is a good model for the useful life of many products after the end of the infant mortality period. Some applications for the exponential distribution are electrical and electronic systems, computer systems, and automobile transmissions. The *normal distribution* is used to model various physical, mechanical, electrical, or chemical properties of systems. Some examples are gas molecule velocity, wear, noise, chamber pressure from firing ammunition, tensile strength of aluminum alloy steel, capacity variation of electrical condensers, electrical power consumption in a given area, generator output voltage, and electrical resistance. The *lognormal distribution* is a positively skewed distribution and can be used to model situations where large occurrences are concentrated at the tail (left) end of the range. Some examples are amount of electricity used by different customers, downtime of systems, time to repair, light intensities of bulbs, concentration of chemical process residues, and automotive mileage accumulation by different customers. The *two-parameter Weibull distribution* can also be used to model skewed data. When $\beta < 1$, the failure rate for the Weibull distribution is decreasing and hence can be used to model infant mortality or debugging period or for situations when the reliability in terms of failure rate is improving or for reliability growth. When $\beta = 1$, the Weibull distribution is the same as the exponential distribution, and all of the previous comments for the exponential distribution are applicable. When $\beta > 1$, the failure rate is increasing, and hence it is a good model for determining wear out and end-of-useful life period. Some of the examples are corrosion life, fatigue life, life of antifriction bearings, transmission gears, and electronic tubes. The *three-parameter Weibull distribution* is a good model when we have a minimum life and the odds of the component failing before the minimum life are close to zero. Many strength characteristics of systems do have a minimum value significantly greater than zero. Some examples are electrical resistance, capacitance, and fatigue strength.

3. SYSTEM RELIABILITY MODELS

To analyze and measure the reliability and maintainability characteristics of a system, there must be a mathematical model of the system that shows the functional relationships among all the components, the subsystems, and the overall system. The reliability of the system is a function of the reliabilities of its components. A system reliability model consists of some combination of a reliability block diagram or cause–consequence chart, a definition of all equipment failure and repair distributions, and a statement of spare and repair strategies (Kapur 1996a). All reliability analyses and optimizations are made on these conceptual mathematical models of the system.

3.1. Reliability Block Diagram

A reliability block diagram is obtained from a careful analysis of the manner in which the system operates. An analysis has to be done of the effects on overall system performance of failures of the various components; the support environment and constraints, including such factors as the number and assignment of spare parts and repairpersons; and the mission for the system.

Engineering analysis on the system has to be done in order to develop a reliability model. The engineering analysis consists of the following steps:

1. Develop a functional block diagram of the system based on physical principles governing the operations of the system.
2. Develop the logical and topological relationships between functional elements of the system.
3. Performance-evaluation studies are used to determine the extent to which the system can operate in a degraded state.
4. Define the spares and repairs strategies (for maintenance systems).

Based on the preceding analysis, a reliability block diagram is developed, which is used for calculating various measures of reliability and maintainability. The reliability block diagram is a pictorial way of showing all the success or failure combinations for the system. Some of the guidelines for drawing these diagrams are as follows:

1. A group of components that are essential for the performance of the system and/or its mission are drawn in series [Figure 5(*a*)].
2. Components that can substitute for other components are drawn in parallel [Figure 5(*b*)].
3. Each block in the diagram is like a switch: it is closed when the component it represents is working and is opened when the component has failed. Any closed path through the diagram is a success path.

The failure behavior of all the redundant components must be specified. Some of the common types of redundancies are:

1. *Active redundancy* or *hot standby:* The component has the same failure rate as if it was operating in the system.
2. *Passive redundancy, spare,* or *cold standby:* The standby component cannot fail. This is generally assumed of spare or shelf items.
3. *Warm standby:* The standby component has a lower failure rate than the operating component. This is usually a realistic assumption.

Some mathematical relationships between the system reliability and the reliabilities of its components are given in the next subsections. In the following, R_s denotes the reliability of the system, and R_i denotes the reliability of the ith component, where $i = 1, 2, \ldots, n$ and the system has n components. In addition, in the following relationships, it is also assumed that all the components work or fail independently of each other.

3.1.1. Series Configuration [See Figure 5(*a*)]

For the static situation we have

$$R_s = \prod_{i=1}^{n} R_i \tag{37}$$

and for the dynamic situation

$$R_s(t) = \prod_{i=1}^{n} R_i(t) \tag{38}$$

The failure rate $h_s(t)$ for the system is given by

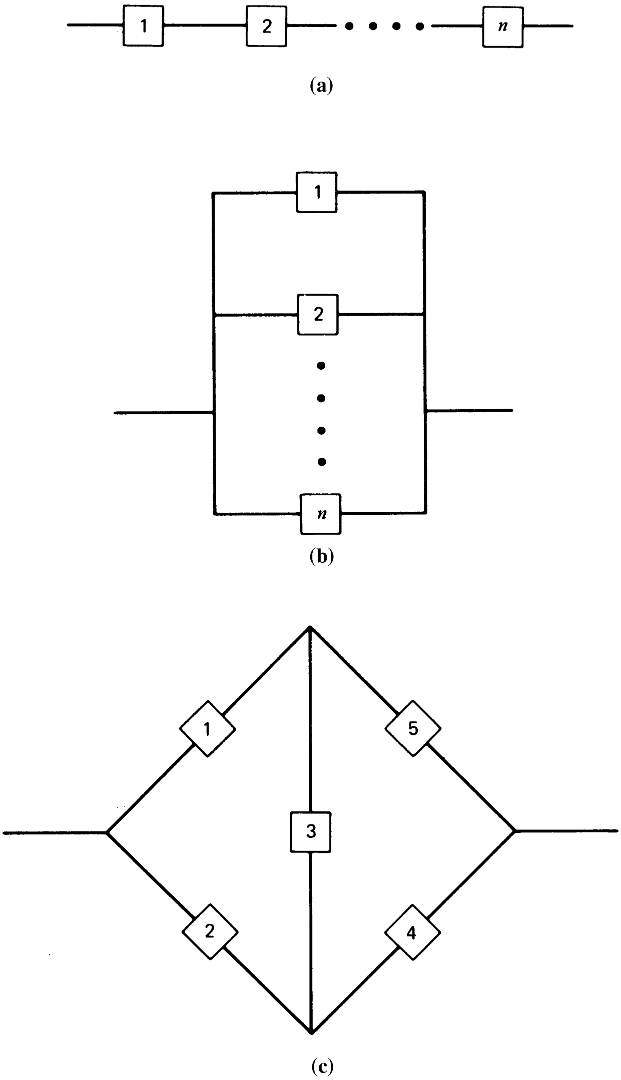


Figure 5 (*a*) Series Configuration. (*b*) Parallel configuration. (*c*) Bridge Structure.

$$h_s(t) = \sum_{i=1}^{n} h_i(t) \tag{39}$$

where $h_i(t)$ is the failure rate of ith component.

If all the components have an exponentially distributed time to failure, we have

$$h_s(t) = \sum_{i=1}^{n} \lambda_i \tag{40}$$

and MTBF for the system is given by

$$\frac{1}{\sum_{i=1}^{n} \lambda_i} \tag{41}$$

3.1.2. *Parallel Configuration* [See Figure 5(*b*)]

For the static case, we have

$$R_s = 1 - \prod_{i=1}^{n} (1 - R_i) \tag{42}$$

and for the dynamic case

$$R_s(t) = 1 - \prod_{i=1}^{n} [1 - R_i(t)] \tag{43}$$

If the time to failures for all the components is exponentially distributed with MTBF θ, then the MTBF for the system is given by

$$\sum_{i=1}^{n} \frac{\theta}{i} \tag{44}$$

where θ = MTBF for every component.

3.1.3. *The* **k**-*out-of*-**n** *Configuration*

In this configuration the system works if and only if at least k components out of the n components work, $1 \le k \le n$. For this case, when $R_i = R(t)$ for all i, we have

$$R_s(t) = \sum_{i=k}^{n} \binom{n}{i} [R(t)]^i [1 - R(t)]^{n-i} \tag{45}$$

If $R(t) = e^{-t/\theta}$, for exponential case, MTBF for the system is given by

$$\sum_{i=k}^{n} \frac{\theta}{i} \tag{46}$$

3.1.4. *Coherent Systems*

The reliability block diagrams for many systems cannot be represented by the preceding three configurations. In general, the concept of coherent systems can be used to determine the reliability of any system (Barlow and Proschan 1975). The performance of each of the n components in the system is represented by a binary indicator variable, x_i, which takes the value 1 if the ith component functions and 0 if the ith component fails. Similarly, the binary variable ϕ indicates the state of the system, and ϕ is a function of $x = (x_1, \ldots, x_n)$.

The function $\phi(x)$ is called the "structure function" of the system. The structure function is represented by using the concept of minimal path and minimal cut. A minimal path is a minimal set of components whose functioning ensures the functioning of the system. A minimal cut is a minimal set of components whose failures cause the system to fail. Let $\alpha_j(x)$ be the jth minimal path series structure for path A_j, $j = 1, \ldots, p$ and $\beta_k(x)$ be the kth minimal parallel cut structure for cut B_k, $k = 1, \ldots s$. Then we have

$$\alpha_j(x) = \prod_{i \in A_j} x_i \tag{47}$$

$$\beta_k(x) = 1 - \prod_{i \in B_k} (1 - x_i) \tag{48}$$

and

$$\phi(x) = 1 - \prod_{j=1}^{p} [1 - \alpha_j(x)] \tag{49}$$

$$= \prod_{k=1}^{s} \beta_k(x) \tag{50}$$

For the bridge structure [Figure 5(c)] we have four minimal paths and four minimal cuts, and their structure functions are given as

$$\begin{aligned} \alpha_1 &= x_1x_5 & \beta_1 &= 1 - (1 - x_1)(1 - x_2) \\ \alpha_2 &= x_2x_4 & \beta_2 &= 1 - (1 - x_4)(1 - x_5) \\ \alpha_3 &= x_1x_3x_4 & \beta_3 &= 1 - (1 - x_1)(1 - x_3)(1 - x_4) \\ \alpha_4 &= x_2x_3x_5 & \beta_4 &= 1 - (1 - x_2)(1 - x_3)(1 - x_5) \end{aligned}$$

Then the reliability of the system is given by

$$R_s = P[\phi(x) = 1] = E[\phi(x)]$$

If R_i is the reliability of the ith component, we have for the bridge structure

$$\begin{aligned} R_s = {} & R_1R_5 + R_1R_3R_4 + R_2R_3R_5 + R_2R_4 \\ & - R_1R_3R_4R_5 - R_1R_2R_3R_5 - R_1R_2R_4R_5 \\ & - R_1R_2R_3R_4 - R_2R_3R_4R_5 + 2R_1R_2R_3R_4R_5 \end{aligned}$$

If all $R_i = R = 0.9$, we have

$$\begin{aligned} R_s &= 2R^2 + 2R^3 - 5R^4 + 2R^5 \\ &= 0.9785 \end{aligned}$$

The exact calculations for R_s are generally very tedious because the paths and the cuts are dependent, since they may contain the same component. Bounds on system reliability are given by

$$\prod_{k=1}^{s} P[\beta_k(x) = 1] \le P[\phi(x) = 1] \le - \prod_{j=1}^{p} \{1 - P[\alpha_j(x) = 1]\}$$

Using these bounds for the bridge structure, we have, when $R_i = R = 0.9$,

$$\begin{aligned} \text{Upper bound on} \quad R_s &= 1 - (1 - R^2)^2(1 - R^3)^2 \\ &= 0.9973 \\ \text{Lower bound on} \quad R_s &= [1 - (1 - R)^2]^2[1 - (1 - R)^3]^2 \\ &= 0.9781 \end{aligned}$$

The bounds on system reliability using the concepts of minimum paths and cuts can be improved. Further details and derivations for coherent systems can also be found in Barlow and Proschan 1975.

4. FAULT TREE ANALYSIS

Fault tree analysis is one of the methods for system safety and reliability analysis (Henley et al. 1992). The concept was originated by Bell Telephone Laboratories as a technique for safety evaluation of the Minuteman Launch Control System. Many reliability techniques are inductive and are concerned primarily with ensuring that hardware will accomplish its intended functions. Fault tree analysis is a detailed deductive analysis that usually requires considerable information about the system. It is concerned with ensuring that all critical aspects of a system are identified and controlled. It is a graphical representation of Boolean logic associated with the development of a particular system failure (consequence), called the top event, to basic failures (causes), called primary events. These top events can be broad, all-encompassing events, such as release of radioactivity from a nuclear power plant or inadvertent launch of an ICBM missile, or they can be specific events, such as failure to insert control rods or energizing power available to ordnance ignition line.

Fault tree analysis is of value in:

1. Providing options for qualitative and quantitative reliability analysis
2. Helping the analyst to understand system failures deductively
3. Pointing out the aspects of a system that are important with respect to the failure of interest
4. Providing the analyst an insight into system behavior

A fault tree is a model that graphically and logically represents the various combinations of possible events, both fault and normal, occurring in a system that lead to the top event. A fault event is an abnormal system state. A normal event is an event that is expected to occur. The term *event* denotes a dynamic change of state that occurs in a system element. System elements include hardware, software, human, and environmental factors. Details about the construction of fault trees can be found in Henley et al. (1992).

5. ALLOCATION OF RELIABILITY REQUIREMENTS

Reliability and design engineers must translate overall system performance, including reliability, into component performance, including reliability. The process of assigning reliability requirements to individual components in order to attain the specified system reliability is called reliability allocation. There are many different ways in which reliability can be allocated in order to achieve this end.

The allocation problem is complex for several reasons, including: (1) the role a component plays for the functioning of the system; (2) the methods available for accomplishing this function; (3) the complexity of the component; and (4) the reliability of the component, which may change with the type of function to be performed. The problem is further complicated by the lack of detailed information on many of these factors early in the system design phase. However, a tentative reliability allocation must be accomplished in order to guide the design engineer. The typical decision process from a reliability allocation standpoint is illustrated in Figure 6. A process such as this attempts to force all concerned to make decisions in an orderly and knowledgeable fashion rather than on an ad hoc basis.

Some of the advantages of the reliability allocation process are:

1. The process forces system design and development personnel to understand and develop the relationships among component, subsystem, and system reliabilities. This leads to an understanding of the basic reliability problems inherent in the design.
2. The design engineer is obliged to consider reliability equally with other system parameters, such as weight, cost, and performance characteristics.
3. Reliability allocation ensures adequate design, manufacturing methods, and testing procedures.

The allocation process is approximate, and the system effectiveness parameters, such as reliability and maintainability apportioned to the subsystems, are used as guidelines to determine design feasibility. If the allocated parameters for a system cannot be achieved using the current technology, then the system must be modified and the allocations reassigned. This procedure is repeated until an allocation is achieved that satisfies the system requirements.

Various allocation algorithms for reliability and availability requirements are available (Von Alven 1964; Kapur and Lamberson 1977).

6. DESIGN FOR RELIABILITY

Reliability is a time-oriented quality characteristic (Kapur 1986) and is defined and evaluated by the customer just like any other quality characteristic. Inherent reliability is a function of system concept and its design. After a design has been completed and released for manufacturing, the maximum reliability level of the system has been determined by virtue of its design. Essentially, the reliability effort is over once the design is released, and all that quality control during manufacturing phase can do is to ensure that this reliability level does not degrade during manufacturing. From a reliability standpoint, quality control during manufacturing is after the fact and thus is too late to consider reliability. Thus, the reliability effort must be an integral part of system design and development because this is where the reliability level is established. Reliability activities that can be performed during system design and development are described briefly in the remainder of this section.

To ensure most economically and effectively the production of a reliable product, the reliability activities must start early in the product-development cycle. However, at this stage, the identification of reliability improvements depends heavily on the experience of the personnel studying the product from blueprints and preliminary system mock-ups because no hard data are available for a quantitative assessment of reliability. To consider reliability early in the design cycle, one must rely on a formalized design review procedure, which is now briefly explained.

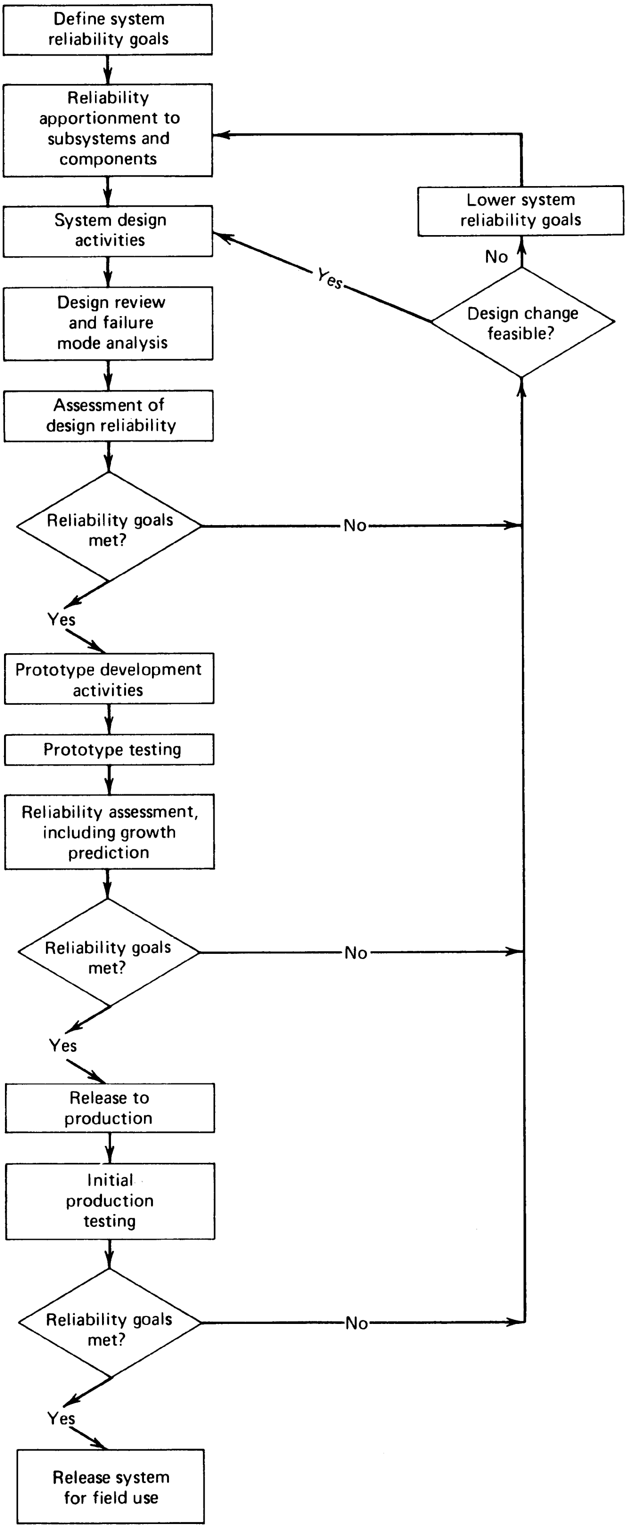

Define system reliability goals
Reliability apportionment to subsystems and components
System design activities
Design review and failure mode analysis
Assessment of design reliability
Reliability goals met?
No
Yes
Prototype development activities
Prototype testing
Reliability assessment, including growth prediction
Reliability goals met?
No
Yes
Release to production
Initial production testing
Reliability goals met?
No
Yes
Release system for field use
Lower system reliability goals
No
Design change feasible?
Yes

Figure 6 Reliability Allocation Process.

6.1. Design Review

The design review, a formal and documented review of a system design, is conducted by a committee of senior company personnel who are experienced in various pertinent aspects of product design, reliability, manufacturing, materials, stress analysis, human factors, safety, logistics, maintenance, and so on. The design review extends over all phases of product development, from conception to production. In each phase, previous work is updated, and the review is based on current information.

A mature design requires trade-offs between many conflicting factors, such as performance, manufacturability, reliability, and maintainability. These trade-offs depend heavily on experienced judgment and require continuous communication among experienced reviewers. The design review committee approach has been found to be extremely beneficial to this process. The committee adopts the system's point of view and considers all conceivable phases of design and system use to ensure that the best trade-offs have been made for the particular situation.

A complete design review procedure must be multiphased in order to follow the design cycle until the system is released for production. A typical example of a review committee, including personnel and their responsibilities, is shown in Table 1. Here the review process has been subdivided into three phases, and each phase is an update of more detailed analysis based on the latest knowledge.

Ultimately the design engineer has the responsibility for investigating and incorporating the ideas and suggestions posed by the design review committee. The committee's chairperson is responsible for adequately reporting all suggestions by way of a formal and documented summary. The design engineer then can accept or reject various points in the summary; however, he or she must formally report back to the committee, stating reasons for his or her actions.

It should be recognized that considerably more thought and detail than the basic philosophy presented here must go into developing the management structure and procedures for conduct in order to have a successful review procedure. It should be noted that this review procedure considers not only reliability, but all important factors in order to ensure that a mature design will result from the design effort.

In the next subsection, attention is focused on a technique that has been proven effective in identifying failure situations early in the design cycle and before product testing.

TABLE 1 Design Review Committee

	Review Phase			
Member	1	2	3	Responsibility
Chairperson	x	x	x	Ensure that review is conducted in an efficient fashion. Issue major reports and monitor follow-up.
Customer and/or marketing representative	x	x	x	Ensure that the customer's viewpoint is adequately presented (especially at the design trade-off stage)
Design engineer (of this product)	x	x	x	Prepare and present initial design with calculations and supporting data.
Design engineer (not of this product)	x	x	x	Review and verify adequacy of design.
Reliability engineer	x	x	x	Evaluate design for maximum reliability consistent with system goals.
Manufacturing engineer		x	x	Ensure manufacturability at reasonable cost. Check for tooling adequacy and assembly problems.
Materials engineer		x		Ensure optimum material usage considering application and environment.
Stress analyst		x		Review and verify stress calculations.
Quality control engineer		x	x	Review tolerancing problems, manufacturing capability, inspection strategies, and testing problems.
Human factors engineer		x		Ensure adequate consideration to human operator, identification of potential human-induced problems.
Safety engineer		x		Ensure safety to operating and auxiliary personnel.
Maintainability engineer		x	x	Analyze for ease of maintenance repair and field servicing problems.
Logistics engineer		x	x	Evaluate and specify logistic support. Identify logistics problems.

6.2. Failure Mode and Effects Analysis

Failure mode and effects analysis is a design-evaluation procedure used to identify all conceivable and potential failure modes and determine the effect of each failure mode on system performance. This procedure is accomplished by formal documentation, which serves (1) to standardize the procedure, (2) as a means of historical documentation, and (3) as a basis for future improvement.

The procedure consists of a sequence of logical steps, starting with the analysis of lower-level subsystems or components. The analysis assumes a failure point of view and identifies all potential modes of failure along with the causative agent, termed the "failure mechanism." The effect of each failure mode is then traced up to the systems level (MIL-STD-1969 1974).

A criticality rating is developed for each failure mode and resulting effect. The rating is based on the probability of occurrence, severity, and detectability. For failures scoring high on this rating, design changes to reduce criticality are recommended. This procedure is aimed at providing a better design from a reliability standpoint.

6.3. Probabilistic Approach to Design

Reliability is basically a design parameter and has to be incorporated in the system at the design stage. One way to quantify reliability during design and design for reliability is the probabilistic approach to design (Kececioglu and Cormier 1968; Kececioglu 1991, 1995; Haugen 1968). The design variables and parameters are random variables, and hence the design methodology must consider them as random variables. The reliability of any system is a function of the reliabilities of its components. To analyze the reliability of the system, we first have to understand how to compute the reliabilities of the components. The basic idea in reliability analysis from the probabilistic design methodology viewpoint is that a given component has certain strength that, if exceeded, will result in the failure of the component. The factors that determine the strength of the component are random variables, as are the factors that determine the stresses or loading acting on the component. "Stress" is used to indicate any agency that tends to induce failure, whereas "strength" indicates any agency resisting failure. "Failure" is taken to mean failure to function as intended; it occurs when the actual stress exceeds the actual strength for the first time.

Let $f(x)$ and $g(y)$ be the probability density functions for the stress random variable X and the strength random variable Y, respectively, for a certain mode of failure. Also, let $F(x)$ and $G(y)$ be the cumulative distribution functions for the random variables X and Y, respectively. Then the reliability R of the component for the failure mode under consideration, with the assumption that the stress and the strength are independent random variables, is given by

$$R = P\{Y > X\} \tag{51}$$

$$= \int_{-\infty}^{\infty} g(y) \left\{ \int_{-\infty}^{y} f(x) dx \right\} dy \tag{52}$$

$$= \int_{-\infty}^{\infty} g(y) F(y) dy \tag{53}$$

$$= \int_{-\infty}^{\infty} f(x) \left\{ \int_{x}^{\infty} g(y) dy \right\} dx \tag{54}$$

$$= \int_{-\infty}^{\infty} f(x) \{1 - G(x) \{ dx \tag{55}$$

For example, suppose that the stress random variable X is normally distributed with mean value of μ_x, and standard deviation (SD) of σ_x, and that the strength random variable Y is also normally distributed with parameters μ_y and σ_y. The reliability R is then given by

$$R = \Phi \left(\frac{\mu_y - \mu_x}{\sqrt{\sigma_y^2 + \sigma_x^2}} \right) \tag{56}$$

where $\Phi(z)$ is the cumulative distribution function for the standard normal variate Z.

The reliability computations for other distributions, such as exponential, lognormal, gamma, Weibull, and extreme value distributions, have also been developed (Kapur and Lamberson 1977). In addition, the reliability analysis has been generalized when the stress and strength variables follow a known stochastic process. The references cited in this subsection also contain simple design examples illustrating the use of the probabilistic approach to design.

7. HUMAN FACTORS IN RELIABILITY

All systems are of, by, and for humans. Human factors therefore become actively important in the system design process and consequently must be weighed against safety, reliability, maintainability, and other system parameters in order to obtain trade-offs to increase system effectiveness. Human interaction with the system of interest consists of:

1. Design and production of systems
2. Operators and repairers of systems
3. Operators and repairers as decision elements

Man–machine interface consists of such aspects as allocation of functions (man vs. machine), automation, accessibility, human tasks, stress characteristics, information presented to the human, and the reliability of interfaces coupled with the decisions on the basis of such information. Both human and machine elements of a system can fail, and their failures have varying effects on the system's performance. Some human errors cause total system failure or increase the risk of such failure. Human factors exert a strong influence on the design and ultimate reliability of a system (Kirwan 1994).

Both reliability and human factors are concerned with predicting, measuring, and improving system effectiveness. When the man–machine interface is complex, the possibility of human error increases, which results in an increase in the probability of system failure. An interesting facet of the human factors–reliability–maintainability relationship is that the system's reliability–maintainability depends on the detection and correction of system malfunctions. This task is generally performed by humans. Thus, the system performance can be enhanced or degraded depending on the human response. The quantification of human reliability characteristics and the development of a methodology for quantifying human performance, error prediction, control, and measurement are given in many sources (Gertman and Blackman 1994; Meister 1996).

Reliability of a system is affected by the allocation of system functions to man, machine, or both. Characteristics tending to favor humans are:

1. Ability to detect certain forms of energy
2. Sensitivity to a wide variety of stimuli within a restricted range
3. Ability to detect signals and patterns in high-noise environments
4. Ability to store large amounts of information for long periods and remember relevant facts
5. Ability to profit from experience
6. Ability to use judgment
7. Ability to improvise and adopt flexible procedures
8. Ability to arrive at new and completely different solutions to problems
9. Ability to handle low-probability or unexpected events
10. Ability to perform fine manipulations
11. Ability to reason instinctively

Characteristics tending to favor machines are:

1. Computing ability
2. Performance of routine, repetitive, and precise tasks
3. Quick response to control signals
4. Ability to exert large amounts of force smoothly and precisely
5. Ability to store and recall large amounts of data for short periods
6. Ability to reason deductively
7. Insensitivity to extraneous factors
8. Ability to handle highly complex operations that involve doing several things at once

8. RELIABILITY MEASUREMENT

Reliability measurement techniques provide a common discipline that can be used to measure, predict, and evaluate system reliability throughout the system life cycles. The two major components of the reliability measurement system are the test program and the data system. Test programs have to be developed throughout the life cycle, and the test effort has to ensure that the reliability goals are met at different stages in the cycle. Procedures for gathering the data generated throughout all the phases must be documented in sufficient detail for complete identification and integration into the data-processing system.

8.1. Test Programs

Figure 7 shows a sequence of different types of tests that may be used throughout the life cycle, consisting of design, development, production, and service/field use phases. Brief descriptions of these tests follow.

8.1.1. Design Support Tests

These tests are used to determine the need for parts, materials, and component evaluation or qualification to meet system performance and other reliability design criteria. Some of the objectives are:

Parts application data
Parts evaluation
Parts qualification
Parts comparative evaluation
Vendor control

8.1.2. Design-Verification Tests

These tests are used to verify the functional adequacy of design and to corroborate preliminary predictions and failure mode and effects analysis that disclose high-risk areas and reliability problems in the proposed design. Design-verification tests fulfill the following essential design phase functions:

Analytical verification
Functional evaluation
Parts and materials definition
Preliminary reliability verification

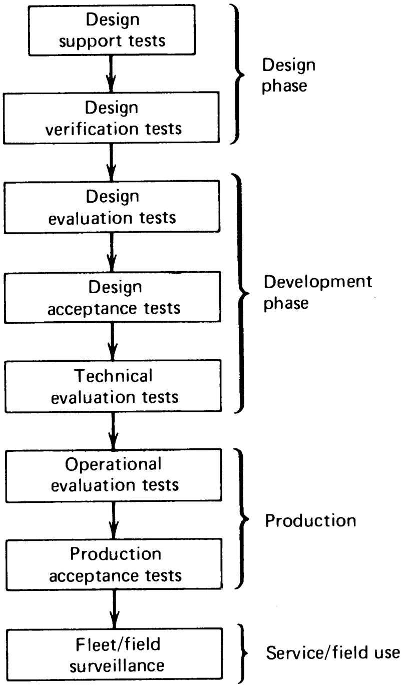

Design support tests
Design verification tests
Design phase
Design evaluation tests
Design acceptance tests
Technical evaluation tests
Development phase
Operational evaluation tests
Production acceptance tests
Production
Fleet/field surveillance
Service/field use

Figure 7 Integrated Test Flow Diagram.

8.1.3. Design-Evaluation Tests

These tests are used to evaluate the design under environmental conditions, verify the compatibility of subsystem interfaces, and review the design from the maintainability point of view. Some of the tests under this category are:

Environmental (evaluation tests)
Longevity (failure rates) tests
Operability tests
Engineering change-evaluation tests

8.1.4. Design-Acceptance Tests

These tests are used to demonstrate that design meets required levels of reliability. Thus, a reliability demonstration test is considered mandatory for design acceptance. Definitions of test requirements are:

1. Define acceptable levels of reliability and decision risks (type 1 and type 2 errors, confidence limits).
2. Define test conditions.
3. Define the specific test plan:
    - **(a)** MTBF tests
    - **(b)** Mission reliability test
    - **(c)** Availability tests
4. Define "failure" and scoring criteria:
    - **(a)** System failures
    - **(b)** Mission failures
    - **(c)** Maintenance actions (chargeable, nonchargeable)
    - **(d)** Criticality factors
5. Determine sample size.

8.1.5. Technical Evaluation Tests

These tests are used to evaluate the technical suitability of a prototype or a preproduction model. It is sometimes practical to integrate technical evaluation with operational evaluation when the earlier design-acceptance test demonstrates complete conformance.

8.1.6. Operational Evaluation Tests

These tests are used to evaluate operational suitability of the production model.

8.1.7. Production Acceptance Tests

These tests are used to determine the acceptability of individual production items in order to ensure production control and critical interfaces, parts, and material quality. Manufacturing operations that result in significant reliability degradation should be carefully studied.

8.1.8. Fleet/Field Surveillance Tests

These tests and evaluation programs during the field use of the product are for the continuing assessment of reliability and quality.

8.1.9. Test Procedures

Any test procedure must consider the following factors:

1. Purpose of test
2. Test items—description and sample selection
3. Test monitoring and review procedures
4. Test equipment requirements
5. Test equipment-calibration procedures
6. Test equipment proofing
7. Environmental conditions to be applied

8. Operating conditions
9. Test-point identification
10. Definition of failures and scoring criteria
11. Procedure for conducting tests
12. Test report procedures and documents

8.2. Reliability Estimation

Reliability measurement tests are used to make estimates of the reliability of a system or a population of items. Parametric and nonparametric estimates are used. Parametric estimates are based on a known or assumed distribution of the system characteristic of interest. The parameters are the constants that describe the shape of the distribution. Nonparametric estimates are used without assuming the nature of the underlying probability distribution. Generally, nonparametric estimates are not as efficient as parametric estimates. Nonparametric reliability estimates apply only to a specific test interval and cannot be extrapolated. Parametric estimates are described in this section when the underlying distribution is exponential and Weibull. The three types of parametric estimates that are frequently used are:

1. Point estimate: a single-valued estimate of a reliability measure
2. Interval estimate: an estimate of an interval that is believed to contain the true value of the parameter
3. Distribution estimate: an estimate of the parameters of a reliability distribution

8.2.1. Reliability Estimation: Exponential Distribution

The two types of test procedures considered here are:

1. Type 1 censored test: The items are tested for a specified time T, and then the testing is stopped.
2. Type 2 censored test: The test time is not specified, but the testing is stopped when a desired number of items fail.

Let us consider the situation when n items are placed on test and the test is stopped as soon as r failures are observed ($r \leq n$). This is type 2 censoring with nonreplacement of items. Let the observed failure times be, in order of magnitude,

$$0 = t_0 < t_1 < t_2 < \dots < t_{r-1} < t_r \tag{57}$$

Then, making the transformation,

$$u_i = \begin{cases} nt_i & i = 0 \\ (n - i)(t_{i-1} - t_i) & i = 1, 2, \dots, r - 1 \end{cases} \tag{58}$$

it is well known that the (u_i, $i = 0, \dots, r - 1$) are independently and identically distributed with common density function

$$\left(\frac{1}{\theta}\right) e^{-u/\theta}$$

It is clear that the total time on test is given by

$$\begin{aligned} V(t_r) &= \text{total time on test} \\ &= \sum_{i=0}^{r-1} u_i \\ &= \sum_{i=1}^{r} t_i + (n - r)t_r \end{aligned} \tag{59}$$

Then

$$\hat{\theta} = \frac{V(t_r)}{r} = \frac{1}{r}\left[\sum_{i=1}^{r} t_i + (n - r)t_r\right] \tag{60}$$

is the minimum variance unbiased estimator of θ. Since $V(t_r) = \Sigma_{i=0}^{r-1} u_i$ and the $\{u_i\}$ are independently distributed with a common exponential density function, it follows that $V(t_r)$ has a gamma distribution with parameters (θ, r). Hence $2V(t_r)/\theta = 2\hat{\theta}r/\theta$ is distributed as χ^2_{2r}.

The 100(1 − α)% confidence limits on θ are given by (Bane and Engelhardt 1991; Kececioglu 1993; Kapur and Lamberson 1977):

$$P\left[\chi^2_{1-(\alpha/2),2r} \leq \frac{2\hat{\theta}r}{\theta} < \chi^2_{\alpha/2,2r}\right] = 1 - \alpha$$

or

$$\frac{2\hat{\theta}r}{\chi^2_{\alpha/2,2r}} \leq \theta \leq \frac{2\hat{\theta}r}{\chi^2_{1-(\alpha/2),2r}} \tag{61}$$

The life-testing procedures are often used in a quality control context in which we wish to detect the deviations of θ below some desired levels, say, θ_0. Then for a significance level of α, the probability of accepting H_0 is

$$P_a = P\left(\frac{2r\hat{\theta}}{\theta_0} \leq \chi^2_{\alpha,2r} | \theta = \theta_0\right) = 1 - \alpha \tag{62}$$

The expected time to complete the test is given by

$$E(t_r) = \theta \sum_{i=1}^{r} \frac{1}{n - i + 1} \tag{63}$$

Let

θ_0 = desired reliability goal for MTBF

$1 - \alpha$ = probability of accepting items with true MTBF of θ_o

θ_1 = alternative MTBF ($\theta_1 < \theta_0$)

β = probability of accepting items with true MTBF of θ_1

With this information, reliability testing consists of putting n items on test and stopping the test when the number of failures is given by the smallest integer satisfying

$$\frac{\theta_1}{\theta_0} \leq \frac{\chi^2_{\alpha,2r}}{\chi^2_{1-\beta,2r}} \tag{64}$$

Thus, when we know θ_0, θ_1, α, and β, we can compute the necessary value for r.

For the type 1 censored test, where r failures are observed on an interval of total test time T, the 100(1 − α)% confidence limits on θ are given by [a modification of Eq. (61)]

$$\frac{2T}{\chi^2_{\alpha/2,2(r+1)}} \leq \theta \leq \frac{2T}{\chi^2_{1-(\alpha/2),2r}} \tag{65}$$

8.2.2. Reliability Estimation: Weibull Distribution

Weibull distribution (Weibull 1961) is probably one of the most widely used distributions in life-testing applications. One of the reasons is the ease with which graphic procedures can be used to estimate the parameters of the Weibull distribution and thus the reliability of the product. Confidence limits can also be easily developed. In addition, various statistical estimation procedures have recently been developed, and these can also be easily used by reliability engineers (Nelson 1990; Abernethy 1996; Kapur and Lamberson 1977).

The density function for the Weibull time to failure random variable is given by [see Eq. (27)]

$$f(t) = \frac{\beta(t - \delta)^{\beta-1}}{(\theta - \delta)^{\beta}} \exp\left[-\left(\frac{t - \delta}{\theta - \delta}\right)^{\beta}\right] t \geq \delta > 0 \tag{66}$$

where

β = shape parameter or the Weibull slope, $\beta > 0$
θ = scale parameter or the characteristic life
δ = location parameter or the minimum life

If the minimum life $\delta = 0$, the cumulative distribution is given by

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\theta}\right)^{\beta}\right] \tag{67}$$

After rearranging and talking twice the natural logarithm, we have

$$\ln\left[\ln\frac{1}{1 - F(t)}\right] = \beta \ln t - \beta \ln \theta$$

or

$$\ln t = \frac{1}{\beta} \ln\left[\ln\frac{1}{1 - F(t)}\right] + \ln \theta \tag{68}$$

Weibull graph paper is constructed by plotting ln t as the horizontal axis vs. $\ln\{\ln(1/[1 - F(t)]\}$ as the vertical axis, and then β is the slope of the straight-line plot. Figure 8 shows such Weibull paper. Various plotting procedures as well as statistical estimation methods with tables are available (Mann et al. 1974). Mann et al. offer point and interval estimation procedures for various other distributions and also tables that can be used for statistical estimation of the parameters of the Weibull distribution. This source is also a good reference for testing reliability hypothesis. The statistical methods discussed in this section can also be applied for accelerated life testing, which consists of a variety of test methods for shortening the life of products or hastening degradation of their performance (Nelson 1990). The aim of such testing is to obtain data quickly that, properly modeled and analyzed, give the proper information on product life under normal operating conditions by the customer.

9. MAINTAINABILITY

Maintainability is one of the system design parameters that has a great impact on the effectiveness of the system (Kececioglu 1995). Failures will occur no matter how reliable a system is made to be. A system's ability to be maintained, that is, retained in or restored to effective usable condition, is often as important to system effectiveness as is its reliability. Maintainability is a characteristic of the system and its design just like reliability. It is concerned with such system attributes as accessibility to failed parts, diagnosis of failures, repairs, test points, test equipment and tools, maintenance manuals, displays, and safety. Maintainability may be defined as a characteristic of design and installation that imparts to a system a great inherent ability to be maintained, so as to lower the required maintenance person-hours, skill levels, tools, test equipment, facilities, and logistics costs and thus achieve greater availability.

9.1. Maintainability Measures

Maintainability is the probability that a system in need of maintenance will be retained in or restored to a specified operational condition within a given period. Thus, the underlying random variable is the maintenance time. Let T be the repair time random variable. Then the maintainability function $M(t)$ is given by

$$M(t) = P[T \leq t] \tag{69}$$

If the repair time T follows the exponential distribution with mean time to repair, MTTR, of $1/\mu$, where μ is the repair rate, then

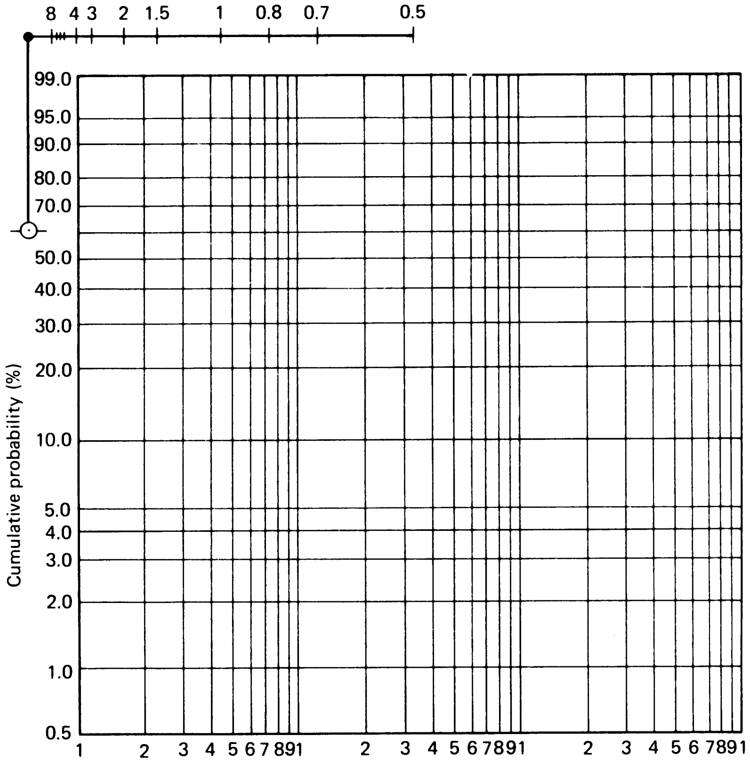


Figure 8 Weibull Probability Paper.

$$M(t) = 1 - \exp\left(-\frac{t}{\text{MTTR}}\right) \tag{70}$$

Various other distributions, such as lognormal, Weibull, and normal, are used to model the repair time. In addition, other time-related indices, such as median (50th percentile) and M_{MAX} (90th or 95th percentile), are used as maintainability measures. The lognormal probability density functions with an MTTR of 15 min, but with different values for SD, are given in Figure 9(*a*), whereas the associated maintainability functions are shown in Figure 9(*b*). From the maintainability function plot, different percentiles, such as the 90th, can be easily read. In other instances, the maintenance person-hours per system operating hours or maintenance ratio MR may be specified and maintainability design goals then derived from such specifications.

The MTTR, which is the mean of the distribution of system repair time, may be evaluated by

$$\text{MTTR} = \frac{\sum_{i=1}^{n} \lambda_i t_i}{\sum_{i=1}^{n} \lambda_i} \tag{71}$$

where n = number of components in the system
λ_i = failure rate of ith repairable component
t_i = time required to repair the system when the ith component fails

In addition, other measures, such as mean active corrective maintenance time and mean active preventive maintenance time, are used to measure maintainability. Some of the components of the corrective maintenance tasks are:

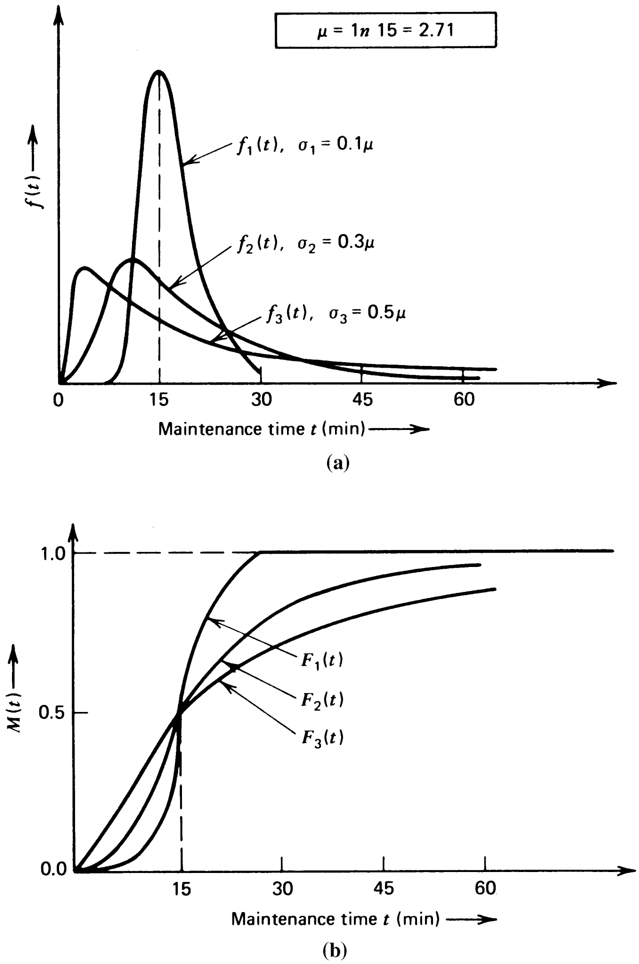


Figure 9 (*a*) Lognormal Probability Density Functions. (*b*) Maintainability Function *M*(*t*) Based on Lognormal Distribution with median equal to 15 min.

Localization: determining the location of a failure to the extent possible, without using accessory equipment

Isolation: determining the location of a failure by the use of accessory test equipment

Disassembly: disassembling the equipment to gain access to the item being replaced

Interchange: removing the failed item and installing the replacement

Alignment: performing any alignment, testing, and adjustment made necessary by the repair action

Checkout: performing checks or tests or verify that the equipment has been restored to a satisfactory operating condition

As an example of MTTR computation, assume that a communication system consists of five assemblies, with the data given in Table 2. Column 2 gives the number of units n_i for assembly *i*. Column 3 indicates the failure rate per thousand hours for each unit. Thus, column 4 gives us the total failure rate for an assembly *i*. Column 5 gives the average time to perform all the maintenance actions discussed previously. Then, MTTR is given by

$$\text{MTTR} = \frac{\sum n_i\lambda_i t_i}{\sum n_i\lambda_i} = \frac{63.5}{161} = 0.394 \text{ hr} \tag{72}$$

TABLE 2 Worksheet for MTTR Prediction

Assemblies	n_I	λ_i ($\times 10^3$)	$n_i\lambda_i$ ($\times 10^3$)	t_i (hr)	Repair Time per 10^3 hr ($n_i\lambda_i\ t_i$)
1	4	10	40	0.10	4.0
2	6	5	30	0.20	6.0
3	2	8	16	1.00	16.0
4	1	15	15	0.50	7.5
5	5	12	60	0.50	30.0
			$\Sigma = 161$		$\Sigma = 63.5$

10. AVAILABILITY

Availability is the vehicle that translates measures of reliability and maintainability into a combined index of effectiveness for a system. It is based on the question "Is the equipment available in a working condition when it is needed?" Availability analysis can be used for trading between and establish requirements for reliability and maintainability.

10.1. Availability Measures

By its very nature, availability measures are time related. The breakdown of total time upon which the availability analyses are based was briefly described earlier in this chapter. The time elements are:

1. Storage, free, and off time
2. Operating time
3. Standby time—availability for operations
4. Downtime, which consists of corrective and preventive maintenance and is also due to administrative and logistics delays

Measures for operational readiness are based on all the time elements. However, the availability measures do not consider the off time, including storage and free time. Achieved availability (A_a) is used for development and initial production testing where the system is not operating in its intended operational environment and is equal to operating test time divided by operating test time plus total preventive and corrective maintenance time (clock time). Excluded are operator before-and-after operating checks and supply, administration, and waiting time. Standby time is excluded both by definition and by environment. Thus,

$$A_a = \frac{\text{OT}}{\text{OT} + \text{TPM} + \text{TCM}} \tag{73}$$

where OT = operating time
TPM = total preventive maintenance time
TCM = total corrective maintenance time

Operational availability (A_o) covers all segments of the time that the system should be operative. Thus, we must consider standby time (ST) as well as administrative and logistics delay time (ALDT). Hence,

$$A_o = \frac{\text{OT} + \text{ST}}{\text{OT} + \text{ST} + \text{TPM} + \text{TCM} + \text{ALDT}} \tag{74}$$

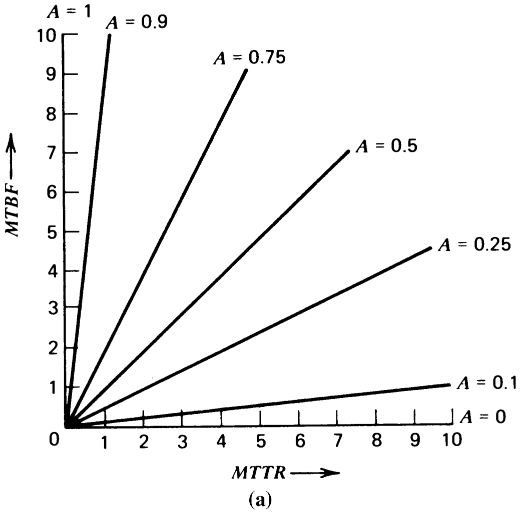
Sometimes there is a need to define the availability with respect to operating time and corrective maintenance when the system is operating in an ideal support environment. This form of availability, called inherent availability (A_I), is useful for determining certain figures of merit for the system per se, such as frequency and type of failure occurrence, repairability (active repair time), and analysis of maintenance actions. Thus, A_I is given by

$$A_I = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} \tag{75}$$

Standby time, delay times associated with scheduled or preventive maintenance, and administrative and logistics downtime are excluded.

10.2. Reliability–Maintainability–Availability Trade-Off

The system availability A_I, is a function of variables of reliability (MTBF) and maintainability (MTTR) as given by Eq. (75). Since MTBF = $1/\lambda$ where λ is the failure rate and MTTR = $1/\mu$ where μ is the repair rate (both valid when the underlying distribution is exponential), Eq. (75) may be rewritten as



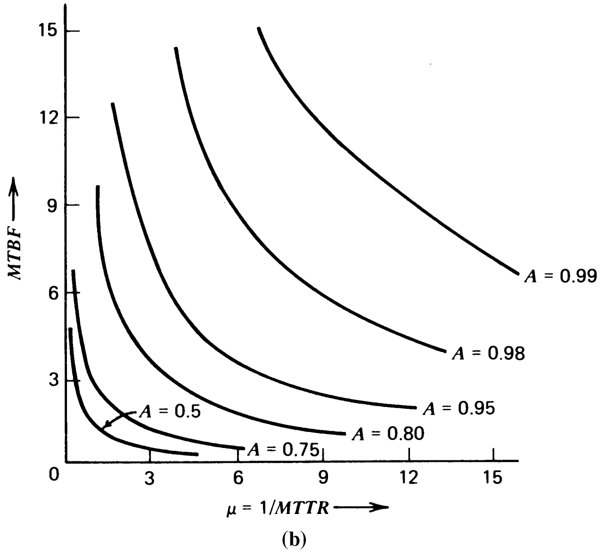


Figure 10 (*a*) Availability (*A*) as a Function of MTBF and MTTR. (*b*) Availability as a function of MTBF and Repair Rate.

$$A_I = A = \frac{\mu}{\mu + \lambda} \tag{76}$$

A generalized plot of Eq. (75) is given in Figure 10(*a*). This shows that to optimize availability, it is desirable to make the ratio of MTBF to MTTR as high as possible. Since increasing MTBF and decreasing MTTR is desirable, the equation for availability is plotted in terms of MTBF and $\mu = 1/\text{MTTR}$, as shown in Figure 10(*b*). Each of the curves representing the same availability is called an isoavailability contour; corresponding values of MTBF and MTTR give the same value of A, all other things being equal. Based on various physical, technological, and economic constraints, trade-off optimization models can be developed. There are practical limits as to how high a value for MTBF can be achieved or how low MTTR can be made. Increasing MTBF may require the redundancy level to be so high that the desired reliability could not be realistically achieved within the state of the art or else the cost would be high. Low values for MTTR would require excellent maintainability design features, such as complete built-in test features, automatic fault isolation, and automatic switchover from a failed to a standby item.

11. RELIABILITY GROWTH

As the product goes through the various steps in the life cycle, its reliability should be estimated and predicted. These values, when plotted at selected points in the life cycle, result in a growth curve, as shown in Figure 11, that reflects the comparative levels of reliability.

Reliability growth represents the effort spent to achieve the reliability potential either during design and development or during production or subsequently during field and operational use. During early development the achieved reliability of a prototype is much lower than its predicted reliability because of initial design and engineering deficiencies as well as manufacturing flaws. Also, the reliability of a fielded system is much lower than its inherent or potential reliability predicted during design and development for the following reasons:

1. Reliability degradation due to manufacturing, assembly, and quality control errors as well as to ineffectiveness of some of the screening tests.
2. Reliability degradation due to interaction of man, machine, and environment. Degradation may be due to rough handling, extended duty cycles, or neglected maintenance.
3. There is degradation due to maintenance activities because excessive handling brought about by frequent preventive maintenance or poor maintenance practices reduces reliability.

However, during all phases there is also reliability growth due to the underlying learning process. There is reliability growth during design and development as a result of an iterative design process. The essential elements involved in achieving reliability growth are:

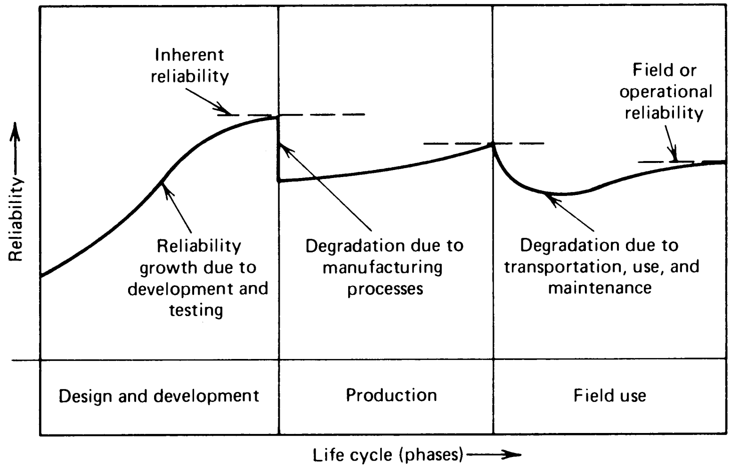

Inherent reliability
Field or operational reliability
Reliability
Reliability growth due to development and testing
Degradation due to manufacturing processes
Degradation due to transportation, use, and maintenance
Design and development
Production
Field use
Life cycle (phases)

Figure 11 Reliability Growth and Degradation During System Life.

1. Detection of failure sources
2. Feedback of problems identified
3. Redesign effort based on problems identified that will potentially be corrected

If the failure sources are detected by testing, then we have the following:

4. Manufacturing or fabrication of prototype hardware.
5. Redesign, detection, and analysis of failure sources serve as verification of redesign effort.

Some of the benefits of reliability growth methodology are:

1. It enables us to take advantage of experience gained in similar programs in the past.
2. It enables us to evaluate better the progress being made by an ongoing program.
3. It enables us to evaluate possible courses of corrective actions if an ongoing program experiences problems.

11.1 Reliability Growth Models

In 1964, Duane of GE (Duane 1964) published a report describing his observations on failure data for five different types of systems during their development programs at GE. His analysis revealed that for these systems the observed cumulative failure rate followed a consistent pattern, approximately a straight line, when plotted on a log–log paper as a function of cumulative test hours. This can be expressed mathematically as

$$\Lambda(t) = kt^{-\beta} \tag{77}$$

where $\Lambda(t)$ = cumulative failure rate at time t
t = cumulative test time
k = a constant ($k > 0$)
β = growth factor ($\beta > 0$)

Let $N(t)$ be the total number of failures accumulated over the cumulative test time t. Then

$$\Lambda(t) = \frac{N(t)}{t} \tag{78}$$

Thus, the cumulative failure rate $\Lambda(t)$ will decrease as reliability grows as a result of the development of corrective effort because fewer failures will be observed. If we take logarithms of both sides of Eq. (77), we have

$$\log \Lambda(t) = \log k - \beta \log t \tag{79}$$

The value of constant k depends on system complexity, design margins, and design objectives for reliability, whereas the value of growth factor β depends on the development effort

Substituting Eq. (78) in Eq (77), we have

$$N(t) = kt^{1-\beta} \tag{80}$$

$$= kt^{\beta'} \qquad (\beta' = 1 - \beta) \tag{81}$$

Talking the logarithm of both sides of Eq. (81), we have

$$\log N(t) = \log k + \beta' \log t \tag{82}$$

Thus, if we plot the cumulative number of failures with respect to cumulative test time t on a log–log paper, we get a straight line with slope β'. It is clear from Eq. (77) that the higher the value of β, the more growth the system has.

The actual failure rate of the system if the design is released after test time t is given by

$$\lambda(t) = \frac{dN(t)}{dt} = k(1 - \beta)t^{-\beta} \tag{83}$$

Thus, if we are given some goal for failure rate $\lambda(t)$, we can compute the total test time required during the development effort using Eq. (83).

12. DESIGN AND MANAGEMENT OF RELIABILITY PROGRAMS

The establishment of an effective product assurance program throughout the system life cycle requires a management with exceptional perception of the assurance sciences. This is because of the many conflicting factors involved when making decisions on the delegation or redistribution of responsibility and authority and also because reliability is at best a poorly understood discipline without a universal approach that applies to every product and every organization. A reliability program extends far beyond the estimation of reliability numbers: it must create an attitude of anticipation of reliability problems and initiate the preplanning necessary to eliminate or reduce the effects of unreliability to an acceptable and planned-for level. Because of the breadth of the subject, a product assurance program must span the total system life cycle; therefore, the program will infringe on several well-established technical management groups, such as design, testing, purchasing, quality control, manufacturing, sales, and service groups.

Unreliability is actually not the result of any one group but is due to the complexity of today's systems and of the organization required to create these systems. During system design, development, and production, anything that can contribute to unreliability should be identified and reviewed. A reliability-management structure must be created to accomplish this task. The question then is, ''Who should do this?'' In a ''perfect'' organization with complete flow of information, the unreliability problems might be taken care of by existing groups. However, such an organization rarely exists in practice; thus, over the years many organizations have found it necessary and advantageous to spend the extra effort to create a reliability group and integrate this group into the existing organization. The reliability group performs essentially an assurance function, facilitating and supporting the design and development process from a reliability point of view.

The establishment and maintenance of a viable reliability program must be done based on management foresight and intuition since the payback is not readily measured in dollars. Unreliability problems uncovered by a reliability group tend to be taken care of outside of the formal organizational communication channels and in general will be attributed to design engineering. Forcing a reliability group to point up all unreliability problems will mean that management is forcing it into a position of accusation, and this will create an intractable climate that may well hinder the product development cycle and the reliability improvement effort. Thus, it is very important to understand the role and the function of the reliability group in the total organization.

12.1. Elements of a Reliability Program

Management and control of system reliability must be based on a recognition of the system's life cycle, beginning at concept and extending through design, production, use, and discarding of the system. One of the objectives is to achieve acceptable levels of operational reliability and maintainability. Achievement of this objective requires numerous tasks. The activities of a reliability program have applications throughout the system life cycle. Some of these applications are: as follows:

Applications during design

Develop safety margins from reliability viewpoint.
Predict component reliability from the data bank of the failure rates.
Compute system reliability from component reliability.
Determine amount of redundancy needed to achieve a reliability goal.
Provide input to human engineering.
Interact with value engineering.
Evaluate design changes.
Perform trade-off analysis.
Compare two or more designs.
Provide guidelines for design review.
Work with cost-reduction programs.

Applications during development of testing

Establish reliability growth curves for the development and testing phase.
Develop guidelines for the amount of testing.

Develop bathtub curve based on the failure rate data and the test data.
Participate in the development of failure definition and scoring criteria document.
Participate in the scoring of the test failures.

Applications during manufacturing

Provide guidelines for manufacturing processes.
Provide input to quality control.
Provide input for guidelines to evaluate the suppliers and vendors.
Develop product burn-in or debugging time.

Applications during Field Use

Establish warranty cost and help reduce it.
Optimize the length of warranty.
Reduce inventory costs.
Develop maintenance procedures, both corrective and preventive.
Provide input to the spare parts-allocation models.
Participate in the collection and analysis of the field data.
Participate in the feedback process to report and correct the field failures.

The reliability group coordinates and directs the overall reliability effort to provide assurance that the optimum reliability has been achieved and that the consequences of unreliability have been considered in the overall plans. Obviously, the reliability group can be effective only if given proper authority, demanding formal sign-off during all critical stages of system development

An acceptable reliability level for a system is really a many-faceted problem that requires many complex trade-offs. Considerations in design, cost, and manufacturability, material availability, maintenance, and serviceability all enter into this problem. For example, a highly reliable design that cannot be manufactured effectively or cannot be maintained may represent an unacceptable situation from the total system point of view. The important thing is to plan for adequate overall system effectiveness, utilizing good knowledge on the actual reliability level of the system.

REFERENCES

Abernethy, R. B. (1996), *The New Weibull Handbook*, 2nd Ed., Gulf, Houston.

Bane, L. J., and Engelhardt, M. (1991), *Statistical Analysis of Reliability and Life-Testing Models: Theory and Methods*, 2nd Ed., Marcel Dekker, New York.

Barlow, R. E., and Proschan, F. (1975), *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart & Winston, New York.

Carrubba, E. R., and Gordon, R. D. (1988), *Product Assurance Principles: Integrating Design Assurance and Quality Assurance* McGraw-Hill, New York (new edition of E. R. Carrubba, R. D. Gordon, and A. C. Spann, *Assuring Production Integrity*, Lexington Books, Lexington, MA, 1975).

Duane, J. T. (1964), ''Learning Curve Approach to Reliability Monitoring,'' *IEEE Transactions on Aerospace*, Vol. 2, pp. 563–566.

Elsayed, A. E. (1996), *Reliability Engineering*, Addison-Wesley Longman, Reading, MA.

Gertman, D., and Blackman, H. S. (1994), *Human Reliability and Safety Analysis Data Handbook*, John Wiley & Sons, New York.

Gumbel, E. J. (1958), *Statistics of Extremes*, Columbia University Press, New York.

Haugen, E. B. (1968), *Probabilistic Approach to Design*, John Wiley & Sons, New York.

Henley, E. J., and Kumamota, H. (1992), *Probabilistic Risk Assessment: Reliability Engineering, Design and Analysis*, IEEE Press, New York.

Kapur, K. C. (1986), ''Quality Evaluation Systems for Reliability,'' *Reliability Review*, Vol. 6, No. 2.

Kapur, K. C. (1996a), ''Technologies of Estimating Reliability at Design Stage,'' in *Handbook of Reliability Engineering and Management*, 2nd Ed., Ireson, G. W., Coombs, C. F., and Moss, R. Y., Eds., McGraw-Hill, New York.

Kapur, K. C. (1996b), ''Mathematical and Statistical Methods and Models in Reliability and Life Studies,'' in *Handbook of Reliability Engineering and Management*, 2nd Ed., Ireson, G. W., Coombs, C. F., and Moss, R. Y., Eds., McGraw Hill, New York.

Kapur, K. C., and Lamberson, L. R. (1977), *Reliability in Engineering Design*, John Wiley & Sons, New York.

Kapur, K. C., and Lamberson, L. R. (1996), "Reliability," in *Mechanical Design Handbook*, H. A. Rothhand, Ed., McGraw-Hill, New York, pp. 8.1–8.23.

Kececioglu, D., and Cormier, D. (1968), "Designing a Specified Reliability Directly into a Component," in *Proceedings of Third Annual Aerospace Reliability and Maintainability Conference*.

Kececioglu, D. (1991), *Reliability Engineering Handbook*, Vols. 1 and Vol. 2, Prentice Hall, Englewood Cliffs, NJ.

Kececioglu, D. (1993), *Reliability and Life Testing Handbook*, Prentice Hall, Englewood Cliffs, NJ.

Kececioglu, D. (1995), *Maintainability, Availability and Operational Readiness Handbook*, Prentice Hall, Englewood Cliffs, NJ.

Kirwan, B. (1994), *A Guide to Practical Human Reliability Assessment*, Taylor & Francis, Bristol, PA.

Klion, J. (1992), *Practical Electronic Reliability Engineering: Getting the Job Done from Requirements through Acceptance*, Van Nostrand Reinhold, New York.

Mann, N. R., Shafer, R. E., and Singpurwalla, N. D. (1974), *Methods for Statistical Analysis of Reliability and Life Data*, John Wiley & Sons, New York.

Meister, D. (1996), "Human Factors in Reliability," in *Handbook of Reliability Engineering and Management*, 2nd Ed., Ireson, G. W., Coombs, C. F., and Moss, R. Y., Eds., McGraw-Hill, New York.

MIL-HDBK-217B (1974), *Reliability Prediction of Electronic Equipment*, Military Standardization Handbook, U.S. Department of Defense, Washington, DC.

MIL-HDBK-217C (1979), *Reliability Prediction of Electronic Equipment*, Military Standardization Handbook, U.S. Department of Defense, Washington, DC.

MIL-STD-1969 (SHIPS) (1974), *Procedures for Performing a Failure Mode and Effects Analysis for Shipboard Equipment*, U.S. Department of the Navy, Naval Ship Engineering Center, Hyattsville, MD.

Nelson, W. (1990), *Accelerated Testing: Statistical Models, Test Plans and Data Analysis*, John Wiley & Sons, New York.

RDG-376 (1964), *Reliability Design Handbook*, Reliability Analysis Center, Griffiss Air Force Base, NY.

United States Army Material Command (AMC) (1968), *Quality Assurance—Reliability Handbook*, AMC Pamphlet No. 702-3, U.S. AMC Headquarters, Alexandria, VA.

Von Alven, W. H., Ed. (1964), *Reliability Engineering*, Prentice Hall, Englewood Cliffs, NJ.

Weibull, W. (1961), *Fatigue Testing and Analysis of Results*, Macmillan, New York.

ADDITIONAL READING

Balakrishnan, N., Ed., *Recent Advances in Life-Testing and Reliability*: *A Volume in Honor of Alonzo Clifford Cohen, Jr.*, CRC Press, Boca Raton, FL, 1995.

Cai, K. Y., *Introduction to Fuzzy Reliability*, Kluwer, Dordrecht, 1996.

Dhillon, B. S., *Design Reliability: Fundamentals and Applications*, CRC Press, Boca Raton, FL, 1999.

Flamm, J., and Luisi, T., Eds., *Reliability Data Collection and Analysis*, Kluwer, Dordrecht, 1992.

Gnedenko, B., and Ushakov, I., *Probabilistic Reliability Engineering*, J. Falk, Ed., John Wiley & Sons, New York, 1995.

Lewis, E. E., *Introduction to Reliability Engineering*, 2nd Ed., John Wiley & Sons, New York, 1996.

Misra, K. B., *New Trends in System Reliability Evaluation*, Elsevier, Amsterdam, 1993.

Onisawa, T., and Kacprzyk, J., Eds., *Reliability and Safety Analysis under Fuzziness*, Physia, Heidelberg, 1995.

SAE International RMS Committee (G-11), *RMS: Reliability, Maintainability and Supportability Guidebook*, SAE, Warrendale, PA, 1995.

Soares, C. G., Ed., *Probabilistic Methods for Structural Design*, Kluwer, Dordrecht, 1997.

Ushakov, I., and Harrison, R. A., Eds., *Handbook of Reliability Engineering*, John Wiley & Sons, New York, 1994.

Vinogradov, O., *Introduction to Mechanical Reliability: A Designer's Approach*, Hemisphere, New York.

CHAPTER 73

Service Quality

LAURA RAIMAN DUPONT
Quality Engineering Consultant

1. **DEFINING SERVICE QUALITY** **1956**
2. **CREATING A SERVICE STRATEGY** **1957**
3. **CREATING SERVICE-DRIVEN LEADERSHIP** **1958**
4. **CREATING A SERVICE-DRIVEN WORKFORCE** **1959**
    - 4.1. Work Systems 1959
    - 4.2. Education and Training 1959
    - 4.3. Employee Well-Being and Satisfaction 1961
5. **SYSTEMS FOR SERVICE DELIVERY** **1961**
6. **MEASURING AND EVALUATING SERVICE QUALITY** **1963**
7. **CONCLUSIONS** **1964**

REFERENCES **1965**

1. DEFINING SERVICE QUALITY

Service quality is hard to define, hard to achieve, but easy to identify when missing. Service encounters happen every day when customers come into contact with an organization, its people, its communications, and the services it provides. These service encounters have been termed the "moment of truth" by Jan Carlzon, CEO of Scandinavian Airlines System (SAS)—the moment at which the representatives of the service provider must prove to their customers that their organization is the best alternative (Carlzon 1987). In order to become a leader in service quality, an organization must adopt a multidimensional perspective to managing service encounters. It must be effective at both designing and delivering services. In other words, an organization must do what the customer wants, and do it the right way. Accomplishing this requires an organization to implement effective approaches in five key areas: service strategy; service-driven leadership; service-driven workforce; systems for service delivery; and measurement and evaluation of service quality. When approaches in these areas are integrated together into an overall template for managing service quality, the result is an organization with a new, customer-oriented culture.

Although specific approaches and methods used in these five areas may differ widely between organizations, there is a vehicle that provides the framework for defining service quality and for guiding an organization in selecting and implementing approaches for managing service quality. That framework is the Baldrige National Quality Program *Criteria for Performance Excellence* (2000). The Baldrige Award was designed to publicize and promote successful quality strategies as a result of the Malcolm Baldrige National Quality Improvement Act of 1987. The Baldrige criteria are designed to help organizations enhance their performance through a focus on two results-oriented goals: delivery of ever-improving value to customers, resulting in marketplace success; and improvement of overall organizational effectiveness and capabilities. For over 12 years, the Baldrige Criteria have been used successfully by all types of organizations to assess and evaluate their operations. Explicit and well-articulated core values and concepts form the foundation of the criteria. The focus on values recognizes the different paths organizations may take to attain performance excellence. The Baldrige criteria provide an organization with a systems perspective for managing operations and for achieving performance excellence.

The Baldrige criteria are presented in seven categories, which are each subdivided into several items (Table 1). The remainder of this chapter describes key approaches in five global areas of service

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

TABLE 1 2000 Baldrige Criteria for Performance Excellence—Categories/Items

1. Leadership
   1.1. Organizational leadership
   1.2. Public responsibility and citizenship
2. Strategic planning
   2.1. Strategy development
   2.2. Strategy deployment
3. Customer and market focus
   3.1. Customer and market knowledge
   3.2. Customer satisfaction and relationships
4. Information and analysis
   4.1. Measurement of organizational performance
   4.2. Analysis of organizational performance
5. Human resource focus
   5.1. Work systems
   5.2. Employee education, training, and development
   5.3. Employee well-being and satisfaction
6. Process management
   6.1. Product and service processes
   6.2. Support processes
   6.3. Supplier and partnering processes
7. Business results
   7.1. Customer-focused results
   7.2. Financial and market results
   7.3. Human resource results
   7.4. Supplier and partner results
   7.5. Organizational effectiveness results

quality. Where there are very clear linkages to the Baldrige criteria, such as in the area of systems for service delivery, the criteria are presented to clarify best practice. In other cases, the criteria will just be referenced. The intent of drawing the Baldrige criteria into the discussion of service quality is not only to provide key ideas and approaches, but also to demonstrate the value of the Baldrige criteria in setting up an integrated approach for addressing service quality, and ultimately performance excellence.

2. CREATING A SERVICE STRATEGY

Service quality happens when an organization has employees who are committed to quality in their own work and are willing to go out of their way to deliver that high level of quality to customers. This will only happen when there is a well-developed service strategy that defines an organization's shared values and organizes all of the other elements of service.

The key step in developing a service strategy involves segmenting customers according to their service expectations. Within each market an organization tries to reach, there are different segments. All of these segments most likely have common core needs and expectations, but once customer segments are defined, an organization can find out exactly what the expectations are for each segment. This allows the organization to develop a service concept that provides it with a competitive advantage in the eyes of its customers. In other words, while an organization is providing a good core service, it may cater to each particular segment in order to differentiate itself from competitors in the eyes of those customers.

The identification of service differentiators can help an organization position itself as a leader. This requires all customer contact points to be carefully analyzed, particularly relative to each key customer segment. For each customer contact point, the organization must analyze what it does right or wrong and compare this to what customers want, need, and expect. It must also analyze what competitors provide and do not provide. Identifying service differentiators for customer segments also allows an organization to identify what employees play a role in each portion of the service-delivery process. This allows salespeople, administrators, production staff, and other employees to adapt their style, behavior, or environment to each key segment.

The right service strategy for a given organization is defined to an even greater extent by considering the strengths of key competitors, the guiding principles and values of the organization, and the organization's market. An effective service strategy provides a concept of service for an organization to rally around and allows leaders to identify conflicts between corporate strategy and customer service. It also allows the organization to design effective ways to measure service performance and service quality, as discussed later. The service strategy also provides an effective means for leadership to choose the optimum mix and level of service for different customer segments.

3. CREATING SERVICE-DRIVEN LEADERSHIP

Strong, effective leadership turns a service strategy into an everyday reality. While a service strategy provides the overall strategy for dealing with customers, customer-contact employees typically must exercise broad discretion when serving customers. Although the strategy is typically supported by service standards, employees must also rely on a strong service culture to guide them in making decisions. That culture takes it tone and its values from an organization's leaders. As W. Edwards Deming clearly pointed out, quality can be no better than the intent of top management. Leaders shape culture, and culture is key to customer service.

Employees at every level in an organization must understand leadership's commitment to customers and service. This requires an organization's senior leaders to set directions and create a customer focus, clear and visible values, and high expectations. A clear, compelling, memorable vision should convey the result to be achieved for the customer and the way in which the results are to be achieved. The leadership category in the Baldrige criteria examines how an organization's senior leaders address values and performance expectations as well as how they focus on customers and other stakeholders, empowerment, innovation, learning, and organizational directions.

An organization's vision is communicated to employees through signals sent by management attitudes, policies, and rewards. Leaders in companies that produce outstanding service incessantly pronounce their beliefs and back up their words with actions. When leaders behave in a manner consistent with the organization's mission and values and focuses on empowering people to make decisions and take risks, it supports a wide variety of desired behaviors and builds self-esteem and teamwork among employees. Leaders also must exemplify for their employees what it means to produce great service. This may include answering calls from customers and fielding complaints one day a month, writing thank-you letters, spending time working in front line jobs, and so on. Their goal is to nurture a service culture that will shape employee behavior more effectively than rules and regulations can. They make service everybody's business and empower employees to make on-the-spot decisions in the customer's interest. By doing this, leadership helps make strategy a day-to-day reality.

TABLE 2 Baldrige Criteria Item 1.1—Organizational Leadership

1.1. Organizational Leadership
Describe how senior leaders guide your organization and review organizational performance.
Within your response, include answers to the following questions:

a. Senior Leadership Direction
- (1) How do senior leaders set, communicate, and deploy organizational values, performance expectations, and a focus on creating and balancing value for customers and other stakeholders? Include communication and deployment through your leadership structure and to all employees.
- (2) How do senior leaders establish and reinforce an environment for empowerment and innovation, and encourage and support organizational and employee learning?
- (3) How do senior leaders set directions and seek future opportunities for your organization?

b. Organizational Performance Review
- (1) How do senior leaders review organizational performance and capabilities to assess organizational health, competitive performance, and progress relative to performance goals and changing organizational needs? Include the key performance measures regularly reviewed by your senior leaders.
- (2) How do you translate organizational performance review findings into priorities for improvement and opportunities for innovation?
- (3) What are your key recent performance review findings, priorities for improvement, and opportunities for innovation? How are they deployed throughout your organization and, as appropriate, to your suppliers/partners and key customers to ensure organizational alignment?
- (4) How do senior leaders use organizational performance review findings and employee feedback to improve their leadership effectiveness and the effectiveness of management throughout the organization?

This also means treating employees as leaders wish employees would treat customers. Leaders must express the same values in their dealings with front-line employees that they want these employees to show in their dealings with customers. Creating a positive climate for customer service means demonstrating concern for employees, enhancing their dignity, and solving their problems quickly and fairly. One of the most frustrating experiences of service employees is to work in an organization that stresses total customer satisfaction and then imposes controls or provides facilities that make it difficult or impossible to deliver it.

Clearly, an organization's leaders play a key role in setting directions, creating and balancing value for all stakeholders, and driving performance. Success requires a strong future orientation and a commitment to both improvement and change. As the Baldrige criteria note, increasingly this requires creating an environment for learning and innovation as well as the means for rapid and effective application of knowledge.

4. CREATING A SERVICE-DRIVEN WORKFORCE

Customers judge service by the quality of their interactions with the people who provide that service. The more contact employees have with customers, the more critical employee behavior is to perceptions of service quality. Therefore, organizations that lead in service quality must pay extraordinary attention to their workforce. Organizations with a strong customer focus achieve this through efforts in three key areas: (1) work and job design, compensation, career progression, and related workforce practices; (2) education and training efforts at all levels; and (3) maintenance of a work environment and an employee support climate that contribute to the well-being, satisfaction, and motivation of all employees.

The entire human resource focus category in the Baldrige criteria directly addresses all three of these areas that are critical to hiring, training, and growing service-driven employees. The criteria focus on how an organization enables its employees to use their full potential in accomplishing organizational objectives and how the organization maintains an environment that promotes performance excellence, full participation, and personal and organizational growth. Accomplishing this requires a close coupling between the leadership system and human resource system to enable defining and implementing the methods and structures needed to lead the organization in its strategic direction. The integration of these two systems sets the stage for effective performance. As leaders change the focus of the organization, employees need to acquire new skills and knowledge to be able to implement the leaders' new direction.

4.1. Work Systems

The work system starts with hiring people whose personalities predispose them to serve customers well. It continues with developing flexible, high-performance work practices tailored to employees with diverse workplace and home life needs. Major challenges in the area of valuing employees include demonstrating leadership's commitment to employees, providing recognition opportunities that go beyond the normal compensation system, providing opportunities for development and growth within the organization, sharing the organization's knowledge so employees can better serve customers and contribute to achieving strategic objectives, and creating an environment that encourages risk taking.

Companies that provide superior service use a wide variety of motivational programs to keep employees' energy flowing. The higher the degree of customer contact a group of employees has, the greater the number and power of motivational programs it can use. Award programs are one formal expression of the encouragement and praise that effective front-line supervisors continually provide. Compensation and recognition might need to be based, totally or in part, on leaders and employees attaining degrees of expertise in skill areas that align with organization objectives. Incentive approaches could include profit sharing and compensation based on acquiring new skills, building existing skills, or demonstrating self-learning. This might also be linked to customer retention or other performance objectives. Compensation and recognition, both monetary and nonmonetary, reward personnel for significant performance contributions that link to achieving the company objectives.

In addition to enabled employees and proper work system design, high-performance work requires ongoing education and training and information systems that ensure proper information flow. To help employees realize their full potential, many organizations use individual development plans for every employee that address individual career and learning objectives. Compensation and recognition approaches might also include profit sharing, team or unit performance, and linkage to customer satisfaction and loyalty measures or other organizational objectives.

4.2. Education and Training

Employee success depends increasingly on having opportunities for personal learning and practicing new skills. Organizations invest in employee personal learning through education, training, and opportunities for continuing growth. Learning opportunities in good service organizations are constant,

TABLE 3 Baldrige Criteria Item 5.1—Work Systems

5.1. Work systems

Describe how your organization's work and job design, compensation, career progression, and related work force practices enable employees to achieve high performance in your operations.

Within your response, include answers to the following questions:

a. Work Systems

(1) How do you design, organize, and manage work and jobs to promote cooperation and collaboration, individual initiative, innovation, and flexibility, and to keep current with business needs?

(2) How do your managers and supervisors encourage and motivate employees to develop and utilize their full potential? Include formal and/or informal mechanisms you use to encourage and support employees in job- and career-related development/learning objectives.

(3) How does your employee performance management system, including feedback to employees, support high performance?

(4) How do your compensation, recognition, and related reward/incentive practices reinforce high performance?

(5) How do you ensure effective communication, cooperation, and knowledge/skill sharing across work units, functions, and locations, as appropriate?

(6) How do you identify characteristics and skills needed by potential employees; how do you recruit and hire new employees? How do you take into account key performance requirements, diversity of your community, and fair work force practices?

intensive, lavish, and universal. Opportunities might include job rotation and increased pay for demonstrated knowledge and skills. On-the-job training offers a cost-effective way to train and better link training to organizational needs. Informal or not, the training is consistent and tightly linked with the company's strategy, culture, and personnel policies and is supported by the design of the service delivery system. Personal learning can result in more satisfied and versatile employees, greater opportunity for organizational cross-functional learning, and an improved environment for innovation.

Education, training, and development require ascertaining both education and training needs, including long-term development of employees' skills and knowledge that will be essential for an

TABLE 4 Baldrige Criteria Item 5.2—Employee Education, Training, and Development

5.2. Employee Education, Training, and Development

Describe how your organization's education and training support the achievement of your business objectives, build employee knowledge, skills, and capabilities, and contribute to improved employee performance.

Within your response, include answers to the following questions:

a. Employee Education, Training, and Development

(1) How does your education and training approach balance short- and longer-term organizational and employee needs, including development, learning, and career progression?

(2) How do you design education and training to keep current with business and individual needs? Include how job and organizational performance are used in education and training design and evaluation.

(3) How do you seek and use input from employees and their supervisors/managers on education and training needs, expectations, and design?

(4) How do you deliver and evaluate education and training? Include formal and informal education, training, and learning, as appropriate.

(5) How do you address key developmental and training needs, including diversity training, management/leadership development, new employee orientation, and safety, as appropriate?

(6) How do you address performance excellence in your education and training? Include how employees learn to use performance measurements, performance standards, skill standards, performance improvement, quality control methods, and benchmarking, as appropriate.

(7) How do you reinforce knowledge and skills on the job?

TABLE 5 Baldrige Criteria Item 5.3—Employee Well-Being and Satisfaction

5.3. Employee Well-Being and Satisfaction

Describe how your organization maintains a work environment and an employee support climate that contribute to the well-being, satisfaction, and motivation of all employees.

Within your response, include answers to the following questions:

- **a. Work Environment**
  How do you address and improve workplace health, safety, and ergonomic factors? How do employees take part in identifying these factors and in improving workplace safety? Include performance measures and/or targets for each key environmental factor. Also include significant differences, if any, based on different work environments for employee groups and/or work units.
- **b. Employee Support Climate**
  - (1) How do you enhance your employees' work climate via services, benefits, and policies? How are these enhancements selected and tailored to the needs of different categories and types of employees, and to individuals, as appropriate?
  - (2) How does your work climate consider and support the needs of a diverse work force?
- **c. Employee Satisfaction**
  - (1) How do you determine the key factors that affect employee well-being, satisfaction, and motivation?
  - (2) What formal and/or informal assessment methods and measures do you use to determine employee well-being, satisfaction, and motivation? How do you tailor these methods and measures to a diverse work force and to different categories and types of employees? How do you use other indicators such as employee turnover, absenteeism, grievances, and productivity to assess and improve employee well-being, satisfaction, and motivation?
  - (3) How do you relate assessment findings to key business results to identify work environment and employee support climate improvement priorities?

organization's future work systems. It also includes short-term needs of both the organization and its employees. Learning is directed not only toward better products and services but also toward being more responsive, adaptive, and efficient, giving the organization and its employees marketplace sustainability and performance advantages.

4.3. Employee Well-Being and Satisfaction

If work systems are to be effective, employees must have a suitable work environment and climate that fulfill their basic needs. Ensuring a safe and healthful environment must be part of an organization's improvement activities. Establishing measures and targets and recognizing that employee groups, might be in very different environments.

An organization must have approaches for enhancing employee well-being, satisfaction, and motivation based upon a holistic view of its entire workforce. A variety of approaches are usually necessary to satisfy a diverse workforce with differing needs and expectations. There must also be approaches for assessing employee well-being, satisfaction, and motivation and relating these assessment findings to key organizational results to set improvement priorities. Some examples of factors to consider in this assessment are effective employee problem and grievance resolution; employee development and career opportunities; work environment and management support; workload; communication, cooperation, and teamwork, job security; and organizational support for serving customers. Measuring how well employees understand the connection between quality and customers can also show what an organization needs to do to achieve an organizational customer focus.

5. SYSTEMS FOR SERVICE DELIVERY

As noted in the criteria's core value of "customer driven," quality and performance are judged by an organization's customers (Criteria 2000). Thus, an organization must take into account all product and service features and characteristics that contribute value to its customers and lead to customer satisfaction, preference, referral, and loyalty. Being customer driven has both current and future components—understanding today's customer desires and anticipating future customer desires and marketplace offerings. Value and satisfaction may be influenced by many factors throughout a customer's overall purchase, ownership, and service experiences. These factors include an organization's relationship with customers that helps build trust, confidence, and loyalty.

Without a clear understanding of customer expectations and priorities, an organization's leaders risk making bad decisions and resource allocations leading to poor quality as perceived by customers.

Information gathered, interpreted, and communicated properly can reduce the likelihood of pursuing inaccurate customer expectation priorities. Customer-driven organizations address not only the product and service characteristics that meet basic customer requirements, but also those features and characteristics that differentiate products and services from competing offerings.

The customer and market focus category in the criteria examines how an organization determines requirements, expectations, and preferences of customers and markets. It also considers how an organization builds relationships with customers and determines their satisfaction. The category looks at how an organization seeks to understand the voices of customers and the marketplace and stresses relationships as an important part of an overall listening, learning, and performance excellence strategy.

The criteria asks an organization to consider its key processes for gaining knowledge about its current and future customers and markets, with the aim of offering relevant products and services, understanding emerging customer requirements and expectations, and keeping pace with changing markets and marketplaces. Listening and learning strategies depend on an organization's key business factors. Some frequently used strategies include focus groups with key customers; close integration with key customers; interviews with lost customers about their purchase decisions; use of the customer complaint process to understand key product and service attributes; won/lost analysis relative to competitors; and survey/feedback information, including use of the Internet. Involving the workforce in identifying customers and their needs and communicating that information throughout an organization reinforces the connection among employees, customers, and quality.

Service organizations often get process specifications by creating customer commitments. Customer commitments are statements of outcomes expected in nonmanufacturing situations. These commitments are designed to satisfy customers' needs and expectations, written to include a little something extra to distinguish an organization from competitors. Customer commitments are based on everyone's inherent general understanding of customer needs, combined with actual knowledge of what the customer wants. For commitments to be as effective as specifications, all employees must know who their customers are; the specific expectations of these customers must be written down; and employees must be empowered by management to do whatever is necessary to meet these expectations.

Customer commitments are frequently translated into service standards. Setting a standard of service that everyone can understand establishes a target. This is critical to success because when written service standards are published, it establishes a goal toward which all employees can aim. People need tangible goals that communicate expectations and let employees know what is expected of them in specific and measurable terms. The process of setting service standards lets all employees, from the top down, know who is responsible for what jobs and reduces conflicts and misunderstandings on who is going to do what.

The second criteria item in the customer and market focus category deals with customer satisfaction and relationships. Customer satisfaction and dissatisfaction results provide vital information for understanding an organization's customers and the marketplace. In many cases, such results and trends

TABLE 6 Baldrige Criteria Item 3.1—Customer and Market Knowledge

---

3.1. Customer and Market Knowledge

Describe how your organization determines short- and longer-term requirements, expectations, and preferences of customers and markets to ensure the relevance of current products/services and to develop new opportunities.

Within your response, include answers to the following questions:

a. Customer and Market Knowledge

(1) How do you determine or target customers, customer groups, and/or market segments? How do you consider customers of competitors and other potential customers and/or markets in this determination?

(2) How do you listen and learn to determine key requirements and drivers of purchase decisions for current, former, and potential customers? If determination methods differ for different customers and/or customer groups, include the key differences.

(3) How do you determine and/or project key product/service features and their relative importance/value to customers for purposes of current and future marketing, product planning, and other business developments, as appropriate? How do you use relevant information from current and former customers, including marketing/sales information, customer retention, won/lost analysis, and complaints, in this determination?

(4) How do you keep your listening and learning methods current with business needs and directions?

TABLE 7 Baldrige Criteria Item 3.2—Customer Satisfaction and Relationships

3.2. Customer Satisfaction and Relationships
Describe how your organization determines the satisfaction of customers and builds relationships to retain current business and to develop new opportunities.
Within your response, include answers to the following questions:

a. Customer Relationships
  (1) How do you determine key access mechanisms to facilitate the ability of customers to conduct business, seek assistance and information, and make complaints? Include a summary of your key mechanisms.
  (2) How do you determine key customer contact requirements and deploy these requirements to all employees involved in the response chain?
  (3) What is your complaint management process? Include how you ensure that complaints are resolved effectively and promptly, and that all complaints received are aggregated and analyzed for use in overall organizational improvement.
  (4) How do you build relationships with customers for repeat business and/or positive referral?
  (5) How do you keep your approaches to customer access and relationships current with business needs and directions?

b. Customer Satisfaction Determination
  (1) What processes, measurement methods, and data do you use to determine customer satisfaction and dissatisfaction? Include how your measurements capture actionable information that reflects customers' future business and/or potential for positive referral. Also include any significant differences in processes or methods for different customer groups and/or market segments.
  (2) How do you follow up with customers on products/services and recent transactions to receive prompt and actionable feedback?
  (3) How do you obtain and use information on customer satisfaction relative to competitors and/or benchmarks, as appropriate?
  (4) How do you keep your approaches to satisfaction determination current with business needs and directions?

provide the most meaningful information, not only on customers' views but also on their marketplace behaviors—repeat business and positive referrals.

Obtaining useful information on customer satisfaction requires carefully designed and developed surveys. Responses must be interpretable and translatable into specifications or strategies for improvement. Surveys should identify and prioritize customer needs and expectations and should rate the importance of the product or service attributes and how well the organization is meeting the attributes. Complaint and satisfaction survey results aggregation, analysis, and root cause determination should lead to effective elimination of the causes of problems and to priority setting for process, product, and service improvements. Successful outcomes require effective deployment of information throughout an organization.

The item also deals with how an organization provides easy access for customers and potential customers to seek information or assistance and/or to comment and complain. Excellent service depends on welcoming customer requests and responding flexibly. Organizations keep customers by remedying commitment failures. To be able to fix a commitment failure on the spot, an organization must anticipate possible failures and devise suitable remedies. These plans must be mapped out in advance, written down, and agreed to by management, and the workforce must be empowered to act on them.

Having an effective, customer-driven system for service delivery is a strategic concept. It is directed toward customer retention, market share gain, and growth. It demands constant sensitivity to changing and emerging customer and market requirements and the factors that drive customer satisfaction and retention. It demands anticipating changes in the marketplace. Being customer driven thus demands awareness of developments in technology and competitors' offerings and rapid and flexible response to customer and market requirements.

6. MEASURING AND EVALUATING SERVICE QUALITY

Measuring and evaluating service quality rounds out the five key areas that that began with creating a service strategy. Through measurement and evaluation, leadership can understand how well their strategies are working and can identify weaknesses in the other elements of service. Measuring and tracking also allows employees at all levels of an organization to increase service quality awareness, understand their performance levels, understand performance relative to competitors and benchmarks, identify strengths and weaknesses, focus efforts, and monitor progress.

Despite the serious difficulties of measuring and controlling the quality of customer service, service leaders have figured out ways of doing it. Measuring service performance or quality is quite different than measuring product quality because service is an experience. Generally, the better measurement systems focus on three different aspects of service: process, product, and customer satisfaction.

Process measures are the more traditional measures, which compare the actual work employees perform in the process of creating service with standards for quality and quantity. These might include measures such as time to answer the phone, time to respond to customer inquiries, and ability to solve a problem without passing a customer on to somebody else. These measures are often associated with the service standards an organization defines. Measuring service standards provides employees with a sense of achievement and accomplishment and motivates them to aim at even higher levels.

Product measures focus on outcomes of a service process and show whether the work has produced the desired results, such as delivering packages when customers want them delivered. Ideally, product measures summarize the effects of the process from the customers' point of view (without actually asking the customer first hand) and are closely linked to customer satisfaction. Examples of such measures include access, courtesy, reliability, and responsiveness.

Satisfaction measures look at the extent to which customers are satisfied with the service they have received. Customer satisfaction must be measured for each area of customer contact that affects the customer's decision to buy. In order to get a full understanding of the benefits and costs of service quality, customer satisfaction surveys should ask how satisfied customers are with a service encounter, what problems were experienced, whether assistance to answer a question or solve a problem was sought and where the customer sought advice, how many people the customer told about the experience, costs incurred by both servers and customers to prevent and correct poor service, and whether the customer intends to purchase the product or service again.

Zeithaml et al. (1990) proposes five attributes of quality service, including reliability, empathy, assurance, responsiveness, and tangibles. By comparing customer perceptions with expectations in these areas, the model provides a two-part measure of perceived quality that links back to market segmentation relative to different service expectations. Whatever measures are used, measurement and evaluation are necessary to establish bases for improvement. The initial fix of a customer commitment failure is really just a temporary fix to keep customers satisfied. Once good data is available, employees can effectively identify and address the real problems behind commitment failures.

The data obtained from service quality measures may be used in a variety of ways: to correct specific performance deficiencies, to identify problems for correction, and to supply data to a variety of economic models on the profit impact of causes of dissatisfaction and the revenue opportunities of sources of satisfaction. Whatever methods are used to obtain the data, service quality results must reach all of the concerned areas of the organization, and they must be evaluated and needed to balance customer satisfaction and company profits. When customer expectations are met or exceeded, they have a tendency to increase, and service-obsessed companies never see the end of programs and actions to improve service.

There is an entire category devoted to information and analysis in the Baldrige criteria, and a core value centered around management by fact, demonstrating the extent to which organizations depend upon the measurement and analysis of performance. The measures and indicators an organization selects should be represent the factors that lead to improved customer, operational, and financial performance. A comprehensive set of measures or indicators tied to customer and/or organizational performance requirements represents a clear basis for aligning all activities with an organization's goals (Criteria 2000).

There are no magic formulas for developing measures, and there is no one recipe for success in monitoring service quality. Lacking good measures, however, an organization cannot assess its progress or adjust to changes in customer expectations. Without effective measures, managers cannot reward employees appropriately, tune strategies and infrastructure to customer needs, or design products and service delivery systems that support outstanding service. Without valid measurement systems, it is impossible to know what actions are required to improve customer service. There must be a continuous, permanent commitment to measuring, evaluating, and acting on actual service delivery.

7. CONCLUSIONS

Organizations that provide outstanding service quality are quite different from their competitors. Leaders act and manage differently, missions are stated differently, employees are treated differently, services are delivered differently, and results are measured and acted upon. Actions are based on totally different assumptions about the way success is achieved. And the results show it, in terms of both conventional measures of performance and the impact these services have on their competitors.

Genuine quality improvements that make a difference to customers and build a stronger, more competitive organization are achievable by any type of organization—manufacturing, service, health care, education, and nonprofit alike. With a framework such as the Baldrige criteria, any organization can systematically identify both strengths and areas for improvement in its own approaches to service quality. Whether being used on an entire corporation, division, or a small service or manufacturing company, the criteria provide a methodical approach for creating a common language and enhancing cooperation and communication with employees and customers alike. Using the criteria can help achieve consensus on what needs to be done and help integrate various quality management and business efforts.

Service quality is definable, achievable, and sustainable when an organization is provided with a systematic approach for self-assessing and improving existing approaches to service delivery. The Baldrige criteria provide a robust, organizational management model, which has been repeatedly used by all types of organizations to gain competitive advantage, enhance their image, improve their overall performance, and ultimately help them achieve excellence.

REFERENCES

Criteria for Performance Excellence, (2000), Baldrige National Quality Program, National Institute of Standards and Technology, Gaithersburg, MD.*

Carlzon, J. (1987), *Moments of Truth*, Harper & Row, New York.

Zeithaml, V. Parasuraman, A., and Berry, L. L., (1990) *Delivering Service Quality: Balancing Customer Perceptions and Expectations*, Free Press, New York.

---

*Individual copies can be obtained from:
Baldrige National Quality Program
National Institute of Standards and Technology
Administration Building, Room A635
100 Bureau Drive, Stop 1020
Gaithersburg, MD 20899-1020
Telephone: (301)975-2036
E-mail: nqp@nist.gov

CHAPTER 74

Standardization, Certification, and Stretch Criteria

Harrison M. Wadsworth, Jr.
Georgia Institute of Technology

1. INTRODUCTION **1966**
- 1.1. Definitions 1966
- 1.2. Reasons for Quality Management System Standards 1967

2. HISTORY OF QUALITY MANAGEMENT SYSTEM STANDARDS **1967**
- 2.1. American Standards 1967
- 2.2. European Standards 1968
- 2.3. International Standards 1968

3. DETAILS OF THE ISO 9001:2000 STANDARD **1969**
- 3.1. Process Model 1969
- 3.2. Scope 1969
- 3.3. Principles of Quality Management 1969
- 3.4. QMS Requirement 1970
- 3.5. Management Responsibility 1970
- 3.6. Resource Management 1971
- 3.7. Product Realization 1971
- 3.8. Measurement, Analysis, and Improvement 1971

4. ISO 9004-21972 **1972**

5. OTHER STANDARDS AND GUIDES IN THE ISO 9000 FAMILY **1972**

6. STANDARDS DEVELOPED BY OTHER ORGANIZATIONS **1973**
- 6.1. ANSI/ASQ/Z1.11, *The Application of ISO 9001 to Educational Institutions* 1973
- 6.2. QS 9000, *Quality System Requirements* 1973
- 6.3. Other Quality Management Requirement Documents 1973

7. REGISTRATION TO QMS STANDARDS **1973**

REFERENCES **1974**

1. INTRODUCTION

This chapter will discuss international and national standards for quality management systems. After defining some terms and concepts, it will consider the need for such standards and how they are used.

The history and evolution of these standards will then be presented, leading up to the current ISO standards, which have revolutionized quality management. The requirements in the latest version (2000) of the QMS standards will be presented. The topic of registration to these standards will be discussed. This includes reasons for registration and means for becoming registered. Similar standards for environmental management, the ISO 14000 family of standards, were discussed in Chapter 39 of this handbook.

1.1. Definitions

The definitions given here are taken from ISO/DIS 9000-2000, which is the latest version of the terminology standard as of the writing of this chapter. This standard will be discussed later. It su-

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

persedes and replaces the previous terminology standard, ISO 8402-1994, and its American counterpart, ANSI/ISO/ASQ A8402-1994.

> A quality management system is defined in this standard as a ‘‘system to establish quality policy and quality objectives and to achieve those objectives’’ (2.2.3).
>
> Quality policy is defined in the standard as the ‘‘overall intentions and direction of an organization related to quality as formally expressed by top management’’ (2.2.4).
>
> Quality objective is defined as ‘‘something sought, or aimed at, related to quality’’ (2.2.5).

Other terms that are useful for the presentation given in this chapter are:

- Quality management: ‘‘coordinated activities to direct and control an organization with regard to quality’’ (2.2.8)
- Quality control: ‘‘part of quality management focused on fulfilling quality requirements’’ (2.2.10)
- Quality assurance: ‘‘part of quality management focused on providing confidence that quality requirements are fulfilled’’ (2.2.11)

1.2. Reasons for Quality Management System Standards

There are several reasons for standards for quality management systems. As indicated, quality assurance involves the demonstration of good-quality products and services. The quality management system is the organization that enables this assurance to be accomplished. Quality management system standards provide a guideline for a manufacturing or service company to determine if proper quality assurance requirements can be met.

The first reason for such standards is to provide these guidelines. While each organization is different and thus has different needs for quality assurance, there are many elements that are fundamental to all organizations. These elements can be standardized and are found in such standards. This provides internal quality assurance.

In addition, quality management system standards provide a means to evaluate an organization’s quality management system and thus its ability to provide high-quality products. This is the provision of external quality assurance by a supplier of a product or service. In an increasing number of situations, customers require, for contractual purposes, the registration or certification of their suppliers’ quality systems. Such registration must use standards for guides.

As we move into a global marketplace, we find the need for standardization increases. Without such internationally recognized standardization we cannot communicate with suppliers or customers in other countries.

These needs for standardization have been recognized by international and national standards writing bodies. The present standards for quality systems and quality assurance are the results of this recognition. More will be said about the development of such standards in the next section.

2. HISTORY OF QUALITY MANAGEMENT SYSTEM STANDARDS

2.1. American Standards

For the marketplace, standardization is a necessity. Product standards, monetary standards, measurement standards, and so forth have been with us for thousands of years. This chapter does not address these standards, however. It addresses generic quality standards that are not product or industry specific. The earliest quality management system standards were developed for contractual purposes. Such standards, for example MIL-Q-9858A (1963) and MIL-I-45208A (1981), were developed by the United States Department of Defense (DoD). These are mandatory standards created for the purpose of assuring the defense procurement agencies of the United States that a supplier has the ability to provide high-quality weapons systems. Other countries developed similar standards for quality management systems.

In the United States, voluntary national standards (as opposed to the mandatory standards) have been published by several organizations. The most prominent publishers of such standards are the American Society for Testing Materials (ASTM), the American Society of Mechanical Engineers (ASME), the Institute of Electrical and Electronic Engineers (IEEE), the American Society for Quality (ASQ), and others. All of these technical societies publish many of their standards under the auspices of the American National Standards Institute (ANSI). ANSI approves such standards, and most of them carry a dual designation of both the writing organization and ANSI. There are currently more than 11,000 ANSI standards, most of which are product-specific standards. ANSI itself consists of individual members, approximately 1,000 companies, 30 government agencies, and 250 professional, technical, trade, labor, and consumer organizations.

ANSI is the sole U.S. representative to international standards writing bodies such as the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), and the Pacific Area Standards Congress (PASC). To facilitate the development of standards dealing with quality assurance and quality control, the American Society for Quality was accredited as a standards-writing body in 1974 by ANSI, with its Standards Committee charged with promoting such standards. Later, in 1978, the ANSI Accredited Standards Committee Z1 was established, with the ASQ holding its Secretariat.

One of the early voluntary quality system standards developed in the United States was ANSI Standard Z1.8-1971, *Specification of General Requirements for a Quality Program.* This was written by the ASQ and later revised as ANSI/ASQC C1 in 1985. A later, much more detailed standard was published by the ASQ as ANSI/ASQC Z1.15-1979, *Generic Guidelines for Quality Systems*. This latter standard was written by a subcommittee of the ANSI Accredited Standards Committee Z1. It was the basis for international standard ISO 9004, published in 1987. In the United States it has been superseded by the American version of ISO 9004, which is designated ANSI/ISO/ASQ Q9004. The third revision of this standard has now been completed and published as ISO 9004-2000. Its American counterpart will be ANSI/ISO/ASQ Q9004-2000. This edition of the standard will be reviewed later in this chapter.

2.2. European Standards

Other single-level standards similar to MIL-Q 9858A and ANSI/ASQC Z1.15 were being developed at the same time by other countries. A partial list of such early standards is:

Country	Standard
United Kingdom	BS4891:1972, *A Guide to Quality Assurance*
France	AFNOR NFX50-110, *Recommendations for a System of Quality Management for the Use of Companies*
France	AFNOR NFX50-111, *Quality Assurance Systems for the Use of Companies*
Germany	DIN 55-355, *Basic Elements of Quality Assurance Systems*

Some standards developed during this time were multilevel standards. That is, they have several levels of quality assurance standards that depend on the requirements placed on the supplier. They were chiefly developed for use in contractual situations. The following are some of these standards:

Country	Standard
United Kingdom	BS 5750-1979, *Specification for Design, Manufacture, and Installation*
Canada	CSA Z299-1978, *Quality Assurance Program Requirements*
Norway	NVS-S-1594, *Requirements of the Contractor's Quality Assurance Program*
South Africa	SABS 0157-1979, *Code of Practice for Quality Management Systems*
Australia	AS 1821-1975, *Suppliers Quality Control System*

2.3. International Standards

In 1979, ISO formed Technical Committee 176, with Secretariat given to Canada. The scope of this Technical Committee, as stated in the 1989 Momento of ISO, is "Standardization in the field of generic quality management, including quality systems, quality assurance, and generic supporting technologies, including standards which provide guidance on the selection and use of these standards." The Technical Committee published six standards in 1986-1987. Since then, other standards have been published, bringing the total to 27 as of 1999. The six original standards were the principal accomplishments of the committee. The first of the six standards published in 1986 was ISO 8402, *Vocabulary*. This standard consisted of 22 terms. It was revised in 1994 with 67 terms. The 2000 edition of the standard, now designated as ISO 9000-2000, has 87 terms. As with the 1994 edition, the terms are divided into 10 sections as follows:

- Seven terms related to quality
- Fourteen terms related to management
- Seven terms related to organization
- Eight terms related to process and product
- Five terms related to characteristics
- Thirteen terms related to conformity
- Six terms related to document
- Six terms related to examination

- Thirteen terms related to audit
- Eight terms related to quality assurance for measurement processes

Many of the standards in the ISO 9000 family will be discontinued following publication of the new edition of the basic standards, ISO 9000, ISO 9001, and ISO 9004. The original 9002 and 9003 standards will be merged with ISO 9001. ISO 9000-1 will be merged with 8402 and designated 9000. ISO 9004-1, 9004-2, 9004-3, and 9004-4 will be merged and designated ISO 9004. ISO 10011-1, 10011-2, and 10011-3 on auditing have been merged with auditing standards for environmental management. The new standard is designated 19011. The remaining members of the ISO 9000 family are as follows:

Core standards
ISO 9000:2000, *Quality Management Systems—Fundamentals and Vocabulary*
ISO 9001:2000, *Quality Management Systems—Requirements*
ISO 9004:2000, *Quality Management Systems—Guidelines for Performance Improvements*
ISO 19011, *Guidelines for Auditing Management Systems*

Other international standards
ISO 10012, *Guidelines for Quality Assurance for Measuring Equipment*
ISO 10015, *Quality Management—Guidelines for Training*

Technical reports
ISO 10006, *Guidelines to Quality in Project Management*
ISO 10007, *Guidelines for Configuration Management*
ISO 10013, *Guidelines for Developing Quality Manuals*
ISO 10014, *Guidelines for Managing the Economics of Quality*
ISO 10017, *Guidance on Statistical Techniques for ISO 9001*

The only standard in this family that may be used for registration of a quality management system is ISO 9001. In previous editions there were three such requirements standards, ISO 9001, 9002, and 9003. The 9001 standard was the most complete in that it covered all parts of a quality management system. ISO 9002 was identical to 9001 except that the design function was not included. Several other functions were omitted from the 9003 standard. The 2000 edition of ISO 9001 allows for the tailoring of the standard to delete requirements such as design when they do not apply to an organization.

3. DETAILS OF THE ISO 9001:2000 STANDARD

3.1. Process Model

This edition of the standard is based on a process approach to quality management. Any activity that receives inputs and converts them to outputs is a process. Most processes are linked in that outputs from one process are often the inputs to other processes. The systematic identification and management of the processes employed within an organization is the process approach.

This edition of ISO 9001 is considered as one part of a consistent pair of standards, the other being ISO 9004:2000. The two standards have identical clause structures, an important property that was missing in earlier editions. ISO 9004:2000 gives guidance on a wider application of a quality management system meant to improve the organization's overall performance beyond that required by ISO 9001:2000. It is not a guideline for implementing ISO 9001:2000 and is not intended for certification or other contractual use. The ISO 9001:2000 standard also has an identical clause structure with ISO 14001:1996 in order to improve the compatibility between registration of quality and environmental management systems.

3.2. Scope

ISO 9001:2000 specifies requirements for a quality management system where an organization needs to demonstrate its ability to consistently provide product that meets requirements of customers and regulatory agencies. It also addresses customer satisfaction through the requirements for continual improvement and the prevention of nonconformities.

3.3. Principles of Quality Management

ISO 9000:2000 lists the following steps that might be used to develop a quality management system:

1. Determine the needs of the customer.
2. Establish the quality policy and quality objectives of the organization.
3. Determine the processes needed to implement the quality objectives.
4. Develop measures for the effectiveness of each process towards the attainment of the objectives.
5. Develop means of preventing nonconformities.
6. Look for opportunities to improve the effectiveness and efficiency of processes.
7. Determine and prioritize proposed improvements.
8. Plan strategies, processes, and resources to obtain improvements.
9. Implement the plan.
10. Monitor the improvements.
11. Assess the results against expected outcomes.
12. Determine follow-up actions.

An organization that adopts the 12 steps outlined above creates confidence in the capability of its processes and provides a basis for continual improvement. This, in turn, leads to increased customer satisfaction and the success of the organization.

3.4. QMS Requirements

Clause 4 of ISO 9001:2000 presents some general requirements of the quality management system. These include requirements dealing with the identification and management of the processes in the system. General documentation requirements are also discussed in this clause. These include the requirements for documented procedures, work instructions, and manuals. The actual, detailed requirements of the standard are stated in four clauses:

- Management Responsibility
- Resource Management
- Product Realization
- Measurement, Analysis, and Improvement

Each of these clauses is discussed briefly in the following sections.

3.5. Management Responsibility

Clause 5 of the standard is entitled ″Management Responsibility.″ It designates a requirement for management commitment to high quality. It states that this management commitment must be highly visible. This is done by the establishment of a quality policy, setting of quality objectives, and a description of the quality system that emphasizes problem prevention rather than dependence on detection after a problem occurs.

Management must develop and state its corporate policy as it relates to quality. This quality policy must be consistent with all other corporate policies. It is the responsibility of management to ensure that its quality policy is understood, implemented, and maintained. The policy contains management's definition of good quality and its goals for quality improvement.

Quality objectives must be explicitly stated. These include the key elements of quality, such as fitness for use, performance, safety, and reliability.

Management's responsibilities also include the organization and operation of the quality system. This responsibility includes the provision of all necessary resources. Management is responsible for seeing that the quality system functions in such a manner that the system is well understood and effective, confidence is provided that products or services satisfy requirements and customer expectations and that its emphasis is on prevention of problems rather than problem detection.

The clause also requires the appointment of a management representative charged with the operation and maintenance of the quality system. The management representative must have direct access to top management. A manual must be developed by management that describes the quality system and includes references to appropriate documented procedures. The manual and the procedures must be controlled using a document control procedure.

Top management shall review the quality management system at prescribed intervals to ensure its continuing effectiveness. These reviews shall include results of audits, customer feedback, process performance and product conformance, status of corrective and preventive actions, and changes that could affect the quality management system.

3.6. Resource Management

Clause 6 of the standard requires the provision of resources needed to maintain the quality management system. These include human resources, requiring the identification of competency needs for personnel, the provision of training to satisfy these needs, evaluation of the effectiveness of this training, and the maintenance of appropriate records of education, training, and qualifications of all personnel. There is also a requirement for the maintenance of a proper work environment and all facilities needed to provide high-quality products and services.

3.7. Product Realization

The last two clauses, 7 and 8, are large and include a number of subclauses. The first subclause of clause 7 deals with quality planning. This requires a statement of the quality objectives for the product or project and the processes needed. These plans are to be recorded in the form of quality plans, and they include any inspection or tests needed to verify the product quality. The second clause, 7.2, requires the organization to determine all customer requirements. It also requires a review of product requirements prior to a commitment to produce a product to ensure the organization is able to meet the requirements. The organization is also required to implement arrangements for adequate communication with the customer, including customer feedback or complaints.

Clause 7.3 deals with the design of a product or service. This clause contains all the requirements of earlier editions of the standard regarding design and development. An appropriate planning activity is required that controls the entire design process. Design input states that inputs must be defined and documented, including all functional, regulatory, and legal requirements. The organization must document that knowledge from prior design activities is used when it is available. Design output must be such that verification against input requirements is feasible. The output must give the necessary requirements for production operations, address product acceptance criteria, and explain how the product may be used safely and correctly. The output documents must be approved before publication.

The standard requires design reviews at suitable stages using a review team made up of representatives of all functions associated with the product. There must be a verification stage during which the output is matched to the design input. There must also be a validation stage during which the final product's performance is compared to requirements. Finally, all design changes must be documented and controlled.

Clause 7.4 covers the control of the purchasing function to ensure that purchased product meets the specified requirements. Criteria for the selection and periodic evaluation of suppliers shall be defined and results recorded. The organization shall identify and implement all activities necessary for the verification of purchased product.

Clause 7.5 is entitled Production and Service Operations. The first subclause in this section requires the control of operations relevant to production and service by making available information including all specifications about the product, work instructions, devices needed for measuring and monitoring, and instructions for release and post delivery activities. The second subclause requires product to be identified and traceable throughout production. Clause 7.5.3 requires the organization to protect and control any customer-owned property used in production. This includes customer-owned tooling, shipping containers, and intellectual property that may be provided in confidence.

Clause 7.5.4 requires the organization preserve the property during all internal processing and delivery to the final destination. Subclause 7.5.5 requires the validation of all production and service processes that cannot be verified by subsequent measurement or monitoring. This includes any processes where deficiencies may not be detected until after the product is put in use. The last subclause, 7.5.6, in this section deals with the control of measuring and monitoring devices. The devices must be controlled, serviced at regular intervals, and protected from damage. The results of calibration must be recorded and corrective action taken whenever a device is found to be out of calibration.

3.8. Measurement, Analysis, and Improvement

Clause 8 contains five subclauses, the first of which, 8.1, states that the organization shall define, plan, and implement the measuring and monitoring activities needed to ensure conformity and improvement. Clause 8.2 on measurement and monitoring requires the organization to measure customer satisfaction, conduct internal audits, and measure and monitor all processes and the product. The internal audits must be performed at regular intervals and must assess whether the quality management system is effective and conforms to the standard.

Clause 8.3 is on control of nonconformity. This is a requirement to identify and control product that is in nonconformance and to take appropriate action to see that such product does not reach the customer. The next subclause requires data dealing with the quality system to be collected and analyzed. These data include customer satisfaction and dissatisfaction, conformance to requirements, characteristics of processes and products, and suppliers.

Clause 8.5 requires the organization to plan for continual improvement of the quality system through the ‘‘use of quality policy, objectives, audit results, analysis of data, corrective and preventive action, and management review.’’ The last two subclauses in clause 8.5 are entitled Corrective and Preventive Action. These two requirements were a single clause in the 1994 edition of the standard. This led to some confusion among users of the standard as to the difference between them. Corrective action includes the correction of problems that have occurred, whereas preventive action deals with potential problems that may have never occurred.

4. ISO 9004-2000

ISO 9004-2000 is a guidance document that provides more information regarding the quality management system. As stated earlier, the clause structure of this standard is the same as that of 9001-2000. The title of this standard is *Quality Management Systems—Guidelines for Performance Improvements.* It is, however, based on the same quality management principles as ISO 9001-2000. The focus of this standard is the improvement of the processes of an organization in order to enhance its performance.

When ISO 9001-2000 and ISO 9004-2000 are used together, the benefits to an organization are likely to be greater than if only one is used. The two standards have identical structures but different scopes. ISO 9004 is not intended to be used as guidance document for compliance to ISO 9001. The purpose of ISO 9001 is to define the minimum requirements needed to achieve customer satisfaction by meeting specified product requirements. The purpose of ISO 9004 is to provide guidance on the application and use of a quality management system to improve the overall performance of an organization.

The clauses of ISO 9001-2000 are included in a box within the clauses of 9004 for immediate reference for users of ISO 9004. The verb used in ISO 9001 is ‘‘shall,’’ whereas that used in 9004 is ‘‘should.’’ The clauses of ISO 9004-2000 are based on the following eight quality management principles: customer focus, leadership, involvement of people, process approach, system approach to management, continual improvement, factual approach to decision making, and mutually beneficial supplier relationships.

A clause in the measurement and monitoring section provides a means for self- assessment of an organization’s quality system. The actual methodology along with appropriate questions to be answered are included in an annex to the standard.

5. OTHER STANDARDS AND GUIDES IN THE ISO 9000 FAMILY

In addition to the three standards discussed above, ISO 9000-2000, 9001-2000, and 9004-2000, there are three other standards, five technical reports, and three brochures in the 2000 edition of the ISO 9000 family. Some of these documents have not been revised to agree with the 2000 editions of the basic standards at the time of the publication of this chapter. The three additional standards are:

- ISO 19011, *Guidelines for Auditing Management Systems*
- ISO 10012, *Quality Assurance Requirements for Measuring Equipment*
- ISO 10015, *Quality Management–Guidelines for Training*

The five technical reports are:

- ISO 10006, *Quality Management–Guidelines to Quality in Project Management*
- ISO 10007, *Quality Management–Guidelines for Configuration Management*
- ISO 10013, *Guidelines for Developing Quality Manuals*
- ISO/TR 10014, *Guidelines for Managing the Economics of Quality*
- ISO/TR 10017, *Guidance on Statistical Techniques for ISO 9001-1994*

The three brochures are:

- *Quality Management Principles and Quidelines on Their Application*
- *Selection and Use of Standards*
- *Small businesses*

Two other standards were formerly in the 9000 family. The first of these is ISO 9000-3:1997, *Quality Management and Quality Assurance Standards*—Part 3: Guidelines for the Application of ISO 9001:1994 to the Development, Supply, Installation, and Maintenance of Computer Software. This standard was transferred to ISO/IEC JTC/SC7, who will update it to correspond to the 2000 edition of ISO 9001. The second is ISO 9000-4:1993, *Quality Management and Quality Assurance*

Standards– Part 4: Guide to Dependability Program Management. This standard has been transferred to IEC/TC 56, Dependability, who is updating it for the 2000 edition of ISO 9001. The following standards in the 9000 family have been dropped:

- ISO 8402:1994, *Vocabulary*—now part of ISO 9000-2000
- ISO 9000-1:1994, *Selection and Use of ISO 9000 Standards*—now part of ISO 9000-2000
- ISO 9002:1994, *Model for Quality Assurance*—now part of ISO 9001-2000
- ISO 9003:1994, *Model for Quality Assurance*—now part of ISO 9001-2000
- ISO 9004-1:1994, *Quality Management and Quality System Elements*—Part 1: Guidelines, now designated as ISO 9004-2000
- ISO 9004-2:1991, *Quality Management and Quality System Elements*—Part 2: Guidelines for Services—now included in ISO 9004-2000
- ISO 9004-3:1993, *Quality Management and Quality System Elements*—Part 3: Guidelines for Processed Materials—now included in ISO 9004-2000
- ISO 9004-4:1993, *Quality Management and Quality System Elements*—Part 4: Guidelines for Quality Improvement—now included in ISO 9004-2000
- ISO 10005:1995, *Quality Management—Guidelines for Quality Plans*—now included in ISO 10013

6. STANDARDS DEVELOPED BY OTHER ORGANIZATIONS

6.1. ANSI/ASQ/Z1.11, *The Application of ISO 9001 to Educational Institutions*

This standard is an American National Standard, published by the American Society for Quality. The standard, now being updated to the 2000 edition of ISO 9001, was written to assist educational institutions conform to the requirements of ISO 9001. The current edition contains the ISO 9001: 1994 standard in the left column of each page, with the corresponding clauses of ANSI/ASQ/Z1.11 on the right. The Z1.11 standard puts the 9001 standard in terms recognizable to persons dealing with education.

6.2. QS 9000, *Quality System Requirements*, 3d Ed., 1998

This document is not really a standard, in that it was not developed using a consensus procedure. It is a set of requirements, in addition to the ISO 9001 requirements, that have been developed by the Automotive Industry Action Group. The actual set of requirements were developed by representatives of Chrysler, Ford, and General Motors in 1994. The second edition was published in 1995 and the third edition published in 1998. The clause structure corresponds to that of ISO 9001:1994. In fact the ISO standard is printed in italics within each clause, while the additional requirements are in roman letters. Since 1994, the document has been printed in five languages in at least 63 countries. The original equipment manufacturers in the automotive industry have required their suppliers to be registered to this document in order to sell their product to them.

The additional requirements of this document consist of additional statistical requirements, such as requirements for statistical process control, gage repeatability and reproducibility studies, and failure mode and effects analyses. Other requirements imposed by the automotive group are also included. Most of the additional requirements are discussed in a set of supplementary manuals entitled *Statistical Process Control Reference Manual, Production Part Approval Process Reference Manual, Measurement Systems Analysis Reference Manual, Advanced Product Quality Planning and Control Plan Reference Manual,* and the *Failure Mode and Effects Analysis Reference Manual.* Organizations wishing to be registered to QS 9000 must be familiar with all of the supplementary manuals as well as the basic set of requirements. QS 9000 is to be replaced in 2000 by ISO/TS 16949, which is the result of cooperation of both European and American automotive industry groups.

6.3. Other Quality Management Requirement Documents

Taking an approach similar to that taken by the automotive industry, the aerospace industry has developed a set of requirements entitled AS 9100 and the telecommunication industry has developed TL 9000. Both of these documents are based on ISO 9001:1994 and contain additional requirements considered necessary for quality management systems in their industry.

7. REGISTRATION TO QMS STANDARDS

Since the original adoption of the ISO 9000 standards in 1987, the registration of companies to the standards has become an accepted practice throughout the world. As of the end of 1998, there were 271,966 certificates of conformance issued in 143 countries on every continent, according to data released by the International Organization for Standardization. The number of certified organizations has increased rapidly over the years, indicating that the interest in registration is increasing. As an indication of the growth rate, the September 1999 issue of *Quality Systems Update* reports there were 27,816 certificates in January 1993; 70,364 certificates in June 1994; 127,353 in December 1995; 162,704 in December 1996; and 223,403 certificates in December 1997. The 1998 data also show there were 7,887 registrations to ISO 14001, the environmental management standard.

These registrations represent every possible industry, from service industries to heavy manufacturing organizations. According to the ISO report, there were 573 registrars listed in 1999. This number had grown from 309 in 1995. During the same period the number of countries with registrars grew from 73 to 93. The directory lists 52 registrars operating in the United States. The number of accreditation agencies that accredit the registrars has increased from 33 in 1995 to 40 in 1999. Usually there is no more than one accreditation agency in any country. Some accreditation agencies operate in several countries.

Of the more than 270,000 registered companies, 36,653 are from the electrical and optical equipment industries, 28,885 from the basic metal and fabricated metal products industries, 20,275 from the machinery and equipment industries, 19,768 from the construction industries, and 16,451 from wholesale and retail trade. More than 61% of the registrations have been issued to organizations in Europe. In the United States, the accreditation agency is the Registrar Accreditation Board, a wholly owned subsidiary of the American Society for Quality and the American National Standards Institute. However, several other accreditation agencies operate in the United States, most notably the Raad voor Accreditatie (RVA) of the Netherlands.

There are several reasons for an organization to consider registration to these standards. The most commonly quoted reason is that their customers require it. In many instances, customers require extensive audits and/or surveys from their suppliers. These are often waived if the supplier is registered to ISO 9001. Of course, the principal reason is that the company will be a better company if it conforms to the standard, even if it is not registered. However, registration is an indication that this conformance has been met.

REFERENCES

American Society for Quality, (ASQ) (1979), ANSI/ASQC Z1.15-1979, Generic Guidelines for Quality Systems, ASQC, Milwaukee.

American Society for Quality, (ASQ) (1996), ANSI/ASQC C1-1996, Specifications of General Requirements for a Quality Program (also designated as Z1.8-1971), ASQC, Milwaukee.

Automotive Industry Action Group (AIAG) (1992), *Statistical Process Control Reference Manual*, 2nd Ed., AIAG, Detroit.

Automotive Industry Action Group (AIAG) (1994), *Advanced Product Quality Planning and Control Plan Reference Manual*, AIAG, Detroit.

Automotive Industry Action Group (AIAG) (1995a), *Measurement Systems Analysis Reference Manual,* 2nd Ed., AIAG, Detroit.

Automotive Industry Action Group (AIAG) (1995b), *Potential Failure Mode and Effects Analysis Reference Manual,* 2nd Ed., AIAG, Detroit.

Automotive Industry Action Group (AIAG) (1995c), *Production Part Approval Process*, 2nd Ed., AIAG, Detroit, 1995.

Automotive Industry Action Group (AIAG) (1998a), *Quality System Assessment*, 2nd Ed., AIAG, Detroit.

Automotive Industry Action Group (AIAG) (1998b), QS9000, *Quality System Requirements,* 3rd Ed., AIAG, Detroit.

International Organization for Standardization (ISO) (1989), ISO Momento, ISO, Geneva.

International Organization for Standardization (ISO) (1999), ISO/TS 16949, *Quality Systems—Automotive Suppliers—Particular Requirements for the Application of ISO 9001:1994,* ISO, Geneva.

International Organization for Standardization (ISO) (2000a), ISO 19011, *Guidelines for Auditing Management Systems,* ISO, Geneva.

International Organization for Standardization (ISO) (2000b), ISO 9000-2000, *Quality Management Systems—Fundamentals and Vocabulary,* ISO, Geneva.

International Organization for Standardization (ISO) (2000c), ISO 9001-2000, *Quality Management Systems—Requirements*, Geneva.

International Organization for Standardization (ISO) (2000d), ISO 9004-2000, *Quality Management Systems—Guidelines for Performance Improvements*, ISO, Geneva.

United States Department of Defense (1963), MIL-Q-9858A, *Quality Program Requirements,* 1963.

United States Department of Defense (1981), MIL-I-45208A, *Inspection Systems Requirements*, 1981.

CHAPTER 75

Design and Process Platform Characterization Methodology

RAJA M. PARVEZ
Lucent Technologies

DONALD FUSARO
Lucent Technologies

1. **INTRODUCTION** **1976**
2. **PRODUCT DEVELOPMENT** **1977**
   - 2.1. Product Development Process 1977
     - 2.1.1. Initial Stage 1977
     - 2.1.2. Specification Alignment Stage 1977
     - 2.1.3. Initial Models Stage 1977
     - 2.1.4. Product Released to Initial Manufacture Stage 1977
     - 2.1.5. Production Ramp Stage 1977
     - 2.1.6. Product Maintenance Stage 1978
   - 2.2. Need for Characterization 1978
   - 2.3. Characterization Overview 1978
3. **DESIGN CHARACTERIZATION** **1978**
   - 3.1. Understanding Customer Requirements 1978
   - 3.2. Design Models 1979
     - 3.2.1. Example: Optical Transmitter Design Model 1979
4. **PROCESS PLATFORM DEVELOPMENT** **1980**
   - 4.1. Process Platform Concept 1980
     - 4.1.1. Platform Definition 1980
     - 4.1.2. Platform Scope 1981
   - 4.2. Process Platform Implementation 1981
     - 4.2.1. Platform Migration 1981
     - 4.2.2. Platform Characterization 1981
   - 4.3. Process Platform Capability 1982
     - 4.3.1. Boundary Conditions 1982
     - 4.3.2. Platform Performance 1982
5. **DESIGN AND PROCESS PLATFORM CHARACTERIZATION METHODOLOGY** **1982**
   - 5.1. Process Definition (Steps 1 and 2) 1982
     - 5.1.1. Step 1: Identify Critical Designs and Processes 1982
     - 5.1.2. Step 2: Develop Process Flow Diagram for Each Design and Process 1983
   - 5.2. Measurement System Characterization (Steps 3–5) 1984
     - 5.2.1. Step 3: Analyze Measurement and Test System (M&TS) Variation 1984
     - 5.2.2. Step 4: Assess Measurement Variation vs. Process Tolerances 1986
     - 5.2.3. Step 5: Establish Statistical Process Control for M&TS Using Standards 1987
   - 5.3. Model Development (Steps 6 and 7) 1987
     - 5.3.1. Step 6: Develop Experimental Plan for Design/Process Model 1987
     - 5.3.2. Step 7: Validate Design/Process Model and Identify Critical Parameters 1990

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

- 5.4. Statistical Process Control (Steps 8–12) 1992
  - 5.4.1. Step 8: Select Statistically Valid Sampling Scheme 1992
  - 5.4.2. Step 9: Establish Statistical Tools to Identify and Control Variation 1994
  - 5.4.3. Step 10: Comprehend Customer and Design Specifications 1994
  - 5.4.4. Step 11: Assess Process Performance and Establish Statistical Limits 1994
  - 5.4.5. Step 12: Develop Corrective Action Plans 1994
- 5.5. Capability Analysis (Steps 13–16) 1995
  - 5.5.1. Step 13: Compute Process Capability 1995
  - 5.5.2. Step 14: Establish Baseline and Set Objectives 1996
  - 5.5.3. Step 15: Determine Gap between Baseline and Objectives 1996
  - 5.5.4. Step 16: Develop and Implement Action Plan to Close Gap 1996

6. LINKAGE OF PRODUCT DESIGN AND PROCESS PLATFORMS 1996

- 6.1. Process Platform Manual 1997
  - 6.1.1. Platform Description 1997
  - 6.1.2. Platform Characteristics 1997
  - 6.1.3. Platform Boundary Conditions 1997
  - 6.1.4. Platform Characterization Model 1997
  - 6.1.5. Platform Capability 1997
  - 6.1.6. Platform Manual Example 1997
- 6.2. Design Manual 1998
  - 6.2.1. Design Description 1998
  - 6.2.2. Design Parameters and Requirements 1998
  - 6.2.3. Design Characterization Model 1998
  - 6.2.4. Design Capability 1998
  - 6.2.5. Design Boundary Conditions 1998
  - 6.2.6. Design Manual Example 1999

7. DEPLOYMENT 1999

- 7.1. Critical Components 1999
  - 7.1.1. Education 1999
  - 7.1.2. Management Commitment 2000
  - 7.1.3. Deployment Infrastructure 2000
  - 7.1.4. Integration 2001
- 7.2. Performance Measures 2002
  - 7.2.1. Metrics 2002
  - 7.2.2. Progress of Culture Change 2003

ADDITIONAL READING 2004

1. INTRODUCTION

Quality is ultimately defined by customers. There is significant literature available on how to measure quality and value as perceived by customers. In virtually every analysis, a major component of customer satisfaction is the ability of the company to provide a competitively priced product into which quality is designed, built, marketed, and maintained. A company-wide system for achieving that objective must be developed and deployed.

This chapter outlines a complete design and process platform characterization methodology and the system for deployment. The underlying principle of this methodology is to provide a vehicle that starts with identification of customer requirements and ends only when the product has been delivered to a customer who is thoroughly delighted.

This chapter is based on an actual corporate deployment. This provides clear evidence that the methodology works. It also demonstrates that this methodology is a critical part of strategic management that will dependably produce superior profits through satisfied customers.

The elements of design and process platform characterization methodology and system include:

- Product development
- Design characterization
- Process platform development
- Design and process platform characterization methodology

- Linkage of product design and process platform
- Deployment process

Each of these will be discussed in detail in subsequent sections.

2. PRODUCT DEVELOPMENT

Traditionally, product development includes initial research, prototype production, design finalization, and transfer to manufacturing. Today's fast-moving market requires a product-development process that is integrated and efficient. This demands up-front planning and execution of design optimization and manufacturing platform-development activities that ensure conformance to end-product requirements.

2.1. Product Development Process

The product development process consists of six stages; initial stage, specification alignment stage, initial models stage, product released to initial manufacture stage, production ramp stage, and product maintenance stage.

2.1.1. Initial Stage

The purpose of the initial stage is to assemble the relevant information for the product-development team, which includes marketing, design, manufacturing, supply chain, operations, and finance, and to make a fact-based go/no-go decision to develop the product. A go decision requires allocating the resources needed to document detailed specifications and plans for the development of the product. The specific goal of this stage is to evaluate the product concept efficiently and objectively as a function of strategic fit, competition, finance, operations technology, market window, and available resources.

2.1.2. Specification Alignment Stage

The primary function of the specification alignment stage is to generate the information that is required to decide whether to allocate resources to design and develop the proposed product. The product-development team develops detailed initial product specifications in response to market requirements. This permits proper project management by creating accountability for completion of the elements that make up this stage. The specification alignment stage is critical to ensure that sufficient lead time is available to develop, acquire, and prove in any new product and process platform facilities needed to ramp the product according to the business plan.

2.1.3. Initial Models Stage

In the initial models stage, product, test programs, software, and associated hardware are designed and developed. Completion is achieved when prototype hardware/software is fabricated and initial evaluation is started. This stage confirms that the initial model results match the initial product specifications. Finally, a ready-for-production checklist is reviewed by the product development team. One of the critical elements of this stage is the development and implementation of design and process platform characterization objectives and plans.

2.1.4. Product Released to Initial Manufacture Stage

The purpose of this stage is to ensure that product is released to production and validation of manufacturability, procedures, documentation, and customer demand is established. This activity verifies compliance of product and system performance, fixes design issues, and prepares for manufacturing ramps. At the end of this stage, the execution of design and process platform characterization plans are well in progress. This includes identification of critical design and process platform performance parameters, estimation of accuracy and precision for test systems, and planning and validation of experiments. In addition, product performance qualification testing, environmental testing, and final system testing progresses during this stage.

2.1.5. Production Ramp Stage

The launch through production ramps is where all the final product qualification, final manufacturing platform tests, design characterization, and process platform characterization requirements are completed. The purpose of this stage is to ensure that all required documentation is in place to efficiently support the release of the product to the market. The product development team also defines and implements plans for stable ongoing manufacturing, which includes validation of design and process models, control of critical platform processes, and surveillance testing. Appropriate product-

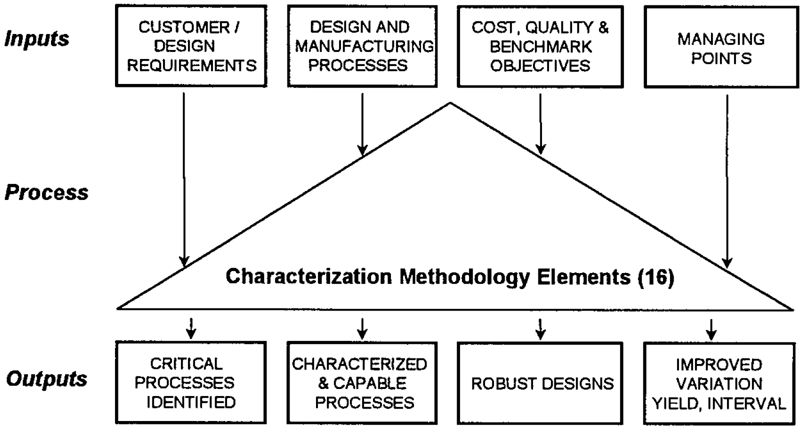

Inputs
CUSTOMER / DESIGN REQUIREMENTS
DESIGN AND MANUFACTURING PROCESSES
COST, QUALITY & BENCHMARK OBJECTIVES
MANAGING POINTS
Process
Characterization Methodology Elements (16)
Outputs
CRITICAL PROCESSES IDENTIFIED
CHARACTERIZED & CAPABLE PROCESSES
ROBUST DESIGNS
IMPROVED VARIATION YIELD, INTERVAL

Figure 1 Characterization Overview.

maintenance plans, business strategies, and product-performance enhancements are established. This is the stage where product robustness and meeting all market and customer requirements are demonstrated.

2.1.6. Product Maintenance Stage

The focus of the product-maintenance stage is to ensure continued product success in the market. The manufacturing and operations teams ensure that product meets all requirements and continues to be robust. This activity ensures stable shipment of product to customers. In addition, periodic quality assurance audits are conducted to determine continued compliance relative to quality system requirements. This is the stage where all required design and platform activities are validated using large samples and appropriate corrective actions and continuous improvement activities are implemented.

2.2. Need for Characterization

Customer requirements and intense competition demand high level of product performance, reliability, and lower costs. This message has profoundly affected corporate America in its strategic goals and objectives as well as day-to-day business policies. During the 1990s we experienced an enormous advancement in all facets of technology. We now see products that have higher performance, more features, and in some cases lower costs. To meet these challenges, it is critical to characterize designs and processes/platforms to optimize the cause-and-effect relationships between customer performance requirements and process tolerances. The absence of this characterization can lead to a lack of consistency, resulting in ineffective planning and eventually higher costs.

2.3. Characterization Overview

Characterization is aimed at improving the overall business by enabling a better understanding of designs and processes. Characterization can be viewed as a process in itself defined by a set of inputs and activities to produce desired outcomes. One such view of a characterization process is shown in Figure 1. It includes a summary of inputs and expected outputs and a characterization methodology. The characterization methodology includes the specific steps for achieving robust designs and processes. The specific steps to the characterization methodology are described in Section 5.

3. DESIGN CHARACTERIZATION

3.1. Understanding Customer Requirements

Customer needs for product performance, reliability and quality are a given in today's market. Customers are no longer bound by national interests as they pursue quality products and services worldwide. To be competitive in this environment, businesses must exceed the needs of their customers by providing products that are high quality, cost effective, and reliable. The standard approaches for

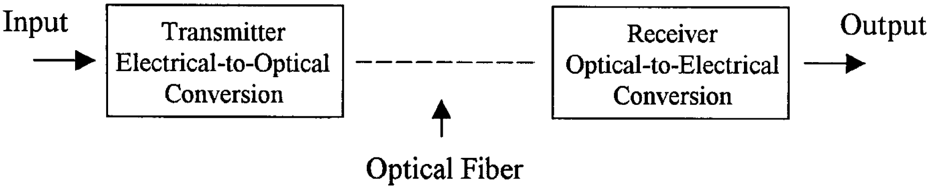


Figure 2 Fiberoptic System.

comprehending customer requirements include review of product specifications and drawings, performance evaluation of similar technology, and simulated testing for customer's application.

Communications with customers is the most important approach for understanding their requirements. This can be initiated by the customer as well as by the supplier. The customer-initiated feedback can be through formal product requests, product returns, field application data, and audits. The supplier-initiated feedback may include product-performance reviews, market trends, technology integration, competitive analysis, and design-characterization models.

3.2. Design Models

Design models are cause-and-effect quantitative models that translate product performance requirements to process requirements and identify key in-process output parameters and associated tolerances. Control of these parameters at in-process steps during manufacturing will generally ensure conformance to end-product requirements. In effect, design models establish the linkage between customer and manufacturing requirements. Design models are typically generated from current knowledge of existing technologies, proven theories, and data from planned experimentation. It is the experimental data that provides the new insight for a specific design. A formal characterization methodology is described in Section 5.

As an example, design models were developed for a fiber-optic system that includes three elements, the *transmitter*, which allows for data input and produces an optical signal, the *optical fiber*, which carries the data, and the *receiver*, which converts optical signal to electrical signal to output the data. Figure 2 depicts the key elements of a fiber-optic system.

3.2.1. Example: Optical Transmitter Design Model

An optical transmitter is the device that generates the signal that is sent through an optical fiber. One of the product parameters that is critical to transmitter performance is optical output power. The optical output power level that is delivered to the fiber is inevitably lower than what the light source generates because of transfer losses. The optical output power of a transmitter is affected by many design parameters, including parameters related to the laser chip, lens, optical fiber, and processing. Figure 3 depicts key elements of a transmitter.

As an example, a design model linking the relationship between key characteristics of these elements and optical output power is:

Optical output power = *f*{laser facet power, laser placement, lens placement,
fiber alignment ...}

In this model, optical output power is impacted by the manufacturing processes, which include

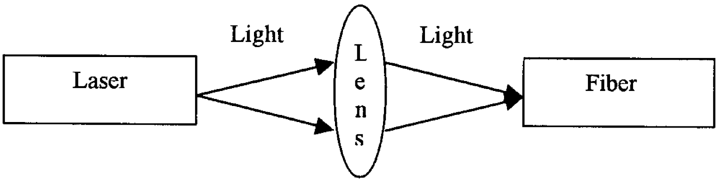


Figure 3 Transmitter Elements.

laser fabrication and the processes associated with laser, lens, and fiber attachment. Hence, control of these processes will ensure conformance to design requirements for optical output power.

4. PROCESS PLATFORM DEVELOPMENT

In addition to characterizing designs, manufacturing processes must also be characterized and controlled to achieve overall business objectives. When a design reuses well-established processes (i.e., platforms), improved time-to-market and reduced development and manufacturing costs can be achieved. This section describes an approach to developing these process platforms.

4.1. Process Platform Concept

All products are built on a series of processes to meet customer requirements and achieve desired performance measures. As products change or new products are introduced, many advantages are gained by the reuse of existing processes. These advantages include minimizing the cost in the development and introduction of new processes. In addition, overall time-to-market can be reduced by the elimination of process-development time in the product-development cycle. Also, requiring new process development as part of new product introduction reduces the probability of overall success of the project due to the additional variable of successful process development to achieve successful product development. However, reusing existing, well-understood processes makes both performance and impact on manufacturing known. The platform concept relies on establishing such processes and taking advantage of their capabilities to achieve new design goals. To introduce the platform concept into a business the following key elements will need to be developed for successful implementation:

- Common definition and understanding of platform concept
- Platform scope based on business plans
- Migration plans
- Characterization plans
- Platform capabilities and requirements

4.1.1. Platform Definition

The definition of platforms to a particular business needs to be jointly developed and communicated throughout all organizations (manufacturing, development, design, and marketing). This definition needs to be tailored to fit the company's infrastructure, systems, and culture for it to be effectively deployed. However, any definition of a platform should include elements of commonality, reusability, characterization, and capability. An example of one such definition is given in Figure 4. The platform definition will provide an organization with the basic framework to determine which processes are to be considered as platforms. In some instance, various levels of platform can be defined to allow for some additional platform flexibility. For example, a level one platform may mean no changes are allowed, a level two platform may allow for changes to process settings, a level three platform may allow for changes to fixturing, and so on.

COMMON:
- Process is used on a variety of product codes and families

REUSABLE:
- Process is used for several generations of product designs

WELL DEFINED:
- Defined Boundary Conditions & Design Guidelines (what it can and can't do)
- Characterized, Stable and Capable Process Inputs and Outputs
- Defined Characteristics, Documented

COMMON CHARACTERISTICS:
- Methods: Common procedures, sequence, recipe (inputs)
- Skills: Defined training requirements (skill needs - operating & engineering)
- Hardware: Machine, Equipment (function & features)
- Materials: Common piece part characteristics, Material composition
- Environment: Controls, Conditions

Figure 4 Process Platform Definition.

4.1.2. Platform Scope

To implement platform processes successfully, an organization needs to develop fundamental business plans in terms of market perspective, technology directions, and existing manufacturing capability. From a market perspective, detailed market plans need to be established that identify direction of future product performance requirements in the markets it serves. From a technology standpoint, technology roadmaps are needed to set design and technology plans that provide design direction to achieve future product performance requirements. From manufacturing, the capability of existing processes and technologies need to be well understood and defined to determine the impact of design changes on process capabilities. All of this information needs to be integrated to develop comprehensive platform plans and objectives.

4.2. Process Platform Implementation

4.2.1. Platform Migration

If the implementation of platforms includes processes already used in manufacturing, then individual platform-specific plans need to be developed to migrate existing products onto the platform. A typical process for the development of these migration plans is shown in Figure 5.

Each product should be evaluated to determine whether it should be migrated to the platform. This evaluation should also consider future products and designs. The evaluation has two key parts associated with it. First, a business case needs to be developed to determine what are the specific expected benefits in terms of cost, interval, capacity improvements vs. the cost of implementation of the platform. Second, there needs to be an assessment to establish the technical feasibility of manufacturing the product using the platform. The technical feasibility must be done from both the design and process viewpoints. This will require that both process and design characterization have been completed. Once both benefits and technical feasibility have been demonstrated, specific experimental and qualification plans can be developed and implemented to migrate products onto the platform.

4.2.2. Platform Characterization

As per the definition, a platform must be well characterized to provide the confidence that the performance achieved from this process is sustainable. This characterization activity will lead to the

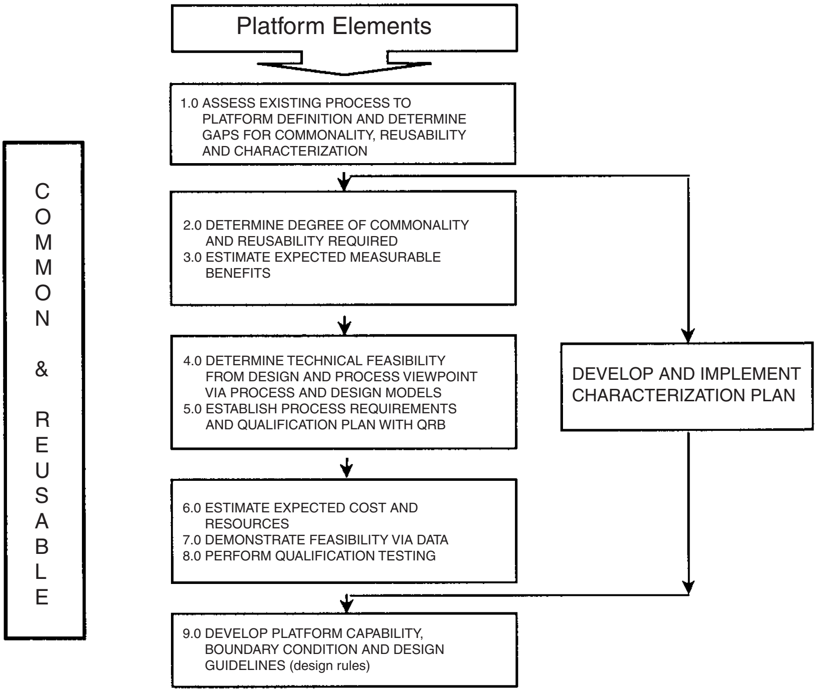

Platform Elements
1.0 ASSESS EXISTING PROCESS TO PLATFORM DEFINITION AND DETERMINE GAPS FOR COMMONALITY, REUSABILITY AND CHARACTERIZATION
2.0 DETERMINE DEGREE OF COMMONALITY AND REUSABILITY REQUIRED
3.0 ESTIMATE EXPECTED MEASURABLE BENEFITS
4.0 DETERMINE TECHNICAL FEASIBILITY FROM DESIGN AND PROCESS VIEWPOINT VIA PROCESS AND DESIGN MODELS
5.0 ESTABLISH PROCESS REQUIREMENTS AND QUALIFICATION PLAN WITH QRB
6.0 ESTIMATE EXPECTED COST AND RESOURCES
7.0 DEMONSTRATE FEASIBILITY VIA DATA
8.0 PERFORM QUALIFICATION TESTING
9.0 DEVELOP PLATFORM CAPABILITY, BOUNDARY CONDITION AND DESIGN GUIDELINES (design rules)
DEVELOP AND IMPLEMENT CHARACTERIZATION PLAN
COMMON & REUSABLE

Figure 5 Platform Migration Methodology.

development of a process model, which is a cause-and-effect quantitative model that establishes the relationships between process requirements and process variables. The process model identifies key process inputs and level of control (i.e., tolerances) needed to achieve desired process performance. An example of such a model is given in Section 5. As with any tolerance, if it is unnecessarily tight it may lead to overall higher processing costs and if it is too loose it may result in producing defective product. A formal methodology to characterize platforms and processes is described in Section 5.

4.3. Process Platform Capability

4.3.1. Boundary Conditions

Upon completion of characterization, a platform's true capability can be determined. This capability provides design organizations with tolerances to design within. These capabilities must be defined relative to appropriate constraints on process conditions. These constraints are termed boundary conditions. Process boundary conditions must be defined that clearly specify any restrictions necessary to achieve the capability and benefits of the platforms. These boundary conditions can include restriction to process settings, fixture requirements, environmental conditions, process methods or sequences, and so on. Thus, these conditions provide design groups with what a process can and cannot achieve.

4.3.2. Platform Performance

For each performance characteristic, the associated process capability needs to be defined. Capability is evaluated in terms of the ability of the process to meet its target and in terms of reproducing results consistently (i.e., process variation). Capability is developed through collection and analysis of performance data after process characterization and control has been established. The resulting measures define the expected performance of a platform. This information is critical input into design in terms of expected product performance.

5. DESIGN AND PROCESS PLATFORM CHARACTERIZATION METHODOLOGY

Each business must have a sound approach for understanding and improving its processes and designs. The design and process platform characterization methodology presented here is one such approach that has been proven successful in modeling, controlling, and improving products and processes. This approach, shown in Figure 6, provides the detailed steps and associated tools to achieving process and design goals.

The methodology consists of five major elements:

- Process definition
- Measurement system characterization
- Design and process characterization (i.e., model development)
- Process control
- Process capability

5.1. Process Definition (Steps 1 and 2)

5.1.1. Step 1: Identify Critical Designs and Processes

Products are built on a series of processes, each with its own impact on the overall performance. The purpose of this step is to determine which processes need to be characterized. A process includes the combination of people, equipment, materials, methods, and environment that produces products or services. Any process whose inputs and outputs are measured, monitored, controlled, or observed should be evaluated. Designs and processes for characterization are typically identified based on their overall impact to the business. A critical design and process is one that has major impact on cost, quality, manufacturability, performance, reliability, or customer satisfaction. Some of the methods and tools available for determining this are given below:

- Product performance analysis
- Design and reliability reviews
- Customer field returns analysis
- Quality function deployment
- Quality improvement methodologies
- Cost of quality
- Fishbone analysis

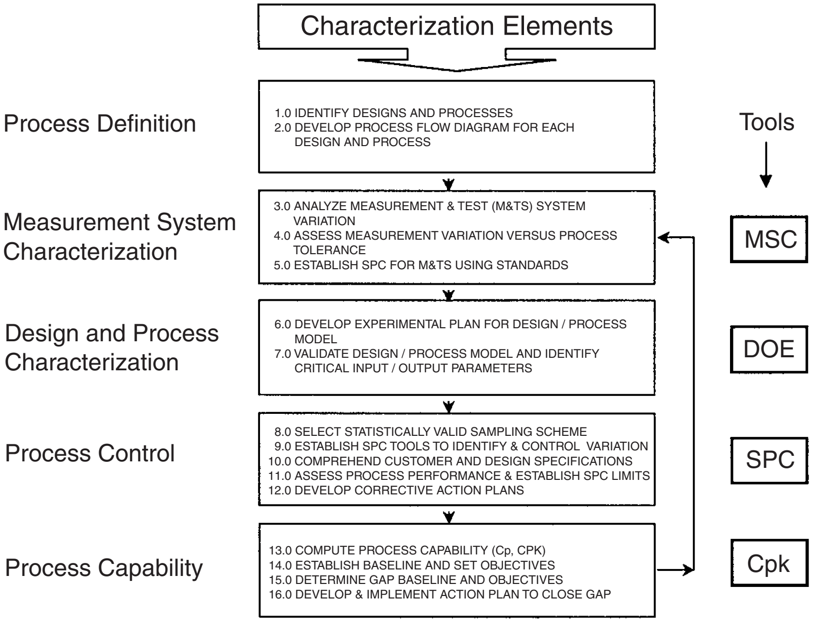

Characterization Elements
Process Definition
1.0 IDENTIFY DESIGNS AND PROCESSES
2.0 DEVELOP PROCESS FLOW DIAGRAM FOR EACH DESIGN AND PROCESS
Tools
Measurement System Characterization
3.0 ANALYZE MEASUREMENT & TEST (M&TS) SYSTEM VARIATION
4.0 ASSESS MEASUREMENT VARIATION VERSUS PROCESS TOLERANCE
5.0 ESTABLISH SPC FOR M&TS USING STANDARDS
MSC
Design and Process Characterization
6.0 DEVELOP EXPERIMENTAL PLAN FOR DESIGN / PROCESS MODEL
7.0 VALIDATE DESIGN / PROCESS MODEL AND IDENTIFY CRITICAL INPUT / OUTPUT PARAMETERS
DOE
Process Control
8.0 SELECT STATISTICALLY VALID SAMPLING SCHEME
9.0 ESTABLISH SPC TOOLS TO IDENTIFY & CONTROL VARIATION
10.0 COMPREHEND CUSTOMER AND DESIGN SPECIFICATIONS
11.0 ASSESS PROCESS PERFORMANCE & ESTABLISH SPC LIMITS
12.0 DEVELOP CORRECTIVE ACTION PLANS
SPC
Process Capability
13.0 COMPUTE PROCESS CAPABILITY (Cp, CPK)
14.0 ESTABLISH BASELINE AND SET OBJECTIVES
15.0 DETERMINE GAP BASELINE AND OBJECTIVES
16.0 DEVELOP & IMPLEMENT ACTION PLAN TO CLOSE GAP
Cpk

Figure 6 Design and Process Platform Characterization Methodology.

Example: Fiber-Alignment Platform

Platform Description:

This process consists of aligning a fiber to the lens for minimizing optical power loss and then welding the fiber to the lens while maintaining fiber alignment. A measure of this fiber alignment is the change in power loss due to welding (i.e., delta insertion loss).

Process	**Inputs**	**Outputs**	**Impact**
Fiber alignment	Weld energy, weld pattern, lens holder roughness, fiber gap, optical fiber, lens, package	Delta insertion loss	Optical output power
Design	**Inputs**	**Outputs**	**Impact**
Transmitter design	Laser facet power, laser placement, lens placement, fiber alignment (delta insertion loss)	Optical output power	System performance

5.1.2. Step 2: Develop Process Flow Diagram for Each Design and Process

For each design and process selected, a process flow diagram should be developed. A process flow diagram will aid in determining the scope of the characterization effort. The flow chart should be developed with inputs from engineering, design, manufacturing, suppliers, and customers to identify all potential design and process characteristics and includes items such as:

- Sequential process steps and/or material flow
- Relationship between process steps
- Rework loops
- List process step, setups, inputs (i.e., temperature, pressure, force)
- List outputs (i.e., epitaxial thickness, pull strength, wavelength, power)
- Decision points
- Process control and yield points

Example

The process flow diagram for the fiber-alignment platform is shown in Figure 7.

5.2. Measurement System Characterization (Steps 3–5)

5.2.1. Step 3: Analyze Measurement and Test System (M&TS) Variation

Once the process characteristics have been identified, it is then critical that appropriate measures of these characteristics be available. If data does not exist, an appropriate measurement method must be developed and evaluated. If data already exists, a thorough study on the measurement system variation in terms of both accuracy and precision needs to be completed to determine the overall measurement system capability. Excessive variation in the measurement and test system (M&TS) will adversely affect the sensitivity of the decision-making process and may be a source of unnatural variation.

To determine statistically the accuracy and precision of the measurement system, calibration and an error of measurement study must be completed.

Accuracy can be defined as the extent to which the average of a series of repeat measurements of the same item agrees with the "true" value. Accuracy is achieved via calibration, where calibration is the combination of checking and adjusting (if needed) M&TS to a known or traceable standard to bring the M&TS within its tolerances for accuracy.

Precision can be defined as the variation observed in individual repeated measurements of the same item. Precision is also referred to as measurement error, repeatability, or reproducibility. The common method of estimating precision is by performing an error-of-measurement study. This study

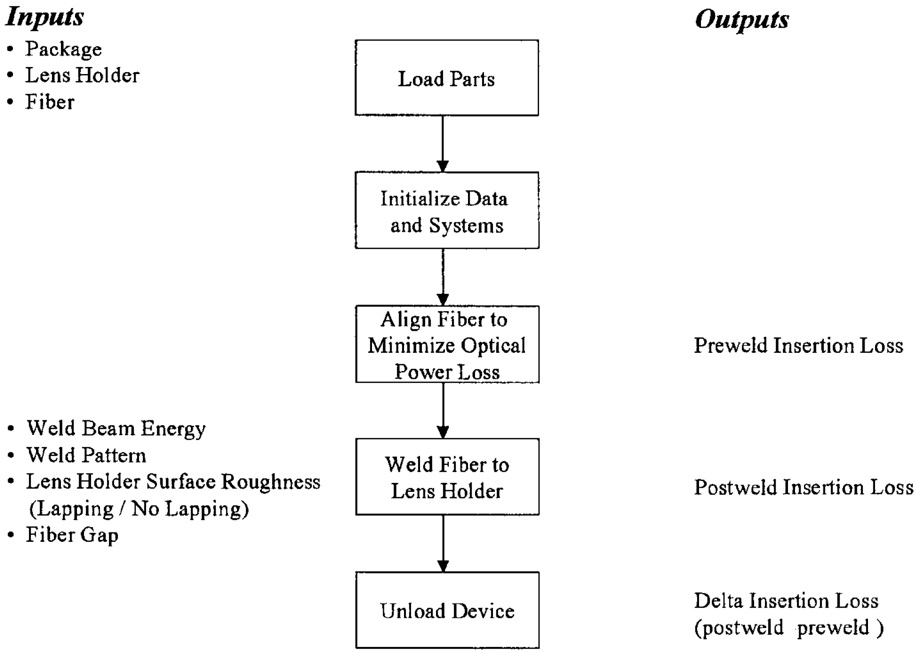

Inputs
• Package
• Lens Holder
• Fiber
Outputs
Load Parts
Initialize Data and Systems
Align Fiber to Minimize Optical Power Loss
Preweld Insertion Loss
• Weld Beam Energy
• Weld Pattern
• Lens Holder Surface Roughness (Lapping / No Lapping)
• Fiber Gap
Weld Fiber to Lens Holder
Postweld Insertion Loss
Unload Device
Delta Insertion Loss (postweld preweld)

Figure 7 Fiber-Alignment Process Flow.

TABLE 1 **Insertion Loss Error-of-Measurement Data.**

Device	Reading 1 (dB)	Reading 2 (dB)	Average	Range
1	0.54	0.52	0.53	0.02
2	0.81	0.85	0.83	0.04
3	0.82	0.81	0.82	0.01
4	0.16	0.18	0.17	0.02
5	0.89	0.85	0.87	0.04
6	0.73	0.72	0.73	0.01
7	0.68	0.68	0.68	0.00
8	0.55	0.55	0.55	0.00
9	0.85	0.83	0.84	0.02
10	0.83	0.84	0.84	0.01
11	0.24	0.25	0.25	0.01
12	0.90	0.91	0.91	0.01
13	0.77	0.80	0.79	0.03
14	0.69	0.66	0.68	0.03
15	0.57	0.54	0.56	0.03
16	0.85	0.85	0.85	0.00
17	0.35	0.34	0.35	0.01
18	0.90	0.88	0.89	0.02
19	0.73	0.79	0.76	0.06
20	0.68	0.70	0.69	0.02

is performed by measuring several devices multiple times each and then estimating the error from the repeated measurements.

If more than one piece of equipment or measurement system is used to make measurements, then it is also necessary to perform appropriate correlation studies to demonstrate the compatibility of the multiple systems.

Example: Fiber-Alignment Platform

An error-of-Measurement study was performed to determine the measurement error of the insertion loss-measurement system. This study included measuring 20 devices repeatedly and creating control charts to evaluate measurement performance. The data and analysis are shown in Table 1 and Figures 8 and 9 respectively.

The X-bar chart in Figure 8 shows the discriminating power of the MT&S by plotting the average (i.e., mean) measurements vs. limits derived from the range of repeated measurements from the same device. Since this chart shows the ability of the MT&S to differentiate between different devices, ideally the data points should be out of control for this chart.

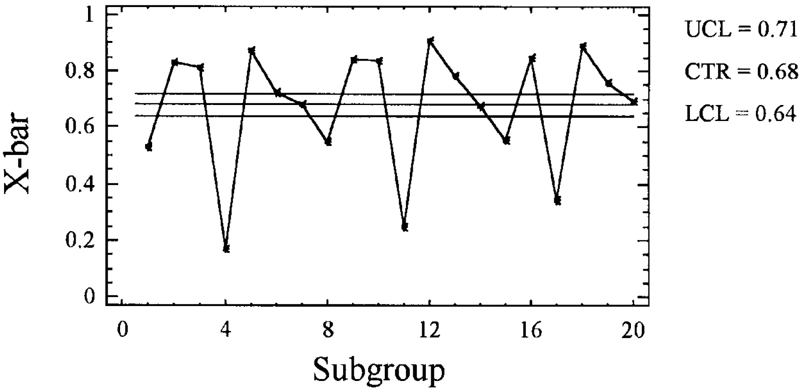


Figure 8 X-bar Chart for Insertion Loss Error of Measurement.

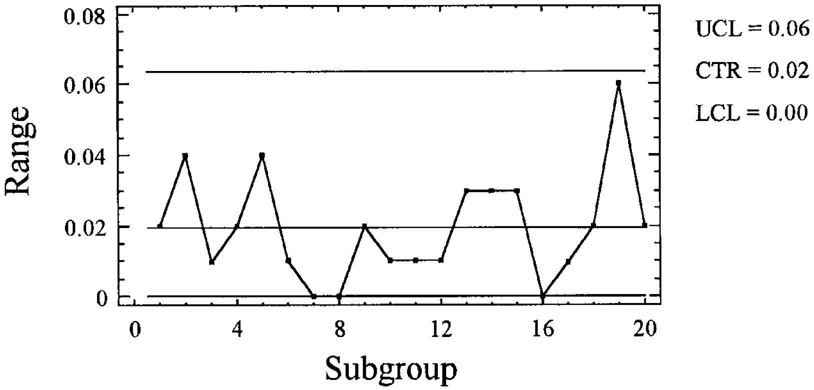


Figure 9 Range Chart for Insertion Loss Error of Measurement.

The range chart in Figure 9 shows directly the magnitude and consistency from device to device of the measurement error associated with this MT&S. The plot shows the difference between measurements made on the same device. Ideally, this chart should have a low centerline and be in control.

Based on the average range (r-bar), the measurement error (σ_m) was then estimated by using the formula as $\sigma_m = r$-bar divided by d_2, where d_2 is the factor that converts the range of a sample of size n from a normal distribution into an unbiased estimate of the standard deviation of that distribution. For a sample size of $n = 2$, $d_2 = 1.128$.

$$\text{Insertion loss-measurement error } (\sigma_m) = 0.02 \text{ dB}$$

5.2.2. *Step 4: Assess Measurement Variation vs. Process Tolerances*

Once the measurement variation has been estimated, one needs to determine its acceptability for its intended use. This is typically done by comparing the measurement system variation vs. process tolerances or product performance. These two methods are described below:

5.2.2.1. Precision to Tolerance (P/T) Ratio This method is used to compare the measurement error relative to process tolerance (i.e., a parameter with an upper and lower specification limit).

$$P/T \text{ ratio} = \left\{\frac{6 * \sigma_m}{\text{USL} - \text{LSL}}\right\} * 100\%$$

A recommended level is that the M&TS have a $P/T < 30\%$. This would be interpreted as meaning that 30% of tolerance interval could be used up by measurement error and still be acceptable.

5.2.2.2. Percent Measurement Error (Percent Error) This method is used to compare the measurement error relative to the total variation in the process and is typically applied for one-sided specification.

$$\text{Percent measurement error} = \left\{\frac{(\sigma_m)^2}{(\sigma_t)^2}\right\} * 100\%$$

where σ_t is an estimate of the variation from the process under study.

A recommended level is that M&TS have percent error < 10% . This would be interpreted as meaning the measurement component to the overall total process variation could be up to 10%.

Note that P/T ratio and % error objectives are only recommendations. These levels should be selected based on individual requirements and applications.

Example: Fiber-Alignment Platform

The following criterion was used to determine acceptability of the insertion loss-measurement system:

$$\text{Percent measurement error} < 10\% \text{ is acceptable}$$

For this example the measurement error was 0.02 dB and an estimate of total error from production data was 0.1 dB:

$$\text{Percent measurement error} = \frac{\{0.02\}^2}{\{0.1\}^2} * 100\% = 4.0\%$$

5.2.3. Step 5: Establish Statistical Process Control for M&TS Using Standards

The purpose of monitoring measurement and test systems is to determine whether the measurement system is stable, where stability is the absence of drift or other changes over time. This will ensure that the estimates of accuracy and precision remain the same over time. In many cases, when the measurement system is consistently stable over a long period of time, a reduced calibration frequency can be implemented.

A method to determine stability is to monitor M&TS using product or traceable standards via statistical process control. Measuring and monitoring standards at a specified frequency allows a measure of stability to be assessed. If traceable standards are not available, then devices that represent the behavior of a product on the M&TS can be used. It is recommended that two active and one reserve standard be specified. The reason for using two active standards is to determine the nature of the problem (i.e., is it due to malfunctioned standard or the M&TS). The active standards should be used to monitor the M&TS performance. The frequency of monitoring active standards depends upon a particular application (i.e., per shift or per day or before use). The reserve standard should be used to replace an active standard that degrades or no longer functions. The reserve standard may be used for confirmation or resolution of a problem. To accomplish these functions, the reserve standard must be measured regularly (typically at less frequency than active standard) to establish a performance history.

Some of the key elements for establishing statistical process control for measurement and test system include:

- Identification of critical parameters
- Determination of measurement error
- Selection of traceable (if available) and/or product standards
- Determination of the sample size and the frequency of measurements for each standard
- Selection of statistical tools (i.e., control charts) to monitor and control variation
- Development and implementation of corrective action procedure for out-of-control conditions

Example: Fiber-Alignment Platform

To assess the stability of the MT&S for insertion loss, two standards were measured and plotted once a day. Individual control charts were implemented to monitor and control this testing process. These control charts are shown in Figure 10.

5.3. Model Development (Steps 6 and 7)

5.3.1. Step 6: Develop Experimental Plan for Design/Process Model

In this step, the designer or process owner develops the hypothesized relationships (i.e., forms the tentative model to be estimated) between design or process inputs and outputs and associated experimental plan.

The two types of models that are developed are the design and process models. Design models are cause-and-effect quantitative models that translate product-performance requirements to process requirements. The design model identifies key in-process output parameters and associated tolerances. For example:

$$\text{Optical output power} = f\{\text{laser facet power, laser placement, lens placement, fiber alignment} \ldots\}$$

Inputs to include in the design model may be identified through:

- Physical (theoretical) relationship
- Previous design work
- Defect analysis via reliability monitoring

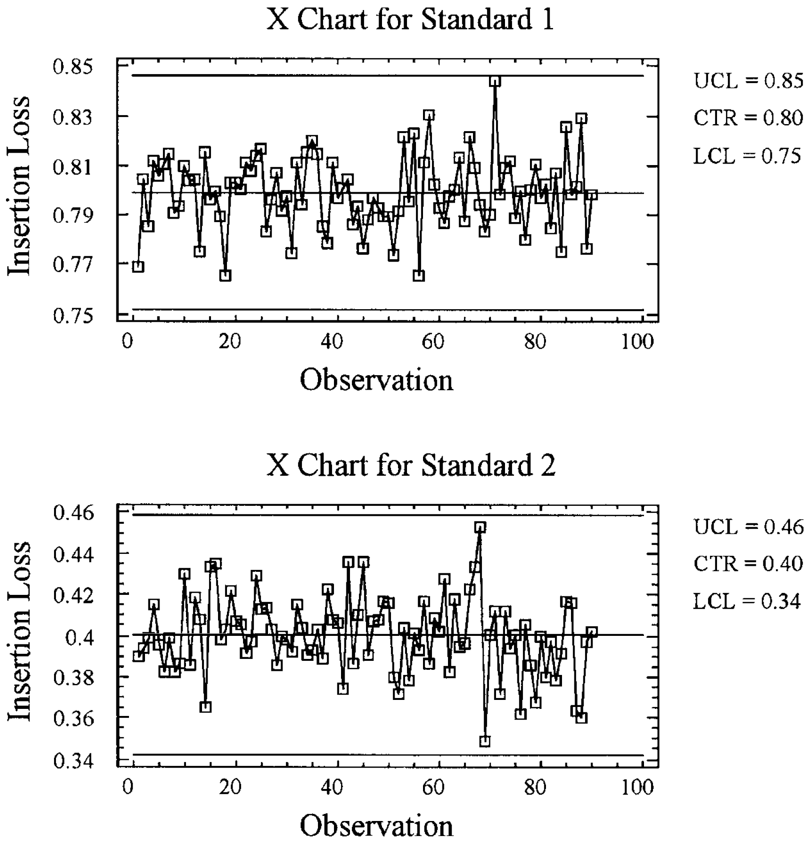


Figure 10 X Charts for Insertion Loss Measurement Standards.

- Device quality issues
- Customer feedback

Process models are cause-and-effect quantitative models that establish the relationships between process requirements and process variables. The process model identifies key process inputs and level of control (i.e., tolerances) needed to achieve desired process performance. For example:

$$\text{Fiber alignment} = f\{\text{weld energy, weld pattern, lens holder roughness, fiber gap} \ldots\}$$

Inputs to include in the process model may be identified through:

- Physical (theoretical) relationship
- Engineering design
- Defect analysis via manufacturing monitoring
- Customer feedback

Once the process or design variables have been identified, a statistically designed experiment should be employed to collect appropriate data to fit the model. This plan includes developing the experimental objective, design matrix, and sample size (i.e., number of experimental replicates).

Example: Process Model

Fiber-Alignment Platform

Experimental objectives:	Determine the relationships and settings for key process variables to meet process objective of delta insertion loss less than 1 dB.
Response variable:	Delta insertion loss (fiber alignment)
Process variables:	Weld energy Weld pattern Lens holder roughness Fiber gap
Experimental design:	2^4 factorial
Experimental matrix:	

Weld Energy (joules)	Weld Pattern (# welds)	Lens Holder Roughness	Fiber Gap (μm)	Delta Insertion Loss (dB)
0.5	2.0	0	2.0	0.466
1.0	2.0	0	2.0	1.141
0.5	6.0	0	2.0	0.134
1.0	6.0	0	2.0	1.783
0.5	2.0	1	2.0	0.346
1.0	2.0	1	2.0	0.945
0.5	6.0	1	2.0	0.269
1.0	6.0	1	2.0	1.803
0.5	2.0	0	8.0	0.519
1.0	2.0	0	8.0	0.907
0.5	6.0	0	8.0	0.440
1.0	6.0	0	8.0	1.632
0.5	2.0	1	8.0	0.397
1.0	2.0	1	8.0	0.971
0.5	6.0	1	8.0	0.232
1.0	6.0	1	8.0	1.285

Example: Design Model

Transmitter Optical Output Power

Experimental objectives:	Determine relationships and settings for key process variables to meet design objective of optical output power between 15–24 mW.
Response variable:	Optical output power
Design variables:	Laser facet power Laser placement Lens placement Fiber alignment (delta insertion loss)
Experimental design:	Replicated 2^4 factorial

Laser Facet Power (mW)	Laser Placement (μm)	Lens Placement (mils)	Fiber Alignment (dB)	Optical Output Power (mW)
20	1.5	1.2	0.6	16.31
30	1.5	1.2	0.6	25.26
20	6.5	1.2	0.6	13.42
30	6.5	1.2	0.6	20.42
20	1.5	5.8	0.6	16.57
30	1.5	5.8	0.6	23.80
20	6.5	5.8	0.6	12.38
30	6.5	5.8	0.6	19.52

Laser Facet Power (mW)	Laser Placement (μm)	Lens Placement (mils)	Fiber Alignment (dB)	Optical Output Power (mW)
20	1.5	1.2	1	15.62
30	1.5	1.2	1	22.94
20	6.5	1.2	1	12.51
30	6.5	1.2	1	19.72
20	1.5	5.8	1	14.77
30	1.5	5.8	1	22.80
20	6.5	5.8	1	12.18
30	6.5	5.8	1	19.34
20	1.5	1.2	0.6	16.13
30	1.5	1.2	0.6	24.21
20	6.5	1.2	0.6	13.88
30	6.5	1.2	0.6	20.47
20	1.5	5.8	0.6	15.58
30	1.5	5.8	0.6	24.10
20	6.5	5.8	0.6	12.84
30	6.5	5.8	0.6	19.35
20	1.5	1.2	1	15.61
30	1.5	1.2	1	23.72
20	6.5	1.2	1	14.47
30	6.5	1.2	1	20.61
20	1.5	5.8	1	16.16
30	1.5	5.8	1	23.26
20	6.5	5.8	1	11.91
30	6.5	5.8	1	19.03

5.3.2. Step 7: Validate Design/Process Model and Identify Critical Parameters

From the experimental data, a mathematical model is developed to identify significant input factors and quantify their impact on the output. From this, appropriate settings for the inputs can be determined as well as the level of control required during manufacturing. The statistical techniques to accomplish this are analysis of variance and regression analyses.

Example: Process Model

Fiber-Alignment Platform

Data from the experiment were analyzed using ANOVA and multiple regression techniques.

The results are shown in Figure 11.

The analysis of this data identifies weld energy and weld pattern as critical process variables and develops the following model:

$$\text{Delta insertion loss} = \{0.354 + 0.320\,(\text{weld energy}) - 0.240\,(\text{weld pattern}) + 0.399\,(\text{weld energy} * \text{weld pattern})\}$$

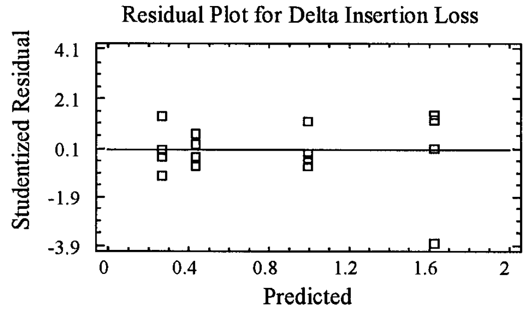
From this model and other processing considerations, the following process settings were implemented.

Parameter	*Setting*
Weld energy	0.6 J
Weld pattern	3
Lens holder roughness	0 (no lapping)
Fiber gap	5 μm

Example: Design Model

```
Analysis of Variance for Delta Insertion Loss
--------------------------------------------------------------------------------
Source                      Sum of Squares   Df    Mean Square    F-Ratio   P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A: Weld Energy                 3.6710600     1    3.67106000     183.57    0.0000
 B: Weld Pattern                0.2223120     1    0.22231200      11.12    0.0207
 C: Lens Holder Roughness       0.0374422     1    0.03744220       1.87    0.2295
 D: Fiber Gap                   0.0158760     1    0.01587600       0.79    0.4137

INTERACTIONS
 AB                            0.63680400     1    0.63680400      31.84    0.0024
 AC                            0.00129600     1    0.00129600       0.06    0.8092
 AD                            0.09765620     1    0.09765620       4.88    0.0781
 BC                            0.00004225     1    0.00004225       0.00    0.9651
 BD                            0.00547600     1    0.00547600       0.27    0.6231
 CD                            0.01276900     1    0.01276900       0.64    0.4605
RESIDUAL                       0.09998970     5    0.01999790
--------------------------------------------------------------------------------
TOTAL (CORRECTED)              4.80072000    15
--------------------------------------------------------------------------------
```

```
Multiple Regression Analysis for Delta Insertion Loss
-----------------------------------------------------------------------------------
Dependent variable: Delta Insertion Loss
-----------------------------------------------------------------------------------
                                             Standard            t
Parameter                     Estimate         Error         Statistic        P-Value
-----------------------------------------------------------------------------------
CONSTANT                      0.353625       0.2654340        1.33225         0.2075
Weld Energy                   0.320000       0.3357500        0.95309         0.3593
Weld Pattern                 -0.240313       0.0593528       -4.04888         0.0016
Weld Energy*Weld Pattern      0.399000       0.0750760        5.31461         0.0002
-----------------------------------------------------------------------------------
R-squared = 94.3644 percent,  R-squared (adjusted for d.f.) = 92.9555 percent
Standard Error of Est. = 0.150152
```



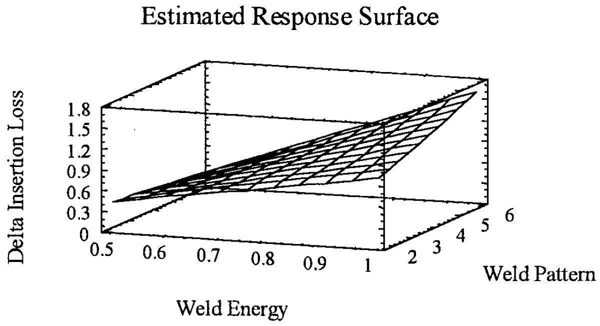


Figure 11 Results of Fiber-Alignment Platform Experiment.

Transmitter Optical Output Power

Data from the experiment were analyzed using ANOVA and multiple regression techniques. The results are shown in Figure 12.

The analysis of this data identifies laser facet power, laser placement, lens placement, and fiber alignment as critical design variables and develops the following model:

$$\begin{aligned}\text{Optical output power} = \{&1.527 + 0.824\ (\text{laser facet power}) - 0.038\ (\text{laser placement})\\ &- 0.025\ (\text{lens placement}) - 1.498\ (\text{fiber alignment})\\ &- 0.021\ (\text{laser facet power} * \text{laser placement})\\ &- 0.034\ (\text{laser placement} * \text{lens placement})\}\end{aligned}$$

Based on the design model, customer requirements, platform requirements, costs, and other design considerations, the following specifications were developed.

Parameter	*Specification*
Laser facet power	22–28 mW
Laser placement	5 μm maximum
Lens placement	5.8 mils maximum
Fiber alignment	1 dB maximum

5.4. Statistical Process Control (Steps 8–12)

Once the critical inputs and outputs have been determined, appropriate statistical process controls can be implemented to control and improve process performance.

In the example, a design model was first developed that indicated that optical output power is a function of laser facet power, laser placement, lens placement, and fiber alignment. To control optical output power, statistical process control must be applied to these critical process output parameters and, where applicable, must be applied to corresponding process input parameters.

Below is the summary of the design model for optical output power and the corresponding process models associated with the critical process output parameters.

- Design model:

$$\begin{aligned}\text{Optical output power} = f\{&\text{laser facet power, laser placement, lens placement,}\\ &\text{fiber alignment} \ldots\}\end{aligned}$$

- Process models:

Fiber alignment = f\{weld energy, weld pattern ...\}
Laser facet power = f\{laser chip slope, epitaxial grating, metal doping composition ...\}
Laser placement = f\{laser chip size, bond temperature, piece part quality ...\}
Lens placement = f\{bonding temperature, ambient conditions, solder oxidation ...\}

Hence, the control of process output parameters and associated input parameters will ensure conformance to optical output power requirements.

5.4.1. Step 8: Select Statistically Valid Sampling Scheme

Based on the process models, appropriate sampling plans for process controls can be developed. A sampling scheme using rational subgroupings is one that provides the valid or right data at minimal cost. The main effort is to ensure that the samples in any one subgroup have been produced under essentially the same conditions. Elements of a rational subgroup include:

- Samples are randomly selected within a subgroup.
- Subgroups are collected at proper frequencies.
- Subgroup size is adequate to detect desired level of change or shift in the process.
- Subgrouping minimizes variation within subgroups.

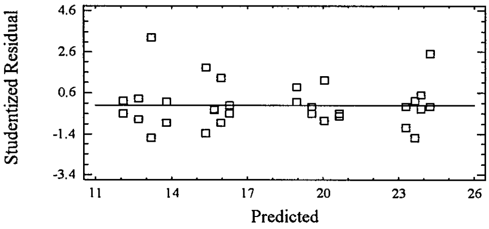
A typical guideline is to collect small samples frequently rather than large samples less frequently.

Example: Fiber-Alignment Platform

```
Analysis of Variance for Optical Output Power
--------------------------------------------------------------------------------------
Source                       Sum of Squares     Df        Mean Square     F-Ratio     P-Value
--------------------------------------------------------------------------------------
MAIN EFFECTS
 A: Laser Facet Power          436.67500        1        436.6750000     1601.71      0.0000
 B: Laser Placement             93.81080        1         93.8108000      344.09      0.0000
 C: Lens Placement               4.28513        1          4.2851300       15.72      0.0007
 D: Fiber Alignment              2.87400        1          2.8740000       10.54      0.0039

INTERACTIONS
 AB                           2.24190000        1         2.24190000        8.22      0.0092
 AC                           0.01087810        1         0.01087810        0.04      0.8436
 AD                           0.10465300        1         0.10465300        0.38      0.5422
 BC                           1.19738000        1         1.19738000        4.39      0.0484
 BD                           0.65265300        1         0.65265300        2.39      0.1367
 CD                           0.00137813        1         0.00137813        0.01      0.9440
RESIDUAL                      5.72524000       21         0.27263100
--------------------------------------------------------------------------------------
TOTAL (CORRECTED)           547.579000          1
--------------------------------------------------------------------------------------
```

```
Multiple Regression Analysis for Optical Output Power
-------------------------------------------------------------------------------------------
Dependent variable: Optical Output Power
-------------------------------------------------------------------------------------------
                                                        Standard              T
Parameter                                Estimate         Error          Statistic      P-Value
-------------------------------------------------------------------------------------------
CONSTANT                                1.5266300     0.97379300         1.567720       0.1295
Laser Facet Power                       0.8235130     0.03400110        24.220200       0.0000
Laser Placement                        -0.0377554     0.19178400        -0.196865       0.8455
Lens Placement                         -0.0245380     0.07391540        -0.331975       0.7427
Fiber Alignment                        -1.4984400     0.45051400         3.326060       0.0027
Laser Facet Power*Laser Placement      -0.0211750     0.00720822        -2.937620       0.0070
Laser Placement*Lens Placement         -0.0336413     0.01567000        -2.146850       0.0417
-------------------------------------------------------------------------------------------
R-squared = 98.8139 percent, R-squared (adjusted for d.f.) = 98.5292 percent
Standard Error of Est. = 0.509698
```

Residual Plot for Optical Output Power



Estimated Response Surface

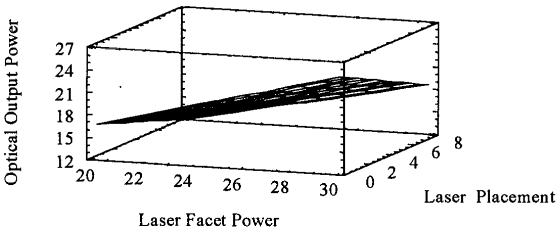


Figure 12 Results of Optical Output Power Experiment.

Based on costs, chosen risk levels associated with detecting assignable cause variation, magnitude of process change to be detected, and rational subgrouping considerations, a sampling scheme to weld and measure the first five parts of each shift was implemented.

5.4.2. Step 9: Establish Statistical Tools to Identify and Control Variation

Many statistical tools have been developed to control critical process parameters. The most commonly used is the control chart, which is an effective way to monitor and control processes and can be defined for both variables and attributes data. The selection of variables data will typically make basic statistical tools more efficient (i.e., lower sample size requirements to achieve necessary confidence levels).

Example: Fiber-Alignment Platform

In order to monitor and control this process, an X-bar and R control chart for fiber alignment (delta insertion loss) was established and data collection and analysis were implemented. These charts were used to identify and address assignable cause variation.

5.4.3. Step 10: Comprehend Customer and Design Specifications

The implementation and validation of design models in manufacturing ensures alignment of specifications. Implementation includes:

- Analyzing design specifications relative to the process capability
- Examining and understanding the reasons for process specifications
- Determining and implementing statistically valid specifications
- Enhancing the specification-alignment process

Example

Based on customer requirements and platform capabilities, the design model was used to align specifications to meet requirements. In this example, based on a customer requirement for optical output power of 15–24 mW, existing fiber-alignment platform requirement of <1 dB, and the design model, the following process tolerances were generated for the remaining processes:

Parameter	*Tolerances*
Laser facet power	22–28 mW
Laser placement	5 μm maximum
Lens placement	5.8 mils maximum

From these process tolerances, the fiber-alignment platform capability was maintained and the ability to meet customer requirement was met at a Cpk of 1.33. Depending on costs and other considerations, additional trade-offs between design and process requirements can be established.

5.4.4. Step 11: Assess Process Performance and Establish Statistical Limits

Assessment of process performance is a systematic procedure that uses statistical tools (usually control charts) to detect and eliminate (via corrective action) the unnatural (assignable) causes of variation until a state of statistical control is reached and hence statistical limits are established. A state of statistical control is a condition describing a process from which all assignable causes of variation have been eliminated via corrective action and only natural causes remain.

Example: Fiber-Alignment Platform

After assignable causes were eliminated, statistical limits were established for the control chart. Figure 13 shows the X-bar and R control charts for the last 90 data points.

5.4.5. Step 12: Develop Corrective Action Plans

Development and timely execution of corrective action procedures for out-of-control conditions is one of the most critical elements of effective implementation of process control. Corrective action procedures must be documented and should include the following information:

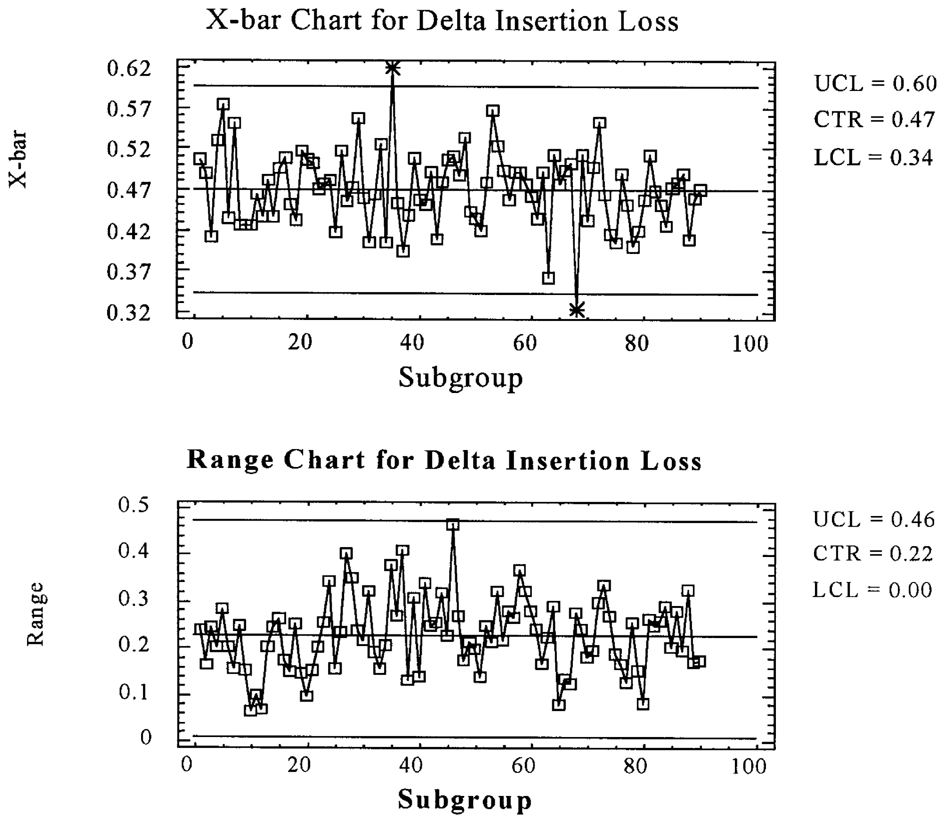


Figure 13 X-Bar and Range Chart for Fiber Alignment Process.

- When to take corrective action (i.e., when control chart signals an out-of-control condition)
- Who should initiate the corrective action procedure (i.e., operator, engineer)
- What corrective action should be performed per each failure mechanism
- What should be the disposition of the suspect devices/lots

It is also important to record properly the corrective action performed for Pareto analysis to further enhance the corrective action procedure.

Example: Fiber-Alignment Platform

Corrective action procedures were documented for this process and included items on verifying process conditions, such as weld beam energy and fiber gap position, establishing failure mode analysis procedures, determining disposition of work in process, and confirming procedures for restarting the process.

5.5. Capability Analysis (Steps 13–16)

Once a state of statistical process control has been achieved, process capability can be estimated.

5.5.1. Step 13: Compute Process Capability (Cp, Cpk)

In order to understand the true process capability for the critical processes, it is important to compute process capability indices using proper statistical procedures. These indices estimate the process's ability to meet design and customer requirements. Two of the most common capability indices are Cp and Cpk, where:

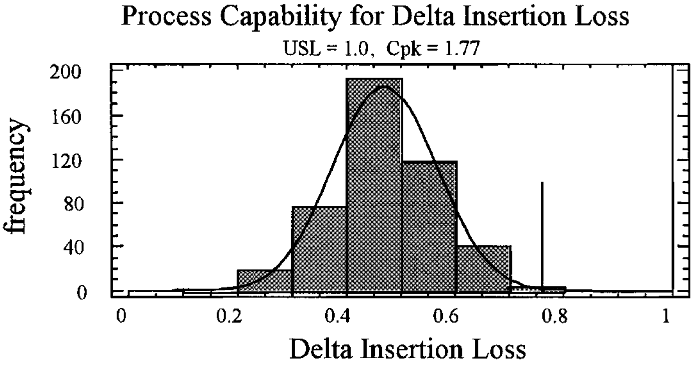


Figure 14 Fiber-Alignment Process Capability Analysis.

$$Cp = USL - LSL/6 \text{ sigma}$$

$$Cpk = \text{Minimum of } \{Cpl, Cpu\}$$

where

$$Cpl = \{\text{mean} - LSL\}/3 \text{ sigma}$$

$$Cpu = \{USL - \text{mean}\}/3 \text{ sigma}$$

A standard method for computing and reporting process capability indices for the critical designs and processes should be developed. Following are some of the key elements that should be considered for a process capability algorithm and report:

- Assessment and identification of statistical outliers from the data set
- Estimation of process capability indices for nonnormal data
- Warning for special conditions (i.e. insufficient data, multimodal distribution and excessive outliers)
- Estimation of Cp and Cpk indices, confidence intervals, summary statistics

Example: Fiber-Alignment Platform

The process capability index (Cpk) was estimated to determine the ability of the process to meet design objectives. Based on last 450 observations, the Cpk was estimated to be 1.77 and is shown in Figure 14.

5.5.2. Step 14: Establish Baseline and Set Objectives

For each process, estimates of process capability and yield for critical processes should be summarized and reported (baseline). Based on overall business objectives, realistic and attainable objectives for process capability and first pass yield should be set.

5.5.3. Step 15: Determine Gap between Baseline and Objectives

A gap analysis should be performed between the current baseline and objectives for each critical process to determine the extent of the improvement plans.

5.5.4. Step 16: Develop and Implement Action Plan to Close Gap

Based on gap analysis, an action plan should be developed and implemented to close the gap and promote continuous improvement.

6. LINKAGE OF PRODUCT DESIGN AND PROCESS PLATFORMS

In order to use the product and platform knowledge gained through characterization effectively, it needs to be documented and available to appropriate functions. This allows an organization to respond to customers more effectively in terms of new product development. One method is the development of platform and design manuals that would contain key aspects of platform elements and design characterization.

6.1. Process Platform Manual

6.1.1. Platform Description

A brief description of the function of the platform process should be included. The description should uniquely identify the scope and purpose of the particular platform.

6.1.2. Platform Characteristics

A list and associated description of the key attributes of the platform needs to be included. Attributes should be defined to document platform components sufficiently. These elements include hardware or equipment sets being used as well as necessary software algorithms. Elements also include what incoming materials are used on the process. Incoming materials may be from internal or external suppliers. When environmental controls are necessary, such as clean-room conditions, that also should be included. Also, items related to carrying out the process, such as methods, skills, and process conditions, should be summarized.

6.1.3. Platform Boundary Conditions

Boundary conditions should be established for each of the key platform characteristics. For each characteristic, the platform developer must determine the limits to which a characteristic can vary and still be manufactured on the platform to meet process and product features. These limits define the boundary conditions. These conditions include tolerance requirements on piece parts and materials, equipment and software specifications, and training requirements. These overall boundary conditions (i.e., what a process can and cannot do) will be the basis of the design guidelines that will be shared with design and marketing organizations.

6.1.4. Platform Characterization Model

For each performance characteristic, a process model must be developed and defined. The process model is a cause-and-effect quantitative model that establishes the relationships between process requirements and process variables. The process model identifies key process inputs and level of control (i.e., tolerances) needed to achieve desired process performance.

6.1.5. Platform Capability

For each performance characteristic, the associated process capability needs to be defined in terms of the ability of the process to meet its target and in terms of reproducing results consistently (i.e., process variation). The following items are typically included in the capability section of the platform manual:

- *Performance parameter*: lists critical platform outputs, as typically defined by design characterization activities, that must be met.
- *Process model*: establishes the relationship between platform inputs and performance parameters (platform outputs).
- *Parameter boundary conditions*: establishes platform requirements in terms of target performance that a platform will achieve. This may be a nominal target, single target, multiple targets, or a range of targets depending on the scope of the platform.
- *Nominal Performance (μ)*: defines actual performance of how well the platform is meeting its target.
- *Capability (1σ)*: defines the process variability associated with respect to each performance parameter.
- *Measurement methods*: describes the measurement method used to obtain performance results.
- *Measurement error*: defines the measurement error (1σ) for each performance parameter.

6.1.6. Platform Manual Example

6.1.6.1. Fiber-Alignment Platform Manual

- *Platform description*: The process consists of aligning a fiber to the lens for minimizing optical power loss and then welding the fiber to the lens while maintaining fiber alignment. A measure of this fiber alignment is the change in power loss due to welding (i.e., delta insertion loss).

Platform/Process Characteristics and Boundary Conditions (Design Rules)

Elements	Platform Characteristics	Boundary Condition
Hardware/equipment	Welder	
Software (algorithms)	Standard	Standard
Incoming materials	Lasers, Package, 3 × 3 Laser Carrier, Lens, Optical Fiber	3 × 3 laser carrier dimensions per drawing C1501 Package foot print dimensions per drawing C1432 Laser chip dimension of 20 by 25 mils Lens diameter (775 to 825 μm) and spherical shape Lens holder size must be 10–12 mm
Elements	Platform Characteristics	Boundary Condition
Environment	Class 10,000 clean room	Class 10,000 clean room
Method	Welding	Welding
Skills	Photonic processor	Photonic processor
Performance level	See below[1]	See below[2]

Platform/Process Capability (Performance Level)

Performance Parameter	Process Model	Parameter Boundary Condition[2]	Nominal Performance (μ)[1]	Capability (1σ)	Measurement Method	Measurement Error (1σ)
Delta insertion loss	Weld energy, weld pattern	1.0 dB maximum	0.47 dB	0.1 dB	Power meter	0.02 dB
Torque	Weld energy, weld pattern, focus, beam balance	10 in./lb minimum	17.8 in./lb	1.2 in./lb	Destructive	$<$ 0.5 in./lb

6.2. Design Manual

As a platform manual defines platform capability and requirements, a design manual defines design capability and requirements. Below is the description of a design manual.

6.2.1. *Design Description*

The design description covers a brief overview of the design scope and intent, including a description of the design construction, physical dimensions, electrical, optical, mechanical, and environmental objectives.

6.2.2. *Design Parameters and Requirements*

This includes a brief description of design parameters and their functions for a given product or family of products. The description may also include any special test conditions or requirements. In addition, specific customer requirements that must be met for each performance characteristic, including reliability, performance, maintainability, producibility, testability, safety, and cost objectives.

6.2.3. *Design Characterization Model*

For each performance characteristic, a design model should be developed and defined. The design models are cause-and-effect quantitative models that translate product performance requirements into process requirements and identify key in-process output parameters and associated tolerances.

6.2.4. *Design Capability*

For each design parameter, its associated capability needs to be defined and documented in terms of current and potential platform capabilities. Including potential capabilities enables improved market plans and offerings to be developed because design possibilities will be known.

6.2.5. *Design Boundary Conditions*

Boundary conditions should be established for each design parameter that define requirements to achieve design performance. These conditions define design limitations based on process and technology capabilities. They also provide marketing with the ability to develop better plans and respond to customer requests in a more timely fashion.

6.2.6. Design Manual Example

6.2.6.1. Transmitter Design Manual

- *Design description*: The transmitter is designed to operate at 2.5 Gb/sec with a NRZ data input format. The elements include housing, electrical interface, optical interface, drive circuitry, temperature control, optical sensors, data buffers, modulator, and attenuator.
- *Design parameters*: The absolute maximum ratings for supply voltage range from –5 to +5 volts with storage temperature range 0–65°C. The electrical and optical requirements include power supply current $<$ 500 mA, optical power output 15–24 mW. The environmental and mechanical requirements include 5 temperature cycles at 0–70°C. The targeted wavelength must be between 1523–1527 nm.

- Design Boundary Conditions
  - Lens must be spherical with a diameter range of 775–825 μ
  - Optical fiber must be single mode and polished at 5° with polishing flatness not to exceed 10 μ
  - Laser far-field angle not to exceed 35°
  - Laser chip facet coating reflectivity should be between 80–88%.

Design Capability

Performance Parameter	Design Model	Design Specifications	Nominal Performance (μ)	Capability (1σ)	Measurement Method	Measurement Error (1σ)
Optical output power	Laser facet power, laser placement, lens placement, fiber alignment	15–24 mW	19.5 mW	1.1 mW	Transmitter test set	0.25 mW
Targeted wavelength	Current, temperature	1523–1527 nm	1525 nm	0.5 nm	Wavelength test set	0.05 nm

7. DEPLOYMENT

Well-defined methodologies with broad objectives do not lead directly to results. They must first be deployed. However, successful deployment is by itself insufficient to guarantee program success. To maximize benefits and ensure longevity, the methodology must become ingrained as part of the culture of the organization. Achieving this is a complex process. It involves, at a minimum, recognition by the user that the methodology is a benefit to them, coupled with incorporation of the methodology into the structure of the business management processes that govern the organization. The following sections describe the deployment of the product design and process platforms methodology and the steps necessary to make it part of the culture of an organization. A case study of one organization is presented as an example.

7.1. Critical Components

In the deployment of the design and process platform characterization program, the critical components of success are, as expected, user education, management commitment, deployment infrastructure, and business process integration. A more subtle and ubiquitous component of success must also be recognized—the celebration of every success, no matter how small.

7.1.1. Education

The foundation upon which any program deployment is built is the knowledge base of the participants. In most organizations, the technical expertise in the use of the principles and the acknowledgment of their utility resides in the quality organization and the population at large has little or no training/recognition of this area. Hence, the initial step in the deployment must be the education of the user community. In addition, for a culture change to occur, the entire populace must speak the same language. Therefore, customized courses must be designed for all levels in the business, from manufacturing operators to the executive leadership team, including the chief operating officer (COO) of the business.

Example

The entire executive management team of the business was required to attend an overview session on the basics of measurement system characterization, design of experiments, statistical process control and process capability analyses. The objective of this course was to provide an overview of the principles of these methods, including how it works, when to use it, and what results to expect. The overall objective was to provide the manager with just enough information so that he or she could ask the right questions of the engineers and in so doing reinforce the objectives of the program.

All R&D and manufacturing engineers were required to attend 48 hours of training in measurement system characterization (4 hours), design of experiments (32 hours), statistical process control (8 hours), and process capability analysis (4 hours). These courses were presented on site by recognized experts in those fields from universities around the country. In addition, these courses were customized to its industry and each participant was required to bring his or her own problem to the class to be worked as an exercise. This brought relevance to the sessions and achieved timely recognition by the participant that “this stuff works.” Over 90% of the engineering staff completed this training within 18 months.

Advanced training was also provided on a voluntary basis. Courses in advanced design of experiments/multivariate analysis (24 hours), regression analysis/ANOVA (16 hours), and DOE, modeling, and SPC using a platform software tool (16 hours) were provided. Seventy percent of the engineering population completed this level of training. These courses were also offered by external technical experts in those fields from the university community.

Finally, the manufacturing associates had required training in statistical process control (8 hours), which was provided by the quality organization. Eighty-five percent of the staff had completed this training in a 12-month period. Advanced training was also made available on a voluntary basis and 60% of the manufacturing staff completed this level of training.

The result of all this education/training was a whole population of believers in these methods and an engineering staff that uses them as a matter of routine.

7.1.2. Management Commitment

A company can undertake such drastic change in quality methodologies only if its executive managers are trained and supportive and are participants in the program. Management commitment can be demonstrated in many ways. However, the effectiveness of any specific approach will depend on the existing culture of the organization.

Example

An approach based on constant recognition of compliance rather than exception-based reporting was used. Teams at various levels of completion presented their results at the quarterly quality leadership council meetings with the executive leadership team and the COO in attendance. Within a 12-month period, all the teams in the business had this opportunity to demonstrate their progress. It was not required that a program be complete, only that it be actively worked in the methodology to get positive reinforcement from upper management.

The reward system for managers and engineers was tied to completion of milestones. This included the annual performance management system and special rewards. Every small success was celebrated, including participation. These were annual celebration banquets held off-site for the entire engineering staff. At this annual event, the teams that had progressed furthest in the program presented their report to their peers and the entire management staff, including the executive team and COO. Constant positive reinforcement with public recognition was a key component of success. The COO was the recognized leader of this activity. He reinforced his leadership role by including references and/or results of this program at every presentation he made to the organization. In addition, he was committed to personal involvement by meeting engineers on a regular basis and asking about their progress and offering any help to eliminate roadblocks.

Finally, to ensure that all levels of management were involved, it was required that first-level managers report the progress of the teams in their area of responsibility at the quality leadership council meetings.

7.1.3. Deployment Infrastructure

The success of every quality initiative is dependent on its acceptance by the receiving organizations. This acceptance is critically dependent on the means of deployment used. Successful deployment of this program cannot result solely from quality consultants assigned to support each functional area. Progress must be driven from within the functional area. Only then can a culture change result.

Example

Champions were designated in each functional area to drive and monitor actively and report progress. These were either the first-level managers or lead engineers in those areas. The methodology was promulgated by the active involvement of these acknowledged and recognized technical leaders. Engineers and managers had been trained and now were expected to apply that training according to the outline of the methodology. They were supported by the quality engineering staff, who provided significant assistance, guidance, and technical leadership initially, which decreased as time passed and competency increased.

The engineering community identified its critical designs and/or processes to which the methodology would be applied. Together with the quality engineering staff, individual customized project plans were developed. Biweekly meetings were held wherein all champions reported their progress and problems. Issues were escalated from these meetings and resolution was swift. In addition, sharing of problems resulted in opportunities for further education of the champions. By the end of the first 12 months, the champions were acknowledged by the engineering and management staff as subject matter experts in the methodology and its application.

7.1.4. Integration

Initially this program must be an adjunct to the quality system that governs the business. It must be viewed as a quality-improvement program and not officially required for design, production, or shipment of product. Since its application initially will be sporadic, management should not wish to raise customer expectations on a program that will not be uniformly applied throughout the business for some time. As customers are apprised of the methodology, they will desire it to be applied to their products.

Example

It took nearly three years of application of this methodology before it was widespread and uniformly applied sufficiently to formally incorporate it into the quality system. The quality system defines two levels of field grade products to customers: model code and full code. The latter has all aspects of the quality system applicable and complete. The former has some aspects of the quality system incomplete but recognized as low risk for customer use. The application of this methodology and its linkage to the quality system via the product level is shown in Figure 15. Review of the qualification review board (QRB) plan before acceptance and progress reporting by quality assurance in the regular monthly executive report are the significant linkages. In addition, the product robustness criteria used for QRB acceptance were also formalized and are shown in Figure 16.

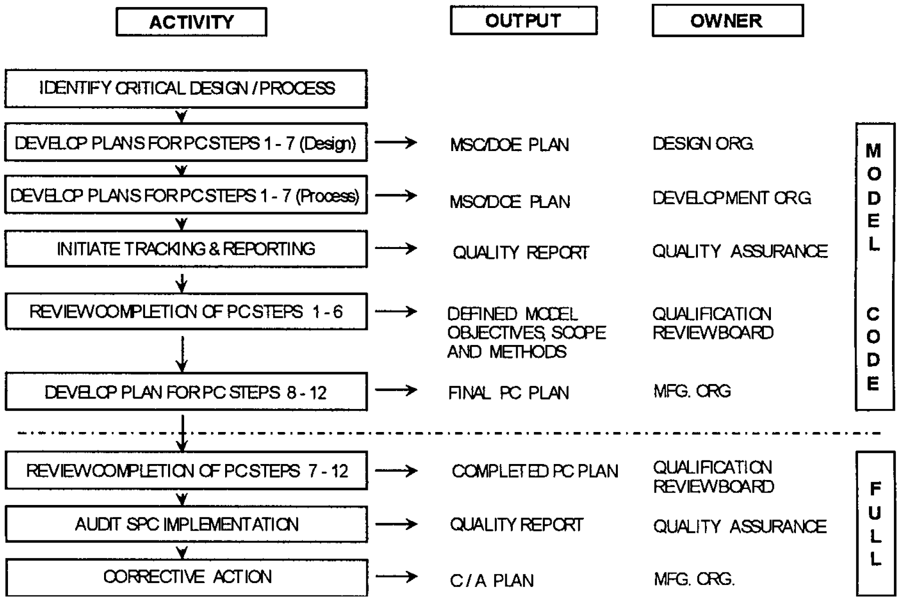

ACTIVITY
OUTPUT
OWNER
IDENTIFY CRITICAL DESIGN / PROCESS
DEVELOP PLANS FOR PC STEPS 1 - 7 (Design)
MSC/DOE PLAN
DESIGN ORG.
DEVELOP PLANS FOR PC STEPS 1 - 7 (Process)
MSC/DOE PLAN
DEVELOPMENT ORG.
INITIATE TRACKING & REPORTING
QUALITY REPORT
QUALITY ASSURANCE
REVIEW COMPLETION OF PC STEPS 1 - 6
DEFINED MODEL OBJECTIVES, SCOPE AND METHODS
QUALIFICATION REVIEW BOARD
DEVELOP PLAN FOR PC STEPS 8 - 12
FINAL PC PLAN
MFG. ORG
MODEL CODE
REVIEW COMPLETION OF PC STEPS 7 - 12
COMPLETED PC PLAN
QUALIFICATION REVIEW BOARD
AUDIT SPC IMPLEMENTATION
QUALITY REPORT
QUALITY ASSURANCE
CORRECTIVE ACTION
C/A PLAN
MFG. ORG.
FULL

Figure 15 Quality System Linkage.

Modeling:	• Completion of Design Models (PC Steps 1-7) • Completion of Process Models (PC Steps 1-7)
SPC:	• Implementation of SPC for outputs/inputs (PC Steps 8-12)
Capability:	• Design Parameters Capable - Goal @ Cpk $\geq$ 1.33 (PC Step 13-16) • Process Parameters Capable - Goal @ Cpk $\geq$ 1.33 (PC Step 13-16)
Platforms:	• Use of Platform Processes

Figure 16 Product-Robustness Criteria.

Included and implicit in these criteria are achievements of minimum yield and variation targets before a design is transferred to manufacture. As with all the metrics in this program, these are customized for each design and process with regard to the complexity and technological capability of the design and process.

Reviews of the activity, required by the quality system, are the critical components to ensuring compliance and guaranteeing longevity of the program. It must be independent of the individuals administering or driving the program and ingrained in the fabric of the quality culture of the organization. It must be linked via appropriate documentation to the quality system, which are then audited regularly. It must simply become part of the way a company does business.

7.2. Performance Measures

Corporations generally set broad quality goals and leave the path to achievement to the discretion of the individual business units. However, these broad goals do not lead directly to results. They must first be deployed to lower levels of the organization via division and subdivision of the objectives until they identify specific activities to be carried out and allocation of responsibility for performing those activities. Execution of the activities incorporated in this program collectively results in improving operational excellence.

In addition, reviews of progress are an essential part of ensuring that goals are being met. The very fact that an executive team reviews progress sends a message to the rest of the organization as to the priority given to the program goals. The metrics chosen to review are critical to the success of the program. They must show steady progress to maintain program momentum, and at the same time they must measure actual operational performance, a much slower improvement process. This is a difficult but achievable combination.

7.2.1. Metrics

The quality adage ‘‘You improve what you measure’’ should be the operational credo. There are two categories of measurements: progress toward compliance with the methodology and operational performance. Progress reviews are also an essential part of ensuring that application of the methodology is resulting in the expected improvements. Reviews should be carried out in two major ways: summarized reports on actual operational performance and reports of program status.

Modeling:	• Percent of Design Output Parameters Modeled • Percent of Process / Platform Output Parameters Modeled
SPC:	• Percent of Process / Platform I/O Controlled
Capability:	• Percent of Design Output Parameters Capable at Cpk $\geq$ 1.33 • Percent of Process / Platform Output Parameters Capable at Cpk $\geq$ 1.33
Education:	• Percent of Engineering & Operating Completed PC Training

Figure 17 Metrics.

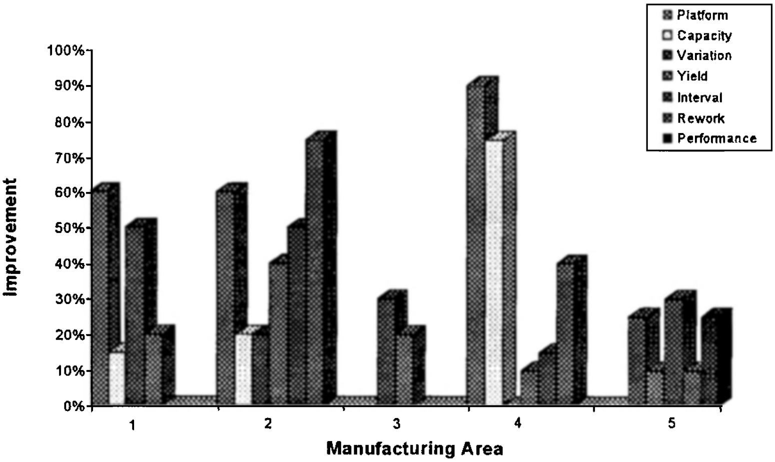


Figure 18 Performance Measures.

Example

The compliance metrics are shown in Figure 17. At the same time, measures were designed for key characterization elements and link to variation, performance, yield, and interval improvements. Results of improvements in operational performance for five manufacturing areas are shown in Figure 18.

Measures of each customized project plan and its specific schedule and reported progress against the steps of the methodology were summarized in a monthly report and also presented at the quality leadership meetings by the first-level manager with the COO and the executive team. In conducting these reviews, it was necessary (albeit difficult) for upper managers to maintain a constructive approach. In general, metrics were used to track progress and define cause for celebration.

7.2.2. Progress of Culture Change

The overall goal is to create a process control-minded culture and exceed customer expectations. On the road to achieving that goal, the foundation and a set of platforms unique to the specific industry should be developed. These goals can be deployed via the design and process platform methodology.

Example

After five years, a fundamental change in the culture of the business was demonstrated. The use of design of experiments is a way of life in this business. Designers and process engineers use it routinely (without quality organization expert assistance) to characterize and control designs and processes. In a three-year period, over 250 DOE were completed resulting in significant yield, variation and cycle time improvements. About 40–50 DOE are in progress at any given time. The designers and process engineers gained sufficient technical expertise to perform the quality engineering function as part of their existing assignments.

The methodology also became an integral part of the product-realization process. Comprehensive implementation of the methodology and achievement of minimum yield and variation targets are required before the design is transferred to manufacturing.

Although there was strong resistance at first, this changed to global acceptance by the technical community as their experience increased. The improvement in operational performance metrics led to a total embracing of the methodology by the management team.

Finally, the keys to success were the active involvement and support of executive management, recognition at executive meetings and peer events, education, and user involvement in the methodology development.

Computer Software

STATGRAPHICS, Manugistics, Rockville, MD.

ADDITIONAL READING

Box, G. E. P., and Draper, N. R., *Empirical Model Building and Response Surfaces,* John Wiley & Sons, New York, 1981.

Duncan, A. J., *Quality Control and Industrial Statistics*, 5th Ed., Irwin, Homewood, IL, 1986.

Feigenbaum, A. V., *Total Quality Control*, 3rd Ed., McGraw-Hill, New York, 1983.

Montgomery, D. C., *Design and Analysis of Experiments*, 4th Ed., Wiley, New York, 1997.

Goff, D. R., Hansen, K. S., and Stewart, J. G., *Fiber Optic Reference Guide*, Focal Press, Boston, 1997.

Grant, E. L., and Leavenworth, R. S., *Statistical Quality Control*, 6th Ed., McGraw-Hill, New York, 1979.

Hecht, J., *Understanding Fiber Optics*, 2nd Ed., Sams, Indianapolis, IN, 1993.

Juran, J. M., *Juran on Quality by Design: The New Steps for Planning Quality into Goods and Services*, Free Press, New York, 1992.

Juran, J. M., and Gryna, F. M., *Quality Control Handbook*, 4th Ed., McGraw-Hill, New York, 1988.

Western Electric Company, Inc., *Statistical Quality Control Handbook*, Delmar, Charlotte, NC, 1982.

IV.F

Quality

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 66

Total Quality Leadership

JOHNSON A. EDOSOMWAN
Johnson & Johnson Associates, Inc.

1. INTRODUCTION AND KEY DEFINITIONS 1793
2. TOTAL QUALITY LEADERSHIP SYSTEM ELEMENTS 1795
3. THE TOTAL QUALITY LEADERSHIP: EIGHT PILLARS OF QUALITY 1796
4. TOTAL QUALITY LEADERSHIP TOOLS AND MODEL 1798
  4.1. PADER Scoring System 1800
5. IMPLENENTATION OF THE TOTAL QUALITY LEADERSHIP PROCESS 1801
  5.1. Step One: Provide Leadership and Vision for Performance Excellence 1801
  5.2. Step Two: Communicate the Continuous Performance Improvement Philosophy 1801
  5.3. Step Three: Establish a Formal Continuous Performance Improvement Focal Point 1802
  5.4. Step Four: Take Immediate Action to Increase the Amount of Interdepartmental Teamwork 1802
  5.5. Step Five: Introduce the Continuous Performance Improvement Plan to the Team 1802
  5.6. Step Six: Implement Comprehensive Performance Measures at All Levels 1803
  5.7. Step Seven: Develop New, Improved Processes to Eliminate Outdated Ones 1803
  5.8. Step Eight: Monitor and Recognize Progress Toward Performance Improvement Goals 1803
6. SELECTED TQL SUCCESS FACTORS AND CONCLUSIONS 1804
REFERENCES 1805
ADDITIONAL READING 1806

1. INTRODUCTION AND KEY DEFINITIONS

Global competition and customer demands for competitive quality products and services require that both public- and private-sector organizations implement results-driven continuous performance improvement approaches and systems. According to Hiam (1993), between 75 and 80% of large companies adopted quality management approaches within the last decade. Malhotra et al. (1994) also point out that quality continues to be a top-ranked strategic issue among top manufacturing executives.

The terms *total quality* (TQ), *total quality management* (TQM), *total quality improvement process* (TQIP), and *total quality leadership* (TQL) have been used interchangeably in the literature and in organizations embarking upon a comprehensive quality-driven performance improvement process. Evans (1992) defines total quality as a people-focused management system that aims at continual increase of customer satisfaction at continual lower real cost. Total quality is a total systems approach (not a separate area or program) and an integral part of high-level strategy; it works horizontally across functions and departments; involves all employees, top to bottom; and extends backwards and forwards to include the supplier and the customer chains. Practitioners and researchers agree that when all the elements of total quality are fully implemented with leadership support and commitment, there are bound to be positive business results. Edosomwan (1994, 1998a) defines total quality lead-

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

ership as a systematic, continuous improvement process that involves measuring, planning, evaluating, and improving all aspects of an organization's performance. It involves a customer- and people-driven process that is led by senior management vision and commitment, with full participation from organizational stakeholders to produce error- and defect-free products and services for end users. The TQL process utilizes quantitative, qualitative, behavioral, technological, technical, and managerial tools and techniques to improve organizational structures, work processes, policies, procedures, products, and service-delivery systems. Edosomwan (1994) points out that total quality leadership encompasses all aspects of the total quality management and improvement process. It is also important to point out that the definition of quality varies quite widely. As shown in Table 1, Edosomwan (1998d) documented 10 approaches and major definitions of quality appearing in the literature.

Various critical dimensions of TQM, TQL, and TQIP are identified in the literature. For example, Whitney and Pavett (1998) note that management support for the improvement process is the genesis, without which little else will happen. Almaraz (1994), Edosomwan (1990, 1998a), Crosby (1979), Juran (1974), Deming (1981), Hackman and Wageman (1995), and Westphal et al. (1997) discuss other important dimensions that are critical to the successful implementation of the total quality

TABLE 1 Ten Key Approaches and Major Definitions of Quality

Approaches	Authors	Definitions
Customer based	Juran 1974	"Quality is fitness for use."
	Edwards 1968	"Quality consists of the capacity to satisfy wants."
	Edosomwan 1988	"Quality is conformance to customer requirements."
Product based	Crosby 1979	"Quality is conformance to requirements."
	Deming 1981	"Quality is durability of product."
Performance based	Edosomwan 1999	"Quality is a measure of product and service delivery system performance."
Management based	Deming 1981	"Eighty percent of the quality problems are caused by organization management system."
	Edosomwan 1998d	"The quality of management drives performance."
System based	JISC 1981	"[Quality is] a system of means to economically produce goods or services which satisfies customers requirements."
Transcendent	Pirsig 1974	"Quality is neither mind nor matter, but a third entity independent of the two, even though quality cannot be defined, you know what it is."
Process based	Deming 1982	"Quality is controlling process variation."
	Edosomwan 1994	"Quality is a measure of process performance."
Value based	Feigenbaum 1983	"Quality means best of certain conditions: (a) the actual use, and (b) the selling price."
Prevention based	Edosomwan 1994	"Quality is preventing errors and defects."
Technology and culture based	Sashkin and Kiser 1993	"Quality means that the organization's culture is defined by and supports the constant attainment of customer satisfaction through an integrated system of tools, techniques and training."

Source: Edosomwan 1998d.

improvement process. These dimensions include, but are not limited to: (1) management practices, (2) customer focus, (3) supplier quality and performance, (4) employee development and training, (5) performance incentives, (6) technology, (7) cultural and relationship factors, (8) workforce attitude, (9) process management, (10) performance measures, (11) strategic planning, (12) problem-solving process, (13) product and service-delivery systems, and (14) labor and management relations. However, implementing the total quality leadership process definitely requires the commitment, support, and dedication of the organization's senior managers and leaders.

2. TOTAL QUALITY LEADERSHIP SYSTEM ELEMENTS

According to Almaraz (1994), implementing the total quality improvement elements and process requires a transformation in the organization's culture, processes, and strategic priorities and in individual attitudes, beliefs, work ethics, and behavior. Edosomwan (1995, 1998a) identifies four total quality leadership (TQL) system areas: the management system, the social system, the technical system, and the behavioral system (see Figure 1).

1. *The management system* encompasses the way that policies, procedures, practices, protocols, and directives are established, enforced, and maintained. The leadership systems of the organization set the tone and vision and provide indicators of what should be done, how it should be done, and what should be accomplished. The management system carries into effect strategies, processes, and project management, and it encompasses the vision, mission, and values of the organization.
2. *The social system* has a significant impact on motivation and the ability to implement new ideas; it addresses organizational culture, structure, rewards, teamwork, values, and the creativity of individuals and groups. The social system is influenced by the values of the founders, leaders, families, peers, and supervisors, as well as group behaviors. Before transformation, the state of the organization is usually influenced by rigid rules and lack of focus on customer requirements, constancy of purpose, and continuous improvement.
3. *The technical system* includes the tools, techniques, and mechanisms necessary to produce excellent products and services. The technical system also involves work processes, technol-

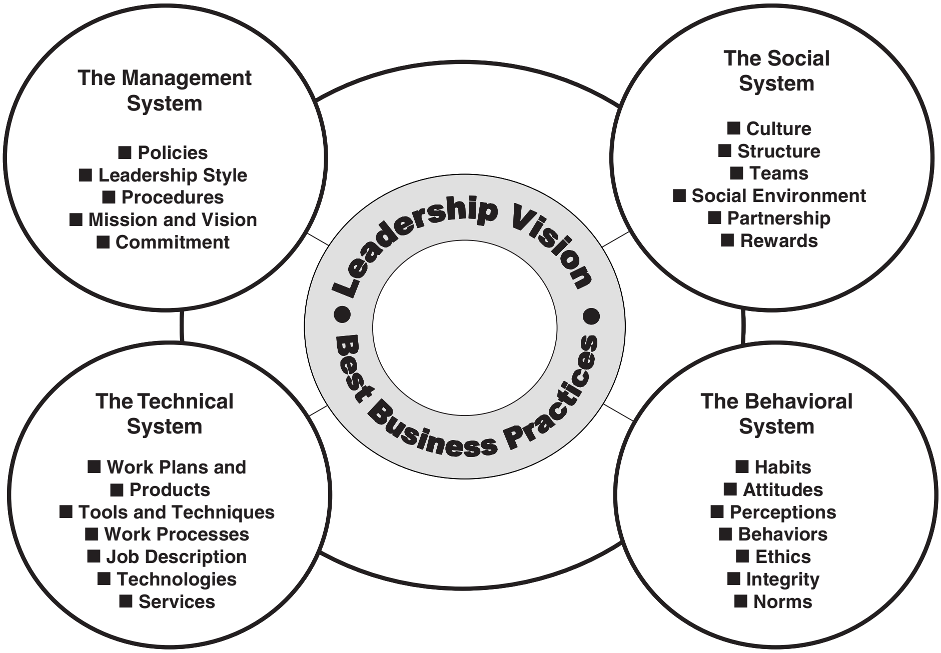

The Management System
■ Policies
■ Leadership Style
■ Procedures
■ Mission and Vision
■ Commitment
The Social System
■ Culture
■ Structure
■ Teams
■ Social Environment
■ Partnership
■ Rewards
Leadership Vision • Best Business Practices •
The Technical System
■ Work Plans and
■ Products
■ Tools and Techniques
■ Work Processes
■ Job Description
■ Technologies
■ Services
The Behavioral System
■ Habits
■ Attitudes
■ Perceptions
■ Behaviors
■ Ethics
■ Integrity
■ Norms

Figure 1 Total Quality Leadership System Areas (From Edosomwan 1995).

ogies, and systems that are utilized in transforming organizational inputs to outputs and outcomes. The job and task descriptions are part of this system. The technical system pertains to measures, which serve as the basis for improvement and planning.

4. *The behavioral system* relates to the fundamentals of the human side of quality, as characterized by the habits, attitudes, work patterns, and behaviors of individuals and groups. Through modification of the elements in the behavioral system, it is possible to implement changes that can lead to significant breakthrough in performance. The behavioral elements are often difficult to change, and when a change is made, it positively influences the speed of organizational process transformation.

3. TOTAL QUALITY LEADERSHIP: EIGHT PILLARS OF QUALITY

Edosomwan (1995, 1998a) identifies eight pillars of quality, as shown in Figure 2 and described below. The eight pillars of quality must be continuously improved throughout the TQL process.

1. *Quality of management* pertains to the quality of organizational policies, procedures, values, and the resource-allocation system. It also addresses the quality of leadership and management decisions, supervision, and guidance provided to the workforce as it relates to the delivery of products and services.
2. *Personal quality* pertains to the quality of personal attributes, trust, and characteristics that enable individuals to maintain excellent work ethics and performance. The critical factors in this area include personal attitude, skills, abilities, work habits, behaviors, integrity, trustworthiness, loyalty, dedication, and commitment to the successful achievement of organizational and personal mission and priorities.
3. *Quality of service* involves several tangible and intangible factors, such as the degree of courtesy provided to internal and external customers. It also includes other factors, such as the appearance of the products and services being offered, responsiveness, and the extent to which customers receive adequate value-added service.
4. *Process quality* pertains to the quality of primary, secondary, and auxiliary work processes that are required to transform organizational inputs to outputs and outcomes. Primary work pro-

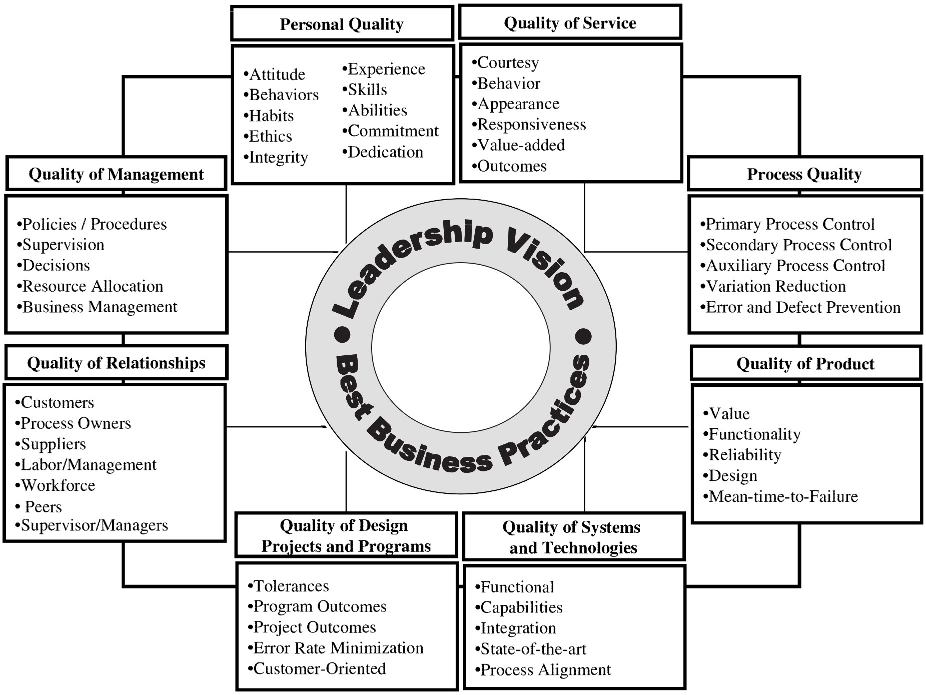

Personal Quality
•Attitude
•Behaviors
•Habits
•Ethics
•Integrity
•Experience
•Skills
•Abilities
•Commitment
•Dedication
Quality of Service
•Courtesy
•Behavior
•Appearance
•Responsiveness
•Value-added
•Outcomes
Quality of Management
•Policies / Procedures
•Supervision
•Decisions
•Resource Allocation
•Business Management
Process Quality
•Primary Process Control
•Secondary Process Control
•Auxiliary Process Control
•Variation Reduction
•Error and Defect Prevention
Leadership Vision • Best Business Practices •
Quality of Relationships
•Customers
•Process Owners
•Suppliers
•Labor/Management
•Workforce
• Peers
•Supervisor/Managers
Quality of Product
•Value
•Functionality
•Reliability
•Design
•Mean-time-to-Failure
Quality of Design Projects and Programs
•Tolerances
•Program Outcomes
•Project Outcomes
•Error Rate Minimization
•Customer-Oriented
Quality of Systems and Technologies
•Functional
•Capabilities
•Integration
•State-of-the-art
•Process Alignment

Figure 2 Eight Pillars of Quality (From Edosomwan 1998d).

cesses have impact on the external customers, while secondary processes have impact on the internal customers. Auxiliary work processes have impact at the job and task levels. Process quality also pertains to how well variation in process variables is managed, controlled, and improved continuously to ensure that the final products and services delivered to the end users are error and defect free.

5. *Quality of relationships* pertains to the quality of interaction and relationship among suppliers, process owners, customers, teams, work units, labor and management, and functional business areas. It also involves the quality of outputs and outcomes produced as a result of a defined relationship among suppliers, process owners and customers.
6. *Quality of design projects and programs* pertains to the quality of products and services at the design, program and project management, and delivery stages. It takes into account design specifications, tolerances, and quality of task outputs and outcomes involved in the design and development cycle of a product or service.
7. *Quality of systems and technologies* pertains to the functional quality of production and service delivery systems and technologies. It addresses the quality of the technology life cycle as well as the operational measures of system and technology performance. Measures such as mean-time-to-design, mean-time-to-failure, mean-time-to-service and -repair, and reliability are utilized to assess the quality of systems and technologies. It also addresses how well state-of-the-art systems and technologies are integrated to deliver error- and defect-free products and services to end users.

TABLE 2 Traditional Organizations vs. Customer-Driven Organizations

Area	Traditional Organizations	Customer-driven Organizations
Product and Service Planning	■ Short-term focus ■ Reactionary management ■ Management by objectives	■ Long-term focus ■ Prevention-based management ■ Customer-driven strategic planning process
Measures of Performance	■ Bottom-line financial results ■ Quick returns on investment	■ Customer satisfaction ■ Market share and profitability ■ Quality and total productivity
Attitudes Toward Customers	■ Customers are irrational ■ Customers create problems ■ Customer concerns are bottleneck to profitability	■ Identify and respond to the voice of the customer ■ Professional treatment and attention to customer needs
Quality of Products and Services	■ Provided according to organizational requirements	■ Provided according to customer needs and priorities
Marketing Focus	■ Seller's market ■ Careless about loss of customers	■ Long-term focus ■ Prevention-based management ■ Customer-driven strategic planning process
Process Management Approach	■ Focus on error and defect detection and risk management	■ Focus on error and defect prevention and total process management
Product and Service Delivery Attitude	■ It is fine for customers to wait for product and services	■ It is best to provide fast-time-to-market products and services just-in-time
People Orientation	■ People are the source of problems and are burdensome to the organization	■ People are the greatest asset and resource to the organization
Basis for Decision Making	■ Product-driven ■ Management by opinion	■ Customer driven ■ Management by facts and data
Attitudes Toward Customers	■ Hostile and careless ■ "Take it or leave it" attitude	■ Courteous and responsive ■ Empathetic and respectful
Improvement Strategy	■ Crisis management ■ "If it is not broken, don't fix it"	■ Continuous process improvement ■ Total performance management
Mode of Operation	■ Career-driven independent workers ■ Short-term planning and profitability	■ Management-supported improvement ■ Teamwork among customers, suppliers, and process owners

Source: Edosomwan 1994.

8. *Quality of product* pertains to product reliability, value, functionality, and dependability. End users also utilize mean-time-to-failure, mean–time-to-repair, and maintenance cycle time and costs to evaluate the quality of products.

The continuous improvement of the eight pillars of quality and the implementation of the TQL system and process enable organizations to move from a tradition-based culture to a customer-driven organization culture. Table 2 summarizes the major differences between traditional and customer-driven organizations.

4. TOTAL QUALITY LEADERSHIP TOOLS AND MODEL

Several models, tools, and techniques are available for improving the quality of products and services and managing the total quality leadership process. Greene (1993) reviewed quality practices in Japan and the United States and claims to have identified about 24 dimensions of total quality management (TQM). Edosomwan (1998a) identified several TQL tools summarized in Table 3 for improving quality and performance at the individual, work-unit, and organizational levels.

The Malcolm Baldrige National Quality Award performance excellence criteria have been widely used as a model for total quality leadership and a tool for achieving performance excellence. The Baldrige criteria and model focus on leadership as a driver of quality. Other aspects of the model include information and analysis, strategic planning, process management, human resource management and development, customer and market focus, and business results. Edosomwan (1998a, b, c, d) developed a more comprehensive performance excellence model, shown in Figure 3, that has been utilized by both public- and private-sector organizations for improving all aspects of performance.

This model is based on augmented, expanded performance criteria, an expanded scoring system, and a defined implementation methodology. The categories and key components of EPEM, described below, retain the components of the Baldrige criteria while providing a more comprehensive, universal criteria and performance-improvement system for both public- and private-sector organizations. There are 10 EPEM categories, as described below:

1. *Leadership* examines how the organization's senior leaders address and communicate the organization's vision, mission, values, and performance expectations, as well as their focus on employees, customers and other stakeholders, empowerment, innovation, learning, organizational direction, and coordination of tasks. Also examined is how the organization addresses its societal responsibilities and community involvement.
2. *Strategic planning* examines the organization's strategy development process, including how the organization develops strategic, tactical, and operational business plans that are aligned with organizational vision, mission, values, resources, budget, measures, and stakeholders' requirements.
3. *Customer and market focus* examines how the organization determines customer requirements, expectations, and preferences of customers and segments. Also examined is how the organization builds relationships with internal, external, and self-unit customers; determines satisfaction levels; and uses information to determine priorities.
4. *Information and analysis* examines the organization's process and approach for collecting, analyzing, and using performance data and other core information across the organization. It also includes the organization's performance measurement system, technology capabilities, data transfer, and information management system.
5. *Human resource development and management* examines how the organization trains and enables employees to develop and utilize their full potential, aligned with the organization's objectives. It also addresses the organization's efforts to build and maintain a work environment of empowerment and a culture and climate conducive to performance excellence, full participation, and personal and organizational growth, providing appropriate recognition and reward.
6. *Process management* examines key aspects of the organization's ability to define, manage, and continuously improve primary, secondary, and auxiliary work processes, including customer-focused design, product and service delivery, and process variation management.
7. *Culture systems and relationship management* examines the organization's ability to create an integrated, cohesive working environment of teamwork and cooperation that is sensitive to the needs and requirements of a diverse employee population. Also examined are the organization's policies, procedures, and actions that are deployed to ensure fair and equitable treatment of stakeholders and the employee population. It examines the organization's culture, environment, and relationships among labor leaders, managers, and labor and their ability to work together in partnership to meet and exceed the organization's goals and objectives.
8. *Technology integration and management* examines the organization's commitment to identification and evaluation of emerging technology related to new products and services, primary

TABLE 3 Selected Examples of Total Quality Leadership Tools

Plan Do Check Act Technique	Provides a methodology to assist individuals and teams in problem solving, planning, work analysis, and implementation of solutions
Personal Performance Improvement Model	Provides a model for assisting individuals in developing and implementing performance-improvement goals
Quality Error-Removal Technique	Provides a team-based approach for analyzing and removing error and defect sources
Statistical Process Control Technique	Provides statistical techniques and tools for controlling managing process variation
Force Field Analysis Tool	Provides a graphical means of understanding, quantifying, and balancing the positive and negative impacts related to goal achievement for individuals and teams
Brainstorming Technique	Provides a process for a group to quickly generate, clarify, and evaluate ideas, suggestions, problems, and issues to stimulate the creativity of individuals in a participative, group problem-solving process
Benefit Energy Dots Tool	Provides a structured process for multivoting and subjective prioritization of ideas and suggestions in a team problem-solving environment
Job Requirement Mapping Tool	Provides an approach and a working tool for defining and understanding the requirements of the job
Problem-Solving Approach	Provides the key steps for successfully resolving any problem by testing, defining and implementing solutions
Process Analysis Technique	Provides a technique for defining the steps of work processes and analyzing them to identify non-value-added steps and opportunity areas
Group Ideas Management Technique	Provides a methodology for generating ideas, solving problems in a team environment, and prioritizing the ideas for the optimal solution
Cause-and-Effect Diagram	Provides steps for constructing a cause-and-effect diagram for identifying root causes of a problem
Pareto Analysis Technique	Provides a graphical representation of identified causes presented in descending order of magnitude or frequency
Error Mapping Technique	Provides an approach for analyzing quantitative error data to identify the sources of errors and defects in work elements, tasks, and processes
Problem Statement Impact Definition Tool	Provides a comprehensive description of the actionable item and stratifies the items which an individual or team intends to analyze for performance improvement; helps individuals specify data and information relevant to a specific problem
Problem-Containment Model	Provides a framework for containing problems from irate individuals, teams, and the work environment
Complaint-Resolution Model	Provides a step-by-step process for resolving internal and external customer complaints
E-Alarmo Technique	Provides an approach for identifying and resolving breaks in customer service and handling irate customers
Process Reengineering Technique	Provides a methodology for reengineering and transforming organizational structures and work processes
Process Element Mapping Technique	Provides a step-by-step approach for mapping process steps, elements, requirements, and owners
PADER Technique	Provides a five-dimension scoring system for evaluating an organization's performance with focus on plan, approach, deployment, evaluation and results

Source: Edosomwan 1998a.

and support process performance, and people management. It assesses the organization's technology utilization, systems integration, and connectivity elements.

9. *Supplier performance management* examines the organization's commitment to and process for ensuring a high-performance supplier base capable of meeting and exceeding specifications and requirements for all goods and services procured by the organization.

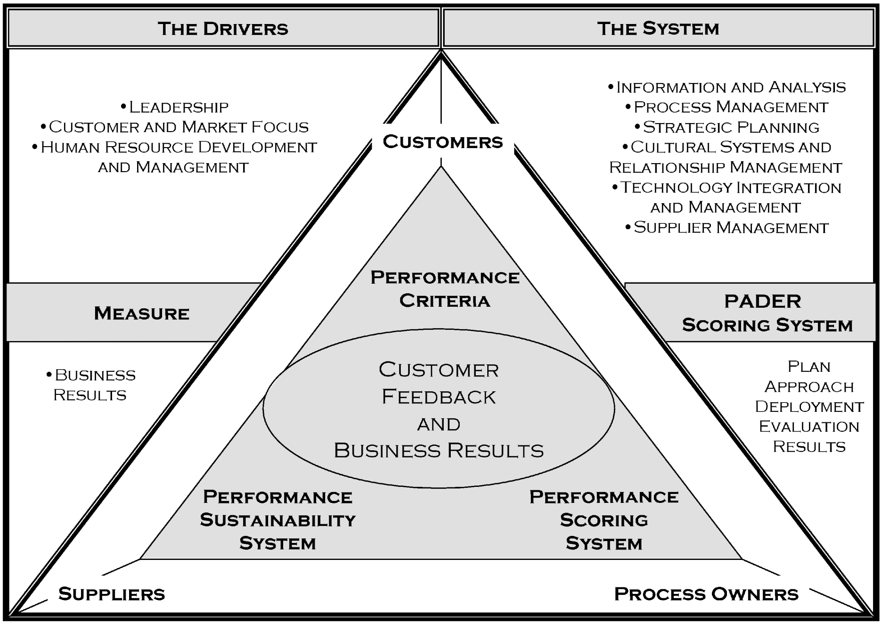

The Drivers
The System
• Leadership
• Customer and Market Focus
• Human Resource Development and Management
• Information and Analysis
• Process Management
• Strategic Planning
• Cultural Systems and Relationship Management
• Technology Integration and Management
• Supplier Management
Customers
Performance Criteria
Measure
PADER Scoring System
• Business Results
Customer Feedback and Business Results
Plan
Approach
Deployment
Evaluation
Results
Performance Sustainability System
Performance Scoring System
Suppliers
Process Owners

Figure 3 Edosomwan Performance Excellence Model (EPEM). (From Edosomwan 1998a, b, c, d)

10. *Business results* examines the key results and trends related to the organization's performance in key business areas: customer satisfaction, employee satisfaction, product and service performance, financial management, marketplace performance, mission accomplishment, human resource results, supplier and partner results, and operational performance. Also examined are performance levels relative to world-class organizations.

4.1. PADER Scoring System

Edosomwan (1995, 1998a, b, 2000), developed the Plan, Approach, Deployment, Evaluation, and Results (PADER) scoring system, shown in Figure 3. The system has five performance-evaluation dimensions. The PADER system is utilized for continuous measurement, evaluation and scoring of the categories of the EPEM model. Each component of the PADER scoring system is described below:

1. *Plan* considers the existence of strategic, tactical, and operational thinking, covering long-range, medium-range, and short-term time horizons, respectively. The success of all aspects of a comprehensive continuous performance improvement system depends heavily on a sound, results-driven plan.
2. *Approach* examines the methods used in the organization. Factors used to evaluate the approach dimension include appropriateness of the methods to the requirements; effectiveness of the methods; degree to which the approach is systematic, integrated, consistently applied, embodies evaluation and improvement cycles, and is based on reliable data and information; and evidence of innovation and effective adaptations of approaches used in other types of applications or organizations.
3. *Deployment* refers to the extent to which the organization's approach is implemented across the organization. The factors used to evaluate deployment include the use of the approach in addressing business requirements and the use of the approach by all appropriate work units.
4. *Evaluation* examines the mechanisms, processes, methodologies, and measures used for evaluating, monitoring, and tracking organizational performance for all category elements on a regular basis.
5. *Results* addresses organizational outcomes. The factors used to evaluate the results dimension include current performance; performance relative to appropriate comparisons to world-class

results or benchmark organization; rate, breadth, and importance of performance improvements; and demonstration of sustained improvement over an extended period of time.

5. IMPLEMENTATION OF THE TOTAL QUALITY LEADERSHIP PROCESS

The successful implementation of TQL requires leaders and managers to possess the characteristics and attributes shown in Table 4. *Leading* TQL involves the thought process and act of creating a quality-driven, executable vision or agenda in a systematic manner for the purpose of achieving desired performance results and outcomes. Organizational leaders create, choose, convince, inspire, direct, cause, and make things happen. *Managing* TQL involves the thought process and act of planning, organizing, directing, evaluating, maintaining, supervising, and producing desired quality results and outcomes through people to accomplish the leadership vision and agenda. Organizational managers implement, administer, supervise, and maintain the process of making this happen. Quality results do not happen by accident. The process must be led, managed, and implemented by customers, process owners, suppliers, and the workforce.

The eight steps shown in Figure 4 and described below are recommended for implementing the TQL process in both public- and private-sector organizations:

5.1. Step One: Provide Leadership and Vision for Performance Excellence

Organizational senior leadership must provide the vision, values, culture, and environment for performance excellence. Senior managers need to be trained on the tools, techniques, and principles for leading an organizational performance improvement process. Comprehensive world-class data and information should be utilized to convince top management of customer-driven performance improvement needs and benefits. Do a thorough analysis of the market requirements. Use competitive data on products and services, market share, cost profile, customer, quality, and performance improvement. Explaining the role of the customer, quality, and successful and unsuccessful customer performance improvement examples from other companies also provides a quick way of helping executives see the benefits of adopting the customer-driven performance improvement philosophy. The message to top management is that improvement of quality, productivity, customer satisfaction, and performance are key to achieving competitive advantage, profitability, efficiency, growth, and effectiveness.

5.2. Step Two: Communicate the Continuous Performance Improvement Philosophy

Once senior management support has been achieved, the next step is to communicate and demonstrate management's commitment to the TQL process. Communication regarding the TQL philosophy should emphasize that everything begins with the customer; that is, the sole purpose of the business

TABLE 4 Attributes of Effective Leaders and Managers

Leaders	Managers
1. Create the vision, agenda, purpose, goals, and objectives	1. Implement goals, objectives, policies, and procedures
2. Lead programs and processes for continuous improvement	2. Administer programs for continuous improvement
3. Ask what and why the organization is now and will be in the future	3. Ask how and when—focused on details
4. Are developers of people and organizations	4. Educate people and motivate them to do their best
5. Lead change, are pace setters and risk takers	5. Recognize and reward in a timely manner
6. Provide trust, hope, results, compassion, and constancy of purpose	6. Appraise and counsel to improve performance
7. Remind people of what is important	7. Maintain the process of getting things done
8. Develop customer- and people-driven agendas and results	8. Are proactive rather than reactive
9. Set policies, procedures, and results	9. Implement policies, procedures, and rules
10. Create and build relationships and trust	10. Maintain positive relationships and trust
11. Innovate new productive systems	11. Maintain new and existing systems
12. Use data- and fact-driven, prevention-based strategies and solutions	12. Use data- and fact-driven, prevention-based management

Source: Edosomwan 1999b.

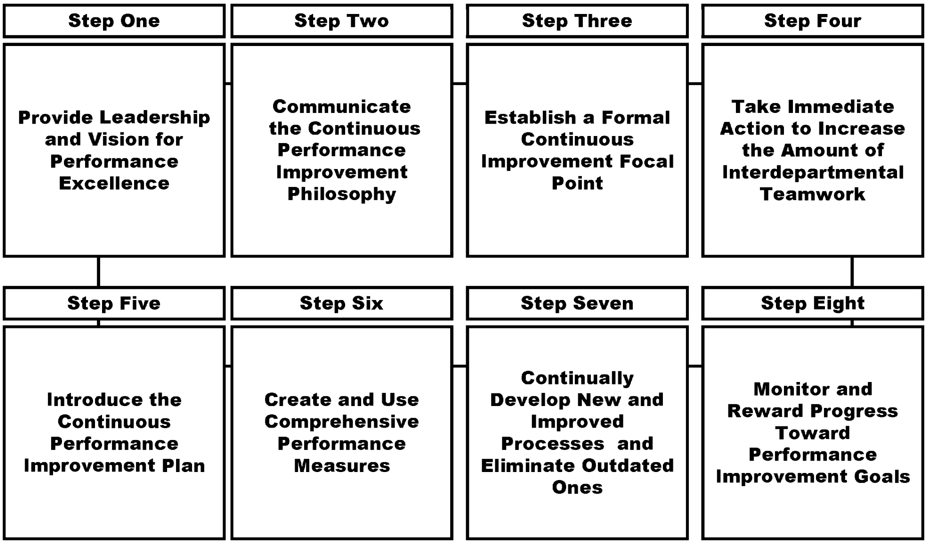

Step One
Provide Leadership and Vision for Performance Excellence
Step Two
Communicate the Continuous Performance Improvement Philosophy
Step Three
Establish a Formal Continuous Improvement Focal Point
Step Four
Take Immediate Action to Increase the Amount of Interdepartmental Teamwork
Step Five
Introduce the Continuous Performance Improvement Plan
Step Six
Create and Use Comprehensive Performance Measures
Step Seven
Continually Develop New and Improved Processes and Eliminate Outdated Ones
Step Eight
Monitor and Reward Progress Toward Performance Improvement Goals

Figure 4 Eight Steps for Implementing Total Quality Leadership Process.

is to attract, serve, and exceed customer's expectations. The commitment to performance excellence means that a total quality culture exists in which everyone constantly considers the quality of their work and how it is reflected in the final output. This also requires that workers be given control, responsibility, and decision latitude for a variety of activities, such as maintenance, information control, quality, process analysis, service, production, and resource allocation. Adopting a TQL philosophy also means focusing on management using facts and data, quality first, proper planning, and effective utilization of human resources.

5.3. Step Three: Establish a Formal Continuous Performance Improvement Focal Point

Recruit seasoned and talented individuals to develop quality-driven performance improvement programs. Individuals charged with this program development must also be given the authority and resources required to institutionalize the program elements. The organization should benchmark successful quality practices implemented in other organizations to avoid potential sources of failure.

5.4. Step Four: Take Immediate Action to Increase the Amount of Interdepartmental Teamwork

People should be trained to work together toward common performance and quality goals while sharing ownership of the final work output. Individuals and work teams should be encouraged to participate in group problem solving, decision making, and incentive systems for rewarding quality excellence. Organizational barriers that hinder teamwork should be eliminated. The method of working should focus on cooperation, not contention. Everyone can make a difference in the quest for continuous performance improvement. Both management and employees should be encouraged to take greater responsibility for their work processes and provide ideas for improving the total process. Managers and supervisors should use a participative management approach to encourage innovative ideas from everyone and put them to work. The environment should allow managers and employees to share in decision making and problem solving and recognition, which are essential for error prevention and innovative improvements.

5.5. Step Five: Introduce the Continuous Performance Improvement Plan to the Team

The organization's quality, productivity, and customer satisfaction strategy should describe the framework, approaches, goals, and objectives that the organization will pursue to achieve continuous performance improvement. At every level of the organization, it is important to have a performance improvement strategy consistent with customer requirements, market demands, operating principles, procedures, and policies. The performance improvement strategy should be the responsibility of the senior management team. Senior management in each operational unit should work with its team to define tactical and operational strategies. The performance improvement strategy should address performance planning, measurement, evaluation, and improvement management. Continuous performance planning involves defining the specific strategies for understanding the market and maintaining an awareness of customer needs, wants, and desires. The plan should define opportunities for supplier and customer involvement in formulating performance improvement needs. This element of the strategy will define the process for continuous improvement and people improvement. Other requirements that may be included in the strategy include updated product specifications such as those for hardware and software; equipment for enhanced measurement, inspection, and testing; process control; and resource management tools. The performance management element should define the strategy for developing the skills and knowledge required by management and employees to do the job right the first time. Plans for education and training should include technical and managerial courses, seminars, and workshops. This element should define communication channels for quality and performance improvement throughout the organization. Such channels of communication should ensure that customers, suppliers, employees, and stockholders are informed about quality goals, objectives, policy, direction, guidance, and performance. The performance management plan should also define specific direction and procedures for monitoring compliance to the organization's quality policy, ensuring successful implementation of the quality goals and objectives. Once developed, the integrated master performance strategy for the organization should be reviewed and revised annually by the organization's senior management committee. This committee should be composed of representatives from all the functional areas of the business, including research and development, manufacturing, marketing, services, and support groups. Once developed, the performance improvement plan should be communicated to everyone with specific milestones and measurements for assessing performance.

5.6. Step Six: Implement Comprehensive Performance Measures at All Levels

Develop and implement comprehensive performance measures for managers and nonmanagers, as well as suppliers and process owners. One of the essential elements of developing a quality-driven organization is deciding how the performance of managers and nonmanagers is evaluated and rewarded. Performance measurements should be more than sales volume, short-term profitability, and rate of return on investment. The focus in developing performance measures for both managers and nonmanagers should include but not be limited to the following areas: short-and long-term profits, customer satisfaction indices, quality and productivity improvement indicators, organizational development, market share profile, rate of return on investment, and product reliability measures. Impediments to successful management of information and measures of performance include poor channels of communication and inadequate data collection processes. Communication channels, such as meetings and electronic data exchange, are recommended at all levels of the organization.

5.7. Step Seven: Develop New, Improved Processes to Eliminate Outdated Ones

Identify the products and services to be provided; define the process sources of variation; and define non-value-added steps with ongoing focus on process measurement, evaluation, control, and improvement. Develop new processes that provide defect-free output and satisfy customer requirements. TQL should encourage broad ownership and total participation by everyone, as well as accountability for results. The commitment from everyone should also include willingness to change unproductive work habits and adopt the attitude of doing the job right the first time. To achieve expected customer satisfaction and quality results, individuals and work teams should be provided with the right training and education. Quality education is essential because it prepares everyone to perform well by providing the knowledge needed to make logical, intelligent decisions. If the right skills are provided, people develop efficient work habits and positive work ethics and attitudes that lead to performance excellence. Training should focus on basic orientation to quality, techniques, and tools for quality improvement, quality leadership, technical skills, process involvement, and teamwork.

5.8. Step Eight: Monitor and Recognize Progress toward Performance Improvement Goals

Publicize successful pilot projects. Reward heroes and consistently performing individuals who have made a difference. A management system that recognizes and encourages ongoing quality improvement efforts must be developed and implemented. In rewarding quality success, place emphasis on accomplishments of teams as well as individuals. Provide recognition and reward in a timely manner. Promotions, pay increases, awards, additional responsibilities, and thanks for a job well done should recognize the accomplishments of teams and individuals. Utilize comprehensive performance measures to evaluate progress of projects, individual and team performance, and overall organizational effectiveness. Continue to build pockets of success through the promotion of successful performance improvement projects.

6. SELECTED TQL KEY SUCCESS FACTORS AND CONCLUSIONS

Edosomwan (1994) presents selected key TQL foundation factors for public- and private-sector organizations aspiring to become benchmarks in performance excellence (see Figure 5). These factors focus on the foundational elements, the key players, the philosophy, the implementation process and levels, the scorecard, and the achievement of desired goals and deliverables. The implementation of the TQL process requires teamwork between all key organizational stakeholders, the philosophy of continuous improvement, an implementation process that focuses on strategic, tactical, and operational issues with ownership for results at the individual, team, work-unit, and organizational levels. Other TQL implementation success factors and lessons learned from world-class organizations by Edosomwan (1994, 1998a, b, c, d) include the following:

1. *Defined leadership expectations:* An organization's senior leaders and managers must ensure the existence of clear quality values, vision, goals, and expectations with defined measures. Systems for achieving organizational goals must be clearly defined. These systems must guide all tasks; activities of the organization to enable all stakeholders contribute to the achievement of the quality goals and objectives.
2. *Employee training, empowerment, and participation:* In world-class quality organizations, employees are partners with management in making decisions about how work is done. People are empowered with the right training, tools, techniques, and authority to deliver error and

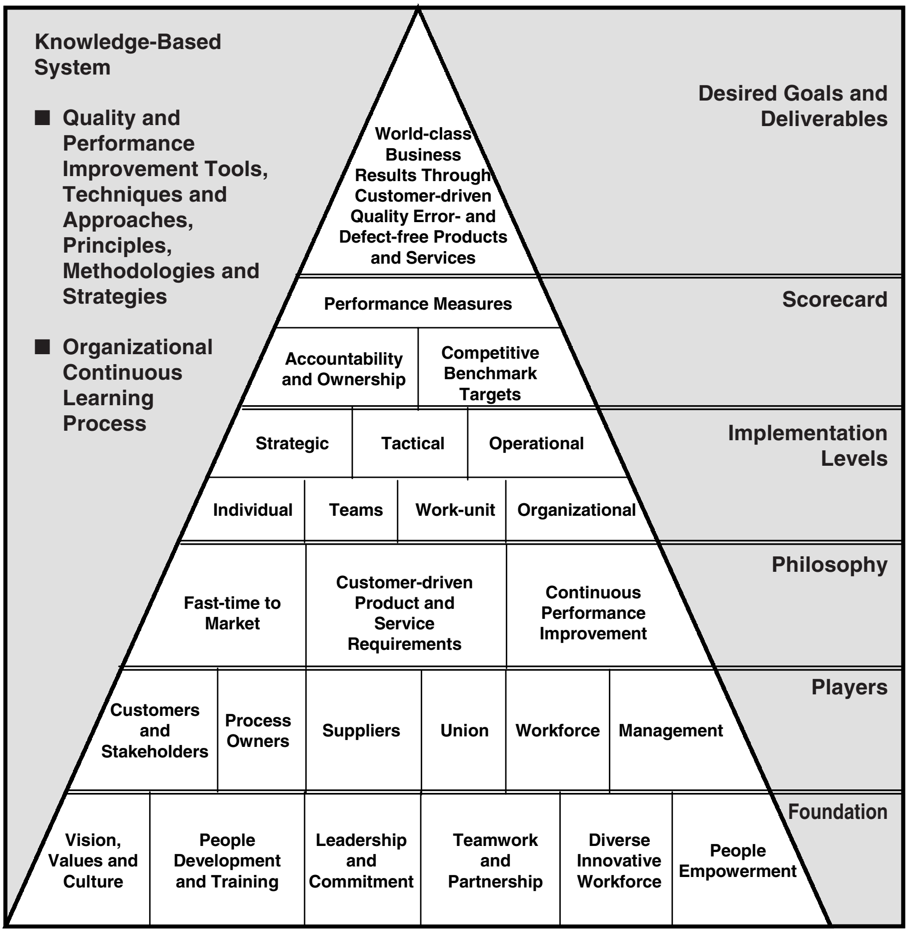

Knowledge-Based System
■ Quality and Performance Improvement Tools, Techniques and Approaches, Principles, Methodologies and Strategies
■ Organizational Continuous Learning Process
World-class Business Results Through Customer-driven Quality Error- and Defect-free Products and Services
Desired Goals and Deliverables
Performance Measures
Accountability and Ownership
Competitive Benchmark Targets
Scorecard
Strategic
Tactical
Operational
Implementation Levels
Individual
Teams
Work-unit
Organizational
Fast-time to Market
Customer-driven Product and Service Requirements
Continuous Performance Improvement
Philosophy
Customers and Stakeholders
Process Owners
Suppliers
Union
Workforce
Management
Players
Vision, Values and Culture
People Development and Training
Leadership and Commitment
Teamwork and Partnership
Diverse Innovative Workforce
People Empowerment
Foundation

Figure 5 Selected Key TQL Factors. (*Source:* Edosomwan 1994, 1998a, b, c, d).

defect free products and services to customers. The organization leadership creates an environment conducive to employee participation in the TQL process.

3. *Continuous improvement and benchmarking:* In world-class quality organization, quality results from a well-executed approach to continuous improvement and learning. The term *continuous improvement* refers to both radical, incremental and breakthrough improvement. The term *learning* refers to adaptation to change, leading to new goals and results-oriented approaches for delivery of products and services to customers. World-class quality organizations also benchmark and implement best business practices to improve performance at the individual, team, work unit and organizational levels.
4. *Partnership for progress and teamwork:* World-class quality organizations promote and implement methods, approaches, and systems for cooperation and suppliers, process owners, and customers. Internal partnering arrangements are also utilized to help individual employees and the workforce achieve defined quality goals and objectives. Also, strong emphasis is placed on team-based problem solving, team-based product design process, self-directed work teams for production, and service-related functions and activities.
5. *Innovation and fast response:* Successful quality organizations promote innovation and risk taking that produces new ideas, suggestions, and breakthrough technologies for performance excellence. They are very responsive to customers. There is continuous focus in product and service-cycle time reduction at all product and service life-cycle stages.
6. *Facts and data-improvement process:* World-class quality organizations utilize reliable data and facts for decisions. Significant investment in information and data gathering systems and tools is made to enable those participating in the quality process to analyze and solve problems with accurate data and information.
7. *Long-range view of the future:* Successful quality organizations usually have a strong orientation toward the future and a willingness to make long-term commitments to customers, suppliers, stakeholders, and employees. They anticipate challenges and put the right strategic, tactical, and operational plans in place to achieve success.
8. *Prevention-based management:* World-class quality organizations utilize highly effective systems and tools to build quality into products and services and the processes through which they are produced. The focus is on error and defect prevention and not on crisis management of problems that could have been prevented through appropriate planning.
9. *Customer-driven quality results:* World-class organizations operate on the philosophy that customers judge quality. Customers' requirements and specifications therefore drive all products and service features that contribute value to customers and lead to customer satisfaction, preference, loyalty, and retention. These organizations also focus on achieving world-class results through defined process for customer segmentation, anticipating demands, problem solving and analysis, performance planning, measurement, evaluating, and improvement.
10. *Diversity as a strength:* World-class quality organizations utilizes the skills, talents, and abilities of a diverse workforce to achieve quality goals and objectives. Employing people of different genders and abilities and with rich variety of educational, cultural, ethnic, linguistic, and physical characteristic enables these organizations to serve diverse customer segments and populations effectively. Diversity is seen by quality organizations as a strength in individual and team capabilities to innovate, solve problems and deliver error and defect-free products and services.

There are enormous benefits from implementing the quality improvement process in an organization. One of the outcomes of the process is a radical shift from a traditional to a customer-driven organization. The new customer-driven organization is responsive and competitive and able to address new market demands and challenges. The customers benefit from improved products and services, and the employees benefit from full-time employment, better wages, job satisfaction, and improved morale. The organizational suppliers and stakeholders benefit from financial gains and business growth.

REFERENCES

Almaraz, J. (1994), "Quality Management and the Process of Change," *Journal Organizational Change Management* Vol. 2, No. 2, pp. 141–149.

Crosby, P. B. (1979), *Quality Is Free: The Art of Making Quality Certain,* McGraw-Hill, New York.

Deming, W. (1981), "Improvement of Quality and Productivity through Action by Management," *National Productivity Review,* Vol. 1, No. 1, pp. 12–22.

Deming, W. (1982), *Quality, Productivity, and Competitive Position,* Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, MA.

Edosomwan, J. A. (1988), *Productivity and Quality Improvement,* IFS, Oxford.

Edosomwan, J. A. (1990), "Implementing Market-Driven Quality and Total Customer Satisfaction Programs," in *ASQC Annual Quality Congress Transactions,* ASQC, Milwaukee.

Edosomwan, J. A. (1994), *Customer and Market Driven Quality Management,* ASQC Press, Fairfax, VA.

Edosomwan, J. A. (1995), *Organizational Transformation and Process Reengineering,* St. Lucie Press, St. Lucie, FL.

Edosomwan, J. A. (1998a), *Quality Begins with Me,* Quality University Press, Fairfax, VA.

Edosomwan, J. A. (1998b), *Customer Satisfaction Management Frontier II,* 2nd Ed., CIC Group, Fairfax, VA.

Edosomwan, J. A. (1998c), *Comprehensive Customer Care and Satisfaction System,* Quality University Press, Fairfax, VA.

Edosomwan, J. A. (1999b), *Winning Leaders and Managers,* Quality University Press, Fairfax, VA.

Edosomwan, J. A. (2000), *The Next Generation Performance Improvement Package,* Johnson & Johnson Associates, Fairfax, VA.

Feigenbaum, A. V. (1983), *Total Quality Control,* McGraw-Hill, New York.

Greene, R. T. (1993), *Global Quality: A Synthesis of the World's Best Management Models,* ASQC Press, Milwaukee, and Business One Irwin, Homewood, IL.

Hackman, R. J., and Wageman, R. (1995), "Total Quality Management: Empirical, Conceptual, and Practical Issues," *Administrative Science Quarterly,* Vol. 40, No. 2, pp. 309–342.

Hiam, A. (1993), *Does Quality Work? A Review of Relevant Studies,* Report #1043, The Conference Board, New York.

Japanese Industrial Standards Committee (JIS) (1981), "Industrial Standardization," JIS Z 8101.

Juran, J. M. (1974), *The Quality Control Handbook,* McGraw-Hill, New York.

Malhotra, M, Steele, D., and Grover, V. (1994), "Important Strategic and Tactical Manufacturing Issues in the 1990s," *Decision Sciences,* Vol. 25, No. 1, pp. 189–214.

National Institute of Standards and Technology (NIST), "Malcolm Baldrige National Quality Award Application Guidelines," U.S. Department of Commerce, Washington, DC, 1999.

Pirsig, R. (1974), *Zen and the Art of Motorcycle Maintenance,* Bantam, New York.

ADDITIONAL READING

Ahire, S. L., Golhar, D. Y., and Waller, M. A., "Development and Validation of TQM Implementation Constructs," *Decision Sciences* Vol. 27, No. 1, 1996, pp. 23–56.

"AMA Survey on Quality and Customer Satisfaction Programs," in *Does Quality Work? A Review of Relevant Studies,* A. Hiam, Conference Board, New York, 1993.

Barnett, C. K., "Organizational Learning and Continuous Quality Improvement in an Automotive Manufacturing Organization," Ph.D. Dissertation, University of Michigan, 1994.

Blackburn, R., and Rosen, B., "Total Quality and Human Resources Management: Lessons Earned from Baldrige Award-Winning Companies," *Academy of Management Executive* Vol. 7, No. 3, 1993, pp. 49–66.

Cole, R. E., "Introduction to the Special Issue on Total Quality Management," *California Management Review,* Vol. 35, 1993, pp. 7–11.

Cole, R. E., "Learning from Learning Theory: Implications for Quality Improvements of Turnover, Use of Contingent Workers, and Job Rotation Policies," *Quality Management Journal,* Vol. 1, No. 1, 1993, pp. 9–25.

Cole, R. E., "Learning from the Quality Movement: What Did and Didn't Happen and Why?" *California Management Review,* 41, 1999, pp. 43–73.

Cole, R. E., "Managing Quality Fads: How American Business Learned to Play the Quality Game," Oxford University Press, New York, 1999.

Deming, W., *Out of the Crisis,* Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, MA, 1986.

Easton, G. S., and Jarrell, S. L., "The Effects of Total Quality Management on Corporate Performance: An Empirical Investigation," *Journal of Business,* Vol. 71, No. 1, 1998, pp. 15–35.

Edwards, C. D. (1968), "The Meaning of Quality," *Quality Progress,* Vol. 1, October, pp. 36–39.

Edosomwan, J. A., *Integrating Productivity and Quality Management,* Marcel Dekker, New York, 1987.

Edosomwan, J. A., *Continuous Improvement Tools and Techniques,* Excellence, Fairfax, VA, 1991.

Edosomwan, J. A., ‘‘Five Initiatives for Improving Your Customer Satisfaction Level,’’ *Quality Observer International News Magazine,* 1991.

Edosomwan, J. A., ‘‘On Becoming a Customer-Driven Organization,’’ *Quality Observer International News Magazine,* 1991.

Edosomwan, J. A., *Understanding and Implementing Total Quality Management,* Excellence, Fairfax, VA, 1991.

Edosomwan, J. A., *The Winning Quality Manager,* Excellence, Fairfax, VA, 1991.

Edosomwan, J. A., *Customer Satisfaction Management Frontier I,* 1st Ed., CIC Group, Fairfax, VA.

Edosomwan, J. A., *Customer-Driven Quality Management,* 2nd Ed., CIC Group, Fairfax, VA, 1997.

Edosomwan, J. A., *The Next Generation Customer Satisfaction Improvement System,* Quality University Press, Fairfax, VA, 1997.

Edosomwan, J. A., *Edosomwan Baldrige-Based Assessment Tool (EBAT-2000) Criteria,* Quality University Press, Fairfax, VA, 1999.

Edosomwan, J. A., Sathe, P., and Hancock, W. H., ‘‘Quality Assurance,’’ in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York, 1991, pp. 2221–2234.

Garvin, D., ‘‘How the Baldrige Award Really Works,’’ *Harvard Business Review,* Vol. 69, No. 6, November–December, 1991, pp. 80–93.

Grant, R. M., Shani, R., and Krishnan, R., ‘‘TQM’s Challenge to Management Theory and Practice,’’ *Sloan Management Review,* Vol. 35, No. 2, 1994, pp. 25–35.

Handfield, R., and Ghosh, S., ‘‘Creating a Total Quality Culture through Organizational Change: A Case Study,’’ *Journal of International Marketing,* Vol. 2, 1994, pp. 7–36.

Hendricks, K. B., and Singhal, V. R., ‘‘Does Implementing an Effective TQM Program Actually Improve Operating Performance? Empirical Evidence from Firms That Have Won Quality Awards,’’ *Management Science,* Vol. 43, No. 9, 1997, pp. 1258–1274.

Juran, J. M., *Juran on Leadership for Quality,* Free Press, New York, 1989.

Juran, J. M., *Juran on Quality by Design,* Free Press, New York, 1992.

Larson, P. D., and Sinha, A., ‘‘The TQM Impact: A Study of Quality Managers’ Perceptions,’’ *Quality Management Journal,* Vol. 2, No. 3, 1995, pp. 53–66.

Lascelles, D., and Barrie, D., ‘‘Quality Management: The Chief Executive’s Perception and Role,’’ *Journal of European Management,* Vol. 8, 1990, pp. 67–75.

Lawler, E. E., III, Mohrman, S. A., and Ledford, G. E., *Employee Involvement and Total Quality Management,* Jossey-Bass, San Francisco, 1992.

Loomba, A., and Johannessen, T. B., ‘‘Malcolm Baldrige National Quality Award: Critical Issues and Inherent Values,’’ *Benchmarking for Quality Management and Technology,* Vol. 4, No. 1, 1997, pp. 59–77.

National Institute of Standards and Technology (NIST) (1992), Malcolm Baldrige National Quality Award Criteria, NIST, Gaithersburg, MD.

National Institute of Standards and Technology (NIST) (1993), Malcolm Baldrige National Quality Award Criteria, NIST, Gaithersburg, MD.

Pavett, C., and Whitney, G. (1997), ‘‘Total Quality Management as an Organizational Change Technique, Predictors and Moderators of Successful Implementation,’’ manuscript.

Peters, T. J., and Waterman, R. H., *In Search Of Excellence,* Harper & Row. New York, 1982.

Powell, T. C., ‘‘Total Quality Management as Competitive Advantage: A Review and Empirical Study,’’ *Strategic Management Journal,* Vol. 16, No. 1, 1995, pp. 15–37.

Schaffer, R., and Thomson, H., ‘‘Successful Change Programs Begin with Results,’’ *Harvard Business Review,* Vol. 70, No. 1, January–February, 1992, pp. 80–89.

Schonberger, R. J., ‘‘Total Quality Management Cuts a Broad Swath through Manufacturing and Beyond,’’ in *Perspectives In Quality: Services and Manufacturing,* T. J. Billesbach, Ed., Mountain Top, Omaha, NE, 1994.

Troy, K., ‘‘Employee Buy-in to Total Quality,’’ in *Does Quality Work? A Review Of Relevant Studies,* A. Haim, Conference Board. New York, 1992.

IV.G

Supply Chain Management and Logistics

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 76

Logistics Systems Modeling

David Simchi-Levi
Massachusetts Institute of Technology

Edith Simchi-Levi
LogicTools Inc.

1. **OVERVIEW** **2007**
2. **INTRODUCTION TO LOGISTICS MANAGEMENT** **2007**
3. **LOGISTICS MODELS** **2008**
   - 3.1. Logistics Network Design and Configuration 2008
   - 3.2. Supply Chain Planning 2009
   - 3.3. Transportation Planning 2011
4. **DECISION SUPPORT SYSTEMS** **2011**
   - 4.1. Input Data 2012
   - 4.2. Analytical Tools 2013
   - 4.3. Presentation Tools 2015
     - 4.3.1. Geographic Information Systems 2016
     - 4.3.2. Integrating Algorithms and GIS 2018
5. **THE IMPACT OF DECISION SUPPORT SYSTEMS ON LOGISTICS MANAGEMENT** **2018**

REFERENCES **2019**

1. OVERVIEW

Information technology is a critical enabler of effective logistics strategies. Indeed, much of the current interest in logistics and supply chain management is motivated by the opportunities that appeared due to the abundance of data and the savings that can be achieved by sophisticated analysis of these data. Thus, in this chapter we focus on models and decision support systems for logistics management that take advantage of the opportunities provided by information technology and allow companies to reduce costs and increase service levels. In particular, we review models and decision support systems for strategic, tactical, and operational decisions and look at the impact of new technology, such as the Internet, on the way decision support systems are used in practice.

2. INTRODUCTION TO LOGISTICS MANAGEMENT

Fierce competition in today's global markets, the introduction of products with short life cycles, and the heightened expectations of customers have forced manufacturing enterprises to invest in and focus attention on their logistics systems. This, together with changes in communications and transportation technologies, such as the Internet, mobile communication, and overnight delivery, has motivated continuous evolution of the management of logistics systems.

In these systems, items are produced at one or more factories, shipped to warehouses for intermediate storage, and then shipped to retailers or customers. Consequently, to reduce cost and improve service levels, logistics strategies must take into account the interactions of the various levels in the logistics network. This network consists of suppliers, manufacturing centers, warehouses, distribution centers, and retail outlets as well as raw materials, work-in-process inventory, and finished products that flow between the facilities.

The goal of this chapter is to present the state of the art in the modeling, analysis, planning, and control of logistics systems, so-called logistics management. But what exactly is logistics management? According to the Council of Logistics Management, a non-for-profit organization of logistics educator and professionals, it is:

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

> the process of planning, implementing and controlling the efficient, cost effective flow and storage of raw materials, in-process inventory, finished goods, and related information from point-of-origin to point-of-consumption for the purpose of conforming to customer requirements.

This definition leads to several observations. First, logistics management takes into consideration every facility that has an impact on cost and plays a role in making the product conform to customer requirements, from supplier and manufacturing facilities through warehouses and distribution centers to retailers and stores. Second, the goal in logistics management is to be efficient and cost effective across the entire system; the objective is to minimize system-wide costs, from transportation and distribution to inventory of raw material, work-in-process, and finished goods. Thus, the emphasis is not on simply minimizing transportation cost or reducing inventories, but rather on a systems approach. Finally, because logistics management evolves around planning, implementing, and controlling the logistics network, it encompasses many of the firm's activities, from the strategic level through the tactical to the operational level.

Indeed, following Hax and Candea's (1984) treatment of production-inventory systems, logistical decisions are typically classified in the following way.

- The *strategic level* deals with decisions that have a long-lasting effect on the firm. This includes decisions regarding the number, location, and capacities of warehouses and manufacturing plants or the flow of material through the logistics network.
- The *tactical level* typically includes decisions that are updated anywhere between once every quarter and once every year. This includes purchasing and production decisions, inventory policies, and transportation strategies, including the frequency with which customers are visited.
- The *operational level* refers to day-to-day decisions such as scheduling, routing, and loading trucks.

In the next section we provide examples of three logistics models that span the three levels of decisions described above. In Section 4 we describe the key components of decision support systems (DSS) for logistics management. Finally, in Section 5 we describe the advantages and impact of using DSS on logistics management.

3. LOGISTICS MODELS

3.1. Logistics Network Design and Configuration

The logistics network consists of suppliers, warehouses, distribution centers, and retail outlets as well as raw materials, work-in-process inventory and finished products that flow between the facilities. In this section we present some of the issues involved in the design and configuration of a logistics network.

Network configuration may involve issues relating to plant, warehouse, and retailer location. As explained earlier, these are strategic decisions since they have a long-lasting effect on the firm. Specifically, in a typical logistics network model, we concentrate on the following key strategic decisions:

1. Determine the appropriate number of warehouses.
2. Determine the location of each warehouse.
3. Determine the size of each warehouse.
4. Allocate space for products in each warehouse.
5. Determine which products customers will receive from each warehouse.

We therefore assume that plant and retailer locations will not be changed. The objective is to design or reconfigure the logistics network so as to minimize annual system-wide costs including production and purchasing costs, inventory holding costs, facility costs (storage, handling and fixed costs), and transportation costs, subject to a variety of service-level requirements.

In this setting, the trade-offs are clear. Increasing the number of warehouses typically yields:

- An improvement in service level due to the reduction in average travel time to the customers
- An increase in inventory costs due to increased safety stocks required to protect each warehouse against uncertainties in customer demand
- An increase in overhead and set-up costs,
- A reduction in outbound transportation costs, i.e., transportation costs from the warehouses to the customers,

- an increase in inbound transportation costs, i.e., transportation costs from the suppliers and/or manufacturers to the warehouses.

In essence, the firm must balance the costs of opening new warehouses with the advantages of being close to the customer. Thus, warehouse location decisions are crucial determinants of whether or not the supply chain is an efficient channel for the distribution of the products. This task requires specialized solvers since it is a complex problem involving large data sets. That is, manual or spreadsheet analysis is not practical to solve most real-life problems. For example, in a simple site-selection problem requiring the identification of 5 optimal warehouse locations from a set of 25 potential sites, 50,000 different combinations must be considered. This is far too many to analyze with a spreadsheet. As Figure 1 demonstrates, the number of combinations grows exponentially as potential sites are added to the analysis.

Figure 2 and Figure 3 present two typical DSS screens that the user would see at different stages of the optimization. The first screen represents the network prior to optimization and the second represents the optimized network.

3.2. Supply Chain Planning

These planning tools determine the appropriate production, transportation, and inventory policies for a set of manufacturing plants, warehouses and retailers. Specifically, given manufacturing, warehouse and retailer locations, production, inventory and transportation costs and capacities, and demand forecasts for each retail outlet, the objective is to determine policies that achieve high levels of customer service with minimal cost.

In this case, we typically refer to three types of models:

1. *Production planning models:* efficient allocation of manufacturing resources over a period of several months to meet demand
2. *Distribution planning models:* efficient allocation of logistics resources over a period of several months to meet demand
3. *Demand planning models:* determine accurate forecasts based on historical data, understanding of customers' buying pattern, economic conditions, etc.

Evidently, these models and decisions are not independent. For instance, the production planning model relies on the quality of the data generated from the demand planning phase. It also depends

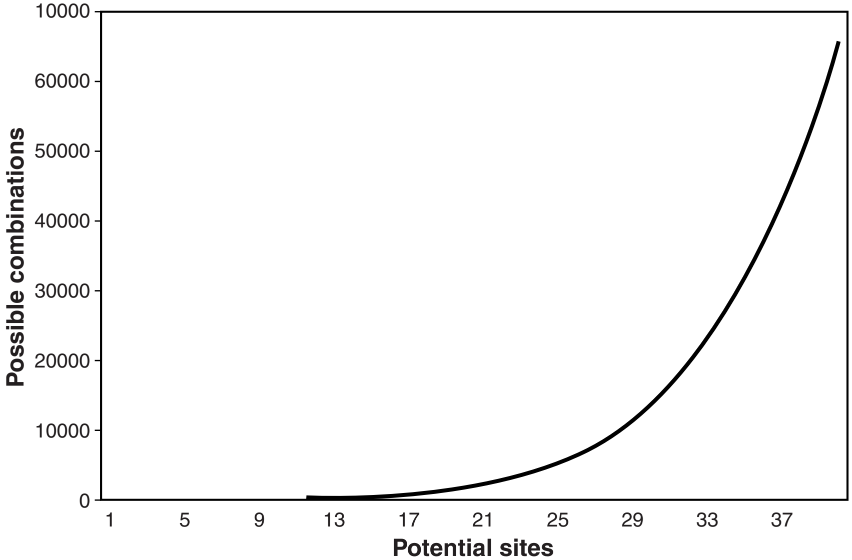


Figure 1 Number of Combinations when Locating Five Facilities in a Given Number of Potential Sites.

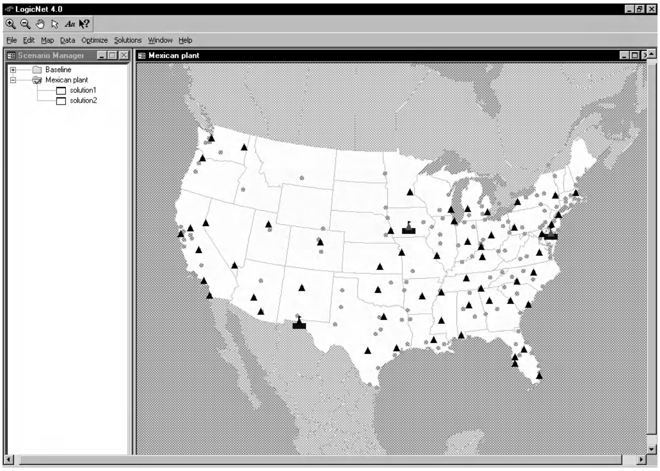

LogicNet 4.0
File Edit Map Data Optimize Solutions Window Help
Scenario Manager
Baseline
Mexican plant
solution1
solution2
Mexican plant

Figure 2 The DSS Screen Representing Data Prior to Optimization.

on the distribution plan since the mode of shipping, inventory plans, and deployment schedule will effect the manufacturing requirements. In addition, in many situations, there is a need to find the optimal trade-off between production timing and inventory levels. Specifically, is it more cost effective to store inventory than to work overtime in the manufacturing plants? A new breed of supply chain coordination DSS is now in the market that optimizes production, inventory, and transportation decisions by taking into account the entire supply chain.

Another key challenge at the tactical level is to take into account the dynamics of the supply chain. Indeed, in recent years many suppliers and retailers have observed that while customer demand for specific products does not vary much, inventory and back-order levels fluctuate considerably across their supply chain. For instance, examining the demand for Pampers disposal diapers, executives at Procter & Gamble noticed an interesting phenomenon. As expected, retail sales of the product were fairly uniform; there is no particular day or month in which the demand is significantly smaller or larger than any other. However, the distributors placed orders to the factory that fluctuated much more than retail sales. In addition, P&G's orders to its suppliers fluctuated even more. This increase in variability as we travel up in the supply chain is referred to as the Bullwhip effect. For more on this effect, see Simchi-Levi et al. (1999).

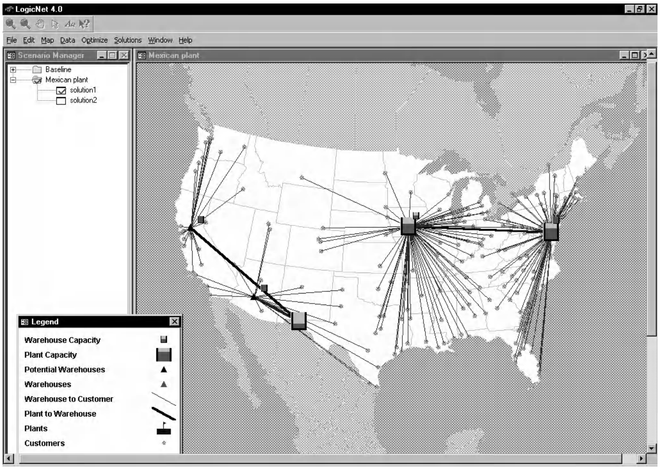

LogicNet 4.0
File Edit Map Data Optimize Solutions Window Help
Scenario Manager
Baseline
Mexican plant
solution1
solution2
Mexican plant
Legend
Warehouse Capacity
Plant Capacity
Potential Warehouses
Warehouses
Warehouse to Customer
Plant to Warehouse
Plants
Customers

Figure 3 The DSS Screen Representing the Optimized Logistics Network.

3.3. Transportation Planning

The transportation component of logistics management is evolving into a complex process where costs are optimized at every step and sophisticated DSS are utilized to achieve efficiencies.

There are a few typical types of DSS used in the transportation planning. Some examples include:

1. *Routing and scheduling:* Streamline daily delivery operations by developing least-cost routes that meet customer delivery requirements. The DSS sums multiple routing passes to determine the best daily plan that meets the specific needs of user's operation. There are many DSS geared towards routing; however, the requirements can differ based on industry and type of transportation used. For instance, garbage pickup routes are different from a soft drink distributor's direct store delivery (DSD) routes. There are also DSS that concentrate specifically on fuel efficiency, driver swaps, and other complex aspects of routing. Figure 4 shows an example of a typical routing software map display.
2. *Mode selection:* Common transportation modes include overnight package, parcel, less-than-truckload (LTL), truckload (TL), and rail carload (CL). Each mode offers different cost and service advantages and disadvantages relating to shipment size, cost, and delivery time. For instance, TL is generally cheaper and faster than LTL but requires large loads. Shipping by air is expensive but may reduce inventory costs considerably.
3. *Carrier selection:* In situations where companies must analyze and negotiate pricing bids from carriers, there are DSS that assist with this process. Inputs are shipping requirements and carrier bid responses that handle selected lanes of freight. The DSS optimally chooses the least-cost set of carriers to fulfill the shipping volumes. A number of operational constraints such as minimum carrier commitment levels, and maximum number of carriers to select can be imposed.

Transportation DSS are evolving to accommodate for real-time needs that are pushed by on-board computers and wireless communication. In addition, data interfaces with enterprise resource planning (ERP) systems are becoming more streamlined and allow optimization of transportation decisions by considering their impact on the entire logistics network. Finally, transportation DSS also have to take into account the recent proliferation of Internet transportation exchanges.

4. DECISION SUPPORT SYSTEMS

In order to select and use decision support systems effectively, it is helpful to understand the essential pieces of a properly configured decision support system. The three major components of a DSS are the input databases and parameters, the analytical tools, and the presentation mechanism:

- The input database is a form of database with the basic information needed for the decision making. This can be a PC-based database extract designed for the specific problem, a data

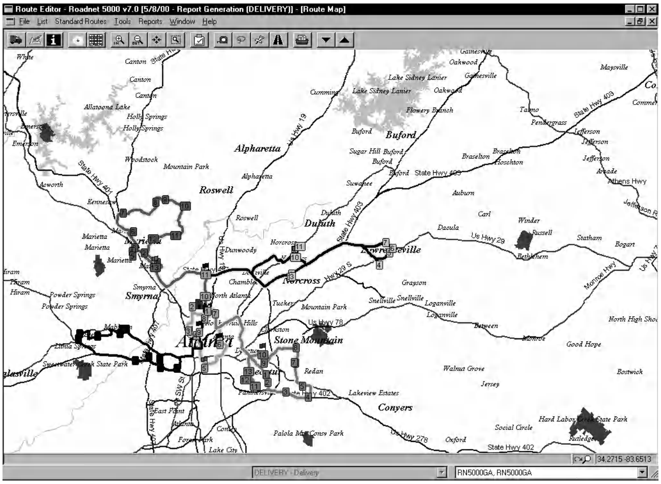

Route Editor - Roadnet 5000 v7.0 [5/8/00 - Report Generation (DELIVERY)] - [Route Map]
File List Standard Routes Tools Reports Window Help
Canton
Allatoona Lake
Holly Springs
Woodstock
Mountain Park
Roswell
Alpharetta
Cumming
Lake Sidney Lanier
Oakwood
Gainesville
Flowery Branch
Buford
Sugar Hill
Suwanee
Duluth
Norcross
Lawrenceville
Marietta
Kennesaw
Acworth
Smyrna
Dunwoody
Chamblee
Tucker
Stone Mountain
Powder Springs
Hiram
Lithia Springs
Sweetwater Creek State Park
Atlanta
East Point
Forest Park
Redan
Lakeview Estates
Conyers
Snellville
Grayson
Loganville
Dacula
Auburn
Carl
Winder
Russell
Bethlehem
Statham
Bogart
Braselton
Hoschton
Talmo
Pendergrass
Jefferson
Arcade
Maysville
Monroe
Between
Good Hope
Walnut Grove
Jersey
Social Circle
Oxford
Rutledge
Hard Labor Creek State Park
Bostwick
North High Shoals
Palola Mtn Consv Park
Us Hwy 29
Us Hwy 78
State Hwy 400
State Hwy 402
Us Hwy 278
34.2715 -83.6513
DELIVERY - Delivery
RN5000GA, RN5000GA

Figure 4 A Typical Routing Software Display.

warehouse with an accumulation of the company's transactions, or distributed databases accessed through a network. This database can also include certain parameters and rules such as the desired service level, hard-coded restrictions, and various other constraints.

- The data analysis usually involves embedded knowledge of the problem while also allowing the user to fine-tune certain parameters. The tools employed are operations research and artificial intelligence-based algorithms, cost calculators, simulation, flow analysis, and other embedded logic procedures. This component is the most complex because there are few off-the-shelf solvers that can deal with the huge variety of problems that companies face.
- Various database and spreadsheet presentation tools can be used to present the results of DSS analysis. Often, however, the output contains too much information, such as lists and tables, that may be too difficult for the decision maker to absorb. Therefore, various data visualization techniques are employed to enable the user to comprehend the vast quantity of output data. For example, location, routing, and sales DSS use geographic information systems (GIS) (see Section 4.3) to display complex geographic data in problems such as site location, routing and supply chain analysis. Similarly, scheduling systems use Gantt charts to display factory schedules, and simulations use animation to illustrate the relationships in the model.

All of these components are markedly affected by the planning horizon of the problem being considered. As we have seen, strategic decisions typically require long-term planning and may involve aggregation of historical data and forecasting considerations, while their analysis and presentation does not need to be particularly fast, since immediate response is not an issue. In contrast, operational decisions typically involve short-term planning, require current data, and demand fast response from the DSS.

4.1. Input Data

As in all kinds of analysis, the data used as input to the DSS is critical to the quality of the analysis. Until fairly recently, acquiring the appropriate information was in itself a complicated feat. By and large, the information-collection battle has now been won in corporate America. Extensive deployment of information systems, bar coding, point of sale, and electronic commerce have provided companies with large databases of business data. These are now often collected into huge data warehouses or smaller data marts that, along with the appropriate tools, facilitate the analysis of the data. In addition, the proliferation of networks and network access tools means that accessing various geographically distributed databases is now feasible.

Depending on the type of analysis, a DSS may require collecting information from various parts of the company. For example, supply chain network design requires both static and dynamic information from different parts of the company. The static data include plant production rates, locations of the plants, warehouses, and customers as well as warehousing costs and transportation costs, and the dynamic data involve forecasts, orders, and current deliveries. This type of information will not usually be found in one database or one department in a company.

In order to evaluate the quality of the data as well as the quality of the models built into the system it may be possible to load the current data and models into the system and see if these correspond to reality. For example, consider a truck-routing decision support system. Ultimately, the goal of such a system is to provide routes that enable the trucks to make deliveries and pickups efficiently at the required times. The model can be tested by loading the current truck routes into the system and observing whether travel times, for example, are the same as the travel times actually experienced by the truck drivers. Similarly, projected costs from the model can be compared with actual financial and accounting records. This process, typically known as model and data validation, is essential to ensure that the model and data are accurate enough. Of course, the meaning of ''accurate enough'' depends on the decisions being made. For more information on this issue, see Simchi-Levi et al. (1999).

In addition, the decision-planning horizon affects the detail of the data required. For strategic planning, it often makes sense to aggregate yearly data, while for short-term planning it may be best to utilize recent raw daily or weekly data. The accuracy of the solution depends on the input data, and therefore the quality of the input data will determine to some extent the tools needed for the analysis.

For instance, consider the logistics network design model discussed previously. A DSS is often used to assist in optimizing the number of warehouses required as well as their size and customer allocation to each warehouse. The DSS uses information about the distribution system to calculate the various costs related to the site selection and customer allocation. The data required for this problem involve the manufacturers, warehouses, and customers and the transportation between them. Since this is a long-term planning tool, yearly demand data and costs are typically used, but sometimes the user may need to determine how to account for seasonality. In addition, in order for this kind of DSS to be utilized successfully, the user needs to break the products into product families and specify inventory policies. This will allow calculation of the warehouse sizes and frequency of deliveries. Some of the required input data are summarized in Table 1.

It is clear that this type of data may not be readily available in the company database. Even if the data are accessible, they may not be in the required format, particularly if the DSS involves geographic display and analysis. As one might expect collecting, tabulating, and verifying the data can take some time.

4.2. Analytical Tools

Another issue that needs to be established when working with a DSS is the measures by which the various solutions will be evaluated. Reducing total cost may be a goal, but in some cases improving customer service level may be more pertinent. DSS interfaces usually allow setting these parameters and indicating the balance required by the user.

Once data has been collected, it must be analyzed and presented. Of course, depending on the DSS and the particular decision being made, there are many different ways to analyze the data. It is important for the decision makers to understand how the DSS analyzes the data, in order to assess the validity and accuracy of the DSS's recommendations. Of course, depending on the analysis, statistics can tell many different stories (see Shenk 1997 for an interesting discussion of these issues). It is up to the decision-maker to determine what analysis is most appropriate.

In what follows we examine common DSS analysis tools and techniques in general.

- *Queries:* Often, vast quantities of data make manual analysis difficult. Simply allowing decision makers to ask specific questions about the data, such as "How many clients do we service in California?" and "How many clients purchased over $3000 of a certain product by state?" often facilitates decisions.
- *Statistical analysis:* Sometimes asking questions is not sufficient. In this case, statistical techniques can sometimes be used to determine trends and patterns in the data. For example, often statistical data such as the average inventory in a warehouse, the average number of stops and length of a route, and the variability of customer demand can be useful to decision makers.
- *Data mining:* Recently, as corporate databases have become larger and more all-encompassing, new tools have been developed to look for hidden patterns, trends, and relationships in the data. Data mining, for example, produced the marketing gem that men tend to purchase beer AND diapers on Friday afternoons.
- *OLAP tools:* Online analytic processing tools provide an intuitive way to view corporate data, typically stored in data warehouses. These tools aggregate data along common business dimensions and let users navigate through the hierarchies and dimensions by drilling down, up, or across levels. OLAP tools also provide sophisticated statistical tools to analyze the data and include presentation tools as well. Mostly they are generic tools, more sophisticated than spreadsheets and easier to use than database tools, for the analysis of large amounts of data.
- *Calculators:* Simple decision support tools can facilitate specialized calculations, such as accounting costs. In many cases, more than simple calculations may not be warranted, especially if the changes are predictable and easy to evaluate. This may be the case with some product

TABLE 1 Input Data for Logistics Network Design

Component	Data
Manufacturer	location production capacity and cost transportation costs to warehouses
Warehouse	location fixed costs variable costs (labor, utilities) inventory turnover transportation costs to retailers
Retailer	location annual demand by product
Product	volume weight holding cost

types for forecasting or inventory management, while for others more sophisticated tools may be needed.

- *Simulation:* All business processes have random components. Sales may take one value or another. A machine may or may not fail. Often these random, or stochastic, elements of a problem make analyzing it very difficult. In these cases, simulation is often an effective tool to help with decisions. In simulation, a model of the process is created on a computer. Each of the random elements of the model (sales, failures, etc.) is specified with a probability distribution. When the model is run, the computer simulates carrying out the process. Each time a random event occurs, the computer uses the specified probability distribution to randomly decide what happens.

  For example, consider a simulation model of a production line. As the computer runs the model, a series of decisions is made. How long does a job take on machine one? On machine two? Does machine three break while job four is being processed on it? As the model runs, statistical data (utilization rates, completion times, etc.) are collected and analyzed. Since this is a random model, each time the model is run, the results may be different. Statistical techniques are used to determine the average outcome of the model as well as the variability of this outcome. Also, varying different input parameters allows different models and decisions to be compared. For example, different distribution systems can be compared utilizing the same simulated customer demand. Simulation is often a useful tool for understanding very complex systems that are difficult to analyze analytically.
- *Artificial intelligence:* Artificial intelligence tools may be employed in the analysis of DSS input data. These may be databases of rules collected from experts that can be applied to specific problems, or online intelligent agents. The former systems are often used to solve technical problems, such as troubleshooting a computer failure or a complex chemical procedure, while the latter are more appropriate for managing different components in the supply chain. Following Fox et al. (1993), we define an agent as a software process whose goal is to communicate and interact with other agents so that decisions effecting the entire supply chain can be made on a global level. Indeed, a number of DSS for supply chain management can be viewed as using intelligent agents to plan and execute different activities in the supply chain. These systems are characterized (see Fox et al. 1993) by the following interrelated issues:

- The activities allocated to each intelligent agent (i.e., software processor)
- The level and nature of interactions between the different agents
- The level of knowledge embedded within each agent

For instance, a real-time supply chain planning tool involves the following components. Intelligent agents are located at each facility and collect information about this facility as well as enable planning and scheduling for the facility. In this case, facilities include manufacturing plants and distribution centers. Each agent interacts with other agents so that they can balance excess capacity at different plants, find missing parts, or coordinate production and distribution. A central planning agent communicates with the agents at each facility to collect status information and relate central planning decisions. The type and level of decisions made by the agents as opposed to human operators, as well as the frequency and level of communications among the agents, depends on the specific implementation.

- *Mathematical models and algorithms:* Mathematical tools, often from the discipline of operations research, can be applied to the data to determine potential solutions to problems. For example, these tools may generate the best set of locations for new warehouses, or an efficient route for a truck to take, an effective inventory policy for a retail store. These algorithms fall into two categories:
- *Exact algorithms:* Given a particular problem, these algorithms will find a solution that is mathematically the best possible solution. In general, these kinds of algorithms may take a long time to run, especially if a problem is very complex. In many cases, it is impossible to find the optimal, or very best, solution. In other cases, it may be possible but not worth the effort. This is because the input data to these algorithms is often approximated or aggregated, so exact solutions to approximate problems may be worth no more than approximate solutions to approximated problems.
- *Heuristics:* These are algorithms that provide good, but not necessarily optimal, solutions to problems. Heuristics typically run much faster than optimal algorithms. Most DSS that use mathematical algorithms employ heuristics. A good heuristic will rapidly give a solution that is very close to the optimal solution. Often, heuristic design involves a trade-off between quality of solution and speed. If is often useful if in addition to the solution, the heuristic

provides an estimate of how far the heuristic solution is from the optimal solution.
See Simchi-Levi et al. (1999) for additional discussion on exact and heuristic algorithms.

The analytical tools used in practice are typically a hybrid of many of the tools described above. Almost all decision support systems will offer a combination of tools, and many will allow further analysis using generic tools such as spreadsheets. Note that some of the tools listed above may be embedded in generic tools, such as spreadsheets.

Of course, most DSS employ analytical tools that have some specific embedded knowledge of the problem being solved. Since these problems are usually complex, the DSS employs its problem knowledge to find efficient solutions.

There are many factors that dictate the appropriate analytical tools selected for a particular decision support system, including:

- The type of problem being considered.
- The required solution accuracy—there may be no need to find the optimal solution.
- Problem complexity—some tools may not be appropriate for very complex problems.
- The number and type of quantifiable output measures.
- The required speed of the DSS. Particularly for operational systems such as lead-time quotation and vehicle routing, speed may be essential.
- The number of objectives or goals of the decision maker. For example, a DSS for truck routing may need to find a solution with the minimum number of vehicles such that total distance traveled is as short as possible.

Applications for DSS are extremely varied, and each problem will typically use a different mathematical tool. In Table 2 we describe some of these applications and the type of tools that would typically be used. Most of these problems are extremely complex, and seemingly similar problems could require a different approach.

Consider the logistics network design problem described in the previous example. For this problem, heuristic approaches as well as optimization based techniques have been developed in the last few years. The choice between heuristics and optimization depends on the complexity of the specific problem as well as on the various modeling issues, such as service level, that the user wishes to consider. For instance, optimization-based techniques may be limited in the size of problem they can handle as well as in the number of parameters and special cases they can consider. Finally, some solvers also combine heuristics and optimization.

4.3. Presentation Tools

These are the tools used to display the data to the decision maker. There are a varied number of formats including:

- Reports
- Charts
- Spreadsheet tables
- Animation

TABLE 2 Applications and Analytical Tools

Problem	Tools Used
Marketing	Query, statistics, data mining
Routing	Heuristics, exact algorithms
Production scheduling	Simulation, heuristics dispatch rules
Logistics network configuration	Simulation, heuristics, exact algorithms
Mode selection	Heuristics, exact algorithms

- Specialized graphic formats, e.g., a layout of a floor plan
- Geographic information systems

The reader is likely to be familiar with most of these items. Reports, charts, and tables are of course very common. Animation is often used as a tool to present output of the simulation models described above. This helps the user verify the validity of the simulation model and understand the simulation results. Specialized graphic formats are, of course, extremely dependent on the nature of the problem being solved. For example, a facility layout DSS may present a suggested floor plan for a new facility.

Of course, particularly in the area of supply chain management, much of the output of DSS is geographic in nature. For example, logistics network design, sales territory analysis, and truck routing software all include geographic-related output. In the last few years, geographic information systems (GIS) have become more and more common as the presentation vehicle of choice for many supply chain management decision support systems. In the following we describe GIS systems in more detail.

4.3.1. Geographic Information Systems

A geographic information system (GIS) is an integrated computer mapping and spatial database management system that provides a broad array of functions for the storage, retrieval, management, analysis, and display of geographically referenced data.

Typical GIS capabilities include:

- Mapping and thematic mapping
- Database management
- Interactive data query
- Spatial data retrieval
- Geographic data manipulation
- Spatial data analysis
- Geocoding
- Geographic data import/export
- Buffering/polygon overlay

The advantage of using a GIS platform is that it combines database, query, and reporting tools as well as geographic display and analysis. In logistics modeling, GIS has the further advantage of allowing automated distance and travel time calculations. Some limited forms of GIS are now included in spreadsheets (Excel 7.0) but are not as extensive as the full-blown packages. These systems originally came only on high-end UNIX workstations, but there are now many excellent, relatively inexpensive systems that run on PC platforms and networks. A major issue in deploying GIS is the availability and quality of the geographic data. Excellent data, with information about the geography, street networks, and census information, are available for the United States at a very low price. However, in other countries the data itself may not exist or may be tightly held so that they can be a major expense and deterrent to the effective use of GIS. Even in the United States the data may not be perfect for every application or may be outdated and need to be upgraded to be used effectively.

Originally, GIS was extensively used in applications such as:

- Market analysis
- Census and demographic data analysis
- Real estate
- Geology
- Forestry

More recently, however, GIS has found application in areas of potentially more interest to the supply chain manager, such as:

- Network analysis—transportation, telecommunications
- Site selection
- Routing
- Supply chain management

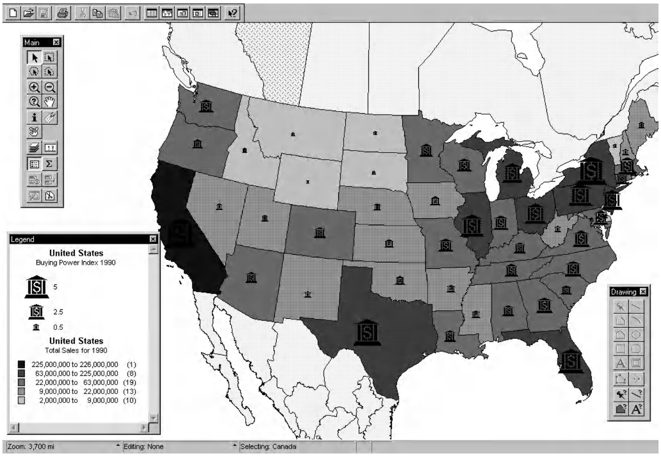

Main
Legend
United States
Buying Power Index 1990
5
2.5
0.5
United States
Total Sales for 1990
225,000,000 to 226,000,000 (1)
63,000,000 to 225,000,000 (8)
22,000,000 to 63,000,000 (19)
9,000,000 to 22,000,000 (13)
2,000,000 to 9,000,000 (10)
Drawing
Zoom: 3,700 mi
Editing: None
Selecting: Canada

Figure 5 A Typical GIS Interface.

Figure 5 presents a typical GIS interface. The screen includes a thematic display of a typical marketing application with data on 1990 U.S. buying power index given by state.

Not surprisingly, there are special considerations when using GIS in logistics modeling. Often, time must be spent geocoding and estimating travel time. Geocoding is the translation of addresses into geographic coordinates. Geocoding requires databases that can assist in the translation. Although widely available in the United States, these data may be hard to come by in other countries. In order to prepare customer data for use by a DSS, this step is required and may be lengthy, depending on the quality of the address data.

In most logistics applications, it is necessary to use the distance between two locations in order to estimate travel time and transportation costs. This can be done in several ways. One is to calculate the straight-line distance between the two coordinates and multiply it by a factor that estimates the circuity of the roads between the two points. This, of course, is a very simplistic approach, and it does not require more information than the coordinates. In this case, the DSS typically apply different factors for different zones. Another way to calculate travel distance is to use the actual road network, identify the best route, and determine the distance. This requires extensive accurate information about the road network, including one-way streets and other details. It is also an extremely time-consuming process to calculate the network even for a moderate-sized problem.

In both types of calculations, assumptions need to be made about the speed of travel in order to estimate the travel time. It is always possible to let the user enter the travel time between each pair of locations in the model, but this is usually not practical in large problems.

Although users of routing systems may demand a road network since it intuitively seems to be a more accurate solution, experience has shown that this approach may not produce significantly better results than using estimates of the distance. This is true even in shorthaul distances and inner city routing, such as school bus routing. See Table 3 for an analysis of road vs. estimated distances.

DSS may come with an embedded GIS or use a commercial GIS as a platform or server for presentation of geographic data. The U.S. geographic data available are mostly based on TIGER/line files. TIGER stands for Topologically Integrated Geographic Encoding and Referencing, for the system and digital database developed at the Census Bureau to support its mapping needs for the

TABLE 3 Road and Estituated Distance

Item	Estimated distance	Road Distance
Data	Cheap	Expensive
Complexity	Low	High
Accuracy	Medium	High
Speed	High	Low

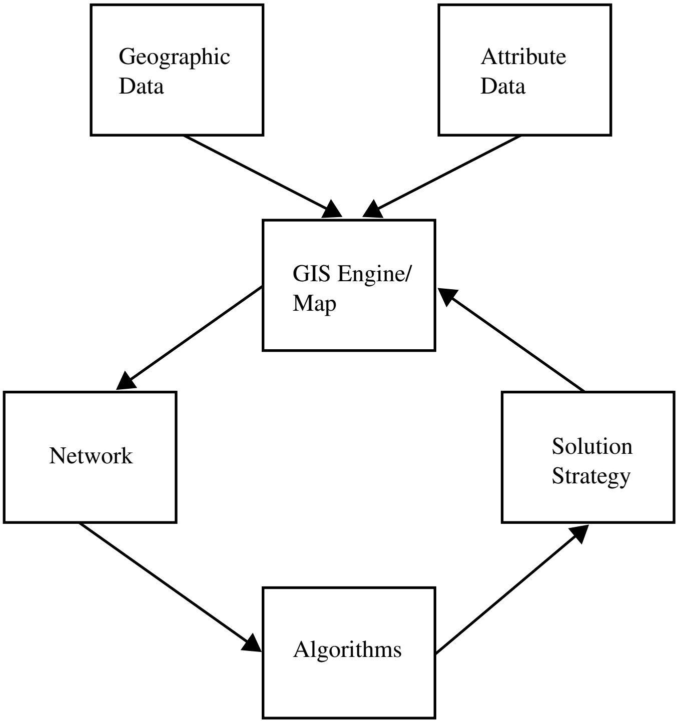


Figure 6 A General Framework for Integrating Algorithms and GIS.

Decennial Census and other Bureau programs. The TIGER/line files are publicly available extracts from the TIGER database. Most GIS vendors allow the user to load these files or distribute their own versions and formats based on this data. There are otherwise no agreed-upon standards for geographic data representation apart from the leading vendors' formats, which can usually be converted from one system to another.

4.3.2. Integrating Algorithms and GIS

As mentioned earlier, GIS has found application in areas important to supply chain management. These include logistics network design, routing, and mode selection. The idea in all these applications is to integrate the GIS with mathematical models and algorithms. Figure 6 provides a schematic representation of such a system.

In such a system, geographic data are provided by the GIS while attribute data, including demand information, costs, production and storage capacities, are downloaded from standard databases. This data is sent to the GIS engine that is the heart of the system. The engine constructs a symbolic network that represents the various relationships among the components of the supply chain. The network is then used by a collection of exact and heuristic algorithms to generate a number of solutions or strategies minimizing various objectives and satisfying all the constraints in the system. These solutions can be viewed, modified, and analyzed by the user so that the most appropriate one is implemented.

What are the advantages of integrating GIS and mathematical models and algorithms?

1. The system allows the user to visualize the data and the model and thus verify that they truly represent the supply chain environment.
2. It provides accurate street-level database (if needed) including one-way streets and turn difficulties.
3. It allows the user to visualize the solution and strategies generated by the system.
4. It allows for sensitivity (what-if) analysis.

5. THE IMPACT OF DECISION SUPPORT SYSTEMS ON LOGISTICS MANAGEMENT

In the last few years, we have seen many companies investing in, and relying on, decision support systems. The reason, of course, is these companies are trying to become best-in-class. As observed recently by PRTM Director Mike Aghajanian, ‘‘For a company with annual sales of $500 million and a 60% cost of sales, the difference between being at median in terms of supply chain performance and in the top 20% is $44 million of additional working capital.’’

Of course, the main reason companies use DSS is to reduce cost and increase service level. Indeed, reducing cost in the supply chain is a key challenge because, as observed by Rick L. Adams, VP Logistics, Grainger Industrial Supply, ‘‘A 5% cost decrease has the same impact on profit as a 30% increase in sales.’’

Thus, it is no surprise that logistics management has been transformed in the last 10 years from a largely manual process to a more automated one. Major advances in computer and communications technology and the introduction of the Internet and e-commerce have effected this trend. These developments provide new opportunities and increase expectations for a fast and flawless logistics process. The Internet also provides new models in how information systems are deployed. For instance, companies may not need to own the sophisticated DSS—they are able to lease them based on their needs. This mode of deployment is now referred to as application service provider (ASP) and is considered one of the most important trends in information systems, especially for mid-sized companies who cannot afford expensive systems.

We summarize the article with major trends that many experts (see, e.g., Shepard and Lupide 1999) envision for the future in applying DSS for logistics management:

1. The utilization of DSS will increase at all levels of decision making. The DSS will provide decision makers with assistance in quickly making effective decisions.
2. DSS will need to handle real-time data and must have a short processing time so that users can respond in the time frame that drives their business. As we have seen in transportation planning, this is an important issue in operational systems.
3. DSS will become better integrated with user's ERP and other management systems. This will allow users to access DSS information seamlessly in order to provide better customer service.
4. Users require better visibility of their data so that systems will provide users with accessible interfaces to their data in various formats and at different aggregation levels. In order for DSS to become effective, they will need to provide this capability or link with tools that already have that capability.
5. Users require collaborative tools—DSS will need to allow for collaboration in the same company and across different companies. One of the first tools in this area is the forecasting portion of collaborative planning forecasting and replenishment (CPFR), which allows partner companies to collaborate on forecasting and utilizes a DSS that assists in finding discrepancies in the process.
6. DSS at various levels will need to become better synchronized so that decisions at the strategic, tactical, and operational levels are all coordinated and accessible. As we have noted, it is difficult to perform efficient production planning without coordinating with demand planning and distribution planning.

This chapter is based on, and borrows extensively from, two chapters from our book *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies,* which was written together with Philip Kaminsky and published by McGraw-Hill in 1999. The chapter is also based on material from *The Logic of Logistics,* by Julien Bramel and David Simchi-Levi, published by Springer in 1997. In both cases this has been done with permission from the copyright holders. Research supported in part by ONR Contracts N00014-90-J-1649, N00014-95-1-0232, and N00014-01-1-0146, NSF Contracts DDM-8922712, DMI-9322828, DMI-9732795, and DMI-0085683.

REFERENCES

Fox, M. S., Chionglo, J. F., and Barbuceanu, M. (1993), ‘‘The Integrated Supply Chain Management System,’’ *Working Paper, University of Toronto.*

Shenk, D. (1997), *Data Smog: Surviving the Information Glut,* HarperCollins, New York.

Shepard, J., and Lapide, L. (1999), ‘‘Supply Chain Planning Optimization: Just the Facts,’’ in *Achieving Supply Chain Excellence through Technology,* Montgomery Research, pp. 166–176 (copies posted at http://www.ascet.com/ascet/wp/wpShepherd.html).

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E. (1999), *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies,* McGraw-Hill, Burr Ridge, IL.

CHAPTER 77

Demand Forecasting and Planning

ANANTH V. IYER
Purdue University

1. **CONSTANT DEMAND FORECAST AND ITS USE IN PLANNING** **2020**
    - 1.1. Inventory Costs 2021
    - 1.2. An Inventory Policy 2021
        - 1.2.1. Example: Choosing an Inventory Policy 2021
        - 1.2.2. Evaluating the Cost of an Inventory Policy 2021
        - 1.2.3. Economic Order Quantity (EOQ) Model 2022
    - 1.3. A Service Application—Training Airline Flight Attendants 2022
    - 1.4. Sensitivity Analysis of the EOQ model 2023
        - 1.4.1 Impact of Using Q Other Than $Q*$ 2023
2. **DEMAND MODELED AS HAVING A DISTRIBUTION** **2023**
    - 2.1. Example—The Fashion Store 2024
    - 2.2. The Single-Period Inventory Model 2024
        - 2.2.1. Suppose Demand Follows a Normal Distribution 2025
3. **UNCERTAIN DEMAND OVER MULTIPLE PERIODS** **2025**
    - 3.1. Impact of Lead Time 2025
    - 3.2. Lead Time and Demand Uncertainty 2026
        - 3.2.1. Example Problem 2026
    - 3.3. A (Q, r) Policy 2026
        - 3.3.1. Another Example Problem 2027
4. **DEMAND AS A MIXTURE OF DISTRIBUTIONS** **2027**
    - 4.1. A Mixture Model of Demand 2027
    - 4.2. Impact of Information on the Demand Model 2028
    - 4.3. Quick Response—Service Commitment 2029
5. **USING AN EXPONENTIAL SMOOTHING FORECASTING MODEL** **2029**
    - 5.1. Sample Calculations for a Demand Forecasting Model 2030
6. **SUMMARY** **2032**

ADDITIONAL READING **2032**

Our goal in this chapter is to provide a number of possible demand models, and illustrate their use in operational decision making. We cover a number of demand models, ranging from the constant demand model to a model of demand as a distribution to use of information over time to generate adaptive estimates of demand. In each case we provide numerical examples to illustrate use of the technique. For additional technical details see Additional Reading at the end of this chapter.

1. CONSTANT DEMAND FORECAST AND ITS USE IN PLANNING

One of the simplest models of demand is to use an estimate of the average demand. This average demand, assuming a constant rate each period, can then be used to understand the effect of production costs or transport costs on inventory levels. Such models are appropriate when we deal with products in situations with predictable demand, that is, low forecast error. In particular, we will focus on the

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

impact of setup costs and their interaction with holding costs. Our focus in this section is on discussing use of a constant demand rate model in an economic order quantity (EOQ) model. This requires us to define some of the costs that are important in understanding the impact of alternative decisions in an inventory system.

1.1. Inventory Costs

There are three main inventory costs we focus on:

1. *Holding or carrying costs:* These costs, expressed as a cost per unit of inventory per unit of time, model the cost associated with storage facilities, handling, insurance, pilferage, obsolescence, opportunity cost of capital, and so on.
2. *Fixed ordering or setup costs:* These costs are incurred each time an order is placed and model the costs to prepare the purchase or production order, costs associated with equipment setups, or costs associated with transportation (a single delivery truck).
3. *Shortage costs:* These are costs associated with loss of demand or penalties associated with delays.

1.2. An Inventory Policy

The context is the following: An inventory manager faces a constant demand rate that has to be satisfied from inventory. Ordering costs entail a fixed cost per order. Inventory held in the system incurs a holding cost. An *inventory policy* determines the quantity and frequency of orders. Each inventory policy has an associated cost to the organization. The two components of an inventory policy are:

- *Review period:* how often inventory levels are monitored and orders placed
- Order quantity: fixed or variable order quantity

We illustrate this concept through the use of an example.

1.2.1. Example: Choosing an Inventory Policy

Consider the problem of managing the inventory of Xerox paper in a warehouse. Although demands from retailers may fluctuate a bit in their demand, your aggregate demand for the item is fairly constant at 100,000 cases for the year. Due to your volume, your supplier has agreed to provide you an everyday low price of $55.00 a case. You calculate that it will cost about $4.00 per case per year to hold each case. Costs associated with each order and delivery charges by your supplier yield a fixed ordering cost of $75.00.

Consider the following inventory policy:

1. Place an order for Q units.
2. When inventory reaches zero, place another order for Q units and repeat.
3. Assume lead time is zero.

Which inventory policy do you recommend?

- Review daily resulting in daily shipments of about 385 cases each.
- Review weekly, resulting in order sizes of about 1923 cases each.
- Place an order each month at about 8333 cases each.
- Place only two orders per year at about 50000 cases each.

1.2.2. Evaluating the Cost of an Inventory Policy

To calculate costs of each policy, let:

d = yearly demand in units
h = holding costs in dollars per unit per year
s = setup or ordering costs in dollars
Q = quantity of each order
Q = uniquely defines this simple inventory policy

$$\text{Total cost (AHO)} = \text{annual carrying costs} + \text{annual ordering costs}$$

$$\text{AHO} = \frac{Q}{2}h + \frac{d}{Q}s$$

So:

d = 100,000 units/yr
h = 4 dollars/unit-yr
s = 75 dollars

$$1.\ Q = 385,\ \text{AHO} = \frac{385}{2}4 + \frac{100{,}000}{385}75 = 20{,}250$$

$$2.\ Q = 1{,}923,\ \text{AHO} = \frac{1{,}923}{2}4 + \frac{100{,}000}{1{,}923}75 = 7{,}746$$

$$3.\ Q = 8{,}333,\ \text{AHO} = \frac{8{,}333}{2}4 + \frac{100{,}000}{8{,}333}75 = 17{,}566$$

$$4.\ Q = 50{,}000,\ \text{AHO} = \frac{50{,}000}{2}4 + \frac{100{,}000}{50{,}000}75 = 100{,}150$$

From these policies the lowest cost is to order weekly with $Q = 1{,}923$.

1.2.3. Economic Order Quantity (EOQ) Model

Consider the following inventory model:
Place an order for Q units.
When inventory reaches zero, place another order for Q units and repeat.

d = yearly demand in units.
h = carrying or holding costs in dollars per unit per year.
s = setup or ordering costs in dollars.

$$Q^* = \text{economic order quantity (EOQ)}$$

$$Q^* = \text{EOQ} = \sqrt{\frac{2ds}{h}}$$

minimizes the total cost function:

$$\text{AHO} = \frac{Q^*}{2}h + \frac{d}{Q^*}s$$

For:

d = 100,000 units/yr
h = 4 dollars/unit-yr
s = 75 dollars

$$Q^* = \sqrt{\frac{2(100{,}000)(75)}{4}} = 1{,}936$$

$$\text{AHO} = \frac{1{,}936}{2}4 + \frac{100{,}000}{1{,}936}75 = 7{,}746$$

$$\text{Total cost} = \text{AHO} + (cD)$$

1.3. A Service Application—Training Airline Flight Attendants

You manage the training department for airline attendants for a major airline. You have to organize and plan the training sessions for all new hires. Each week personnel sends you a list of the new hires that you must schedule into the next available training session—they are on the payroll and basically idle until they can be trained and put into service. The yearly demand for new attendants is fairly constant at 1200, and personnel is always interviewing and hiring to keep up with the demand. The new hires, on average, cost the company about $40,000 per year in salary and benefits.

You wonder if there is a way to think about the costs in the system and better manage them. First you realize that personnel and training should coordinate better. Why should personnel put someone on the payroll before they can be scheduled for training? After all, a new hire won't argue about a week here or there on a start date. This seems like an easy thing to accomplish, but it certainly puts the burden on you to schedule the sessions and keep personnel abreast of your plans. Next you wonder how to best plan the training sessions—how often should you hold sessions and how many trainees should you put in each class. You gather the following data: The cost of the teachers, the conference rooms, and so on combine for a fixed cost of $10,000 per training session.

How should you plan the sessions to minimize total cost? Note that given the data, we see that $s = 10{,}000$, $d = 1{,}200$ and $h = 40{,}000$. Thus, the optimal training batch size is

$$\sqrt{\frac{2 \times 10{,}000 \times 1{,}200}{40{,}000}}$$

which is rounded up to 25 per batch. This corresponds to training approximately once a week. The cost associated with this decision

$$\frac{10{,}000 \times 1{,}200}{25} + \frac{40{,}000 \times 25}{2} = 980{,}000$$

1.4. Sensitivity Analysis of the EOQ Model

While the EOQ model provides quick estimates of optimal behavior, it requires knowledge of the costs described earlier. It also requires us to have a good estimate of demand rate. However, the benefit of the model is that it is robust to variation in the assumptions. This is indicated as follows:

*1.4.1. Impact of Using a Q Other Than Q**

Suppose, in practice, the actual Q used in a situation is different from Q^*. We will examine the impact of this difference is Q on average annual costs AHO(Q). We define the average annual costs as

$$AHO(Q) = \frac{sd}{Q} + \frac{hQ}{2}$$

and

$$Q^* = \sqrt{\frac{2sd}{h}}$$

(In this analysis, we thus drop the term cd from both expressions.) Note that

$$\frac{\text{AHO}(Q)}{\text{AHO}(Q^*)} = \frac{1}{2}\left\{\frac{Q}{Q^*} + \frac{Q^*}{Q}\right\}$$

The equation above is obtained by substituting the optimal Q^* in the cost expression for AHO(Q^*)

Note that if Q varies between $Q^*/2$ and $2Q^*$, the impact on average annual cost is less than 25%. Similarly, if Q varies between $0.8Q^*$ and $1.25Q^*$, the average annual costs increase by no more than 2.5%.

These results suggest that the EOQ model is robust, that is, variations in decisions result in small variation in their cost impact as long as we are around the optimal EOQ model.

2. DEMAND MODELED AS HAVING A DISTRIBUTION

In the earlier section, we modeled demand as a constant rate. Often, however, demand is not very predictable but has a significant amount of randomness. To understand the effect of demand forecast error, we first focus on problems where decisions regarding inventory are made once for an entire period. Examples of products that might require inventory decisions that cover demand over a single period include

- Newspapers for which the period refers to one day
- Fruits/flowers for which the period refers to one week

- Fashion products for which the period refers to a season (three months)
- Hotel room reservations for which the period refers to one day
- Airline reservations for which we refer to a particular flight

To illustrate the basic ideas, we start with a numerical example.

2.1. Example—The Fashion Store

The Fashion Store sells fashion items. The store has to order these items many months in advance of the fashion season in order to get a good price on the items. Each unit costs Fashion $100. These units are sold to customers at a price of $160 per unit. Items not sold during the season can be sold to the outlet store at $75 per unit. If the store runs out of an item during the season, it has to obtain the item from alternative sources and the cost including air freight to Fashion is $190 per unit.

Fashion wants help in choosing the initial order quantity to minimize costs to run the store.

Historical data from comparable items over the last few years have generated 100 demand observations as follows:

86	94	90	86	82	84	91	76	85	83	92	82	89	88	79	83	83	85	89	90
73	84	86	90	90	92	83	91	85	85	82	81	81	76	81	81	78	85	84	82
88	86	85	88	86	89	87	84	83	79	90	87	83	87	82	81	85	84	87	89
82	80	92	85	88	85	83	87	84	84	86	80	87	80	89	79	83	80	86	87
81	93	91	89	80	86	87	86	88	84	81	84	84	82	77	93	94	97	87	75

A frequency table of these points and the corresponding cumulative probabilities for each demand value are as follows:

Demand (D)	Frequency	P(demand) P(D)	Cum. Prob. P(Demand ≤ D)
73	1	0.01	0.01
75	1	0.01	0.02
76	2	0.02	0.04
77	1	0.01	0.05
78	1	0.01	0.06
79	3	0.03	0.09
80	5	0.05	0.14
81	7	0.07	0.21
82	7	0.07	0.28
83	8	0.08	0.36
84	10	0.10	0.46
85	10	0.10	0.56
86	9	0.09	0.65
87	8	0.08	0.73
88	5	0.05	0.78
89	6	0.06	0.84
90	5	0.05	0.89
91	3	0.03	0.92
92	3	0.03	0.95
93	2	0.02	0.97
94	2	0.02	0.99
97	1	0.01	1.00
	100	1.0	

The sample mean is 85 units and the standard deviation is 4.43 units.

2.2. The Single-Period Inventory Model

Given the problem described earlier, it turns out that the optimal decision is described using two cost parameters:

1. C_e = cost per unit of an excess item at the end of period
2. C_s = cost per unit of an item short

The optimal service level defined as the probability that all demand is satisfied in a period immediately from primary stock is described as

$$\text{ser}^* = \frac{C_s}{C_s + C_e}$$

From our example we have:

$$C_e = 25$$

$$C_s = 90$$

So,

$$\text{ser}^* = \frac{C_s}{C_s + C_e} = \frac{90}{90 + 25} = 0.782$$

From frequency table we get for a service level of 0.782,

$$r^* \approx 89$$

2.2.1. Suppose Demand Follows a Normal Distribution

Example 1: Fashion Store. If we approximate demand by a normal distribution with mean $m = 85$ and standard deviation $\sigma = 4.43$, we get

$$\text{ser}^* = \frac{C_s}{C_s + C_e} = \frac{90}{90 + 25} = 0.782$$

so

$$r^* = 85 + Z_{0.782}(4.43) = 85 + 0.76(4.43) = 88.45 \approx 89$$

Example 2: Fashion Store. Approximate by Normal with $m = 85$ and $\sigma = 4.43$

Suppose the Fashion Store wants to provide a service level of 90%. What level of inventory is required?

$$r_{0.90} = 85 + Z_{.90}(4.43) = 85 + 1.29(4.43) = 90.7 \approx 91$$

Therefore, if Fashion has an inventory of 91 items with probability 0.90 all the demand in a period is satisfied from primary stock.

3. UNCERTAIN DEMAND OVER MULTIPLE PERIODS

In this section we return to making decisions over continuous time. Demands each period are modeled as following a distribution. We first consider the case where we can only make decisions once each period. The costs consist of the costs of holding inventory each period and the costs of shortage each period. We include the case where there may be a fixed lead time L for the supplier to deliver an order. The basic idea is that in the presence of uncertainty and lead time for delivery, orders have to be placed well in advance of inventory running out in order to guarantee a high level of availability.

3.1. Impact of Lead Time

If we consider the problem in the earlier sections with constant demand and just add in a lead time L between order placement and delivery, note that orders should be placed when the inventory level hits dL. This level is called the reorder level. The order size remains Q as calculated by the EOQ model.

An interesting issue is the fact that in the absence of demand forecast error, lead time has no impact on costs. This is seen by the fact that while the order trigger times are affected by lead time

L, the physical inventory levels and the rate of order placements are unaffected by L. Since these two parameters affect costs in this model, costs are unaffected by lead time.

3.2. Lead Time and Demand Uncertainty

In the presence of lead time and demand uncertainty, the reorder level requires us to decide how much of the demand should be satisfied from stock. This factor can be expressed as *service level.* We express service level as the *probability* of being in stock. This probability is a number that the customers (generating the demand) can use to do their own planning. This probability will depend (intuitively) on the industry and on the extent of competition faced by the company.

If demand follows a normal distribution, then the reorder level requires us to merely generate a Z value that corresponds to the desired cumulative probability.

Then set the reorder level as

$$r = (dL) + (Z\sigma\sqrt{L})$$

where σ refers to the standard deviation of demand per unit time and d is the mean demand per unit time.

3.2.1. Example Problem

The Reliable Hardware Store sells electric pumps. The supplier lead time is one week. Each pump costs Reliable \$100. Annual inventory carrying cost is 25% of the investment. If Reliable stocks out, it will fill demand by buying the required pumps elsewhere at an additional cost of \$20, that is, Reliable's cost would be \$120. Reliable's planned probability of in stock is 91.2%.

Demand recorded for the last 100 weeks shows the following data:

86	94	90	86	82	84	91	76	85	83	92	82	89	88	79	83	83	85	89	90
73	84	86	90	90	92	83	91	85	85	82	81	81	76	81	81	78	85	84	82
88	86	85	88	86	89	87	84	83	79	90	87	83	87	82	81	85	84	87	89
82	80	92	85	88	85	83	87	84	84	86	80	87	80	89	79	83	80	86	87
81	93	91	89	80	86	87	86	88	84	81	84	84	82	77	93	94	97	87	75

The mean demand is 85 units and the standard deviation is 4.43 units. The same data can be plotted as a histogram or as a line graph using a spreadsheet.

Note that when Reliable reorders, there should be sufficient inventory on hand to cover demand during the lead time of one week.

3.3. A (Q, r) Policy

We now formalize our multiperiod inventory system. By convention, some parameters are usually given in annual terms (the EOQ parameters), while other parameters are stated in smaller period terms (the lead time parameters).

s = ordering cost
d = average annual demand
h = annual holding cost per item
L = fixed lead time in periods
ser = planned service level
Z_{ser} = the Z value that generates the required in stock probability
m = mean demand during a period
σ = standard deviation of demand during a period

Then we first calculate:

$$Q = \sqrt{\frac{2ds}{h}}$$

And then the reorder point:

$$r = (mL) + (Z_{\text{ser}}\ \sigma\sqrt{L})$$

$$\text{Average physical inventory level} = \frac{Q}{2} + r - (mL)$$

3.3.1. *Another Example Problem*

Thus, for the problem faced by Steco, a retailer of titanium rods. Weekly demand for these rods follows a normal distribution with a mean of 100 units and a standard deviation of 5 units per week. These rods cost \$5 each and the cost of holding a rod in inventory for a year is 20% of its cost. The cost to Steco to place an order for replenishment is \$25/order. Delivery lead time is one week. Management wants a less than 6% probability of stocking out. Assume 50 weeks per year.

Develop a (Q, r) policy for Steco.

Answer:

Ordering cost $= s =$ \$25/order
Holding cost $= h =$ \$1/unit/year
Annual demand $= d = 100 \times 50 = 5000$ units/year

$$Q = \sqrt{\frac{2sd}{h}} = 500$$

$$\begin{aligned} m &= 100, \\ \sigma &= 5, \\ L &= 1, \\ \text{ser} &= 0.94, \text{ and} \\ Z_{0.94} &= 1.56 \text{ (from the normal table)} \\ r &= 108 \end{aligned}$$

Thus, Steco should reorder Q units whenever the pipeline inventory level falls to below r units.

Steco's average inventory level $= Q/2 + r - (mL) = 258$ units

4. DEMAND AS A MIXTURE OF DISTRIBUTIONS

We now consider an alternative representation of demand that will permit us to incorporate information collected to lower demand uncertainty. The role of the decision maker is to use the best available assessment of the weights to be assigned to each demand model in order to model the demand at that point in time. We do this by representing demand as a mixture of a number of possible demand models. This in effect means that the model of demand will change over time, reflecting the link between information and its role in reducing demand uncertainty. We illustrate this idea using an example problem.

4.1. A Mixture Model of Demand

Consider a retailer (ASSORT) who sells women's dresses. Analysis of the historical data has indicated that at the end of the season, some dresses follow a demand whose distribution is uniform between 1 and 5. Other (in-fashion) dresses follow a demand whose distribution is between 4 and 8 units. About eight months in advance, when the order is placed, the best estimate is that demand for a particular dress will be either of these two distributions with a 50% probability.

These dresses are bought for \$100 and sell for \$200. If ASSORT runs out of dresses, the future profit impact is estimated to be \$200. Dresses that are held through the fashion season incur a holding cost of \$20. Those dresses that do not sell during the season are sold to an outlet store for \$40.

Under this system, what will be ASSORT's order size and associated expected profit if orders have to be placed eight months in advance?

Answer: Given these costs, note that ASSORT will estimate the optimal service level to be

$$\text{Service level} = \frac{r + g - c}{r + h + g - s}$$

where $r = 200$, $c = 100$, $h = 20$, $s = 40$, $g = 200$. Thus, the optimal service level is 78.9%.

The demand distribution at this point in time is

Demand	Probability
1	0.1
2	0.1
3	0.1
4	0.2
5	0.2
6	0.1
7	0.1
8	0.1

With this demand distribution, the inventory of dresses purchased is 6 units. Associated with this inventory purchased, ASSORT's expected profit is as follows:

$(-100*6)$ + (purchase costs)

$((0.1*200*1) + (0.1*200*2) + (0.1*200*3) + (0.2*200*4) + (0.2*200*5) +$
$(0.1*200*6) + (0.1*200*6) + (0.1*200*6))+$ (expected revenue)

$((0.1*20*5) + (0.1*20*4) + (0.1*20*3)+$
$(0.2*20*2) + (0.2*20*1)) +$ (expected salvage − Holding Costs)

$((-0.1*200*1) + (-0.1*200*2))$ (expected penalty costs)

$= 216$

The associated manufacturer revenue is $100 * 6 = \$600$.

4.2. Impact of Information on the Demand Model

The retailer has heard of a new scheme called quick response (QR). Under this scheme, the manufacturer has to receive the order only four months in advance. This enables the retailer to collect data regarding sales of similar products. These similar product sales enable the retailer to further refine the demand distribution estimates. What will be the impact of QR on the retailer?

Answer: Under QR, the retailer observes a draw from the demand distribution. Depending on the value of this observed demand, ASSORT will adjust demand estimates as follows:

Demand	Probability
$1 \le \mathrm{d}1 \le 3$	$P(1) = 1, P(2) = 0$
$4 \le \mathrm{d}1 \le 5$	$P(1)=1/2, P(2) = 1/2$
$6 \le \mathrm{d}1 \le 8$	$P(1) = 0, P(2) = 1$

Thus, the observed demand changes the weights we would place on each of the two demand distributions.

As before, we can then derive the optimal inventory policy and associated expected profit for the retailer as follows:

Demand ($d1$)	Probability	Inventory	Expected Profit
$1 \le \mathrm{d}1 \le 3$	0.3	4	144
$4 \le \mathrm{d}1 \le 5$	0.4	6	216
$6 \le \mathrm{d}1 \le 8$	0.3	7	444
			Total expected profit 262.8

Thus, the retailer's profit increases from 216 to 262.8, an increase of 22%.

What is the expected quantity purchased from the manufacturer?

$$(0.3 * 4) + (0.4 * 6) + (0.3 * 7) = 5.7$$

Thus, manufacturer revenues decreases to \$570, a drop of 5%.

4.3. Quick Response—Service Commitment

Suppose the retailer commits to a service level of 100% in return for the manufacturer providing QR. What will be the impact on the system?

Answer: We will have to change the inventory purchased after observing demand and thus get the following results:

Demand (*d*1)	Probability	Inventory	Expected Profit
$1 \le d1 \le 3$	0.3	5	140
$4 \le d1 \le 5$	0.4	8	170
$6 \le d1 \le 8$	0.3	8	440
			Total expected profit 242.0

The associated manufacturer revenue is

$$100\ [(0.3 * 5) + (0.4 * 8) + (0.4 * 8)] = 710$$

Thus, the manufacturer and the retailer are better off than under the old system.

Where are the retailer's increases in expected profit coming from?

Answer: Consider the service level in the old system and the new QR system with a 100% service level.

Demand (*d*1)	Probability	Old System Service Level	QR System 100% Service Level
$1 \le d1 \le 3$	0.3	100%	100%
$4 \le d1 \le 5$	0.4	80%	100%
$6 \le d1 \le 8$	0.3	60%	100%

By choosing the assortment of dresses closer to the season, ASSORT faces a lower forecast error. This enables the retailer to have fewer stockouts, the manufacturer to have higher revenue, and the customer to have a higher service level. All of this is accomplished without a decrease in retailer profits.

Note that providing a 100% service level in the old system would have decreased retailer pro?ts to \$ 170 because of the increased holding and salvage related costs.

Thus, QR allows the customer service level to be increased without decreasing retailer profits. This example thus illustrates the benefit of using a mixture of demands to represent demand for a product. As information is collected, the weights associated with the possible distributions can be adjusted, thus providing a convenient way to incorporate information into the forecasting process. It also shows the benefit of members of a channel working together to permit information to be incorporated into a decision-making process.

5. USING AN EXPONENTIAL SMOOTHING FORECASTING MODEL

We now illustrate the role of another classic demand-forecasting model—the exponential smoothing model. The exponential smoothing model works as follows: Given a demand forecast from previous periods and an observation this period, and a parameter α, the exponential smoothing model is that

$$\text{Demand forecast} = (\alpha \text{ observed demand}) + ((1 - \alpha) \text{ previous forecast})$$

We now illustrate the role of the exponential smoothing model in a supply chain. The model will allow us to understand how a supply chain reacts to partial information in adjusting its decisions. We start with a two-level supply chain involving a retailer and a manufacturer with a lead time L. The system operates as follows:

1. Retailer receives deliveries and then faces demand. Thus, the retailer has either too much or too little inventory at the end of the period.
2. Retailer updates demand forecast based on observed demand. This forecast follows an exponential smoothing model with parameter α.

Thus,

$$\text{Demand forecast} = (\alpha \text{ observed demand}) + [(1 - \alpha) \text{ previous forecast}]$$

3. Given a total lead time of L, the base stock level is set as $(L + 1)$ demand forecast
4. Retailer places an order to bring pipeline inventory level up to the base stock level.
5. Orders placed by the retailer reach the manufacturer after an information lead time.
6. The manufacturer receives deliveries and then faces demand. Thus, the manufacturer has either too much or too little inventory at the end of the period. Manufacturer shipments reach the retailer after a delivery lead time.
7. Manufacturer updates demand forecast based on observed demand. This forecast follows an exponential smoothing model with parameter α. Thus,

$$\text{Demand forecast} = (\alpha \text{ observed demand}) + [(1 - \alpha\ ) \text{ previous forecast}]$$

8. Given a total lead time of L, the base stock level is set as $(L + 1)$ demand forecast
9. Manufacturer places an order to bring pipeline inventory level up to the base stock level.
10. Manufacturer orders are filled after a production lead time.

5.1. Sample Calculations for a Demand Forecasting Model

We illustrate the model in the earlier section for a four-level supply chain consisting of a retailer, a wholesaler, a distributor, and a manufacturer. There is a 4-period lead time between successive levels, consisting of 2 periods to transmit an order upstream and 2 periods to deliver downstream between successive levels. Manufacturing lead time is 2 periods. Demand is 4 units at the retailer for the first 10 periods and goes to 8 units from then on. In this section, we summarize the steps in generating values that reflect orders placed at each of the four stages of the beergame in response to demand changes. We will consider the case where each level uses $\alpha = 0.2$. The steps in calculating the order size are outlined below. We will assume that each level uses a base stock level $= (L + 1 + 2) \times$ demand forecast. The parameter L refers to the lead time for order delivery. The value 2 refers to 2 periods of safety stock (assumed).
Retailer—period 11:

1. Observe demand of 8 units.
2. Update demand forecast as follows:

$$\text{Demand forecast} = (0.2 \times 8) + [(1 - 0.2) \times 4] = 4.8$$

3. Base stock level $= (4 + 1 + 2) \times 4.8 = 33.6$.
4. Pipeline inventory level after satisfying period 11 demand $= 28 - 8 = 20$
5. Order placed in period 11 $= 33.6 - 20 = 13.6$.

Similarly, for period 12, the calculations are as follows:

1. Observe demand of 8 units.
2. Update demand forecast as follows:

$$\text{Demand forecast} = (0.2 \times 8) + [(1 - 0.2) \times 4.8] = 5.44$$

3. Base Stock Level $= (4 + 1 + 2) \times 5.44 = 38.08$.
4. Pipeline inventory level after satisfying period 11 demand $= 33.6 - 8 = 25.6$.
5. Order placed in period 11 $= 38.08 - 25.6 = 12.48$.

We now consider the orders placed by the wholesaler in period 13.
Wholesaler—period 13:

1. Observe demand of 13.6 units.
2. Update demand forecast as follows:

$$\text{Demand forecast} = (0.2 \times 13.6) + [(1 - 0.2) \times 4] = 5.92$$

3. Base stock level = $(4 + 1 + 2) \times 5.92 = 41.44$.
4. Pipeline inventory level after satisfying period 13 demand = $28 - 13.6 = 14.4$
5. Order placed in period 13 = $41.44 - 14.4 = 27.04$.

Similarly, for period 14, the calculations are as follows:

1. Observe demand of 12.48 units.
2. Update demand forecast as follows:

$$\text{Demand forecast} = (0.2 \times 12.48) + [(1 - 0.2) \times 5.92] = 7.232$$

3. Base stock level = $(4 + 1 + 2) \times 7.232 = 50.624$.
4. Pipeline inventory level after satisfying period 14 demand = $41.44 - 12.48 = 28.96$.
5. Order placed in period 14 = $50.624 - 28.96 = 21.664$.

We now consider the orders placed by the distributor in period 15.
Distributor—period 15:

1. Observe demand of 27.04 units.
2. Update demand forecast as follows:

$$\text{Demand forecast} = (0.2 \times 27.04) + [(1 - 0.2) \times 4] = 8.608$$

3. Base stock level = $(4 + 1 + 2) \times 8.608 = 60.256$.
4. Pipeline inventory level after satisfying period 15 demand = $28 - 27.04 = 0.96$.
5. Order placed in period 15 = $60.256 - 0.96 = 59.296$.

Similarly for period 16, the calculations are as follows:

1. Observe demand of 21.664 units.
2. Update demand Forecast as follows:

$$\text{Demand forecast} = (0.2 \times 21.664) + [(1 -0.2) \times 8.608] = 11.2192$$

3. Base stock level = $(4 + 1 + 2) \times 11.2192 = 78.5344$.
4. Pipeline inventory level after satisfying period 16 demand = $60.256 - 21.664 = 38.592$.
5. Order placed in period 16 = $78.5344 - 38.592 = 39.9424$.

We now consider the orders placed by the factory in period 17.

Factory—period 17:

1. Observe demand of 59.296 units.
2. Update demand forecast as follows:

$$\text{Demand forecast} = (0.2 \times 59.296) + [(1 - 0.2) \times 4] = 15.0592$$

3. Base stock level = $(2 + 1 + 2) \times 15.0592 = 75.296$.
4. Pipeline inventory level after satisfying period 17 demand = $20 - 59.296 = -39.296$.
5. Order placed in period 17 = $75.296 - (-39.296) = 114.592$.

Similarly, for period 18, the calculations are as follows:

1. Observe demand of 39.9424 units.
2. Update demand forecast as follows:

$$\text{Demand forecast} = (0.2 \times 39.9424) + [(1 - 0.2) \times 15.0592] = 20.0358$$

3. Base stock level = $(2 + 1 + 2) \times 20.0358 = 100.179$.
4. Pipeline inventory level after satisfying period 16 demand = $75.296 - 39.9424 = 35.3536$.
5. Order placed in period 16 = $100.179 - 35.3536 = 64.8256$.

Note that the effect of the information and delivery lags between the stages is to increase a customer demand increase from 4 to 8 units in period 11 to a wholesale order change from 4 to 27.04 in period 13 to a distributor order increase from 4 to 59.296 in period 15 and a factory increase in its brewed cases from 4 to 114.592. Each decision reflects a rational choice given the parameters. This increase in variance or orders as we go up a supply chain is referred to as the bullwhip effect. Demand updating is only one possible reason for the bullwhip effect.

6. SUMMARY

In this chapter we have provided a quick review of four possible approaches to forecast demand and its use in planning. The constant demand model allows for a quick analysis of the effect of ordering costs in a system. The models of demand as a distribution permit details of lead time and demand uncertainty to be included. The modeling of demands as a mixture of distributions enables us to consider the role of information acquired over time. Finally, the exponential smoothing model shows how demand forecast updating can create large swings upstream in a supply chain.

ADDITIONAL READING

Iyer, A. V., and Bergen, M. E., ‘‘Quick Response in Manufacturer–Retailer Channels,’’ *Management Science,* Vol. 43, No. 4, 1997, pp. 559–570.

Lee, H. L., Padmanabhan, V., and Whang, S., ‘‘The Bullwhip Effect in Supply Chains,’’ *Sloan Management Review,* Vol. 38, No. 3, Spring 1997, pp. 93–102.

Nahmias, S., *Production and Operations Analysis,* 2nd Ed., Richard D. Irwin, Homewood, IL, 1993.

Zipkin, P. H., *Foundations of Inventory Management,* McGraw-Hill, New York, 2000.

CHAPTER 78

Advanced Planning and Scheduling for Manufacturing

KENNETH MUSSELMAN
Frontstep, Inc.

REHA UZSOY
Purdue University

1. INTRODUCTION 2034

2. THE PLANNING AND SCHEDULING FUNCTIONS 2034
  2.1. Planning 2034
  2.2. Scheduling 2035

3. RELATIONSHIPS BETWEEN PLANNING AND SCHEDULING 2036
  3.1. Function 2036
  3.2. Treatment of Capacity 2037
  3.3. Representation 2038

4. PLANNING ALGORITHMS 2038
  4.1. Infinite Capacity Algorithms: Material Requirements Planning 2039
  4.2. Finite Capacity Algorithms 2042
    4.2.1. Extensions to MRP 2042
    4.2.2. Optimization Approaches 2043
    4.2.3. Artificial Intelligence Approaches 2044
    4.2.4. Congestion Models 2044
    4.2.5. Detailed Scheduling as Part of the Planning Process 2044

5. ADVANCED PLANNING AND SCHEDULING 2045
  5.1. Advanced Planning 2045
  5.2. Advanced Scheduling 2046
  5.3. Order Promising 2046

6. APS IMPLEMENTATION ISSUES 2046
  6.1. ERP Integration 2047
  6.2. Timing, Access, and Quality of Data 2047
    6.2.1. Timing 2047
    6.2.2. Access 2047
    6.2.3. Quality 2049
  6.3. Business Process Reengineering 2049
  6.4. A Well-Defined Manufacturing System 2049
    6.4.1. As-Manufactured, Indented Bills of Material 2050
    6.4.2. Accurate Routes 2050
    6.4.3. Supportive Operational Buffer Times 2050
    6.4.4. Consistent Workcenter Definitions 2050
    6.4.5. Representative Purchasing Lead Times 2050
    6.4.6. Appropriate Workcenter and Material Constraints 2050
    6.4.7. Representative Scheduling Rules 2050
  6.5. Usual Suspects 2051
  6.6. Implementation Strategies 2051

7. CONCLUSIONS 2052

REFERENCES 2052

ADDITIONAL READING 2053

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

The problem of how to allocate a company's resources and material effectively among competing activities over time to optimize the company's market and financial positions is encountered in all industries producing goods or services. Unless a company has significant excess capacity and high inventories, decisions have to be made as to which orders it will accept, which it will turn away, and which products and customers will be given priority over others—in other words, how its available inventory and production capacity will be allocated among revenue-generating activities.

Several factors combine to make this a difficult task. The company must respond to often rapidly changing market conditions and technological developments. A number of different, often conflicting objectives, such as filling customer orders and maintaining low inventory levels and lead times, must be traded off against each other. Different amounts of variability in the production processes and customer demand must be managed. However, there is considerable evidence from various industries that effective execution of this task can provide a significant competitive advantage.

Today a number of trends are combining to render the area of production planning rather more active than it has been for several decades. Increasing competitive pressures have forced companies to forgo the expensive luxury of large amounts of excess capacity and high inventories, making effective allocation of manufacturing capacity and coordination of production activities throughout the supply chain a critical component of market success. The strong demand from industry for these services is demonstrated by the rapidly increasing number of software products and consulting companies specializing in this area that have emerged in the last five years. The explosive expansion of information technology fueled by Moore's law, manifested in the Internet, better databases, and faster, cheaper computers, has rendered feasible a whole range of solutions that could not even be imagined 10 years ago.

In this chapter we shall examine developments in the area of production planning, focusing on the case where a single factory is considered. However, much of the discussion and many of the modeling issues remain valid when production systems involving multiple plants are considered. We do not, however, make any attempt to consider the larger supply chain, which by its nature must consider such aspects as transportation, warehousing, and interactions with other companies. The issue of supply chain management is a fast-growing area of both research and practice and is discussed in more detail in Chapter 82, as well as in Tayur et al. (1998).

We begin by discussing the relationship between the production planning and scheduling functions and their importance to the manufacturing firm. In this context we discuss the effects of congestion on the shop floor and the relationship between workload and lead times, which is fundamental to the relationship between planning and scheduling. We introduce the topic of production planning by discussing at some length the well-known and widely used Material Requirements Planning (MRP) algorithm (Orlicky 1975) and its extensions. This approach is widely used in industry, and much of the current effort in developing advanced planning and scheduling (APS) systems is aimed at remedying its various deficiencies. After defining our view of APS, we present a range of production planning algorithms that have been proposed in industry and academia over the last several years, discussing their strengths and weaknesses. Finally, we identify a number of issues that in our experience must be addressed to implement an APS system successfully.

2. THE PLANNING AND SCHEDULING FUNCTIONS

Generally speaking, planning and scheduling jointly determine how, when, and in what quantity products will be manufactured or purchased. In essence, planning establishes *what* should be done and scheduling determines *how* to do it. The conventional approaches to both these functions are explained below, together with fundamental issues associated with both.

2.1. Planning

Planning determines when to manufacture and purchase parts and how many in order to satisfy future demand for end products. The process is externally focused since the demand comes from both actual and anticipated customer orders. It is controlled by higher-level attributes, such as end-item due dates and order types, and takes an aggregate view of the production process. This aggregation takes several forms: individual machines are aggregated into workcenters for the purpose of representing capacity, time is aggregated into discrete buckets, and the flow time for a number of operations required to produce a component or subassembly is often aggregated into a single lead time. The details associated with how work actually flows through the plant, such as the specific timing of individual operations and production sequences on individual machines, are left unresolved. The time frame, or horizon, over which the plan is made is normally on the order of weeks or months. The result of the process is a time-phased projection of inventory levels, production quantities, and workcenter requirements to satisfy independent demand.

In this incarnation, the production plan serves a number of functions. It represents a decision as to how the company's manufacturing capacity will be allocated among competing products and customers and hence the point where the company's strategy is turned into concrete actions visible by employees and customers. Secondly, it provides management with some visibility into the future status of production, allowing them to identify at least some problematic situations such as mismatches between demand and capacity in time for remedial action to be taken. It thus provides critical information for a number of activities, such as negotiating due dates with customers and deciding the timing and quantity of raw material purchases from suppliers.

It is important to note that the production plan is by nature somewhat tentative, being subject to significant uncertainty. In many companies, at least part of the demand considered in the plan is based on forecasts rather than firm orders and is hence subject to varying degrees of change over time as orders are modified, added, or canceled by customers. Even when demand uncertainty is minimal, the actual execution of the plan on the shop floor is subject to random disruptions such as machine failures, quality problems, and unexpected rush orders. In many companies, production plans are developed and used on a rolling horizon basis, with decisions in the early periods being considered binding but those further out being revised and altered as new information on the realizations of demand and production become available.

It is also important to note that the nature of the planning problem may differ quite substantially depending on the nature of the company's business. At one extreme is the make-to-stock environment, where demand is relatively high and stable. The lead time to produce an end item is sufficiently high that the company maintains substantial finished goods inventory to meet customer orders and produces to replenish this inventory based on demand forecasts. Another situation is a make-to-order environment, where customized items are produced upon receipt of the order. Here the planning system must allow the company to assess intelligently whether or not customer orders can be completed by the requested date and allow detailed coordination of material and other resources to achieve this. Assemble-to-order systems are intermediate in nature, where a number of basic subassemblies are produced to stock and then combined in different ways in response to customer orders. A company's planning and scheduling needs may differ quite substantially depending on the environment in which it is operating. Hendry and Kingsman (1989) discuss the needs of make-to-order companies and the relevance of many planning and scheduling approaches such as kanban (e.g., Monden 1983) and theory of constraints (Goldratt and Fox 1986) in this environment.

An important role in production planning belongs to the master production schedule (MPS), which specifies the quantity of each final demand item required in each time period and drives the requirements planning calculation of how many of each component and subassembly to produce to meet this demand over time and thus the scheduling system that moves the work through the individual operations to meet this plan. Originally MPS was treated as an exogeneous input to the system that developed the quantities and timing of releases to the shop floor. The goal of many APS systems, especially transaction-oriented systems, is to integrate planning and scheduling decisions to a much higher degree. Hence, much of the functionality of the MPS is fulfilled by the plan developed by the APS system.

The MPS, or the planning function that fulfills this role, is a crucial element of the production-planning process for several reasons. First, if the MPS is not realistic with respect to the various constraints such as production and supplier capacity and material availability faced by the manufacturing organization, it is unlikely that the best APS software or shop-floor scheduling package can save the situation. Secondly, the MPS is where a crucial set of decisions determining how the company's limited production capacity will be allocated among competing orders and customers is made. These decisions involve trade-offs whose effective resolution requires an understanding of the strategic and tactical goals of the company as well as negotiation among various functional groups within the company. For example, the manufacturing organization is likely to prefer an MPS that allows them to use equipment efficiently by having long production runs with few setup changes. On the other hand, sales and marketing are likely to push for an MPS that emphasizes delivery to key customers, as well as perhaps to customers who are getting ready to take their business elsewhere. Hence effective, thoughtful procedures for developing an MPS are critical to the company's long-term performance. However, in practice we find that in many cases companies develop their MPS using the intuition and knowledge of a few key employees and simple spreadsheet-based tools for data management. There is currently far more art than science to the development of an MPS, a situation that renders this area attractive for future research.

2.2. Scheduling

While the production plan lays out what mix and quantity of products the company is expected to produce over a certain time horizon, the schedule describes the detailed execution of this plan, giving a step-by-step work list in the form of a dispatch list or a specification of the times at which every operation should start and end. In contrast to the production plan, whose focus is on independent demand for end items, the scheduling process is internally focused, driven by the need to ensure that all the components, subassemblies, and assemblies needed to produce the end items are completed according to the plan as efficiently (in terms of resource usage) as possible. Inputs to the scheduling process are product definition information (e.g., routes), facility information (e.g., workcenter availability), and shop-floor status. Shop-floor status defines the current state of production, identifying what orders are still open, their current locations (e.g., what machines are working on them), and the yield or scrap rates for each of the operations associated with these orders. Depending on the capabilities of the particular scheduling algorithm used, the process can also address work order resequencing for more efficient workcenter processing and improved plant throughput. The process is controlled by lower-level attributes, such as manufacturing order due dates and work order selection logic at a workcenter. It also takes a more detailed view of the production process, working with individual resources (e.g., machines) vs. workcenters, operational level routes for parts vs. fixed lead times, shift times vs. planned hours, and often even continuous time vs. discrete time buckets. The time horizon is typically short, usually on the order of a shift, a day, or a week. The usefulness of the schedule generally decreases rapidly in the future since responses to shop floor disruptions and changes to the production plan will result in significant revisions.

In summary, planning focuses on allocating production resources and material to various end products or customers, while scheduling focuses on how to meet component level deliveries without sacrificing efficiency. In general, one of these functions tends to dominate, in the sense that its decisions are considered to be more important to satisfy and the other is forced to adhere to them. Many of today's APS systems have their roots in the attempt to remedy the deficiencies of traditional planning systems by integrating planning and scheduling more closely.

3. RELATIONSHIPS BETWEEN PLANNING AND SCHEDULING

It should by now be apparent that planning and scheduling are tightly intertwined. Planning starts with high-level demand and produces a "schedulable" plan. Scheduling then generates a time-sequenced allocation of individual resources to tasks over time that efficiently supports this plan. Closing the loop, planning then honors these allocations as it replans. In other words, planning affects scheduling and vice versa.

Effective planning leads to effective scheduling. No amount of clever scheduling can overcome the effects of a poor plan. Another way of saying this is that if a plan calls for the plant to build the wrong things at the wrong time, efficiently executing this plan may still leave the plant in considerable trouble. If the plan is so tightly determined that it does not allow the scheduling procedure to adapt work order (job) sequences to the detailed needs of the shop floor, such as sequence-dependent setup times, then the task of scheduling is obviated and processing efficiency is likely reduced. At the other extreme, a plan that fails to constrain the scheduling task properly risks a decline in on-time performance in the name of efficiency. Planning needs to provide the shop floor enough direction to maintain overall order performance without unduly limiting opportunities for efficient workcenter processing. Recall that planning is where we prioritize orders and customers; the scheduling function does not usually have access to the right information to make these decisions in the company's best interest. However, it is unfortunately quite common to see these decisions being made by scheduling personnel due to dysfunctionalities in the planning system such as inaccurate data on workcenter capabilities.

Just as proper planning leads to good scheduling, proper scheduling leads to good planning. Efficiently rearranging the work on the floor allows production gains to be made that translate into more available capacity in which to plan the work. This also offers the added benefit of closer promise dates. Conversely, poor scheduling can undo the work of a good plan when the inefficiencies incurred on the floor cause exceptions to the plan. These exceptions are often addressed by ad hoc remedies on the shop floor based on local objectives such as machine utilization, resulting in a reduction in global performance.

In spite of this close relationship between planning and scheduling, inherent differences exist between them. These manifest themselves in many ways, including their function, their treatment of capacity, and their representation of the manufacturing process. An interesting discussion of the relation between planning and scheduling can be found in Pritsker and Snyder (1997).

3.1. Function

In manufacturing there is a basic dualism at play between synchronization and sequencing. Synchronization is the process of prioritizing independent demand from customer orders and demand forecasts and deriving all dependent demand for the necessary components, materials, and subassemblies accordingly. Workcenter reservations, allocation of work-in-process to specific orders (pegging), and purchase requisitions are made in support of this coordinated plan. This is an order-driven process and is the primary function of planning. Sequencing, on the other hand, which is the primary function of scheduling, is workcenter driven in that it locally ranks demand on a time-phased basis and projects order completions accordingly.

Under APS, planning, with its synchronization emphasis, dominates and sequencing (or scheduling) is subservient to it. The operations to manufacture a part are initially placed in the plan to support a global criterion, say customer order due date, and then the sequencing of these operations is done to adhere to the plan as closely as possible.

3.2. Treatment of Capacity

Since the basic problem of production planning is that of allocating manufacturing capacity among various customers and products over time, it would seem natural to assume that such a basic notion as capacity should be well understood. However, once one tries to get specific, it turns out that capacity is a remarkably elusive concept. Elmaghraby (1989) gives an insightful discussion of this issue. In this section we first discuss some basic aspects of capacity and its relation to the outputs of the planning process, such as resource utilization, batch sizes, and lead times.

Despite the difficulty of achieving a rigorous definition of capacity, it is widely accepted that the ability to produce a given set of orders on a given set of equipment by specified due dates is affected by product mix, shop-floor scheduling decisions, and the stochastic nature of events on the shop floor. A basic driver of shop-floor dynamics is the phenomenon of congestion, in which the rate of response of the manufacturing system degrades as more work is introduced into the system, even when the demand rate for the system is well below its nominal capacity. This is due to variations in the arrival rates of jobs at workcenters causing short-term saturation, where the workcenter is unable to process all jobs arriving within a short period of time and queues form. These variations arise from the myriad random events that determine the outcome of most manufacturing processes, such as inherent variability in processing and setup times, machine failures, and quality problems.

A fundamental relationship that holds in a broad range of environments states that the average time to process an order at a workstation is a highly nonlinear function of the workload in the system and that as the workload approaches a nominal capacity both the mean and variance of the lead time increase exponentially, as illustrated in Figure 1. A second fundamental relationship is Little's law (Hopp and Spearman 1996), which states that the average work-in-process inventory (WIP) level and the average time to process a job through the system (i.e., lead time or cycle time) are directly proportional.

The most important aspect of congestion for planning purposes is that a plan that is feasible with respect to manufacturing capacity when aggregated over a specified time period may not be capacity feasible at all times during that interval. To see this, consider a situation where the planning time period is a day and we have a single machine that is available for eight hours. In this sense, placing four orders, each of which requires two hours of processing, on this machine on this day is perfectly feasible. However, it may well be that all four jobs arrive at this workcenter at 10 a.m. This causes significant short-term congestion, which renders it impossible to complete all orders in the time period assigned.

Another area in which this shop-floor dynamic is becoming better understood is that of the effect of batch sizes, which has been studied by Karmarkar (1987). This work shows that initially, increasing batch sizes lead to rapid reductions in flow times as excessive setups are eliminated, essentially increasing the capacity of the system. However, as the batch size continues to increase, the flow time begins to increase linearly with batch size due to the additional queueing and processing time incurred by the larger batch sizes.

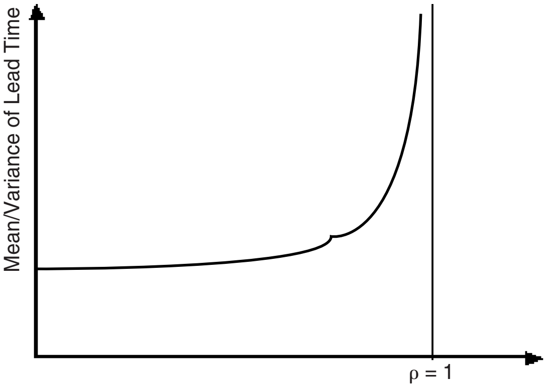


Figure 1 Relationship between Utilization (ρ) and Mean and Variance of Lead Time.

The important aspect of these relationships with regard to planning and scheduling is that both the workload of a given workcenter in a given time and the batch size used in a manufacturing facility are often outputs of the planning process. Hence, the planning process defines to a large extent the basic performance we can realistically expect from the shop floor. The specific scheduling algorithm we use will clearly affect this, at least in the sense that clever job sequencing will make the best possible use of available capacity. However, the high-level response of the system, in the sense of performance measures such as the average lead time, are defined to a large degree by the planning process and cannot be fundamentally altered by scheduling.

Another difference between planning and scheduling stems from their different views of how much capacity should be committed. In planning, it is often desirable to leave some capacity idle to allow time for the workcenters to handle contingencies that may arise, such as failures or rush orders. This is especially true for bottleneck or near-bottleneck workcenters that have demonstrated unreliable performance or whose cycle times vary widely. In planning for these workcenters, it is often advantageous to load the workcenter to a level below that of its nominal availability, yielding a more sparsely populated plan that can tolerate a certain degree of disruption before overall order performance begins to degrade. The extra capacity made available by this undercapacity approach to planning allows the scheduling algorithm, which tries to use the workcenter's full capacity, an opportunity to adjust to the disruptions encountered on the floor. However, while the advantages of this undercapacity planning approach may be significant, too much idle capacity in the plan may well lead to the company making inefficient use of its capacity, quoting overly conservative due dates and hence losing competitive advantage.

3.3. Representation

The advent of the computer and the powerful information technology tools it has brought with it have created many new possibilities for the planning and scheduling process. One area where this is evident is in the level of detail at which the manufacturing process is represented for planning and scheduling. Many more factors can be considered simultaneously with a computerized approach than with a manual one.

Yet even with this ability, planning and scheduling, in practice, still differ in how they represent the manufacturing process. Planning, which coarsely sets the boundaries within which to schedule, can take liberties in its representation. Average setup times, for example, are usually sufficient when generating a plan. The intricate logic that can accompany a setup time calculation is typically superfluous when the objective is to develop a synchronous, as opposed to an executable, plan. Appropriate decisions as to which resources must have their capacity explicitly modeled and which others are nonconstraining and can be treated as having infinite capacity can often improve execution speed without significantly compromising the quality of the plans generated.

This is not to say that detail is not appropriate at the planning level. When capacity is a major determinant of performance, incompatibilities can be introduced as a result of relaxing too many constraints. Consideration of tooling and overlapped operations, for example, may be necessary during planning to obtain a realistic picture of workcenter capacity and provide appropriate goals to the scheduling function. In some industries, such as integrated steel mills, the dynamics of the shop floor affect the capacity of the shop to such a degree that in order to be viable a plan must consider detailed shop-floor dynamics. Again, a well-constructed plan gives scheduling the opportunity to refine the sequence of operations to improve workcenter efficiencies without sacrificing, and hopefully enhancing, global performance to plan. If the plan is wildly inconsistent with the realities of the shop floor, then scheduling may be forced to make radical changes, which in turn cause major adjustments to the plan. This can be viewed as the scheduling function usurping some of the functionality of a dysfunctional planning system, which it may well not have sufficient information to perform adequately.

4. PLANNING ALGORITHMS

We can classify planning algorithms on two basic characteristics. The first of these is how manufacturing capacity is modeled. This yields two basic classes of algorithms: those that consider capacity within a given time period to be unconstraining, that is, essentially infinite, and those that recognize some limitation on the amount of available capacity. We will refer to these two classes of algorithms as infinite capacity and finite capacity, respectively.

The second classification concerns how the algorithm models the timing of production events. Some algorithms do not consider the issue of congestion at all, essentially assuming that the time for a task to be processed at a workcenter is independent of its workload, that is, is a property of the product being manufactured. Note that this does not necessarily imply that the underlying model assumes infinite capacity, only that it is incapable of modeling the congestion effects discussed above. A second set of algorithms considers congestion effects explicitly. We will refer to these two classes of algorithms as noncongested and congested, respectively. While noncongested algorithms work well when the production system is at low utilization, the ability of this model to predict job completion times accurately deteriorates rapidly as system utilization increases.

We will begin with a discussion of material requirements planning, which is the most widely used infinite capacity algorithm today. We will then introduce a variety of finite capacity algorithms, beginning with enhancements to the basic MRP algorithm and continuing through fundamentally different approaches such as optimization and artificial intelligence techniques.

4.1. Infinite Capacity Algorithms: Material Requirements Planning

Material requirements planning (MRP) was developed in the 1960s to apply the computational power of computers to production and inventory management problems. The basic paradigm adopted was that manufacturing was fundamentally a problem of coordinating the complex material flows involved in producing large, assembled products with deep bills of material. This basic logic, with a number of extensions and additions, is the common ancestor of most of today's enterprise resource planning (ERP) systems (Ptak and Schragenheim 1999). As such, MRP is prevalent in industry and continues to be the basic planning mechanism for many manufacturing companies as well as the driver for a multibillion-dollar software and consulting industry.

We will first briefly describe the basic MRP procedure and then discuss its inherent strengths and weaknesses. Details of this procedure and its many variants and enhancements can be found in texts such as Vollmann et al. (1988) and Nahmias (1993). Efforts to remedy these deficiencies will then lead us to a discussion of several alternative approaches to production planning that form the kernel of several of today's most successful APS systems.

The MRP planning procedure has three main inputs:

1. *The master production schedule (MPS)* or some other planning document that specifies how much of each end product is required in each time period t, over some specified planning horizon involving T periods. In most practical applications, the basic time period is a week, although longer periods of a month or so may be used for periods far in the future where there is more uncertainty in the demand process (e.g., the MPS is based more on forecasts than on firm customer orders). The relationship of the MPS to the production plan was discussed in Section 2.
2. *The bill of material (BOM)*, which specifies the structure of each product in terms of the components, subassemblies, and assemblies that constitute it. This structure is usually represented graphically as a tree whose root node represents the complete product, leaf nodes purchased components or raw materials, and intermediate nodes subassemblies and assemblies. For the sake of brevity we shall refer to these items collectively as modules. An example of such a BOM tree is shown in Figure 2. For each node (i.e., module) the BOM specifies which items are combined to make it and the quantity of each such item. This formalization allows MRP to draw a distinction between two types of demand. Independent demand originates with customers outside the company and is typically for end items. Dependent demand, on the other hand, refers to the demand for the BOM items used to produce the end item. Once the BOM and the independent demand are given, the dependent demand can easily be computed—if we know we need to build four cars, we know we will need four engines, four back seats, and 16 wheels, for instance. This distinction is of great importance since it allows us to focus on forecasting and managing the independent demand with confidence that the dependent demand can be generated easily if the BOM and estimates of independent demand are correct.

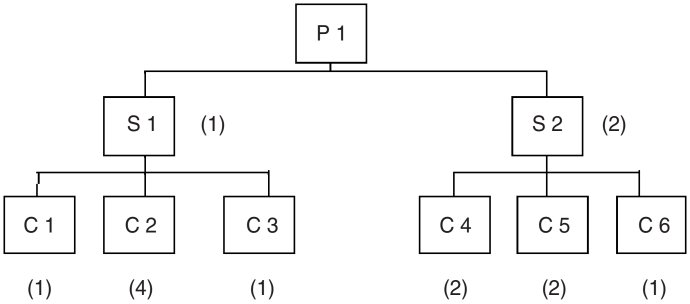


Figure 2 Example BOM Tree.

3. *Inventory status:* This is usually a database specifying the amount of each module on hand or on order. For parts that are on order, it will usually specify when the parts are expected to arrive.

The MRP algorithm combines these three inputs to generate planned order releases for all items in the BOMs of products that occur in the MPS. This will specify how many of each BOM item are needed in each time period covered by the MPS. We will assume that the nodes of the BOM have been indexed such that no node with a lower index occurs at a lower level of the tree than a node with a higher index. Well-defined algorithms to generate this type of indexing, known as level coding, exist. The MRP algorithm can be stated as follows:

For each independent demand item in each time period t, perform the following:

- *Step 1 (initialization):* Set $i = 0$, where node 0 denotes the root node of the BOM tree. Let d_{it} denote the number of item i required in period t.
- *Step 2 (time-phased order releases):* For each descendant j of node i in the BOM tree, perform the following:
  - Calculate the gross requirement $g_{jt} = a_{ij}d_{it}$, where a_{ij} is the number of units of module j required to form module i.
  - Calculate the net requirements for j as $n_{jt} = g_{jt} - I_{it}$, where I_{it} denotes the projected inventory of module j at the beginning of period t. I_{jt} is actually made up of inventory carried over from the last period and items on order expected to arrive in that period. For the sake of brevity, details of how the inventory status is updated can be found in any of several books on this subject (e.g., Nahmias 1993; Vollmann et al. 1988).
  - Generate the planned order release $p_{j,t-L_j} = n_{jt}$, where L_j denotes the number of periods after an order is placed that will elapse before the order is filled. We assume that the L_j are known constants that depend only on the item being ordered. Note that L_j may represent either the production lead time for a part manufactured in house or the supplier lead time for a part or material ordered from a supplier.
  - Once the total quantity of each module to be started in each period has been determined, we can combine the requirements for a number of periods into a release for a single period. This procedure, known as lot sizing, can serve to reduce the amount of setup time required for production changeovers, or may be dictated by process concerns.
- *Step 3 (stopping criterion):* Mark node i as expanded. Select the unexpanded node in the BOM tree with the lowest-level code and go to step 2. If no unexpanded nodes exist, stop.

The basic idea of the above procedure is almost absurdly simple: starting with the requirements for the end product for a given period of the MPS, subtract the estimated amount in inventory in that period to determine how many we actually need to make (the net requirements). Net requirements at one level of the tree become the gross requirements for the next level down. Once we have the net requirements for a given module in a given period, we then schedule the planned order release L_j periods earlier, such that the material will be where it is needed at exactly the right time. This is referred to as backward planning, where the order release date is obtained by working backward from the date the material or product is required and subtracting an estimate of the time needed to complete it.

To illustrate the operation of this algorithm, consider the product whose BOM is given in Figure 2, and let inventory availability and lead times be as listed in Table 1. In addition, we expect 400 units of P1 to become available in week 6, 100 units of S2 in week 4, and 400 units of C4 in week 5.

Based on this information, we can show the results of the MRP calculations in Table 2. Considering the end item P1, we will need 400 units in period 6, but we currently have none available and none expected to become available. Hence, our net requirement for period 6 is 400 units. Since the

TABLE 1 Inventory Status and Lead Times for Example

Item	Lead Time	On Hand
P1	1	–
S2	1	10
C4	2	50

TABLE 2 MRP Calculations for Example

Product P1	0	1	2	3	4	5	6
Gross requirements							400
On hand							
Scheduled receipts							
Net requirements							400
Order releases						400	
Assembly S2	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Gross requirements						800	
On hand	10	10	10	10	10	110	
Scheduled receipts					100		
Net requirements						690	
Order releases					690		
Component C4	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Gross requirements					1380		
On hand	50	50	50	50	50		
Scheduled receipts					400		
New requirements					930		
Order releases			930				

lead time is one period, this means that the order must be released at the beginning of period 5 for the end items to be ready when needed at the end of period 6. Since each unit of P1 requires two units of S2, this release causes a gross requirement of 800 units for S2 in period 5. Netting out the 110 units on hand at the beginning of the period, this yields a net requirement for S2 of 690 units, released as an order in period 4. Since, again, each unit of S2 requires 2 units of C4, we obtain a gross requirement of 1380 units of C4 in period 4, resulting in an order for 930 units placed in period 2.

It is interesting to examine the MRP algorithm from a number of different perspectives. First of all, its basic goal appears to be to achieve just-in-time material flow—if the lead times L_j are accurate, material will arrive exactly in the time period in which it will be used. From an optimization standpoint, it attempts to minimize deviation from the MPS (in terms of items completed after their request date) subject to the constraints of the lead times, the BOM structure, and the initial inventory status. Clearly, if the MPS is unrealistic, the backward scheduling procedure in step 2 may indicate a need to release an order in a time period that is already in the past. In this situation, the basic MRP logic offers no help—it is up to the user to revise the MPS to achieve a feasible situation. Many commercial systems provide the option to plan problematic orders forward in time, assuming that required modules at the lowest level of the BOM are started immediately and using the lead times to work forwards to an earliest achievable completion time. In this context, however, it should be noted that neither the backward planning algorithm used by MRP nor the forward algorithm used to remedy deficiencies are rigorously correct. Hence, there may exist a pattern of releases that renders the MPS feasible even though the MRP calculations show them to be otherwise. Moreover, we cannot be certain that if the MRP system claims a plan to be feasible it will indeed turn out to be so when executed on the shop floor.

The basic MRP logic, although widely used and still actively promoted in industry, has a number of fundamental flaws that seriously limit its usefulness as a planning tool. The most important is the treatment of manufacturing capacity. The only reflection of manufacturing capacity in the MRP algorithm is the lead times L_j. These are viewed as being known constants that are an attribute of the module j only. Specifically, they are assumed to be independent of the product mix in the shop and the loading or utilization of the shop at the time the order is released. As we discussed in Section 3.2, the lead time is a function of how the shop is loaded relative to its capacity at the point in time the order is released. Hence, while the fixed lead time assumption may be reasonably accurate for shops at low levels of utilization, as utilization increases, congestion will become more significant. Both the mean and the variance of the time in system at heavily loaded workcenters will increase, rendering the fixed lead times an increasingly inaccurate representation of the actual situation. Inaccurate lead times, in turn, will result in the release of orders to the shop that the factory will not be able to complete on time.

In many environments where the shop is heavily loaded and cannot achieve the lead times quoted in the MRP system, manufacturing will often try to make the case that the lead times should be extended to allow them more time to get work through the plant. However, when the lead times are extended, we are essentially releasing orders to the shop earlier than we used to, and hence we actually increase the number of orders on the floor. Congestion increases, the actual lead times again increase, and manufacturing goes back to planning to ask for another lead time extension.

Another disadvantage of the MRP algorithm is that when a capacity infeasible plan is recognized, it does not offer any help as to how to repair the problem. In most cases the user has two options—go with the infeasible plan and hope for the best (an option often adopted when the time available for plan revisions expires), or examine the MPS and try to move production requirements between periods so that the MRP algorithm can generate a feasible release plan. The latter is difficult for even an experienced planner to do, especially for a number of products with large, complex BOMs.

On the positive side, MRP correctly distinguishes between dependent and independent demand and provides a useful framework for requirements planning, the calculation of production requirements for dependent demand items given the BOM structure. MRP is also easy to understand, and its adoption forces companies to systematize a great deal of data about their operations, which is a beneficial exercise in and of itself, regardless of what planning algorithm the data are used in. For better or worse, MRP has become a de facto industry standard against which alternative approaches must be measured.

4.2. Finite Capacity Algorithms

The basic source of many of the problems identified with MRP lies in its fundamentally material-centric view of production planning. Other than the extremely indirect representation through the lead times, MRP does not try to represent manufacturing capacity at all. One would expect the infinite capacity approach to work well when capacity is plentiful and the main concern is to coordinate the flow of work through the factory. However, in environments where capacity is expensive and highly utilized, we would expect the infinite capacity model to be increasingly inaccurate in its predictions of job completion times. It did not take long for these problems to be noted in practice, and a number of extensions to the basic MRP procedure have been developed over the years to provide some form of capacity check on MRP calculations. A variety of *finite capacity* algorithms have evolved out of different attempts to address these deficiencies. These algorithms take various approaches to modeling capacity and to generating the actual production plan but are mostly noncongested. We shall present a broad overview of the basic approaches that have been developed in academia and industrial practice.

4.2.1. Extensions to MRP

The two best-known approaches to adding some capacity checks to the basic MRP calculations are rough-cut capacity planning and capacity requirements planning. Both of these approaches have essentially the same philosophy: to estimate the amount of capacity required at each workcenter in each time period and notify the user of any violations. It is up to the user to decide how to modify the MPS to obtain a capacity-feasible order release scheme. They differ in the amount of data required and the accuracy of the capacity profile generated.

Rough-cut capacity planning (RCCP) is intended to be performed on the MPS before the actual MRP run is made. In this approach, we associate a bill of resources with each item in the MPS. These data specify how much time on the various types of resources a given MPS item requires. The RCCP procedure then multiplies each entry in the bill of resources by the number of units required by the MPS in a given time period to estimate the total workload implied for each resource. Note that this calculation is performed before the MRP run, so the timing of planned order releases is unavailable. Hence, lead time information is not used in RCCP. By the same token, neither does it consider the amount of available inventory in estimating this workload. Hence, the accuracy of the workload predictions made by RCCP is often quite poor, although it may allow the user to identify gross capacity violations before the actual MRP run is made. Its chief virtue is that its data requirements are modest and the computations simple.

Capacity requirements planning (CRP) is performed after the MRP run and essentially converts the planned order releases into a capacity profile. This approach in its pristine form considers lead times for individual processing steps rather than for the production of the BOM item and thus requires a lot more data than RCCP. It is also substantially more time consuming to perform. However, being done after the MRP run, it considers both lead times and inventory status. Unfortunately, the arguments above against constant lead times also hold here, rendering the capacity profile generated increasingly inaccurate at high utilization levels.

Another approach to enhancing MRP is the capacitated MRP (MRP-C) approach proposed by Tardif and Spearman (1997). This approach has a number of advantages: it specifically pinpoints the reason for the infeasibility as being due to the current WIP distribution, which does not allow the work in process to be finished in time to meet demand, or lack of capacity at a specific point in time. The basic idea of this approach is that capacity is explicitly considered while doing the calculation itself. Hence, net requirements are calculated in a given time period and are checked against available capacity to determine whether they can actually be manufactured in that period. If they cannot be produced, then the current plan is infeasible and action must be taken to render it feasible, such as delaying demand or adding overtime. An interesting aspect of this approach is that it identifies infeasibilities as being due to either poor positioning of the WIP in the line or mismatches of demand and capacity in time. The authors show that if their algorithm generates a feasible plan, that plan will minimize the total inventory over the planning horizon. Similarly, they provide algorithms to delay demand or add overtime in a manner that minimizes inventory and lateness costs. This algorithm is of considerable interest in that it integrates the two steps of generating material requirements and checking capacity feasibility that are performed serially under conventional MRP approaches. The fact that it identifies the source of infeasibilities and gives guidance as to what to do about it is also important.

4.2.2. Optimization Approaches

There is a long history of optimization models for production planning, almost all of them based on linear programming. The capacity of each workcenter is recognized to be finite in each time period. The planning problem is then that of assigning orders to time periods, subject to a subset of the relevant resource and relational constraints. Those most commonly considered are workcenter capacity, that is, the number of hours available in each time period and the routing of the products through those workcenters. In most cases the lead times at the workcenters are assumed to be independent of workload. Hence these models are classified as finite capacity noncongested. The objective is usually to maximize some combination of revenue and costs.

A simple model of this form to illustrate the basic types of issues that can be modeled is given below. Let the decision variables x_{it} denote the amount of product i to be produced in period t. We will let I_{it} denote the amount of inventory on hand at the end of period t and S_{it} the amount of backlog at the end of period t. The product j produced in period t is assumed to become available τj periods later, that is, we assume a fixed lead time of τ_j for product j. The model will also determine the amounts of regular and overtime labor to be used, which will be denoted by LR_t and LO_t respectively. Costs are also incurred when we increase and decrease our labor force, and are proportional to the amount of the increase or decrease. Denoting the amount of labor force increase or decrease in period t by λ_t^+ and λ_t^-, respectively, we can then state a basic model as follows.

$$\min \sum_{t=1}^{T} \left[\sum_{i=1}^{n} (p_{it}x_{it} + h_{it}l_{it} + w_{it}S_{it}) + C_{Rt}LR_t + C_{ot}LO_t + c_{lt}\lambda_t^+ + c'_{lt}\lambda_t^- \right]$$

subject to

$$NI_{it} = NI_{i,t-1} + x_{i,t-\tau_1} - c^i_{it} \text{ for all } i, j, t$$

$$NI_{it} = I_{it} - S_{it} \text{ for all } i, t$$

$$LR_t = LR_{t-1} + \lambda_t^+ - \lambda_t^- \text{ for all } t$$

$$LO_t - LU_t = \sum_{i=1}^{n} m_i x_{it} - LR_t \text{ for all } t$$

All variables are assumed to be nonnegative. LU_t is a slack variable denoting the amount of excess labor available in period t. Note that LU_t and LO_t cannot both be positive in the same period. The first set of constraints ensures that the inventory levels are consistent across periods, where NI_{it} is the net inventory of product i in period t. The second set of constraints defines the net inventory to be either positive or negative, since both variables on the right-hand side cannot be positive in any optimal solution. The third set of constraints models the evolution of the labor level over time, while the fourth set indicates the relationship among overtime, undertime, and the amount of regular labor on hand.

While this particular model views the production facility as a single stage whose production capacity is limited by the available workforce, it is easy to extend this approach to situations with multiple products following different routings through multiple-stage production systems. This basic model has been extended in many directions, such as the inclusion of variables and constraints related to working capital, marketing, and promotion decisions (Shapiro et al. 1993).

This model considers capacity at an aggregate level, in the sense of recognizing that the number of hours available on a given resource in a given time period is limited, but the modeling of congestion is still inadequate. As in the MRP algorithm, we are assuming that lead times (the τ_j in the above model) are known a priori and independent of the loading of the shop in that time period. While this may well be an acceptable approximation in lightly utilized facilities, it rapidly degenerates as the level of resource utilization increases. A number of authors, notably Leachman and his coworkers (Hackman and Leachman 1989; Leachman 1993) have significantly extended the capability of linear programming models in this regard, allowing lead times to vary over time as the workload in time periods changes. Hung and Leachman (1996) use an iterative approach where the results of the linear program are fed into a simulation model of the production facility to obtain updated lead time estimates, which are then input into the linear program for another run.

The model as stated here is also more suitable for MPS generation in that it deals with end-item demand translated into demand for workcenter capacity. Optimization models that explicitly address the multilevel nature of the BOM have been proposed (e.g., Billington et al. 1983) but are generally mixed-integer programs that are significantly more complex than the model above.

A number of authors, such as Graves (1986) and Karmarkar (1989), have attempted to develop models of manufacturing capacity that reflect the effects of congestion using the idea of clearing functions. As the name implies, a clearing function represents the amount of inventory at a workcenter that can be moved out of it in a given period. While a number of different forms for such clearing functions have been suggested in the academic literature, they remain an unexplored possibility in terms of industrial implementation. A particular area of research is how to obtain such clearing functions empirically for specific industrial scenarios.

4.2.3. Artificial Intelligence Approaches

Another approach that appears to be used in some current planning systems is to model the capacity of at least a subset of near-bottleneck workstations in an aggregate manner, as a total number of available hours per planning period. A plan is constructed by calculating the amount of time each operation of the order will require on each workcenter and assigning each operation to a specific planning period. The lead time between nonbottleneck operations is modeled as infinite-capacity, the idea being that these stations have such low utilization that the infinite capacity model is a reasonable approximation. These procedures tend to use quite sophisticated heuristic search procedures to assign orders to periods. Examples of such algorithms are the ReDS system (Hadavi et al. 1989) and the system developed at Texas Instruments described by Fargher and Smith (1994). Smith (1993) presents an excellent review of artificial intelligence approaches to production planning and scheduling problems, and Zweben and Fox (1994) provide a number of case studies.

4.2.4. Congestion Models

All the approaches considered so far have been based on an aggregate model of manufacturing capacity, in the sense that capacity constraints are checked only at an aggregate level over a specified time period. This clearly does not prevent infeasibilities due to mismatches between the arrival and completion of tasks at the workcenters. A number of procedures have been developed to address this problem by making a detailed scheduling model an integral part of the planning algorithm. These models address the congestion effect directly by modeling the operation of the shop in considerable detail. We shall discuss two basic flavors of this approach, asking the reader to bear in mind that in many implementations the distinction between them may become blurred, with any particular procedure exhibiting characteristics of both groups.

4.2.5. Detailed Scheduling as Part of the Planning Process

As discussed above, a major difficulty in most finite-capacity production planning models is that of modeling the nonlinear effects of congestion at the shop floor, caused by variability in the arrival patterns of jobs to the workcenters. A number of planning systems in both the research community and industry have attempted to address this by making the generation of a detailed schedule of shop-floor operations a part of the planning process. The basic idea is to generate a plan using some planning procedure and then try to construct a detailed schedule that meets the deadlines given by the plan. If such a schedule can be generated successfully, it shows that the plan is capacity feasible. In general, one would not expect the schedule generated during the planning process actually to be executed, as both plan and schedule will be revised. Rather, the function of the schedule is to verify that the proposed plan is indeed capacity feasible in the sense that at least one schedule that satisfies its demands can be constructed.

This approach has been illustrated in a research environment by Dauzère-Pérès and Lasserre (1994), who use an optimization algorithm for production planning. Once a plan has been developed, they use a sophisticated optimization-based scheduling heuristic to build a schedule that meets the deadlines set by the plan. In the event that the deadlines cannot be met, they revise the plan and iterate until a satisfactory solution is reached.

While intuitively attractive, this approach has a number of problems. While the idea of using a scheduling algorithm to verify the capacity feasibility of a proposed plan is attractive, the correctness of the conclusion depends on the quality of the scheduling algorithm used. A simple dispatching heuristic may well be unable to find a schedule that satisfies the plan that a more sophisticated exact solution procedure could identify. Moreover, the essentially discrete nature of the scheduling problem and its well-known computational intractability render the task of generating high-quality schedules very time consuming. This is an important drawback in an environment where the plan must be revised and schedules generated repeatedly before a satisfactory solution is reached.

A common approach widely discussed in industry is to feed the output of an infinite capacity planning system into a detailed discrete-event simulation model of the plant and use the simulation to determine whether the proposed start times will allow all orders to be completed by their due date. If some orders are identified as being late, the plan is modified and the simulation rerun until an acceptable result is achieved. While conceptually attractive, this approach has a number of difficulties. First of all, the time and effort involved in developing and maintaining a detailed simulation model of the facility may be significant. Secondly, the time required to obtain results from the simulation, especially if several replications must be run to obtain statistically valid results, may be quite substantial. Thirdly, if the simulation identifies an infeasibility, the user is often reduced to manual intervention and experimentation to identify a feasible plan, which may be difficult to achieve in the time available for the decision to be made. Finally, this approach requires some fairly detailed assumptions as to how the shop does the scheduling—in effect, the simulation model must generate the schedule as executed on the shop floor to ensure the capacity feasibility of the plan driving the schedule. All in all, this approach may work well in relatively simple manufacturing environments, but is unlikely to be practical in complex multistage environments where the time to make multiple simulation runs is substantial and the effects of changes in the schedules are hard to predict. However, despite the drawbacks of this basic approach, the APS procedures described in the following section are often quite successful using variants of this approach that use scheduling ideas to maintain records of how much capacity is available at each critical resource in a given time period and use this information to drive planning procedures.

5. ADVANCED PLANNING AND SCHEDULING

APS is the process of simultaneously coordinating material and capacity constraints at the operational level to best meet market demand. APS offers the planning function a detailed representation of the production process that was formerly found only in more advanced scheduling systems. As for the scheduling function, it links work orders to customer orders, permitting direct tracking of their progress. Arguably its most important advantage over traditional planning approaches is that material and capacity are *simultaneously* considered as elements that may constrain production. This ensures that the material plan, as it is being generated, is in agreement with the capacity schedule down to the level of individual resources such as machines, and the capacity schedule is in agreement with the material plan throughout all BOMs. This stands in marked contrast to the conventional MRP approach of independently planning material and then subsequently checking this plan against capacity to identify violations.

The computational complexity of this task is quite daunting—to achieve this, one needs to consider detailed scheduling information as well as customer information, essentially creating a detailed schedule in parallel with a production plan to ensure that the plan is indeed achievable. The need to do this for all levels of the potentially complex BOM adds to the difficulty. The computational complexity of scheduling problems on their own is well recognized (Pinedo 1995). Hence, many APS systems do not try to develop plans that are optimal in a rigorous sense. Instead, they are focused on transaction processing, determining at the time the order is placed whether the order can be completed on time or what the earliest possible completion date is if the original request date cannot be met. A complete resynchronization of the plan is then done periodically to optimize resource usage and material plans.

APS systems also differ in how they approach the generation of plans and schedules. In some systems, planning is done periodically in batch mode, with integrated plans being developed for a set of products and workcenters. On the other hand, others are focused on transaction processing, where a plan is constructed and capacity allocated incrementally as orders arrive.

In its pristine form, APS integrates three key processes: advanced planning, advanced scheduling, and order promising.

5.1. Advanced Planning

The goal of advanced planning is the synchronization of constrained material and resources to independent demand. Its purpose is to create a plan that is feasible with respect to all resources required (machines, material, tooling etc.) with sufficient operational slack to permit resequencing of work orders to enhance production efficiency. The independent demand comes from several sources, including customer orders, demand forecasts, master production schedules, transfer orders (i.e., orders from other plants), and the company's policies on safety stock. The advanced planner also considers the work order schedule already released to the shop floor as presented by the advanced scheduler. For each end-item demand, a complete requirements explosion is done using that item's BOM and backward scheduling dependent demand based on component routes, available resource capacity (not simply workcenter capacity), and available and projected inventory. Some advanced planning engines are refined enough to consider resource and material requests at their point-of-use, which can be helpful in environments where long operation times create extended delays for material usage. Typically, the APS horizon is similar to that of conventional planning tools, ranging from several weeks to months.

5.2. Advanced Scheduling

The second APS component, advanced scheduling, involves the detailed sequencing of operations and material in support of the aforementioned plan. Its purpose is to provide properly sequenced work orders, under possibly more refined constraints than those considered in the plan (e.g., sequence-dependent setups, maintenance schedules, more detailed machining constraints, additional operator restrictions) while still attempting to hold to the plan dates. It serves to efficiently load the workcenters and present a more discriminating schedule to the advanced planner. While advanced planning produces a detailed allocation of resources and material to orders, some applications require further refinement, especially where work sequencing can significantly affect workcenter production rates. Advanced scheduling produces a schedule constrained by both material and capacity. This schedule serves as a projection of what the shop floor should be doing and is used in the advanced planner as a basis for component supply. Exceptions to the advanced plan are identified for resolution in the advanced scheduler or for adjustments to the advanced plan. The advanced scheduling horizon tends to be short, as with conventional scheduling tools, but may need to be extended to support better the needs of the advanced planning process.

The reader should keep in mind that one may not need to run an advanced schedule to produce what the shop floor needs for execution. The main difference between the advanced plan and the advanced schedule is not necessarily the detail used in the representation, but rather the amount of emphasis placed on sequencing. If it is enough to determine what needs to be produced over a given time frame (e.g., a shift) without regard to sequence, then the advanced plan may be sufficient. If, on the other hand, the sequence in which this work is done can significantly affect the workcenter's production rate, an advanced schedule with a more refined sequenced execution list is required. This type of situation often arises in manufacturing systems where the decisions at various stages of production are tightly coupled and setup times between products are significant.

5.3. Order Promising

The third component of APS is order promising, which lies at the center of the transaction-based aspect of APS systems. This component is designed to suggest realistic promise dates for customer orders. The process, sometimes referred to as capable-to-promise (CTP), involves testing the customer's request date for feasibility and, if the date cannot be met, calculating the earliest date that it can be met. This is done based on available and projected inventory and available resource capacity.

CTP functions at two levels: disruptive and nondisruptive. Nondisruptive promising uses available capacity and material to determine the order's projected completion date without altering the planned completion times of orders currently in the system. Disruptive promising, on the other hand, reallocates capacity and material to determine the feasibility of meeting a particular date (presumably earlier than is possible under nondisruptive promising) and identifies what orders are affected as a result. In either case, CTP differs significantly from conventional available to promise (ATP), which considers only the uncommitted inventory balance by period. With CTP, the process extends through the bill of materials and part routings to examine the potential for manufacturing the item if it is not available.

The transaction focus of APS systems renders their speed of execution critical to their effectiveness. Taking advantage of the many advances in hardware and software technology over the last decade, APS systems have execution speeds orders of magnitude faster than those of traditional MRP. Most APS engines have their data downloaded, either in batch or transaction mode, to a dedicated server that is architected to run memory-resident programs and databases. Under this scheme, they are able to deliver real-time order promising and make multiple runs to test various actions in an effort to further improve the plan.

6. APS IMPLEMENTATION ISSUES

Compared to MRP alone, APS has a broader functional impact, is less tolerant of inaccurate and incomplete information, requires more data and in greater detail, and affects more people more directly every day. Beyond this, it causes a cultural change. These factors make APS implementations more challenging than MRP. Nevertheless, remaining cognizant of the value received will enable the technical and organizational hurdles that will inevitably appear throughout the course of the APS implementation to be overcome. Areas in which to be particularly vigilant are described below.

6.1. ERP Integration

The comprehensive nature of today's APS algorithms drives the need for copious amounts of data—data that typically reside in an ERP system. This means that attaining the full benefits of APS is largely predicated on how well it is integrated with ERP. When done well, both systems benefit.

The importance of the ERP system is seen in two key roles it plays. First, ERP provides the necessary infrastructure, holding and managing information about orders, parts, resources, and status. It houses the modules that feed and are fed by the APS system, including forecasting, customer order management, product definition, inventory management, purchase order management, work order reporting, dispatching, and costing. When APS is planning and scheduling, it needs to know what orders to consider, what jobs are finished, what work is in process, what purchased materials are coming in and when, and what capacity is available. Secondly, ERP functions as an execution system, firming and releasing manufacturing orders, cutting purchase orders, and communicating schedules. The orders to be considered are the purview of APS; the actions themselves are the purview of ERP.

This need for ERP integration does, however, pose a number of challenges. First of all, the full benefits of APS can only be realized through a two-way exchange. Closed-loop functionality gives APS the data it needs to plan and schedule and gives ERP the information it needs to take appropriate action. The amount of data transferred and the numerous ERP modules affected require a complex web of communications, as illustrated in Figure 3. Another challenge is the need to disable specific functionality. Some traditional ERP modules, being superseded by APS, are no longer relevant and require circumvention. This must be done with ERP vendor expertise, for these modules often contain utilities that are still germane to the process. This can raise issues of functional ownership.

Finally, determining the type of APS system to be adopted, stand-alone or preintegrated, also deserves some thought. Stand-alone systems, using straightforward handshakes through specially written programming utilities, are less intrusive, require less vendor involvement whenever customized changes are required, and offer well-targeted value. Preintegrated systems, on the other hand, provide a broader range of functionality and are in proven agreement with the ERP system (and in some cases actually share the same database). They may also offer a better upgrade path as improvements in both systems are made. Preintegrated systems also lower the cost and time to implement, thus improving the return on investment.

6.2. Timing, Access, and Quality of Data

6.2.1. Timing

In general, communications between ERP and APS are asynchronous. This means that data are sent to the associated system and no immediate response is required. This is particularly true in systems where the APS engine is transaction oriented. In these systems, updated information is sent on a continual basis in anticipation of a new plan or schedule being generated. Then, at the time of execution, APS knows the current conditions without requiring a time-consuming download, as may be the case with batch-oriented systems. The timing of a full replan or reschedule is often dependent on when the shop floor has consistently reported across the facility (e.g., end of shift) or when a major disruption has occurred (e.g., machine failure). More frequent planning and scheduling is possible with today's technology but is seldom realized given the requirement for the reported information to be consistent.

In the case of order promising, the demands on this timing change. Here, when an order is entered, it triggers the APS engine for the expected completion date and an immediate response is expected. This is a synchronous process. Moreover, as new orders are entered, they need to be promised based on the latest information, including those orders that have just been accepted. Thus, the planning system, upon order acceptance, must immediately reserve materials and capacity so that all future promises can reflect the impact of even the most recently accepted orders.

6.2.2. Access

Planning and scheduling often differ in their data requirements. Capturing some of the subtleties of the production process may be pertinent to scheduling because it seeks to refine the sequence of events, whereas planning may not need to be as precise and thus may not require this same level of detail. In practice, the information required to portray these scheduling subtleties properly usually resides outside the ERP system. For example, traditional ERP systems typically define the manufacturing process at the workcenter level. The scheduling engine, on the other hand, may be forced to examine the detailed differences among the various machines within a workcenter. When more refined information such as this is required, it is imperative that the scheduling system have access to sources of information outside the traditional footprint of ERP or at least be able to circumvent, as necessary, the ERP data definition.

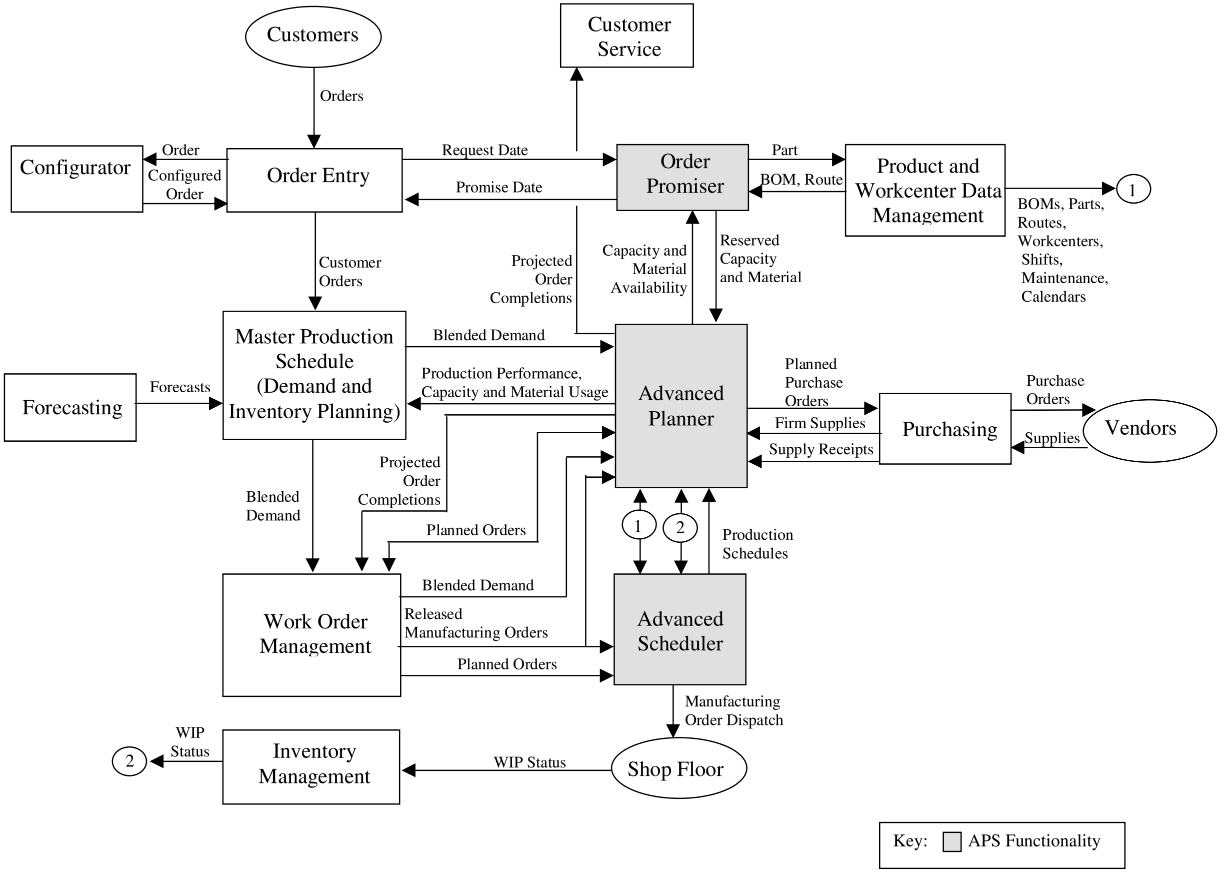

Customers
Customer Service
Orders
Configurator
Order
Configured Order
Order Entry
Request Date
Promise Date
Order Promiser
Part
BOM, Route
Product and Workcenter Data Management
1
BOMs, Parts, Routes, Workcenters, Shifts, Maintenance, Calendars
Customer Orders
Projected Order Completions
Capacity and Material Availability
Reserved Capacity and Material
Master Production Schedule (Demand and Inventory Planning)
Blended Demand
Production Performance, Capacity and Material Usage
Forecasting
Forecasts
Advanced Planner
Planned Purchase Orders
Firm Supplies
Supply Receipts
Purchasing
Purchase Orders
Supplies
Vendors
Projected Order Completions
Blended Demand
Planned Orders
1
2
Production Schedules
Blended Demand
Work Order Management
Released Manufacturing Orders
Planned Orders
Advanced Scheduler
Manufacturing Order Dispatch
WIP Status
2
Inventory Management
WIP Status
Shop Floor
Key: ▢ APS Functionality

Figure 3 APS/ERP Integration.

6.2.3. *Quality*

Data quality strongly influences APS results. This is especially true of shop-floor data, which are not always available at the level, at the time, or in the form most appropriate for planning and scheduling.

Work-in-process (WIP) accuracy is particularly problematic. The data-collection system may be, for example, recording only an order's closed operations. Time remaining on partially completed operations may not be known, thus forcing the APS engine to replan and reschedule the operation as if nothing has been done. The significance of this obviously depends on the operation's duration. Similarly, backflushing, which records inventory changes based on assumptive issuing at the end of the process vs. discrete issuing at time of use, can also pose a problem, for it fails to record an order's intermediate progress. If an order's route requires key workcenters for extended lengths of time, this method of capturing status can lead to an erroneous representation of important near-term capacity. Fortunately, advances in data-collection systems have made it easier and more cost effective to track WIP.

Reporting frequency can also adversely affect results. Work accomplished between reports is not recognized until reported. The significance of this misrepresentation increases as the reporting interval widens. More frequent reporting is desirable.

However, frequent reporting does not guarantee valid and consistent status information. False reporting and an inability to hit an order's exact quantity levels at an operation can also mislead the APS engine. As a result, some filtering and interpretation of the shop-floor information may be required. Moreover, where consistent information is a problem (e.g., a downstream station reporting work on an order for which an upstream station has yet to acknowledge), the frequency of the APS resynchronization is forced to match those points in time when the information is most likely to be in agreement across the facility, such as the end of a shift.

Startup conditions (i.e., order and machine status) form yet another critical data source for the planning and scheduling process. As the APS engine is being asked to provide nearer-term dispatch lists (e.g., the next eight hours), the status of resources and WIP at the beginning of the planning and scheduling horizon takes on added importance. This is particularly true in applications where sequence-dependent setups are consequential, for failing to capture a machine's initial state can dramatically alter the resulting sequence. A false start can have a lasting impact.

6.3. Business Process Reengineering

Business process reengineering (Hammer and Champy 1994) is the single biggest factor affecting the success of APS and is what ultimately governs the company's true return on investment. Technology alone cannot guarantee APS success. While APS affords a company the opportunity to improve its order fulfillment process dramatically, it can only happen if the business processes change to accommodate and exploit it. There needs to be a conceptual match between these business processes and the APS system. When the current processes are based on antiquated planning and scheduling practices, changes are in order.

Unfortunately, these changes are not always easily understood and take time to implement. The new planning and scheduling paradigm of APS cuts across organization boundaries and threatens the company's traditional, and now obsolete, mode of operation. The real challenge becomes changing the way people currently think and operate. Confidence is required to move forward. It takes a strong sense of need to implement the procedural changes required. Enlistment of top-level management may be necessary to overcome the pushback that occurs when people are asked to change how they work.

Consequently, the implementation team should take time, in the early stages of the project, to understand better the business processes that will need to be changed in sales, customer service, purchasing, engineering, production planning and scheduling, and manufacturing. Sales will be challenged with sobering promise dates and with reconciling, at the time of order entry, a customer's request with the realities of the production plan. Customer service will wrestle with realistic (and variable) customer order projections. Purchasing will be immediately impacted by sales, and conversely, sales by purchasing. Engineering will be directed to keep BOMs more up to date and more in line with how items are actually built and could be asked to support a configurator to enable order-entry clerks to take direct advantage of the order promise capability. Planning and scheduling will be challenged with having a more comprehensive and integrated view of the production process and with needing improved communications with various departments to better induce the necessary and frequent changes. Manufacturing will be asked to be more disciplined (i.e., follow the schedule and report status) and more flexible (i.e., heedful of new orders and changes to existing orders). In addition, the dynamic reallocation of work will be commonplace.

6.4. A Well-Defined Manufacturing System

Importance must also be placed on properly representing the manufacturing process in the APS engine. This means the manufacturing data need to be brought up to a level of accuracy commensurate with the APS task. As the manufacturing data deviate from this ideal, the effectiveness of the APS engine (or any other planning and scheduling tools) diminishes accordingly. The data that are most relevant to this process include:

6.4.1. As-Manufactured, Indented Bills of Material

If the item to be produced is not correctly defined, the APS engine is not able to plan and schedule it properly. The issue is usually not whether it is defined (unless in a configure-to-order environment), for most manufacturers have BOMs. Rather, the issue is *how* it is defined. Does an item's BOM reflect the sequence in which that item is to be manufactured, or does it simply represent a listing of what components are required or how the item was engineered? Do the bill's components mirror only those items of importance for planning and scheduling, or do they include a complete listing of all the items in the bill, regardless of their relevance?

6.4.2. Accurate Routes

An item's route directly controls when the requisite resources and materials used in its production are engaged. Obviously, correct identification of these resources and materials is important. This does not mean, however, that every resource or material used in the item's production needs to be identified; only if it supports the representation for purposes of planning and scheduling. Readily available material and secondary as well as tertiary resources at an operation are often not necessary. Furthermore, the operation times themselves are important, for they not only hold the resources for the correct apportionment of time, but they also influence when the other associated operations in the route request their resources and materials.

6.4.3. Supportive Operational Buffer Times

The APS plan may need to allow for the resequencing of work at an operation. When this is the case, operational or resource buffer times are used. This gives a certain degree of latitude to the advanced scheduler for work order resequencing. It is important to set these resequence buffers large enough to allow for some scheduling adjustment, but not so large as to misrepresent the item's lead time.

6.4.4. Consistent Workcenter Definitions

Often, workcenter definitions in ERP reflect a costing orientation more than a planning and scheduling orientation. This can substantially limit the options available to the APS engine. For example, instead of recognizing the more complete set of machines that can actually do a particular operation, the route may only identify a specific (e.g., least-cost) machine, thus misrepresenting the true scheduling options available and potentially artificially extending the time it should take to produce the item. When this is the case, circumvention of the ERP workcenter definition may be required.

6.4.5. Representative Purchasing Lead Times

Purchasing lead times are often inflated to trigger early action by purchasing. In a materially constrained APS application, this can cause overly conservative projections. Flagging an item for early purchase is different than constraining on it. Under a nonconstraining paradigm like MRP, cautious lead times are acceptable, for the purpose is to expose potential problems. In a constraining paradigm like APS, more aggressive (i.e., shorter) lead times generally present a more realistic picture of what is possible.

6.4.6. Appropriate Workcenter and Material Constraints

Bottleneck resources and critically scarce materials need to be modeled in a way that reflects their limited availability, for they are largely responsible for setting the manufacturing flow rate. Failure to represent these limitations properly can render the resulting plans and schedules useless. Nonbottleneck resources and noncritical materials, on the other hand, need not be represented to this same level of detail. This can save on execution time and obviate the need for their precise representation in the model.

6.4.7. Representative Scheduling Rules

Seldom does a manufacturer operate according to one rule (e.g., due date). Rather, each machine or workcenter has its own individualized set of rules (e.g., highest priority, then dynamic slack, and lastly minimum setup). Taken collectively, these rules define the company's manufacturing strategy. Correctly defining these rules and capturing them in the scheduling engine not only ensures valid schedules but also helps APS gain acceptance by those responsible for executing these schedules.

6.5. Usual Suspects

Some APS systems are particularly challenged to represent or work with certain aspects of a manufacturing process adequately. This is not to say it cannot be done, just that it may require additional thought, necessitate extensions to the software, or run counter to their planning and scheduling paradigm. Examples of these challenges include batching, outsourcing, nonsequential operations, flexibility, and early order-release strategies.

Batching, in a discrete process, refers to the combining of like orders for the purposes of processing them together through a portion of their routes. This often surfaces in applications where there are painting or heat-treat operations. The difficulty lies not only with synchronizing these orders appropriately but also with maintaining their unique identities (for these orders may be routed separately after their batched sequence has completed), establishing the batch criteria (e.g., item attributes, timing constraints), associated batching rules (e.g., number in a batch, sum of the batched items' attributes), and gaining access to the requisite information to form the batch.

Outsourcing refers to the practice of having certain operations (or parts) done outside the plant. From a planning perspective, this can be effectively captured. The appropriate delay is simply factored into the item's route. The problem arises when a particular operation or set of operations is done outside the plant before the item reenters for its final set of operations. Precise scheduling of these remaining operations is predicated on having accurate status information from the outside source.

Routing operations that can be done in any order are defined as nonsequential. To bring some structure to the process, an ERP system assumes a specific operational sequence. Since the APS system receives its routing information from the ERP system, the same implied sequence is applied in APS. To function otherwise would require additional information outside the traditional bounds of the ERP system. Nonsequential operations also give rise to significant combinatorial issues.

Most often, the issue of flexibility arises with operators. As more latitude is given to dynamically reassigning people on a line, the demands on the planning and scheduling system increase accordingly. The rules on when to engage a change in assignment and what to change to can be quite complex. Moreover, depending on the number of people assigned, the station times on the line may change. Capturing these dependencies in the planning and scheduling model can be an arduous task and should only be considered when commensurate value results.

Releasing work orders as late as possible is in the best interest of the manufacturing facility and APS. Later releases commit manufacturing at the last possible moment and give planning more flexibility to adjust to ever-changing conditions. Keeping the planning engine in control for as long as possible increases the ability to satisfy "drop-in" demand. Unfortunately, many ERP systems require the order to be released before it can be changed. This forces orders to be released early and prevents the planning engine from making last-minute adjustments, since manufacturing is now committed to these orders.

6.6. Implementation Strategies

APS implementation can be a daunting task. The sophistication of the software, the effort to get the necessary data to effectively drive the system, and the need to make fundamental procedural changes across the organization present significant challenges. Moreover, the people involved, their understanding of the business needs, and their willingness to change vary. This makes every APS implementation a unique endeavor. Aside from general project-management guidelines, little is offered to address the challenges that are faced.

Two vastly different APS implementation approaches have been advanced: big bang and evolutionary. With the big bang approach, the strategy is to leap to a complete implementation as quickly as possible. Emphasis is placed on time to value more than fidelity. The approach initially uses the data as it is currently defined rather than going through an extensive modification process to better support the requisite APS functionality. This makes it most appropriate in those environments where the manufacturing data are well defined. With this approach, reliance on the APS system occurs with deliberate speed, immediately exposing opportunities for improvement and abruptly forcing people to work within the new system to effect change. Because of this, procedures need to be in place to quickly and regularly maintain the APS data. This approach usually results in a faster return on investment, but with considerably more trepidation.

With the evolutionary approach, the APS system is initially set up to honor just the BOM constraints, leaving resources and purchased parts unconstrained. If good routing data are not initially available, standard manufacturing lead times are invoked. Under this configuration, the APS system mimics an order-based MRP system. Then, as routes are introduced, key workcenters are constrained, giving an improved representation of the process. As understanding is gained, more constraints are added, such as other workcenters or key purchased parts, until a realistic representation is achieved. This approach, which tends to be more palatable in a proven MRP environment, provides for a more gradual and systematic implementation but extends the time to reach full-value APS functionality and runs the risk of stopping short of this goal.

7. CONCLUSIONS

We have attempted to present a review of the basic planning algorithms, their strengths and weaknesses, and the role of APS in integrating the planning, scheduling, and demand management functions more closely. The issues to be addressed in implementing APS systems have also been addressed. The field of production planning and control has been extremely active in the last several years, and it promises to remain a challenging and interesting area for both researchers and practitioners in the decades to come.

REFERENCES

Billington, P. J., McClain, J. O., and Thomas, L. J. (1983), ‘‘Mathematical Approaches to Capacity-Constrained MRP Systems: Review, Formulation and Problem Reduction,’’ *Management Science*, Vol. 29, pp. 1126–1141.

Dauzère-Pérès, S., and Lasserre, J.-B. (1994), *An Integrated Approach in Production Planning and Scheduling*, Lecture Notes in Economics and Mathematical Systems, Springer, Berlin.

Elmaghraby, S. E. (1991), ‘‘Manufacturing Capacity and its Measurement: A Critical Evaluation,’’ *Computers and Operations Research*, Vol. 18, pp. 615–627.

Fargher, H. E., and Smith, R. A. (1994), ‘‘Planning in a Flexible Semiconductor Manufacturing Environment,’’ in *Intelligent Scheduling*, M. Zweben and M. Fox, Eds., Morgan Kaufmann, San Francisco.

Goldratt, E., and Fox, R. (1986), *The Race*, North River Press, Croton-on-the-Hudson, New York.

Graves, S. C. (1986), ‘‘A Tactical Planning Model for a Job Shop,’’ *Operations Research,* Vol. 34, pp. 522–533.

Hackman, S. T., and Leachman, R. C. (1989), ‘‘A General Framework for Modeling Production,’’ *Management Science*, Vol. 35, pp. 478–495.

Hadavi, K., Shahraray, M. S., and Voigt, K. (1989), ‘‘ReDS: A Dynamic Planning, Scheduling and Control System for Manufacturing,’’ *Journal of Manufacturing Systems*, Vol. 9, pp. 332–344.

Hammer, M., and Champy, J. (1994), *Reengineering the Corporation: A Manifesto for a Business Revolution*, Harper Business, New York.

Hendry, L. C., and Kingsman, B. G. (1989), ‘‘Production Planning Systems and Their Applicability to Make-to-Order Companies,’’ *European Journal of Operational Research*, Vol. 40, pp. 1–15.

Hopp, W., and Spearman, M. L. (1996), *Factory Physics*, Irwin, Chicago.

Hung, Y-F., and Leachman, R. C. (1996), ‘‘A Production Planning Methodology for Semiconductor Manufacturing Based on Iterative Simulation and Linear Programming Calculations,’’ *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, pp. 257–269.

Karmarkar, U. S. (1987), ‘‘Lot Sizes, Lead Times and In-Process Inventories,’’ *Management Science*, Vol. 33, pp. 409–423.

Karmarkar, U. S. (1989), ‘‘Capacity Loading, Release Planning and Master Scheduling with WIP and Lead Times,’’ *Journal of Manufacturing and Operations Management*, Vol. 2, pp. 105–132.

Leachman, R. C. (1993), ‘‘Modeling Techniques for Automated Production Planning in the Semiconductor Industry,’’ in *Optimization in Industry*, T. A. Ciriani and R. C. Leachman, Eds., John Wiley & Sons, New York.

Monden, Y. (1983), *Toyota Production System*, Industrial Engineering and Management Press, Norcross, GA.

Nahmias, S. (1993), *Production and Operations Analysis*, 2nd Ed., Richard D. Irwin, Homewood, IL.

Orlicky, J. (1975), *Material Requirements Planning: The New Way of Life in Production and Inventory Management*, McGraw-Hill, New York.

Pinedo, M. (1995), *Scheduling: Theory, Algorithms and Systems*, Prentice Hall, Englewood Cliffs, NJ.

Pritsker, A. A. B., and Snyder, K. (1997), ‘‘Production Scheduling Using FACTOR,’’ in *The Planning and Scheduling of Production Systems*, A. Artiba and S. E. Elmaghraby, Eds., Chapman & Hall, London.

Ptak, C. A., and Schragenheim, E. (1999), *ERP: Tools, Techniques and Applications for Integrating the Supply Chain*, St. Lucie Press, St. Lucie, FL.

Shapiro, J. F. (1993), ‘‘Mathematical Programming Models,’’ in *Handbooks in Operations Research and Management Science,* Vol. 4, *Logistics of Production and Inventory*, S. C. Graves, A. H. G. Rinnooy Kan, and P. Zipkin, Eds., North-Holland, Amsterdam.

Smith, S. F. (1993), "Knowledge-Based Production Management: Approaches, Results and Prospects," *Production Planning and Control* Vol. 3, pp. 350–380.

Tardif, V., and Spearman, M. L. (1997), "Diagnostic Scheduling in a Finite-Capacity Environment," *Computers and Industrial Engineering*, Vol. 32, pp. 867–878.

Tayur, S., Magazine, M., and Ganesham, R., Eds. (1998), *Quantitative Models for Supply Chain Management*, Kluwer, Dordrecht.

Vollman, T. E., Berry, W. L., and Whybark, D. C. (1988), *Manufacturing Planning and Control Systems*, 2nd Ed., Richard D. Irwin, Homewood, IL.

Zweben, M., and Fox, M., Eds. (1994), *Intelligent Scheduling*, Morgan Kaufmann, San Francisco.

ADDITIONAL READING

Baker, K. R., "Requirements Planning," in *Handbooks in Operations Research and Management Science, Logistics of Production and Inventory*, S. C. Graves, A. H. G. Rinnooy Kan, and P. Zipkin, Eds., North-Holland, Amsterdam, 1993.

Bermudez, J., "Advanced Planning and Scheduling Systems: Just a Fad or a Breakthrough in Manufacturing and Supply Chain Management?" Report on Manufacturing, Advanced Manufacturing Research, Inc., Boston, December 1996.

Johnson, L. A., and Montgomery, D. C., *Operations Research in Production Planning, Scheduling, and Inventory Control*, John Wiley & Sons, New York, 1974.

Leachman, R. C., Benson, R. F., Liu, C., and Raar, D. J., "IMPReSS: An Automated Production Planning and Delivery Quotation System at Harris Corporation—Semiconductor Sector," *Interfaces*, Vol. 26, pp. 6–37, 1996.

Venkataraman, R., "Frequency of Replanning in a Rolling Horizon Master Production Schedule for a Process Industry Environment: A Case Study," *Production and Operations Management*, Vol. 5, pp. 255–265, 1996.

CHAPTER 79

Transportation Management and Shipment Planning

Jeffrey H. Fischer
United Parcel Service

1. **INTRODUCTION** **2054**
2. **THE OPTIMIZATION PROBLEM** **2054**
3. **SUPPLY CHAIN MANAGEMENT AND THE TRANSPORTATION ELEMENT** **2055**
4. **TECHNOLOGY REQUIREMENTS** **2056**
    4.1. The Need for Information 2056
    4.2. Information Exchange 2057
5. **TRANSPORTATION MANAGEMENT SYSTEMS: SUPPLY CHAIN'S FINAL STAGE** **2057**
    5.1. Pickup, Delivery, and Routing 2058
    5.2. Case Study: Electronics Industry 2059
    5.3. The Traveling Salesman Problem 2060
    5.4. The Vehicle Routing Problem 2062
    5.5. Other Vehicle Routing Problems 2062
6. **SHIPMENT PLANNING** **2063**
    6.1. Tactical and Operational Considerations 2064
    6.2. The Total Shipping Solution 2065
    6.3. Case Study: Manufacturer of Medical Instruments 2065
7. **LOCATION PROBLEMS** **2067**
8. **SUMMARY** **2068**

REFERENCES **2068**
ADDITIONAL READING **2069**

1. INTRODUCTION

Uncertainty breeds inventory. Managers involved in transportation often have to make planning decisions, like routing, that directly affect the movement of raw materials or finished goods. These decisions often affect other components in the supply chain network, in which case the transportation management team cannot afford to make an incorrect decision. Consequently, any mistakes not only jeopardize other elements within the system but also lead to customer dissatisfaction created by the delay in the delivery times (Quinn 1998).

Beyond routing decisions, however, effective transportation management will assist with solving common transportation and shipping problems by generating various scenarios and simulations in order to arrive at optimal or best solutions for shipment planning, the selection of distribution center sites, and the allocation of resources to critical system components.

2. THE OPTIMIZATION PROBLEM

Often, optimization problems seek a solution where decisions need to be made in a constrained or limited resource environment. The majority of supply chain optimization problems require matching demand and supply when one, the other, or both may be limited. By and large, the most important limited resource is the time required to procure, make, or deliver something. Since the rate of procurement, production, distribution, and transportation resources is limited, demand cannot be im-

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

mediately satisfied. There is always some amount of time required to satisfy demand, and this may not be quick enough unless supply is developed well in advance of demand. In addition to time, other resources, such as warehouse storage space or a vehicle's capacity, may be constrained in meeting demand. All of these factors drive inventory levels, which in turn drive costs.

In achieving optimization, decision variables that are within the control of the planner, such as when to manufacture an order or when and how much of a raw material needs to be ordered, must be balanced with the inherent constraints or limitations that are placed upon supply.

Constraints, such as a supplier's capacity to produce raw materials or components or a customer's distribution center's capacity to handle and process receipts, can be considered either hard or soft constraints. Hard constraints, such as the number of working hours in a shift or the maximum capacity of a transportation vehicle, must be adhered to or satisfied. Soft constraints, on the other hand, can be relaxed or violated. Examples of soft constraints include customer due dates and facility storage limitations. Customer due dates can be modified or product may be temporarily allocated in a warehouse, making constraints less harsh. However, there are cost penalties if a soft constraint is not adhered to, permitting constraints to be weighted by their relative significance. For example, missing a customer due date carries more important consequences than cluttering a warehouse aisle (Lapide and Shepherd 1999).

3. SUPPLY CHAIN MANAGEMENT AND THE TRANSPORTATION ELEMENT

At the center of today's supply chain optimization technology are complex algorithms that can examine millions of variables and solve increasingly complex problems in ever-shortening time frames, enabling solutions in a matter of hours rather than days.

The traditional trade-off in supply chain management has been the maintenance of costly buffers of inventory vs. the ability to meet complex customer prerequisites. Reduce safety stocks and costs will be reduced, but customer service may suffer. Due mainly to advanced planning and scheduling systems and improved forecasting applications, production planners now have the opportunity to reduce reliance on safety stock—while still meeting customer demand—by trading inventory for information.

In transportation management, a similar trend is occurring, but rather than inventory buffers, logistics managers are doing away with time buffers. By using information technology in the elaborate mix of transportation modes, carriers, and shipment consolidation possibilities, manufacturers are obtaining more accurate estimates of the time and cost it will take to deliver goods throughout their supply chains. Transportation management applications are being used to better plan and execute shipments. The software lends visibility, consistency, and economy to the handling of complex variables. Some manufacturers have even begun to integrate their transportation and order management systems, giving transportation optimization an up-front role in supply chain dynamics.

Whether approached on a strategic level or shorter-term tactical and operational levels, transportation management, using new technology, is trimming time and cost. And time, according to most experts, is one of the most precious commodities in today's supply chains. Shortened product life cycles necessitate time-based competition throughout the supply chains (Michel 1997).

For some time now, optimization techniques have been used to solve for least-cost shipping configurations. The classic transportation problem was to solve for the best combination of routes that fulfilled all the demands, subject to all the availability and, naturally, at the least cost. With a considerable number of possible routes, the problem was too complex to solve by hand, and therefore linear programming and network algorithms provide quicker solutions to the problem.

Although times have changed, these methods are every bit as applicable today as they have ever been. While the nature of the model constraints is considerably different and more complex, optimization modeling aids in filling the supply chain more effectively. Thus, transportation vendors have been customizing their delivery systems to meet a more stringent set of customer requirements.

In our global economy, customers are demanding items having exact options, in exact quantities, of zero defect, to be delivered precisely at specific locations, on certain production lines, and at exact times. In light of this new paradigm, we are still confronted with managing transportation costs. As the marketplace demands a far more flexible delivery system, both shippers and carriers are hard pressed to balance these demands against a complicated set of constraints. Fortunately, through mathematical modeling, all the competing requirements in arriving at not only a feasible but also an efficient delivery program can be evaluated and studied.

Linear programming (LP), commonly used to solve a variety of industrial and scientific problems by arriving at an optimal solution, has been around since the 1940s. The early applications for LP that yielded the largest benefits involved creating schedules for massive capital investments such as rail, bus, and airline schedules. With ever-increasing competitive markets, however, additional requirements have been added. Linear programming still remains an effective technique to solve a variety of industrial applications problems (Lustig 1999).

Specific to transportation and logistics issues, some applications include:

1. Transportation and distribution:
   - Shipping plans: Determine optimal shipping assignments from manufacturing facilities to distribution centers or from warehouses to consumers (e.g., customer direct).
2. Site selection:
   - Facilities: Establish the optimal location of a plant or distribution center with respect to total transportation costs between various alternative locations and existing supply and demand sources.
3. Scheduling:
   - Shifts: Solve for the minimum-cost assignment of workers to shifts, subject to varying demand.
   - Vehicles: Allocate available vehicles to jobs and determine the number of trips to make, subject to vehicle size, availability, and demand constraints.
   - Routing: Solve for the optimal routing of a product through a number of sequential processes, each with its own unique capacities and characteristics.
4. Production Planning:
   - Production: Solve for minimum-cost production scheduling for an established workforce, taking into account inventory carrying and subcontracting costs.
   - Production and workforce: Solve for minimum-cost production scheduling, accounting for hiring and layoff costs as well as inventory carrying, overtime, and subcontracting costs, subject to various capacity and policy constraints.
   - Staffing: Determine the appropriate staffing levels for various categories of workers, subject to various demand and policy constraints.

4. TECHNOLOGY REQUIREMENTS

A general trend exists toward increased systems integration within the supply pipeline in order to create and provide better information faster. In turn, this has decreased standard transactional costs but has also led to a fundamental restructuring of industry practices for distributing and supporting goods and merchandise. Over the last few years, decision points, such as supplier selection, price, quantity, routing, and delivery, have required greater coordination throughout the supply chain. Hence, these critical activities have become more and more integrated systems themselves in order to govern the flow of physical goods between shipper and consumer (Lewis and Talayevsky 1997). According to Donald J. Bowersox, the John H. McConnell Professor of Business Administration at Michigan State University, "technology serves as the primary enabler to facilitate supply-chain-wide integration while simultaneously allowing key business relationships to be conducted on an exclusive enterprise-to-enterprise basis."

4.1. The Need for Information

As cycle times are reduced and more efficient inventory processes are embraced, transportation buyers have become more increasingly concerned with the location of a shipment in the logistics pipeline than with the shipment itself. Providing information on a shipment, including its contents, its current location, its destination, and its expected time and date of arrival, is critical in transportation planning.

This desire to have timely and accurate shipment information has transportation providers investing millions of dollars each year on high-tech bar coding, communications, and networking equipment. This desire has also made information one of the most important factors in the transportation equation.

The factors that have made shippers demand more information on their shipments reflect major shifts in business practices, new shipping patterns, and the availability of new and more affordable technology.

Advanced manufacturing research (AMR), a market-analysis company that specializes in supply chain technology, estimates that there will be a 48% compound annual growth rate for supply-chain management software until 2003. That will put annual sales of these integrated suites at nearly $19 billion. Transportation management systems, with 1998 sales of $314 million, are expected to reach $1.9 billion by 2003 (Forger 1999).

Simply stated, businesses do not operate the way they used to. Instead of stockpiling finished goods in warehouses, shippers are adopting just-in-time (JIT) and lean manufacturing strategies, which operate with little or no inventory. And this has had a significant impact on transportation management, shipment planning, and the information associated with it.

4.2. Information Exchange

In order for shippers really to improve their operations, they must be willing to share critical information, such as production, supply, and cycle time data, with their transportation providers and other supply chain partners in order to make the entire process more effective.

Sharing this type of information with suppliers allows for many new transportation and distribution alternatives, allowing carriers to reroute loads in transit, consolidate shipments for more efficient distribution, and merge shipments so they arrive to customers as a single order. In the information technology age, sophisticated shippers will know not only where to get accurate data on their shipments but also how to leverage that data to improve operations along their companies' supply chains. This information is being used by shippers and transportation suppliers in the following manner (Minahan 1997):

By shippers:

- Process orders
- Tender freight
- Shop for rate and schedule data
- Generate, transmit, and file shipping documents
- Manage inventory and multiple-point distribution
- Trace shipments
- Measure carrier performance
- Identify supply chain weaknesses
- Process and pay freight bills
- Budget and manage costs

By carriers:

- Receive freight bookings
- Construct rate quotes
- Issue bills of lading
- Track and manage equipment
- Plan routings
- Determine load sequencing
- Manage documentation
- Trace shipments in transit
- Monitor equipment utilization
- Respond quickly to failure situations
- Coordinate consolidated loads and multiple-point distribution
- Confirm pickup and delivery
- Generate performance, accounting, and other reporting
- Issue freight bills

In the context of industrial engineering, many times industrial engineers will be charged with the development, integration, and execution of the complex systems used in supporting these activities.

5. TRANSPORTATION MANAGEMENT SYSTEMS: SUPPLY CHAIN'S FINAL STAGE

The real potential of transportation management systems (TMS), beyond operational efficiencies, is the substantial cost savings that it is capable of generating for shippers. Recognizing the enormous logistics costs that are transportation related, transportation management is as complex and difficult as any other problem associated with an organization's business environment.

Transportation management, an integral part of a firm's logistics strategy, involves purchasing, monitoring, and controlling freight transportation services (Temple, Barker, and Sloane 1982). Considering that, on average, 3.5% of a manufacturer's sales costs and 40–60% of total logistics costs are devoted to the movement of products, transportation management is essential in today's business environment. Therefore, incorporating TMS into a supply chain management strategy is also essential

(Weil 1998). The potential savings from identifying, for example, shipment inefficiencies, excess labor, and other unnecessary costs on a regular basis can be substantial.

The identification of cost-savings opportunities occurs primarily because the system automates the shipping and carrier selection process. In addition, TMS functionality includes load planning, rating, pickup scheduling, shipment consolidation, freight payment, and claims management. With this type of real-time information available, TMS introduces flexibility into a company, allowing the shipping department to make last-minute, but accurate, decisions as priorities and carrier costs shift (Forger 1999).

Standard software packages (see Table 1) are available that directly reduce operating costs by optimizing shipment plans, including freight consolidation, mode/carrier selection, and dedicated fleet routing and scheduling. Other benefits include improved service due to more accurate and timely shipments and the automation of manual processes. The best transportation management software, however, has strong strategic and tactical planning modules, which allow extensive "what-if" capabilities to optimize the design of a transportation network. They also aid the planner in the determination of fleet size, the design of fixed/master routes, consolidation strategies, optimal shipment size/frequency, and territory design.

5.1. Pickup, Delivery, and Routing

A primary transportation management concern is the determination of how to utilize a given fleet of vehicles efficiently. To minimize total cost, whether small-parcel, less-than-truckload (LTL), ship-

TABLE 1 Leading Transportation Management Products

Software Provider	Product	Functions	Platforms
CAPS Logistics Inc.	TransPro	Freight consolidation and mode/carrier selection	Windows, Windows NT
i2 Technologies Inc.	Rhythm Transportation Optimizer	Load consolidation, routing, and carrier selection	Windows NT; Unix version due shortly
Manugistics Inc.	Transportation Management	Plans and optimizes shipments for multipoint distribution; includes freight payment to facilitate Web-based carrier tenders	Windows NT, Unix
McHugh Software International	McHugh TMS	Mode/carrier selection, electronic load tendering, carrier assignment, Web tracking, and rating/auditing	Windows NT, Unix
Optum Inc.	Optum SCE Transportation	Optimizes transportation for timely delivery by the most efficient carrier	Windows, Unix
Provia Software Inc.	FreightLogic (formerly from Pinnacle Distribution)	Optimizes order processing to plan most economical loads	For hosted model, PC with Internet connection; for in-house model, Windows NT server
Sabre Inc.	OptiBid	Solicits carriers, analyzes bids	Client/server technology that runs on Windows NT
	OptiFlow	Freight consolidation and routing and scheduling	Unix workstations
	OptiMatch	Evaluates and processes real-time load demand data to recommend mode and carrier	Dedicated networked workstations

Source: Tausz 1999.

ments that are typically too large for the package companies and too small for truckload (TL), or carriers that transport trailers direct from origin destination, determining the pickup and delivery sequence of shipments assigned to each vehicle is subject to a variety of constraints (e.g. vehicle capacity and pick-up/delivery times).

This problem can be modeled as a vehicle routing problem (VRP) with numerous side constraints. Recognizing that the VRP is notoriously hard, the size and scope of the real-world data sets can make it impractical to just formulate the problem as an integer program (IP) and use an advanced IP solver to get an optimal solution. Therefore, practitioners usually seek solution techniques that yield acceptable solutions within a reasonable time frame (see Chapter 30). Some of these techniques include:

- Route-building heuristics select arcs greedily in a sequential manner until a feasible solution has been formulated.
- Route-improvement heuristics start with a feasible solution and seek a minor change that reduces cost while maintaining feasibility.
- Mathematical programming-based heuristics solve to optimality some mathematical programming approximation of the problem using several techniques (e.g., Lagrangian relaxation algorithm and column generation).
- Artificial intelligence/self-adaptive methods start with initial feasible solutions, then repeatedly make a local change to the current solution (such as swapping shipments between vehicles). In turn, each new solution is accepted if it satisfies certain criteria. Whereas traditional local improvement methods accept a local change if it strictly decreases the cost and stop when such a change does not exist, taboo search and annealing, for example, allow the selection of nonimproving solutions under certain conditions.

Regardless of the chosen technique, it is important to understand the business rules concerning fleet and vehicle utilization. Therefore, observing current procedures and obtaining real data early in the development process will allow the business rules to be incorporated directly into the model as constraints or applied in a preprocessing step to drastically reduce the problem size. This will make the problem more manageable and help ensure that the plans the transportation management software generates can be implemented in practice (Ergun 1998).

5.2. Case Study: Electronics Industry

An electronics manufacturer had 400 transportation providers within its United States distribution network. Only 42 carriers delivered 98% of the volume. Of this select group, United Parcel Service was the primary carrier handling outbound ground movements, inbound shipments, and less-than-truckload freight to their customers or from suppliers. Due to either the physical product or package constraints, there were few exceptions that prevented them from having a minimum number of carriers within their distribution network. That network consisted of two plants (Syracuse, NY, and Salt Lake City, UT) and four distribution centers (Atlanta, GA, Chicago, IL, Dallas, TX, and San José, CA).

Since transportation represented more than 40% of the company's total logistics expenditures, they began to aggressively pursue opportunities to create additional value through their service providers.

Their primary goal, from a logistics perspective, was to maintain or expand delivery coverage while the number of distribution centers would be potentially reduced to zero. Ultimately, all customers would be served from only the two plants, Syracuse and Salt Lake City, within two days. The offer of a two-day delivery for all products to all customers would be a first within their industry and would give them a tremendous competitive advantage by freeing up cash and generating sales.

Serving customers within two days, however, required an entirely new operating plan for both UPS and the electronics firm. Based upon 18 months of historical distribution data (e.g., traffic lanes, product distribution by customer, mode usage, etc.), various cost scenarios were determined by specific transportation provider, origin, destination, ZIP codes, weight, number of packages, and so on. Every element related to transportation that could be measured was analyzed and associated with a specific product and product group. From this, a plan was developed that met the objectives to lower inventory, reduce overall logistics costs, and improve customer service levels.

The preliminary analysis indicated that a reengineered network and operation, customized to the customer's specific characteristics and requirements, could function with two sites that could deliver 35.5% of the shipments the next day, followed by 63.6% and 0.9% by the second and third day, respectively.

To implement the second-day coverage, the planning team developed a master operating plan divided into phases over a few years. Within each phase were various scenarios that addressed package characteristics, overall volume, pull times, hub sorts, and so on. For example, the plan called for direct loads to be built for final destination hubs, bypassing intermediate hubs, reducing handling and processing time. This reduced costs and increased the geographic coverage. As additional volume entered the system, more direct loads were made. The plan was highly customized with the manufacturer's and the carrier's requirements, which incorporated shipment origin data, destination ZIP codes, volume adjustments, trailer departure times, linehaul transit times, and hub sort start and stop times.

Their annual logistics costs, with an average inventory of $500 million, was $186 million at the beginning of the project. This was composed of:

- $38 million in transportation
- Inventory carrying costs (including cost of capital, depreciation, obsolescence, damage and shrinkage) at 25% the value of the inventory or $130 million
- Fixed distribution expenses added another $18 million

As mentioned before, management wanted to decrease total logistics costs by reducing warehouses. A consolidated network would reduce inventory levels and other associated costs. Transportation costs could increase since it was estimated that there would be a trade-off in this case between less critical next-day air shipments and stock transfer shipments but more direct and smaller customer shipments.

Although logistics problems such as this one are complex because there are so many possible combinations of the underlying variables, a good solution was found. With two-day service recognized as the only acceptable service level, the key criterion was known. Incorporating the parameter to ship using mostly ground transportation established the overall cost objective. Utilizing optimization-based software to determine the best routing within the UPS ground network identified the necessary geographic coverage. Anything beyond two-day ground capability was supplemented with two-day air shipping (meeting the criterion of eliminating next-day air shipments).

The project was successfully implemented and has achieved the original goals. The reduction of warehouses and the associated inventory resulted in significant savings. With reliable transportation services established throughout the redesigned distribution network, inventory levels dropped to $385 million. In turn, the logistics budget was reduced by approximately 28% to $134 million:

- $32 million in transportation
- Inventory carrying costs (including cost of capital, depreciation, obsolescence, damage, and shrinkage) at 25% the value of the inventory, or $96 million
- Fixed distribution expenses adding another $6 million

The keys to successful implementation of this type of project include:

- *The approach to implementation:* Use a team-based concept for all project implementations, from inception to completion. To accomplish this, immediately identify the most qualified individual to be the project manager and begin formulating the project-implementation schedule itself. The schedule, which will be a mutually agreed-upon timeline, will incorporate the strategies of all parties.
- *The implementation team:* Experienced managers who represent critical areas that will be impacted by any potential changes. For example, industrial engineering, logistics planning, information technology, and finance are all essential functions to be represented in a cross-functional team.
- *Milestones:* Establish significant tasks that must be accomplished and acknowledged by the team before proceeding with other critical assignments. Validation of data and requirements or system testing or training of management and staff, for example, are important steps as the project proceeds. In addition, discussing and validating the original project charter is also critical in order to stay on the intended course.
- *Critical dependencies:* Since each party recognizes that the project must adhere to an aggressive timeline, all critical points must be responded to quickly. To complete the implementation process on time, both parties must reach consensus and respond as soon as possible.
- *Contingency planning:* Although there is an established implementation timeline, adhering to it is challenging. There is always the probability that an issue or a variety of circumstances can delay or jeopardize the final implementation date. Therefore, if changes occur that can potentially affect the optimum design plan, for example, dates should be extended past the date originally set forth in the beginning of the project.

5.3. The Traveling Salesman Problem

The traveling salesman problem (TSP) (Lawler et al. 1985), the classic problem in which a mythical traveler must find a minimum-length cycle through a set of nodes in a completely connected graph, has an important place in computational complexity theory. But more significant for the transportation and logistics industry is that it has an important place in a continuously expanding field of operations research: vehicle routing.

As an illustration of a TSP application, suppose that a traveler starting from Chicago must visit several U.S. cities exactly once and return to Chicago. A solution to this problem is shown in Figure 1.

A computer scientist would call the TSP a ''hard'' problem because of the long computer times needed to solve large TSPs optimally. Yet transportation dispatchers, faced with the real problem of routing large fleets, would find that their problems include factors that the TSP does not account for, including:

- *Route capacities and times:* Each cycle is constrained by the available space in a vehicle, which might be measured by cubic volume, weight, number of pieces, or floor space (if items are not stackable). The cycle is also constrained by the time in a driver's day, which itself depends on safety regulations, company work rules, and whether the driver has already handled other loads.
- *Time windows:* To remain competitive, companies are much more responsive to their customers' needs and deliver shipments exactly when customers want them. These requirements are stated as time windows, either as rigid lower and upper bounds or as soft time windows, which can be violated with penalty.
- *Dynamic routing:* Planning out the stops in advance is not always possible. Often all stops aren't known until the vehicles are in the field, partway into their routes, forcing them to double back or circulate to finish their tour. In addition, travel times and costs are time dependent because vehicles confront the commuter rush hours on the road, as well as at the loading and receiving docks.
- *Randomness:* Finally, nothing is certain. A shipment that is supposed to have 10 pieces turns out to have 25, and immediately there's not enough space to finish the route. Or perhaps a driver is detained with paperwork and doesn't have time to visit the rest of the stops or get back to the airport gateway in order to connect with the next flight out.

Thankfully, industrial engineers and software developers are aware of these real constraints, which are handled by various extensions of the TSP (Ball et al. 1995; Golden and Assad 1988). For example, route capacities and times are considered by the vehicle routing problem (VRP), while time window constraints are included in the TSP with time windows (TSPTW) and the VRP with time windows (VRPTW). Advances in telecommunications make it now possible to implement models that take into account the dynamic aspects of vehicle routing, while time-dependent vehicle routing problems take into account time-dependent travel times and costs. The stochastic aspects of cost, time, demand

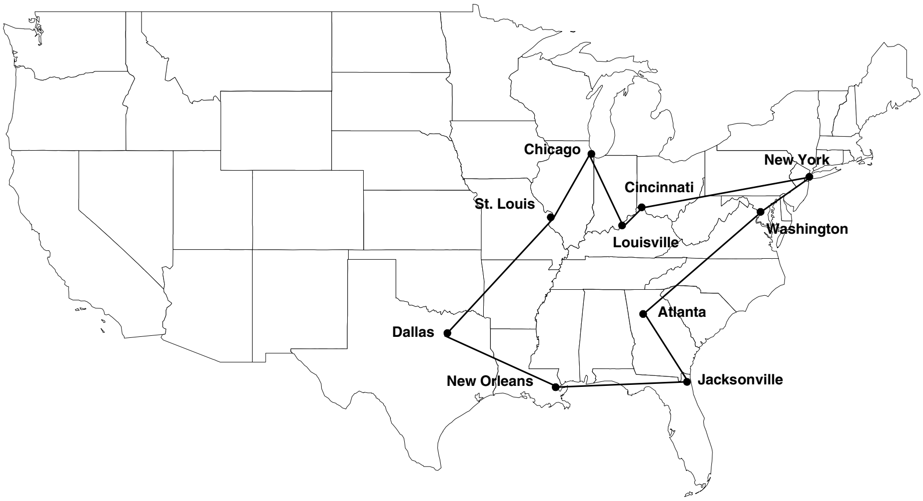


Figure 1 A TSP Tour through 10 U.S. Cities.

size, and even the presence or not of a customer are considered in the stochastic versions of the problems.

Since these problems are very difficult to solve optimally, TMS packages employ a blend of heuristic and optimization algorithms to assist dispatchers in routing their vehicles. In addition to creating more efficient routes, meaning fewer miles, fewer labor hours, and fewer vehicles, not to mention getting stops served on time, today's transportation management software is being integrated into the overall supply chain process to manage the movement of goods from source to destination, tracking the productivity and quality of drivers and generating information for planning purposes, all in a paperless environment.

5.4. The Vehicle Routing Problem

The vehicle routing problem (VRP) (Christofides 1985) is a capacitated version of the TSP. A fleet of vehicles is available at one or more terminals to serve a set of defined stops. A shipment size is associated with each stop, and a cost is associated with the movement between each pair of stops (and between a stop and a terminal). The goal is to deliver the shipments to all the stops at minimum total cost in a set of cycles without violating vehicle capacity. The VRP formulation matches well local pickup and delivery problems where the pickup stops are known before the vehicle starts on the route.

Solving the VRP or its variants may necessitate actually solving additional problems. First, the input to the problem needs to be obtained, such as the distances, travel times, or costs between each pair of stops. This can be achieved either by approximate calculations or by using geographic information systems (GIS). Using approximation, the coordinates of each stop are determined and the distances in a straight line are computed. Since a vehicle cannot drive straight from point to point, the distances must be adjusted upward by about 15% to approximate actual road mileage.

When GIS is used, distances between stops are derived from shortest path algorithms applied to very large networks rather than by simple algebraic calculations. GIS distances are more exact than approximations (especially when stops are separated by bodies of water or mountains), but they are not without flaws. Errors in the input data can occur, and sometimes only an experienced truck driver knows that a fast car route is impossible for a semi-tractor negotiating sharp turns. Nevertheless, GIS are an invaluable source of point-to-point distance and time data, particularly for shorter-length trips, where accuracy becomes even more important. GIS is fast enough to be practical, but GIS data are much more expensive to acquire than by using simple approximations.

From the distances between each pair of stops, travel times are computed assuming specific speeds. Costs are computed assuming vehicle and driver costs particular to each application.

Solving the VRP determines which vehicle serves which stops and in what sequence. There are different solution methodologies for solving the VRP, either optimally or heuristically. Since optimal algorithms can solve only small problems, emphasis is given to heuristics algorithms that aim at finding near-optimal solutions.

Some algorithms assign stops to vehicles and determine the stop sequence concurrently. Other algorithms use the so-called cluster first, route second approach, which consists of the following two steps. First a service area is partitioned into smaller regions, where each region represents a feasible collection of stops for a single route. These regions can overlap, especially when time windows are involved or when requests for pickups arrive dynamically throughout the day. Determining the sequence of stops in a single region amounts to solving a TSP (or a TSPTW, if time windows need also to be satisfied).

If the same 10 cities were considered as in the TSP example but a capacity constraint was added that necessitated the use of three vehicles, the VRP solution could look as in Figure 2.

While the previous description represents a generic implementation, specific vehicle routing applications may have their own individual characteristics, requiring that transportation management and shipment planning software be customized to reflect the operating environment, customer needs, and the characteristics of the transportation mode (Hall and Partyka 1997).

Chapter 30 presents a detailed overview of the VRP and its applications in transportation.

5.5. Other Vehicle Routing Problems

The VRP matches quite well local pickup and delivery routing in the trucking industry. Long-distance truck routing, however, is much more focused on crew-assignment issues, along with balancing interregional freight flows. The problem is much more closely related to transshipment problems than to the VRP, where the objective is to balance the flow of equipment and drivers in and out of terminals while minimizing empty mileage. This must be solved within the context of routes that can take up to several days to complete, requiring driver or equipment exchanges or possibly sleeper/driver teams that operate almost continuously.

Vehicle routing can be divided into three primary categories: service vehicles, passenger vehicles, and freight vehicles. Service vehicles usually do not move things or people from place to place

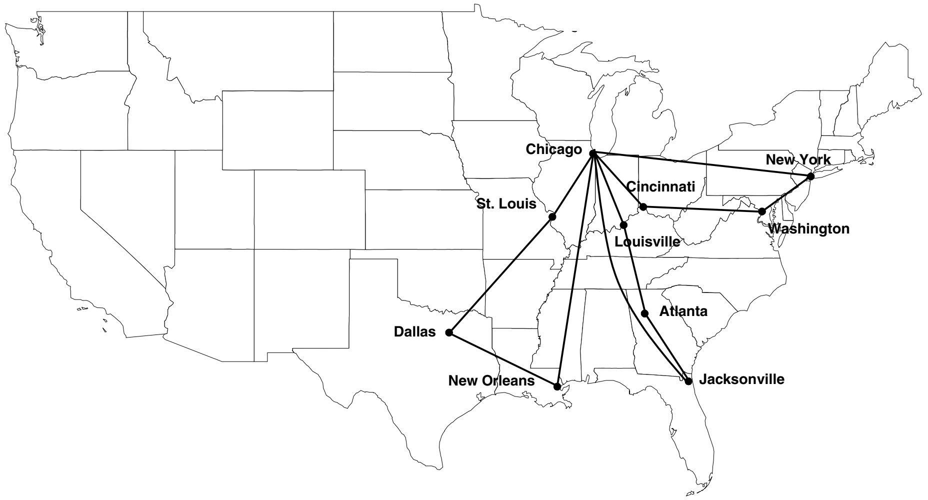


Figure 2 A VRP Solution Using Three Vehicles.

(snow-salting vehicles are an exemption to this) but are used to support jobs in the field (e.g., vans for the copier repair technician or trucks fixing potholes). When service routes are not constrained by shipment sizes and vehicle capacity, the route lengths are constrained only by the time in the driver's day or shift. Local pickup and delivery in the package transportation industry is also dominated by the duration of driver shifts rather than capacities, while time windows are of primary importance to this industry.

In contrast to service vehicles, passenger fleets or carriers carry something from one place to another. Buses and vans, for example, carry people, so the lengths of their routes are constrained by the number of seats, or possibly by a combination of standing and sitting room. Freight vehicles, including ships, trucks, rail, and air, are also capacity limited. When a shipment or passenger is carried from one place to another, it may be transported through a logistics network of terminals connected by a variety of different routes. For example, the less-than-truckload (LTL) industry specializes in shipments that are typically too large for the package companies (such as United Parcel Service) and too small for truckload (TL) carriers, which transport trailers direct from origin to destination.

LTL shipments are commonly handled in at least two, and probably up to four or five, terminals. First the shipment is picked up from the origin and taken to a local end-of-line terminal. Next it may be shuttled to a regional consolidation center, where it is consolidated with shipments originating throughout the general metropolitan area or geographic region. Next the combined load is transported on a long-haul route to another consolidation center, near the final destination. From there, the reverse takes place. The shipment is shuttled to an end-of-line near the destination and finally on a delivery route to the customer's destination.

Each segment of the LTL shipment's journey demands a different kind of route and has a different routing challenge associated with it. Local pickup is a dynamic problem that requires flexibility to serve shipments as they are called in. Shuttle routing necessitates careful driver scheduling to ensure that workshifts are fully utilized. Linehaul routing requires balancing interregional flows to minimize empty equipment miles and reduce driver and fleet downtime. In the end, only the final segment of the trip—the delivery route—closely matches the VRP formulation, while the rest of the problems are solved by a variety of network design and crew-scheduling problems that often need to be customized.

6. SHIPMENT PLANNING

Shipment planning starts with the receipt of orders in the planning system. Transportation management actually begins far in advance of individual shipments. It starts with the configuration of a transportation network.

Recognizing that transportation management is a multidimensional discipline, one must look horizontally at all domains of control that comprise the supply chain and vertically at strategic planning, tactical planning, and execution. This establishes the business model for a supply chain, including all the physical locations, constraints, and processes. With strategic development modeling tools, a manufacturer can provide a better shipment forecast, by transportation lane, for carriers. With a more accurate model, collaboration between manufacturers and carriers will improve, leading to forecasting tighter carrier commitments and even the planning of sales and promotions around low-cost transportation lanes.

Transportation management systems are able to balance carrier rate, mode, and shipment consolidation variables because they contain carrier rate and equipment data and feature planning algorithms that suggest the fastest or lowest-cost options. Although the planning capabilities of software programs may vary, one of the most basic capabilities is consolidation of small orders into larger shipments. Effective consolidation planning means that multiple orders can be combined to form a full-truckload (TL) shipment rather than having orders go out individually as more costly, less-than-truckload (LTL) shipments.

More advanced shipment planning options in TMS include continuous moves—routing options that seek to keep trucks loaded on all transit legs—as well as sequential loading of shipments to shorten transit times and minimize handling.

Another shipment-planning option built into some transportation management systems is pooling. In pooling, small orders for multiple destinations are combined together and delivered in a full truckload to a cross-docking or distribution facility within the same geographical area as the orders' final destinations. At the pooling point, the freight is separated into small shipments or individual packages and routed to their individual destinations. Other desirable planning features within a TMS and shipment-planning program include the ability to meet short delivery windows, avoiding traffic gridlock at shipment destinations, and restocking options in which replenishment of a shipper's warehouse can ride for free when consolidated with customer orders.

Ultimately, the goal is to understand the trade-off between cost and customer service. Transportation management software strikes a balance between what customers want and what the carriers can deliver (Michel 1997).

6.1. Tactical and Operational Considerations

Transportation management systems can be categorized into three fundamental types:

1. Network planning and modeling applications
2. Transportation resources planning and management (TRPM) applications, which perform tactical planning
3. Transportation administration and management systems, which are operational execution applications

Typical TRPM systems perform some of the same operational tasks as transportation administration and management systems but have the ability to plan and execute enterprise-wide plans rather than single business unit plans. Such systems, therefore, must consider inbound, outbound, and replenishment demands throughout a global supply chain network.

For instance, one inbound and outbound optimization planning software package has the ability to manage shipments from multiple origins to multiple destinations and builds and consolidates loads as orders are imported into the system, using a library of transportation algorithms. Combined with a costing module, it rates, ranks, and selects carriers based on customer needs.

After the system selects the best transportation mode, it determines the best travel path, which loads to deliver first, and which orders should go on the trucks first—an important consideration since customers do not want to hold inventory, making scheduling much more significant due to the pressure to meet on-time delivery (Dilger 1998).

Roadnet Technologies, a leading provider of routing, loading, and planning and dispatch software for the transportation industry, has helped users streamline their supply chains by optimizing their transportation operations with Roadnet 5000 and Territory Planner. (See the Roadnet Technologies website, www.roadnet.com.)

Territory Planner strategically plans delivery and route sales territories. This analytical tool can streamline a company's operation and suggest routes that are in line with the way a shipper does business. Similar tools save reroute time, reduce transportation costs, and improve customer service. Similarly, Roadnet 5000 routes and schedules delivery vehicles by considering the parameters of a company's operation. The consolidated routes that are created provide a competitive advantage by improving driver performance and information management.

Ideally, the optimal TMS will permit transportation and logistics personnel to configure the parameters or rules for processing shipments in relation to the firm's entire customer base (e.g., optimal parcel carriers for each region of the country, discounts applicable to each carrier and region of the country, and carriers available for expedited service, including next-day and two-day delivery). Subsequently, the TMS manages the dynamic characteristics and unique variables of each customer order (e.g., final destination, shipment weight, and delivery date requirements) to make an intelligent rating decision within seconds. In turn, the system alleviates manual ship-rate shopping and guarantees that the shipper is using the least-cost carrier defined within the dynamic parameters of the order.

6.2. The Total Shipping Solution

If shippers incorporate TMS as part of their total supply chain execution solution, they will be able to achieve a strategic advantage and improve supply chain performance. Properly integrated, transportation management software, which can cost between $300,000 and $1 million (Cooke 1998), can enhance numerous areas within the supply chain. Some common areas for improvement include (Weil 1998):

- Order entry / customer service:
  - Real-time rating and routing information with customers on the phone
  - Real-time tracing and tracking of the shipment, including details and value of individual packages
  - Guarantees on customer carrier preference
  - Guarantees that the customer's delivery date will be complied with while still providing a cost-effective order of shipping
- Purchasing:
  - Inbound freight expense and shipment delivery analysis, including back-hauling capabilities or preferred carrier delivery service
- Invoicing:
  - Transportation charge line items automatically added to invoices
  - Ability to configure, maintain, and invoice customer program costs
- Shipping:
  - Increased parcel processing throughput and accuracy via bar code data and scanning to capture package details, package weight, and package tracking numbers
  - Automate carrier and shipment documentation generation, including bills of lading with preassigned carrier freight bill numbers, unique customer reference numbers, shipment labels complete with carrier tracking numbers, retail vendor-compliant data, international documentation, and EDI advanced shipping notice (ASN) bar codes.
- Accounts payable:
  - Automation of freight payment and matching processes
  - Introduction of self-invoicing practices that place claim maintenance in the hands of the carrier, not the shipper.

6.3. Case Study: Manufacturer of Medical Instruments

While many shippers have talked about making changes in their shipping and distribution practices for years, the supply chain service division within a medical instruments organization acted on its beliefs and made changes. In 1998, the company decided to review how small shipments were processed within their distribution centers. Everything had to be considered, from order picking to paying the freight bill. Adding value, taking the cost out of the supply chain, and the desire to improve their logistics network mandated an audit of their entire logistics system.

Their existing routing guidelines and system indicated that small shipments, primarily small parcel freight, were very cost effective based on current transportation rates. However, when all aspects of the shipment process (e.g., picking and packing process, label printing) were incorporated in the analysis, smaller shipments appeared to be less cost effective when compared to, for example, LTL orders. Thus, the planning team reviewed the shipping areas, compared processing times, and identified the costs of each. In addition, their primary carrier, along with the third-party warehouse provider, evaluated proposed warehouse layouts and procedural changes to streamline the shipping process.

Observations and studies were made of the distribution center's activities to quantify current practices. Any assumptions or atypical occurrences observed were evaluated on their individual merit in order not to adversely affect the study. To deliver better alternatives, accurate information was essential for the activity-based costing models utilized in the study. In addition, all exceptions had to be considered to maintain the integrity of the information being supplied to the manufacturer.

In addition, the team identified all the existing procedures inherent to processing a small parcel shipment as well as an LTL shipment. This comparison, which included all aspects of the process, from order entry to packing and transporting the merchandise, was conducted so that each mode of transportation could be effectively evaluated on its true value. In other words, actual costs would be clearly identified and allocated to processing orders. Since the manufacturer's goal was to improve the overall process within the existing cost structure, they recognized that requesting carriers to lower their transportation prices would produce only minimal gains and not enhance their value to their customers within the market. Focus on transportation rates had been the previous and predominant belief regarding reducing overall costs. Thus, the weight break between small parcel shipments and LTL shipments had been established only to reflect that element. All costs associated with the shipping process were to be included in order to illustrate effectively the actual expenditures. In turn, the weight break for small parcel shipments would be raised from 100 to 150 pounds.

The team's success was successful because they had planned prior to redesigning the procedures within the facility. Relevant data on the products were first gathered, such as:

- Sizes and weights of the products being handled
- Anticipated throughput requirements
- Weighing requirements
- Manifest requirements
- The current and projected packaging and labeling requirements (such as compliance labeling) for the products handled
- The packaging material(s) specified by the customer
- How the products were shipped (pallet loads or loose cartons)
- The company's experience base with various packaging and unitizing methods
- The projected number of inventory turns per year
- EDI or ASN requirements
- Any special handling requirements (e.g., DOT restrictions)

The study revealed that some orders took longer and were actually more costly when processed as a LTL shipment rather than as a small parcel shipment. The primary difference was the consolidation of the cases to a master identification number and application of the shipping labels. While a small parcel shipment underwent a similar procedure, the existing shipping system streamlined several of the steps that were otherwise manually entered for a LTL shipment. Completing the bill of lading, pick/pack time within the warehouse, stretch wrapping, and moving the pallets also contributed to additional time for the LTL order. Additional costs were also realized with the expenditures for pallets, stretch wrap, shipping documents, and administrative costs related to freight payment.

From this base, the planning team considered alternatives that ranged from simple procedural changes that could be implemented immediately to extensive automated sortation designs that required time and capital. With many considerations, the team recognized that the order-selection process was the most labor-intensive activity performed in the customer's warehouse. As a result, it offered the greatest opportunity for improvement. The team identified and implemented several ideas to reduce the order-picking costs:

- Separating broken-case picking from full-case picking to eliminate the need for the selector to change materials-handling equipment.
- Clearly marking the pick-slot numbers in one place eliminated confusion and reduced errors.
- Sequenced slot numbers in a logical pick path reduced travel distance and time.
- Translating quantities ordered into pick quantities on the pick document to eliminate errors from miscalculations (e.g., if the product is packaged in cartons of 12 units, the pick sheet needs to indicate to select 10 cartons, not 120 units).
- Establishing selectors, where appropriate, to pick with labels rather than from a picking sheet. For example, a selector using labels attaches them to each carton picked for an order. The labels indicate one of four in an order, so the selector knows an order is complete when the labels are gone.

The manufacturer quickly benefited from this basic process change and was now well positioned to explore more elegant alternatives to meet future business needs.

This medical instruments company has enjoyed reductions in both the costs of processing and transporting shipments. With all elements assessed, a better conclusion could be drawn in order for the shipment planning process not to affect the outbound shipments adversely. If the manufacturer's original conclusions have been acted upon, the weight break level between small parcel shipments and LTL shipments would have been lowered even further. With a detailed analysis performed, a new, higher weight break was established. In conjunction with the existing transportation rates, the average cost per shipment was lowered.

7. LOCATION PROBLEMS

The transportation models discussed in this chapter assume that the location of the facilities involved is given. Obviously, the location of various facilities plays an important role in the total transportation costs incurred. The location of factories, warehouses, and distribution centers plays a major role in the quality of service and competitiveness of a manufacturer, while transfer terminals and depot facilities greatly influence the cost structure and effectiveness of a transportation company.

It is to be expected, therefore, that an extensive body of work exists dealing with optimization models that are used to find optimal facility locations (Daskin 1995; Mirchandani and Francis 1990). These optimization models try to provide answers concerning:

- The number and size of facilities
- Where the facilities should be located
- How demand for the facilities is allocated among them to minimize the cost or maximize the profit of satisfying the demand for a commodity

The problems usually involve fixed costs for locating the facilities and distribution costs for transporting the commodities between facilities and customers that are distance related.

A large class of facility location models assumes that facilities can be located on a network composed of nodes and links. Travel can occur only on the links of this network. This is to be contrasted with planar models, which can locate facilities anywhere on the plane. Often, facilities are characterized by capacities (e.g., warehouses) or throughput (e.g., transfer terminals). We present next a qualitative overview of some useful network location models, drawing primarily from Daskin (1995).

- *Set-covering problems:* The set covering problem finds a set of facilities of minimum cost from a finite set of candidate facilities (each with a given cost) so that every demand node is covered. A node is considered covered if at least one facility is located within a given distance of the node. The set-covering problem does not account for possible congestion in the facilities since it does not consider the number of demand nodes that are served by each facility or the size of the demand of each node.
- *Center problems:* The vertex P-center problem finds the locations of P facilities on the nodes of a network that minimize the maximum distance between a demand node and the nearest facility to the node. A better solution can be obtained if facilities are allowed to also be located on the links of the network, resulting in the absolute P-center problem. Center problems are appropriate for locating emergency services like fire fighting, emergency medical vehicles, etc.
- *Median problems:* The P-median problem finds the location of P facilities on a network that minimize total cost, where the cost of serving demands at a node is represented by the product of the demand at the node and the distance between the node and the nearest facility. It can be shown that at least one optimal solution of the P-median problem locates facilities only on the demand nodes of the network. Median problems are appropriate for locating nonemergency services like transportation terminals, post offices, etc.
- *Facility-location problems:* The uncapacitated facility-location problem finds the location of facilities (that have no capacity limitations) so that the total cost of locating the facilities and the operating costs of transporting a commodity between the facilities and clients are minimized. When each candidate facility has a capacity indicating the maximum demand that it can supply, the problem becomes the capacitated facility-location problem. The model is used to locate plants, warehouses, transportation terminals, etc. and is most appropriate for the private-sector type of problem, where both the costs of locating the facilities and the operating costs are borne by the same organization and can be made comparable.
- *Location/routing problems:* The facility-location problems described previously assume that each customer is served on an individual route from the facilities being located. This is sometimes inappropriate, especially for the transportation industry. Location/routing problems refer to problems that involve locating a number of facilities from a candidate set of facilities and establishing delivery routes so that the combined total cost is minimized. The decisions involved in such a problem may include: (1) determining the number and location of the facilities, (2) allocating customers to facilities, (3) assigning customers to routes, and (4) determining the sequence of serving the customers on each route. Location/routing problems are extremely difficult to solve. Problems need to be modeled as location/routing problems only if the stra-

tegic, long-term decision of locating facilities has to be made jointly with the tactical, short-term vehicle routing decisions. Otherwise, the problem can be broken into two separate problems, a location problem and a routing problem.

The problems presented above can be extended further when the facilities are not all similar but are organized hierarchically, resulting in hierarchical facility-location problems. Similarly, when multiple, and sometimes conflicting, objectives are present, multiobjective facility-location problems are obtained. Finally, many models exist that deal with the location of undesirable facilities (e.g., hazardous waste dumps) where instead of wanting to minimize, we want to maximize some measure of the distance between the demand nodes (e.g., population centers) and the facilities.

8. SUMMARY

Transportation management system software is finally getting the attention and recognition it deserves from logistics professionals. Software that aids shipment tendering and carrier selection has become an essential front-line tool in the battle to cut supply chain costs and bolster efficiency. More manufacturers now view TMS as a strategic extension of their enterprise resource planning (ERP) system and no longer delegate it to second-tier status.

The transportation component used to be seen as the stepsister of information technology systems. Only recently, the action was focused on warehouse management systems (WMS) to reduce inventories. Now it is transportation that is getting attention from manufacturers as the last corporate savings frontier.

Since becoming more sophisticated in the past few years because of real-time optimization capabilities (determining the best tender, given pricing and volume discounts, delivery schedules, and consignee), TMS packages will continue to be connected to a company's other business systems.

Interfacing transportation applications with an organization's other business enterprise systems has typically had a quick payback for most organizations. With annual logistics savings of 3–12% in usually less than one year, effectively implementing a TMS (Tausz 1999) for $300,000 and $1 million, as mentioned earlier, is a very economical decision.

Advances in operations research techniques and computing power have revolutionized transportation and shipment planning. Gone are the days of dividing the country into separate regions and assigning a freight planner to each one to find manually the best driver-to-load pairing within the region. Now many carriers are relying on global optimization systems that generate the best possible system-wide matching between drivers and loads, introducing new options to planners that were almost impossible to find using the traditional methods.

As supply chains and logistics cycles become more complex and diverse, transportation and distribution planners are forced to consider multiple points of origin and destination throughout the world. Reduced product life cycles stimulate time-based competition, and customers require better delivery service and inventory-reduction plans. In turn, managers are looking to optimize their transportation and logistics network. Thus, they will continue to define and deploy solutions aimed at finding the most economical means of transporting both inbound and outbound product via shipment and load planning, freight management, consolidation or pooling processing, accounting and analysis, and mileage and location tracking.

REFERENCES

Ball, M., Magnanti, T., Monma, C., and Nemhauser, G., Eds. (1995), *Handbooks in Operations Research and Management Science,* Vol. 8, *Network Routing*, North-Holland, Amsterdam.

Christofides, N. (1985), "Vehicle Routing," in E. Lawler, J. Lenstra, A. Rinnooy Kan, and D. Shmoys, Eds., *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*, John Wiley & Sons, New York, pp. 431–448.

Cooke, J. A. (1998), "Software Is IT!," *Warehousing Management Online,* May (accessed October 8, 1999).

Daskin, M. (1995), *Network and Discrete Location: Models, Algorithms and, Applications*, John Wiley & Sons, New York.

Dilger, K. A. (1998), "Strategic Moves,'' *Manufacturing Systems Online,* February (accessed November 26, 1999).

Ergun, O. (1998), "Tips for Embedding Operations Research into Transportation Management Software,'' *OR/MS Today Online,* December (accessed January 2, 2000).

Forger, G. (1999), "The Brave New World of Supply Chain Software," *Modern Materials Handling Online,* October 1 (accessed November 1, 1999).

Golden, B. and Assad, A., Eds. (1988), *Vehicle Routing: Methods and Studies*, North-Holland, Amsterdam.

Hall, R. W. and Partyka, J. G. (1997), ‘‘On the Road to Efficiency,’’ *OR/MS Today Online,* June (accessed January 2, 2000).

Lapide, L., and Shepherd (1999), ‘‘A Primer on Optimization,’’ Achieving Supply Chain Excellence through Technology, www.ascet.com/wp/wpLapide.html.

Lawler, E., Lenstra, J., Rinnooy Kan, A., and Shmoys, D., Eds. (1985), *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*, John Wiley & Sons, New York.

Lewis, I., and Talalayevsky, A. (1997), ‘‘Logistics and Information Technology: A Coordination Perspective,’‘ *Journal of Business Logistics*, Vol. 18, No. 1, p. 143.

Lustig, I. J. (1999), ‘‘Optimization: Achieving Maximum ROI within the Supply Chain,’’ Achieving Supply Chain Excellence through Technology, March, www.ascet.com/wp/wpLustig.html.

Michel, R. (1997), ‘‘Precision Movement,’‘ *Manufacturing Systems Online,* November (accessed November 26, 1999).

Minahan, T. (1997), ‘‘Information Emerges as Transportation’s Fifth Mode,’’ *Purchasing Online*, July 17 (accessed July 29, 1999).

Mirchandani, P., and Francis, R. (1990), *Discrete Location Theory*, John Wiley & Sons, New York.

Quinn, F., J. (1998), ‘‘Building a World-Class Supply Chain,’’
Logistics Management Online, June (accessed August 30, 1999).

Tausz, A. (1999), ‘‘Transportation Software Bypasses Secondary Status,’’ *Managing Automation Online,* December (accessed June 4, 2000).

Temple, Barker, and Sloane, Inc. (1982), *Transportation Strategies for the Eighties*, National Council of Physical Distribution Management, Oak Brook, IL, p. 6.

Weil, M. (1998), ‘‘Moving More for Less,’’ *Manufacturing Systems Online,* September (accessed January 2, 2000).

ADDITIONAL READING

Cooke, J. A. (1999), ‘‘A Tool for All Tasks,’’ *Logistics Management Online,* February (accessed November 26, 1999).

Optimal Planning Techniques website, www.optimalpt.com.

CHAPTER 80

Restructuring a Warehouse Network: Strategies and Models

Hokey Min
University of Louisville

Emanuel Melachrinoudis
Northeastern University

1. BACKGROUND 2070

2. WAREHOUSE-RESTRUCTURING STRATEGIES 2071

- 2.1. Warehouse-Centralization Strategy 2071
- 2.2. Warehouse-Decentralization Strategy 2071
- 2.3. Profile Analysis for a Strategic Choice 2072

3. MODEL DEVELOPMENT 2072

- 3.1. A Case Scenario of Warehouse Network Restructuring 2072
- 3.2. Model Formulation 2074
  - 3.2.1. Indices 2074
  - 3.2.2. Model Parameters 2074
  - 3.2.3. Decision Variables 2075
  - 3.2.4. Mathematical Formulation 2075
- 3.3. Model Application 2075

4. DECISION SUPPORT SYSTEM FRAMEWORK 2079

- 4.1. Data Management Subsystem 2079
  - 4.1.1. Cost Data 2079
  - 4.1.2. Traffic Data 2079
  - 4.1.3. Market Data 2079
  - 4.1.4. Local Incentive Data 2080
- 4.2. Model Management Subsystem 2080
- 4.3. Dialogue Management Subsystem 2080

5. CONCLUDING REMARKS AND FUTURE RESEARCH DIRECTIONS 2081

REFERENCES 2081

1. BACKGROUND

With the advent of the supply chain concept and the growing popularity of electronic commerce (EC), which transcend geographic and business functional boundaries, the traditional roles of warehouses are changing. Traditionally, warehouses have played three main roles in supporting logistics systems: a storage role, which makes long production runs more economical and bridges temporal gaps between demand and supply by making products available for customers on a timely basis; a consolidation role, which reduces total transportation cost by aggregating small orders into large shipments; and a customization (postponement) role, which reduces inventory carrying cost and enhances customer services by delaying the final stage of production and distribution until customers actually demand specific types of products (see, e.g., Maltz 1998). The relative importance or focus of warehousing roles to the entire supply chain network may vary from one warehouse to another due to dynamic shifts in the firm's business environments and strategic priorities. Such shifts include changes in supplier and customer bases, distribution networks, economic outlooks, corporate mergers/ acquisition, downsizing, and government legislation.

For instance, many e-businesses (dot-com ventures) are concerned with filling their customers' orders and keeping customers' deliveries up to speed. Consequently, they tend to decentralize ware-

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

housing operations and increase the number of regional warehouses, which enhances the firms' responsiveness to their online customers. Amazon.com, for example, recently has increased the number and capacity of its warehouses from 2 to a total of 15 facilities. The decentralization of a warehouse network, however, can be costly due to the duplication of safety stocks, reduced freight consolidation opportunities and greater warehousing costs. In an effort to control cost, some companies may downsize or reengineer their corporate structures that involve the consolidation and phase-out of some existing warehouses. According to Ballou and Masters' survey (1993) of 200 logistics executives, 65% of the respondents indicated that they intended to review their current warehouse network and consider restructuring it in the near future. The strategy of consolidating and centralizing warehouses can help the firm save transportation, inventory, and warehousing cost due to economies of scale. Indeed, Ballou (1999) observes that restructuring a warehouse network could generate annual savings of 5–10% of total logistics costs. In the next section, we will elaborate on the pros and cons of different warehouse-restructuring strategies.

2. WAREHOUSE-RESTRUCTURING STRATEGIES

2.1. Warehouse-Centralization Strategy

One of the most noticeable trends in today's warehousing operations is increased inventory velocity. According to the recent study conducted by Speh (1999), the surveyed firms showed the average increase of inventory turns by 30% from 1995 to 1998 and expected to improve inventory turns by 27% between 1998 and 2000. A vast improvement in inventory turns can be attributed to increasing adaptation of more effective inventory management practices such as cross-docking, cycle counting, improved forecast, radio frequency, automated ID systems, and warehouse management systems. Due to significant strides in increasing inventory turns, many firms no longer require a large number of stocking points and consequently allow them to consolidate existing stocking points into fewer locations. In general, warehouse-centralization strategy involves consolidation of regional warehouses into a smaller number of master stocking points and the subsequent phase-out of redundant warehouses. Along with better capacity (or asset) utilization and higher throughput of centralized warehouses, such a strategy often brings substantial amount of savings in warehousing and inventory carrying costs due to the reduced number of warehouses and aggregated inventory. For instance, the square-root rule of inventory consolidation allows the company to estimate the amount of savings in inventory investment as a result of warehouse consolidation. Assuming that the firm relies on economic order quantity (EOQ) rules for inventory management and all stocking points of the firm carry the same amount of inventory, the simplest form of the square-root rule can be mathematically expressed as (Ballou 1999, pp. 352–354):

$$I_T = I_i \sqrt{n}$$

where I_T = the optimal amount of inventory to stock, if consolidated into one location in dollars, pounds, cases, or other units of measurement
I_i = the amount of inventory in each of n locations in the same units of measurement as I_T
n = the number of stocking locations before consolidation

Another benefit includes reduced material-handling costs resulted from bulk storage and picking at centralized locations. Similarly, transportation cost can be reduced due to increased opportunities of large-volume shipments and the subsequent negotiation leverage for better freight rates. In addition, central administrative costs can be reduced through less effort being spent in managing fewer warehouses. On the other hand, warehouse-centralization strategy lengthens lead time and consequently causes deterioration of customer services. With fewer warehouses to serve markets, the accessibility of centralized warehouses to large segments of customer bases and major distribution hubs is critically important. Also, to offset increased distances between centralized warehouses and customers, this strategy may necessitate the more direct shipment of products from master stocking locations to end customers.

2.2. Warehouse-Decentralization Strategy

Forrester Research, Inc. expects that business-to-consumer online sales will reach $184.5 billion in 2004 and business-to-business e-commerce will grow to $1.33 trillion by 2003 (see, e.g., Massie 2000). The explosion of e-commerce will change the way the company distributes its products due to an increasing number of small, unpredictable orders for individual customers. E-commerce requires that warehouses deal with unit quantities of inventories rather than pallet loads of inventories and subsequently accommodate high frequency of small parcel delivery services. Because the role of warehouses has become more of flow-through transshipment facilities intended for quick order fulfillment and product return than traditional storage facilities, their locations need to be dispersed in wider geographical areas. In other words, warehouses that support e-commerce transactions are likely to be decentralized and located in proximity to each segmented market.

In general, warehouse-decentralization strategy aims to shorten customer response time and improve order-fill rates by positioning inventory locations at the lowest downstream of the supply chain (i.e., end customers). Such a strategy may make sense, particularly when online sales require direct-to-home or store delivery services on a quick-response basis. Despite its merits, the warehouse-decentralization strategy leads to increases in inventory carrying cost and warehousing cost due to a larger number of stocking locations. It can also increase transportation cost due to smaller, more frequent shipping requirements.

2.3. Profile Analysis for a Strategic Choice

To choose a warehouse-restructuring strategy properly between centralization and decentralization, the company may conduct a profile analysis. The profile analysis is intended to provide the management team with a tool to identify the supply chain needs and the degree of fit between the company's network-restructuring decision attributes (e.g., level of investment, time scales, market positioning) and the available restructuring strategy. The profile analysis is composed of four steps (Hill 1994):

1. Select the appropriate aspects of products (e.g., types of product, product range, life cycles) and markets (e.g., sales/promotional tools, delivery service requirements), logistics (e.g., volume shipment via consolidation), investment and cost (e.g., level of investment, level of inventory, shipping cost), and infrastructure (e.g., corporate culture and management styles).
2. Display the trade-offs of strategic choices (centralization vs. decentralization).
3. Develop the profiles of products and targeted market segments to see alignment between those profiles and the strategic choice.
4. Illustrate the degree of consistency between the characteristics of products/markets and the strategic choice. The straighter the profile, the more consistent the chosen strategy is with the characteristics of products/markets.

The excellent example of the profile analysis for the choice of logistics strategy can be found in the recent study conducted by Pagh and Cooper (1998).

3. MODEL DEVELOPMENT

3.1. A Case Scenario of Warehouse Network Restructuring

This section describes a case study of a firm that plans to restructure its warehouse network and reduce total logistics costs. The firm (called Beta hereafter) plans to consolidate 23 warehouses across the United States and Canada into a smaller number of warehouses, while offering a majority of its customers next-day delivery services. Considering hours-of-service regulations stipulated by the Federal Highway Administration (FHA), a majority of Beta's customers should be within 10 hours of driving time from nearest warehouses. Beta's primary mode of transportation is either less-than-truckload (LTL) or truckload (TL) carriers, and consequently compliance with such regulations is important for Beta's distribution operations. However, since Beta's restructuring plan entails the phase-out of some existing warehouses, truck delivery time to a group of isolated customers will likely increase. Therefore, the restructuring plan must smooth away transition to the centralized warehouses in such a fashion that it minimizes the total length of out-of-service dates and any potential disruption of supply chain activities during transition.

Beta has its main manufacturing plant in Terre Haute, Indiana, and currently operates 23 regional public warehouses to serve a total of 281 customers scattered around the United States and Canada (see Figure 1). Beta primarily manufactures and distributes rolls of films and related materials used for packaging such as clear wrappers on the outside of cigarette packages throughout North America. Beta's outbound distribution activities are supported by a single-echelon (or one-tier) warehouse network where all products move from a manufacturing plant through regional warehouses to end customers. To avoid any duplicated distribution efforts and redundant inventory investment with the preexisting warehouses in 14 different U.S. states and 3 Canadian provinces, Beta prefers to maintain a total of less than 15 regional warehouses. The rationale may be that some warehouse locations are within 100 miles of each other and a large number of volume customers are heavily concentrated in certain states and provinces such as Ohio, Wisconsin, Georgia, Illinois, Texas, North Carolina, New Jersey, California, and Manitoba (see Figure 2). Each state or province generates more than 1.2 million pounds of outbound shipping volume. Customers there account for more than 56.6% of the current outbound shipping volume. Seizing increasing market expansion opportunities on the West Coast and the Mexican and Canadian borders, Beta prefers to keep at least one regional warehouse for each of California, Texas, Winnipeg, and Montreal. Beta hopes that its restructuring plan can be

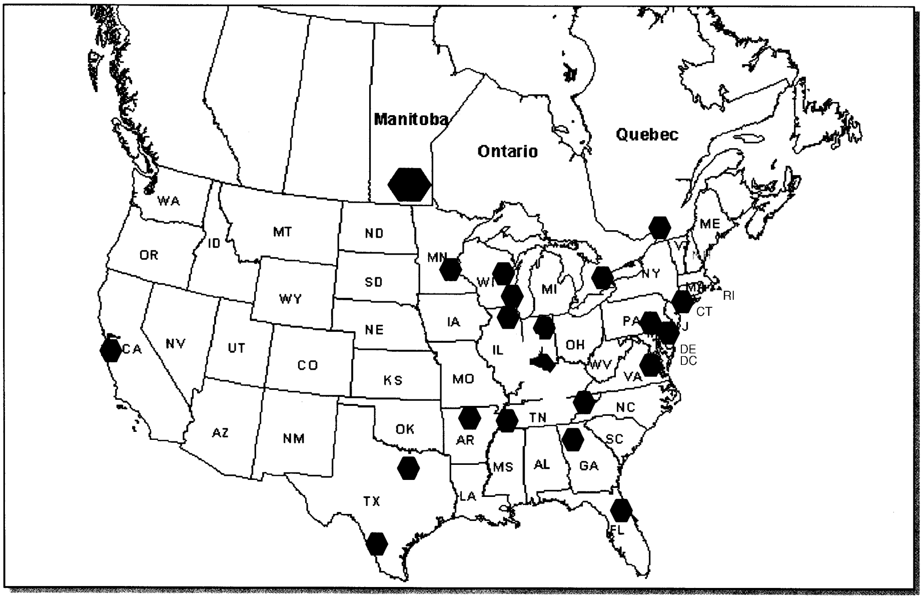

Manitoba
Ontario
Quebec
WA
MT
ND
MN
OR
ID
SD
WI
MI
NY
ME
VT
NH
MA
RI
CT
WY
IA
PA
NJ
NE
IL
IN
OH
DE
DC
CA
NV
UT
CO
KS
MO
WV
VA
KY
NC
TN
OK
AR
SC
AZ
NM
MS
AL
GA
TX
LA
FL

Figure 1 Existing Locations of Regional Warehouses.

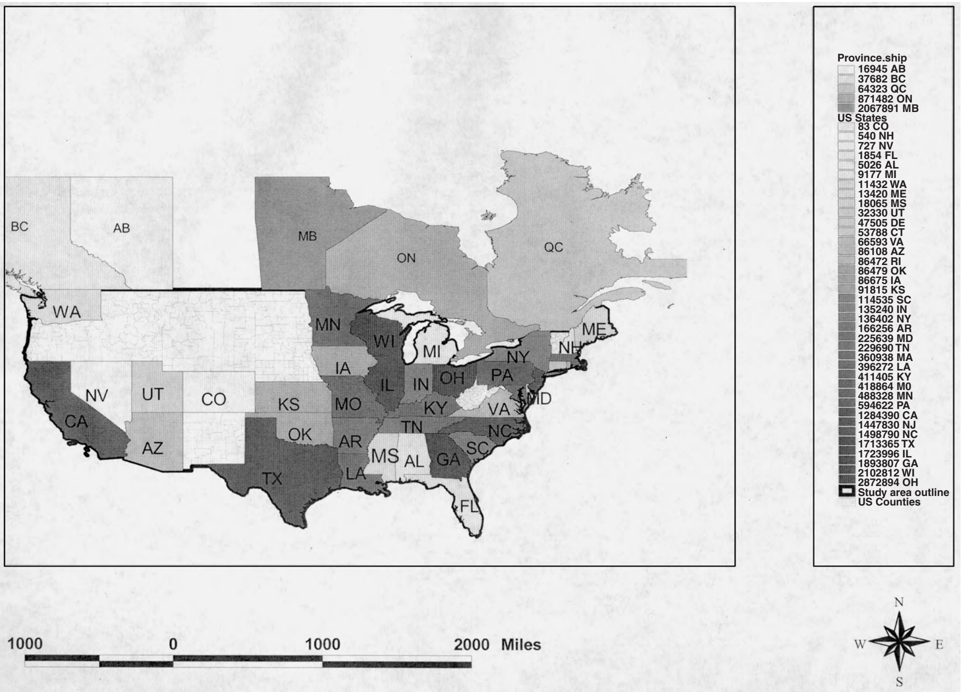

BC
AB
MB
ON
QC
WA
MN
WI
MI
NY
NH
ME
IA
IL
IN
OH
PA
MD
NV
UT
CO
KS
MO
KY
VA
CA
OK
AR
TN
NC
AZ
SC
MS
AL
GA
TX
LA
FL
Province.ship
16945 AB
37682 BC
64323 QC
871482 ON
2067891 MB
US States
83 CO
540 NH
727 NV
1854 FL
5026 AL
9177 MI
11432 WA
13420 ME
18065 MS
32330 UT
47505 DE
53788 CT
66593 VA
86108 AZ
86472 RI
86479 OK
86675 IA
91815 KS
114535 SC
135240 IN
136402 NY
166256 AR
225639 MD
229690 TN
360938 MA
396272 LA
411405 KY
418864 M0
488328 MN
594622 PA
1284390 CA
1447830 NJ
1498790 NC
1713365 TX
1723996 IL
1893807 GA
2102812 WI
2872894 OH
Study area outline
US Counties
1000
0
1000
2000 Miles
N
W
E
S

Figure 2 Customer Demand in Shipping Volume.

consistent with the corporate goals of increasing inventory turns and reducing transportation and warehousing costs.

The warehouse network-restructuring problem (WNRP) facing Beta differs from the classical warehouse-location problem in that the former is primarily concerned with determining which warehouses to retain and which warehouses to phase-out among the *existing* locations, whereas the latter is primarily concerned with selecting the optimal site among the alternatives of *new* locations. For excellent discussions of the classical warehouse-location problem, see Baumol and Wolfe (1958), Khumawala (1972), and Meidan (1978). On the other hand, the problems are similar in that both are influenced by the same attributes (or factors), such as warehouse operating cost, transportation cost, delivery access time, and proximity to major customer bases and transportation infrastructure.

In WNRP, each of Beta's consolidated facilities is expected to meet the current demand of all the existing customers and serve its customers within 10 hours of truck-driving time. Some of the existing warehouses, such as two in Terre Haute and one in Indianapolis, Indiana, that Beta considers eliminating may be redundant with their nearest warehouses due to close geographical proximity. But there is no guarantee that elimination of those sites will bring the substantial logistics cost savings without disrupting Beta's supply chain operations. For example, a regional warehouse that gives faster delivery advantages may turn out to be most costly because of its higher inventory taxes. Indeed, the local tax on inventories may vary significantly from one state to another. On the other hand, a warehouse that incurs the lowest cost and provides the best tax incentive packages may be distant from Beta's major customer bases and transportation infrastructure such as break-bulk terminals and major highways.

To deal with this dilemma, systematic decision-aid tools are needed that consider a multitude of conflicting factors affecting the restructuring plan and analyze trade-offs among them. Such decision-aid tools include various mathematical programming techniques such as integer programming (see, e.g., Bradley et al. 1977) and scoring methods such as the analytic hierarchy process (see Cohon 1978; Hwang and Masud 1979; Saaty 1980; Steuer 1986; Harker 1989; Vargas 1990 for excellent discussions of scoring methods). There are comparative advantages and disadvantages associated with the aforementioned decision-aid tools in terms of ease of use, data requirements, computational difficulty, and sensitivity analysis capability.

Considering that Beta's main objective of its restructuring plan is the maximization of a potential cost saving accrued from centralization of warehouses, we propose a single-objective, mixed-integer programming model as our decision-aid tool. The proposed model is designed to find the optimal number of warehouses in the restructured network under capacity limits and service requirements.

3.2. Model Formulation

Under the above scenario, the WNRP addresses the following issues:

1. Which warehouses to retain and which warehouses to eliminate in such a way that Beta's restructured warehouse network minimizes total cost associated with Beta's distribution operations while meeting current customers' demand and delivery service requirements
2. Which customers (or markets) to be served by which consolidated warehouses
3. How to evaluate the sensitivity of restructuring decisions with regard to changing priorities of Beta's restructuring plans

To address the above issues systematically, we develop a mixed-integer programming model that is formulated as follows.

3.2.1. *Indices*

k = index for manufacturing plants
i = index for warehouses
j = index for customers

3.2.2. *Model Parameters*

c_{ki} = cost of shipping unit product from manufacturing plant k to warehouse i
s_{ij} = cost of shipping unit product from warehouse i to customer j
v_i = variable cost of operating warehouse i
f_i = fixed cost of maintaining warehouse i
q_i = capacity of warehouse i
d_j = demand of customer j
t_{ij} = truck delivery time (in minutes) from warehouse i to customer j
τ = maximum daily hours of service regulated by FHA
$C(i) = \{j \mid t_{ij} \leq \tau\}$

$D(j) = \{i \mid t_{ij} \leq \tau\}$

M = is a large number

3.2.3. Decision Variables

x_{ij} = amount of products shipped from warehouse i to customer j

y_{ki} = amount of product supplied by plant k to warehouse i

$$z_i = \begin{cases} 1 & \text{if warehouse } i \text{ remains open} \\ 0 & \text{if warehouse } i \text{ is phased out} \end{cases}$$

3.2.4. Mathematical Formulation

$$\text{Minimize} \sum_k \sum_i c_{ki} y_{ki} + \sum_i v_i \sum_{j \in C(i)} x_{ij} + \sum_i \sum_{j \in C(i)} s_{ij} x_{ij} + \sum_i f_i x_i - \sum_i f_i (1 - z_i) \tag{1}$$

Subject to:

$$\sum_{j \in C(j)} x_{ij} \leq q_i \qquad \forall i \tag{2}$$

$$\sum_k y_{ki} = \sum_{j \in C(i)} x_{ij} \qquad \forall i \tag{3}$$

$$\sum_{i \in D(j)} x_{ij} = d_j \qquad \forall j \tag{4}$$

$$x_{ij} + y_{ki} \leq M z_i \qquad \forall k, i, j \in C(i) \tag{5}$$

$$x_{ij}, y_{ki} \geq 0 \qquad \forall k, i, j \in C(i) \tag{6}$$

$$z_i = (0, 1) \qquad \forall i \tag{7}$$

The objective function (1) minimizes total logistics costs composed of shipping costs and warehousing costs while maximizing cost savings resulted from the closure of redundant warehouses. Constraint (2) ensures that the total amount of products shipped to a group of customers does not exceed the capacity of a warehouse serving them. Constraint (3) ensures that the total amount of products supplied by all the manufacturing plants to each warehouse matches the total amount of products shipped from that warehouse to its customers. In other words, inbound shipping volume for each warehouse should be equivalent to its outbound shipping volume. Constraint (4) requires that customer demand be satisfied. Constraint (5) states that unless the warehouse remains open, it cannot serve its customers. Constraint (6) ensures the nonnegativity of decision variables x_{ij}, y_{ki}. Constraint (7) ensures the binary integrality of decision variables z_i.

3.3. Model Application

To demonstrate how the proposed WNRP model works and verify its usefulness, the model was applied to the real-world problem facing Beta. As explained earlier, the Beta management team intended to reduce the number of regional warehouses it was currently operating but had no idea of the ideal number of warehouses to retain. They considered keeping in the range of 7 to 14 warehouses. Although the smaller number of warehouses will reduce Beta's total logistics costs, Beta's management team would like to know the optimal number of warehouses sufficient accommodate current customer demand. Furthermore, Beta would like to offer consistent delivery services to geographically dispersed customers. Therefore, heavy concentration of consolidated warehouses in certain regions is not desirable for Beta's commitment to next-day delivery services. Regardless of cost-saving opportunities, Beta would like to maintain at least one regional warehouse in California and one in Canada.

Given the current warehouse network, however, we (analysts) discover that 11 customers isolated from other clusters of customers cannot be served within 10 hours of delivery time by existing locations of regional warehouses. Herein, we estimated delivery time using actual driving distance between different locations as opposed to Euclidean (straight-line) distance. Most of these isolated customers are located on the West Coast, such as Colorado, Utah, Arizona, Washington, and British Columbia. Since Beta currently has no plan either to relocate its warehouses or to open new warehouses, our task is to find the optimal number and geographic locations of consolidated warehouses that can best serve the remaining customers. In so doing, we run the model specified in Section 3.2

using LINGO's modeling language (LINGO 1995). LINGO is nonprocedural mathematical programming software that provides an environment where analysts can develop, run, and modify mathematical models interactively. Unlike other conventional software, it requires only *what* the modeler wants, rather than *how* it should find the solution (LINGO 1995). The model was run on a TD- 260 personal computer (PC). The total number of feasible warehouse–customer pairs considered by LINGO was 3625. The baseline LINGO model resulted in 23 binary integer variables, 3648 noninteger variables, and 3952 constraints. The execution time of the baseline model and the subsequent runs ranged from 50 seconds to approximately 3 minutes.

The optimal baseline solution suggests to retain consolidated warehouses in four different locations: Paterson, New Jersey; Winnipeg, Manitoba; Memphis, Tennessee; and San Leandro, California (see Figure 3). This solution makes sense in that it covers geographically dispersed areas while preventing any warehousing locations from being too close to each other. It is also congruent with the Beta's management team's wish that the restructured warehouse network include at least one Canadian and one West Coast location. However, in reality, a dramatic reduction in the number of warehouses from 23 to 4 may create confusion among Beta's common carriers, employees, and utility companies (e.g., phone companies) during transition. In addition, such an attempt will likely disrupt customer services during transition because it entails suspension of delivery services, reroutes of shipping, reorganization of load plans, and transfer of some inventory items from phased-out warehouses to consolidated warehouses.

To provide Beta's management team with greater sets of alternatives that allow them to keep the disruption risk to a minimum, we added a constraint that sets the total number of consolidated warehouses prior to running the model. This constraint is mathematically expressed as:

$$\sum_i z_i = p \tag{8}$$

where p = the desired number of consolidated warehouses. The above constraint is common in the prototypical p-median problem (see, e.g., Church 1974; Krarup and Pruzan 1983 for a detailed discussion of the p-median problem). For instance, by setting $p = 5$, we identified five locations of

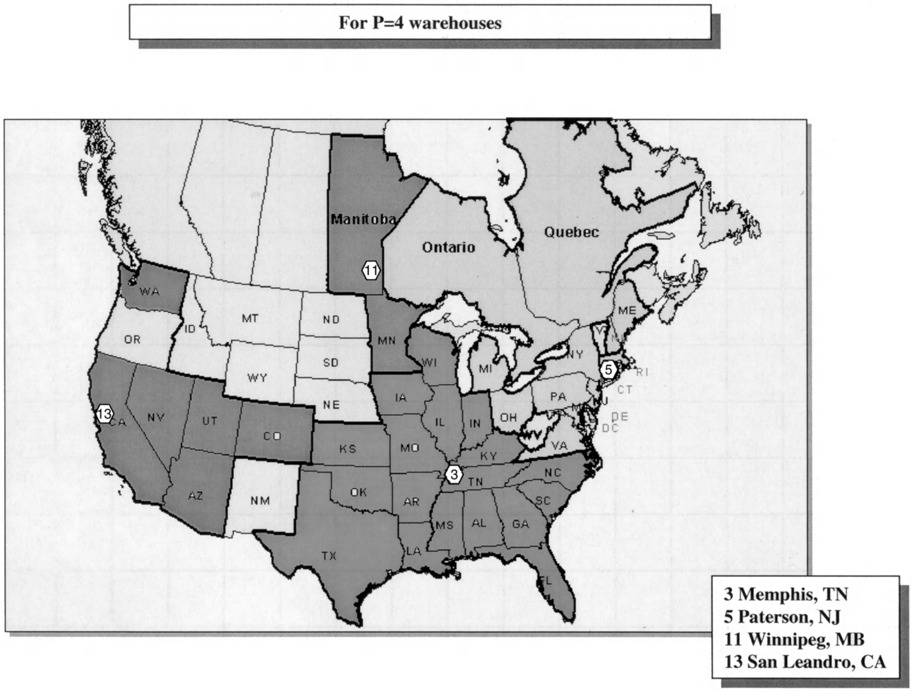


Figure 3 Locations of Consolidated Warehouses and Their Market Boundaries.

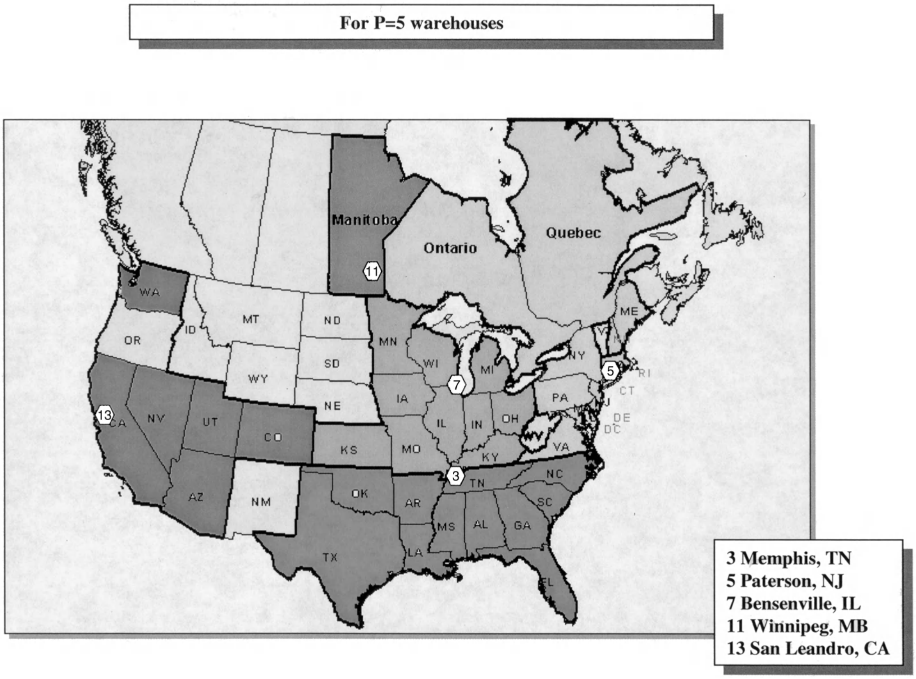


Figure 4 Locations of Consolidated Warehouses and Their Market Boundaries.

consolidated warehouses: Paterson, New Jersey; Winnipeg, Manitoba; Bensenville, Illinois; Memphis, Tennessee; and San Leandro, California (see Figure 4). When we set $p = 6$, the optimal solution suggested retaining six locations: Paterson, New Jersey; Winnipeg, Manitoba; Bensenville, Illinois; Atlanta, Georgia; Dallas, Texas; and San Leandro, California (see Figure 5). Finally, when $p = 7$ was set, we identified seven locations of consolidated warehouses: Paterson, New Jersey; Winnipeg, Manitoba; Toronto, Ontario; Bensenville, Illinois; Atlanta, Georgia; Dallas, Texas; and San Leandro, California (see Figure 6). It is interesting to note that these locations are either at the center or in the vicinity of the center of concentrated customer demand locations. As shown in Figures 3, 4, 5, and 6, among the eight states that originate customer shipping volume in excess of 1.2 million pounds, only three (Ohio, Wisconsin, and North Carolina) do not have consolidated warehouses. However, these states are still within next-day delivery areas from nearest consolidated warehouses such as the ones in Paterson, New Jersey; Atlanta, Georgia; Toronto, Ontario; and Winnipeg, Manitoba.

More interestingly, with the exception of warehouses located in Atlanta, Georgia, and Paterson, New Jersey, all the other warehouses to be retained for consolidation (i.e., Winnipeg, Manitoba; Toronto, Ontario; Bensenville, Illinois; Dallas, Texas; San Leandro, California) were substantially underutilized under the current warehouse network. According to the warehouse-utilization ratio (the number of pallets in the warehouse at the end of a year divided by the theoretical capacity of the warehouse) used by Beta, the Atlanta-based warehouse and the Paterson-based warehouse have an 11% and a 9% utilization ratio, respectively. On the other hand, those in Winnipeg, Toronto, Bensenville, Dallas, and San Leandro have 5% or less utilization ratios. Despite the sustained growth of customer demand in Texas and Mexican border areas, the Dallas-based warehouse has been virtually unused, as evidenced by a less than 1% utilization ratio. Such underutilization may have resulted from redundant warehouses located in Laredo, Texas, and Little Rock, Arkansas, which are within a 500-mile radius from Dallas. If Beta decided to reduce the number of warehouses to 7 from 23, each of the 7 consolidated warehouses would enjoy the average of an approximately 14% utilization ratio, which is a 10% increase from the current average utilization ratio of slightly over 4%.

As shown in Figure 6, the Paterson-based warehouse will serve customers located in Maine, Massachusetts, Rhode Island, New Hampshire, Connecticut, New Jersey, New York, Pennsylvania, Delaware, Maryland, and Virginia. The Winnipeg-based warehouse will serve customers located in Minnesota, Manitoba, and Alberta. The Toronto-based warehouse will serve customers located in

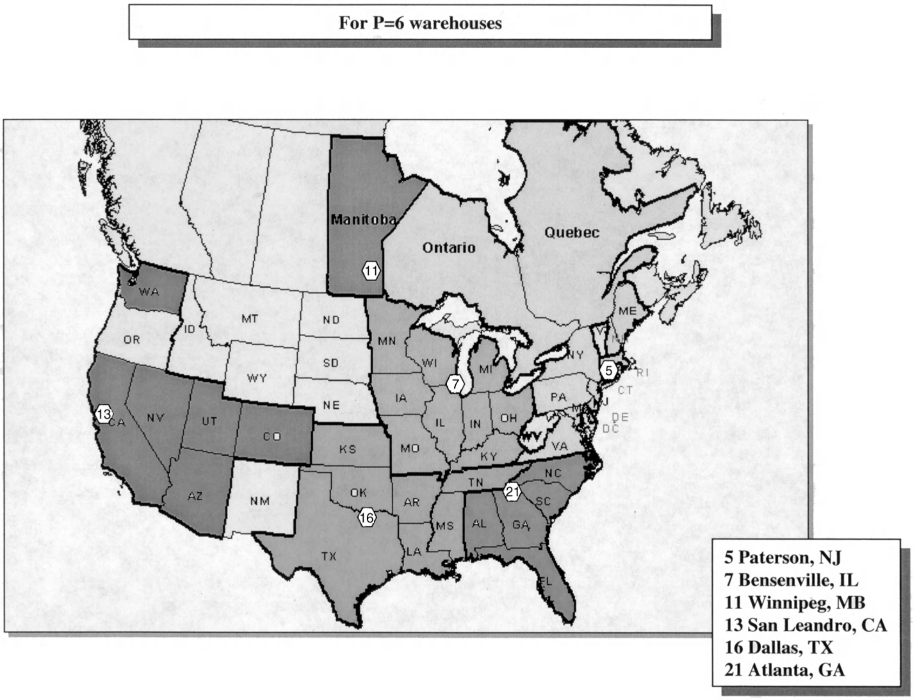

For P=6 warehouses
Manitoba
Ontario
Quebec
11
13
5
7
16
21
5 Paterson, NJ
7 Bensenville, IL
11 Winnipeg, MB
13 San Leandro, CA
16 Dallas, TX
21 Atlanta, GA

Figure 5 Locations of Consolidated Warehouses and Their Market Boundaries.

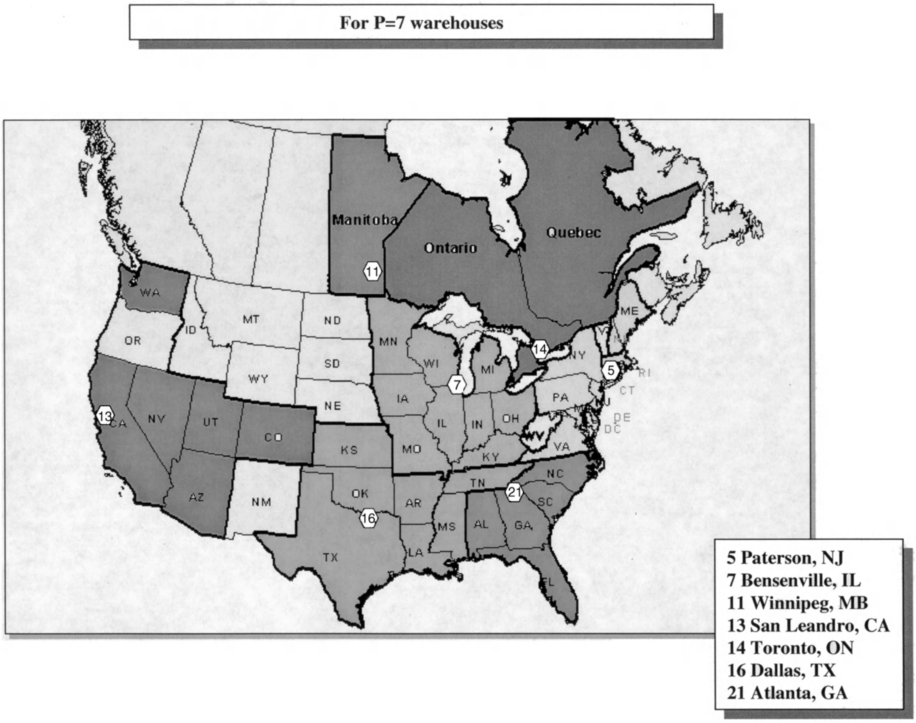

For P=7 warehouses
Manitoba
Ontario
Quebec
11
14
13
5
7
16
21
5 Paterson, NJ
7 Bensenville, IL
11 Winnipeg, MB
13 San Leandro, CA
14 Toronto, ON
16 Dallas, TX
21 Atlanta, GA

Figure 6 Locations of Consolidated Warehouses and Their Market Boundaries.

upstate New York, Ohio, Michigan, Quebec, and Ontario. The Bensenville-based warehouse will serve customers located in Kentucky, Ohio, Indiana, Michigan, Iowa, Wisconsin, Minnesota, Illinois, Missouri, and Kansas. The Atlanta-based warehouse will serve customers located in North Carolina, South Carolina, Georgia, Florida, Alabama, Tennessee, Kentucky, and Louisiana. The Dallas-based warehouse will serve customers located in Tennessee, Mississippi, Missouri, Kansas, Louisiana, Arkansas, Oklahoma, and Texas. The San Leandro-based warehouse will serve customers located in Colorado, Utah, Arizona, Nevada, California, Washington, and British Columbia. Notice that some warehouses cover overlapped states such as New York, Kentucky, Tennessee, and Louisiana due to heavy geographical concentration of customers in those states. Another thing to note is that, unlike other consolidated warehouses serving at least 30 different customers, the Winnipeg-based warehouse is allocated to serve only three customers. The possible explanation for such an aberration is that it has a huge customer in Winnipeg with its annual order volume in excess of 2 million pounds. That is to say, the Winnipeg-based warehouse is mainly dedicated to the one major customer and consequently its location is contingent upon the continuation of a business relationship with that customer.

4. DECISION SUPPORT SYSTEM FRAMEWORK

Even though the proposed model described in the previous section was designed to help Beta's management team find the minimum cost solution to the problem of determining which warehouses to retain and which warehouses to eliminate, it cannot capture all the complexities and dynamics of the WNRP facing Beta. As such, the proposed model should be embedded within the decision support system (DSS) framework, where the decision maker (e.g., Beta's management team) can freely add, delete, update, and modify the model objective functions, parameters, and constraints. Following three paradigms suggested by Sprague and Carlson (1982), this DSS framework has three components: data-management, model management, and dialogue management.

4.1. Data Management Subsystem

A model is only as good as the quality of the data that support it (Napolitano 1998). To enhance data quality and avoid data redundancy, we developed a database that contains three data sources: external, internal, and governmental. External sources include public data files available from local chambers of commerce, regional economic development agencies, *Site Selection Handbook*, *Industrial Development Magazine*, and websites such as Lycos Roadmaps. Internal sources include the Beta's order history files (e.g., sales shipment data, warehouse shipment data, ZIP code data, statistics reports), inventory record files, accounting data files, bills of lading, and internal customer policy manuals. Government sources include legal documents, regulatory guidelines, and reports issued by federal (e.g., Department of Transportation, Federal Highway Administration, Surface Transportation Board) and state agencies. In addition to raw data that can be obtained from the above sources, Beta may create more specific data categories that are relevant to WNRP. These categories are as follows.

4.1.1. Cost Data

Cost is one of the primary concerns of the WNRP decision. The total logistics costs involved in WNRP encompass annual rental fees for public warehousing, administrative expenses for general office personnel, inventory carrying cost (e.g., local inventory property taxes and insurance premiums on inventory), information-processing cost (e.g., maintenance of warehouse management systems), equipment-handling cost, cost of purchased labor, cost of unloading/loading vehicles, cost of palletizing/sorting, inbound/outbound shipping expenses, freight penalty incurred from delivery route changes, inventory transfer cost, cost of moving value-added services (e.g., packaging, labeling), out-of-service cost, and cost of changing service standards. For additional details of warehousing cost elements, see Speh (1990) and Ackerman (2000).

4.1.2. Traffic Data

Since Beta deals primarily with relatively heavy and bulky products using LTL and TL carriers, it needs to look for a warehouse location that can not only handle the large volume of traffic but can also provide easy access to transportation infrastructure and traffic-related services. Important traffic concerns include proximity to major interstate highways, any overhead construction impairing truck movement, curfew restrictions on hours of traffic operations, close distance to break-bulk terminals, and the availability of freight forwarding and brokerage services. In addition, Beta requires that the warehouse be within a one-day delivery range (approximately a 500-mile radius) of most of its major customers, including food packaging companies in the United States.

4.1.3. Market Data

The profitability of Beta depends heavily on the market potential of the area it serves. In a broad sense, the market potential of a chosen trading area is dictated by its business climate and competition level. Since Beta's business efforts are geared toward business-to-business transactions, an explicit measure of the local business climate is not readily available. As an indirect barometer for the local business climate, however, we can consider Beta's customers' customers' (e.g., grocery shoppers) aggregated effective buying power and buying power index in the area (metropolitan city or county) that surrounds the warehouse location. Herein, effective buying power is expressed as net gross personal income (= gross personal income − personal taxes − nontax payments). Buying power index is a measure of market ability to buy, expressed as a percentage of U.S. totals. In addition, Beta's past sales volumes and forecasted future business growth in Mexican, Canadian, and West Coast markets can be taken into consideration.

Under the premise that a majority of customers would gravitate toward the Beta's warehouse closest to their market centers, we used proximity to existing customers as part of a surrogate measure of Beta's competitive position in the trading area. Finally, since short inbound delivery to the warehouse can reduce order cycle time for the entire supply chain and subsequently help serve customers better, proximity to Beta's suppliers' locations can be considered as an indicator of Beta's competitiveness in the trading area. Given that the amount of rates (e.g., class rates) for the transportation of freight increases as mileage scales (e.g., rate basis numbers) increase, proximity to Beta's suppliers' locations affects inbound shipping cost.

4.1.4. Local Incentive Data

To increase the return on investment from the consolidated warehouse, local incentives should be taken into consideration. According to the recent Conway Data global survey of development organizations (Venable 1996), local incentives such as labor availability (e.g., annual unemployment rate), labor quality, tax breaks, and loans were considered one of the five most important location factors by thousands of economic development executives. In particular, the Beta management team is greatly concerned with ongoing labor shortage problems. For the last several years, high employee turnover and the subsequent labor shortage have been a key concern in labor-intensive warehousing operations because labor shortages can disrupt continuous distribution of products to customers. In addition, since Beta's product is an important component of distribution packages, which are regarded as a main source of waste, Beta should carefully review local environmental regulations before determining the consolidated warehouse location. Finally, Beta should look for a broad range of tax incentives (e.g., enterprise zone incentives, job creation tax credits or super tax credits for capital investment and job creation) around the consolidated warehouse.

4.2. Model Management Subsystem

The main focus of the model management subsystem is the incorporation of the WNRP model (problem generator) into the DSS. Although the WNRP model can be regarded as a "black box" whose algorithm and solution procedures need not be understood by the Beta's management team, it should be integrated with a problem solver such as LINGO. LINGO is standard, commercially available software designed to solve the mixed integer programming problem. To further enhance user-friendliness, we also integrated a geographic information system (GIS) into the model management subsystem. GIS simplifies the data display mechanism by separating data presentation from data storage.

Whereas the WNRP model represents esoteric numbers and symbols, GIS creates a wide variety of visual aids in the form of sound, images, and maps, illustrated in Figures 2, 3, 4, 5, and 6. In general, GIS allows Beta's management team to visualize how distinctive geographic information (e.g., customer shipping volume) from one location is from that from another by superimposing the information on a map. By recognizing geographical differences between one location and another, Beta's management team can choose the warehouse network that is most suitable for their needs. One of the most important advantages of combining the WNRP model with GIS is the enrichment of data quality, because GIS helps analysts visualize database errors that may otherwise go undetected. Particularly, GIS is capable of segmenting customer markets and measuring future market shares. Such an ability can aid the Beta's management team in developing the warehouse network that best serves customers.

4.3. Dialogue Management Subsystem

At best, the model is an abstraction of real-world situations. Consequently, it cannot capture reality without running it more than one time (Dyer and Mulvey 1983; Min 1989). Thus, the model should enable Beta's management team to evaluate "what-if" scenarios associated with shifts in Beta's management philosophy (e.g., a shift from cost minimization to quick-response services) and competitive positions (e.g., a shift from domestic to global operations). In other words, the model's successful implementation depends on its flexibility for contingency planning. To enhance the model flexibility, the results of model runs should be reported in user-friendly formats. These formats include standardized reports such as tables summarizing cost-saving opportunities and figures depicting maximum service radius as a function of the number of warehouses.

5. CONCLUDING REMARKS AND FUTURE RESEARCH DIRECTIONS

The warehouse is in the middle of the supply chain and therefore dictates the efficiency and effectiveness of the supply chain. With the increasing importance of a warehousing role in the supply chain, the warehouse network restructuring strategy can all but determine the success and failure of supply chain operations. This chapter introduces two distinctive warehouse network restructuring strategies: warehouse centralization and decentralization. These two strategies have their pros and cons, depending on the company's strategic focus. This chapter also develops a mathematical model that aims to provide a minimum-cost solution for the real-world warehouse restructuring (centralization) problem. Despite numerous merits, the proposed model points to a number of directions for future work:

1. The model can be expanded to include the element of risk and uncertainty involved in the warehouse-restructuring problem and can be tested for the expanded time periods.
2. The future research theme should include multiobjective treatments of the WNRP that explicitly analyze the trade-offs among cost, traffic access, market potential, and local incentives.
3. The multicommodity problem, which considers both slow-moving and fast-moving products, may be studied in the future.
4. The future network-restructuring problem should look into the possibility that the company will not only phase out some of the redundant warehouses but also relocate others to serve customers better.
5. Pure WNRP research can also continue by incorporating the combined usage of private, public, and contract warehouses into the mathematical modeling process.
6. The multiechelon hierarchical network configuration, which considers the options of both direct shipment from manufacturing plants to customers and indirect shipment through either master distribution centers or regional warehouses, may be an intriguing subject for further studies.

Acknowledgments

The authors wish to thank the senior management team and the vice president of the anonymous company for providing them with valuable data and assistance in the case study reported in this paper.

REFERENCES

Ackerman, K. B. (2000), *Warehousing Profitably: A Manager's Guide*, Ackerman, Columbus, OH.

Ballou, R. H. (1999), *Business Logistics Management: Planning, Organizing, and Controlling the Supply Chain*, Prentice Hall, Upper Saddle River, NJ.

Ballou, R. H., and Masters, J. M. (1993), "Commercial Software for Locating Warehouses and Other Facilities," *Journal of Business Logistics,* Vol. 14, No. 2, pp. 71–107.

Baumol, W. J., and Wolfe, P. (1958), "A Warehouse Location Problem," *Operations Research,* Vol. 6, pp. 252–263.

Bradley, S. P., Hax, A. C., and Magnanti, T. L. (1977), *Applied Mathematical Programming*, Addison-Wesley, Reading, MA.

Church, R. L. (1974), *Synthesis of a Class of Public Facilities Location Models*, Ph.D. Dissertation, Johns Hopkins University.

Cohon, J. L. (1978), *Multiobjective Programming and Planning*, Academic Press, New York.

Dyer, J. S., and Mulvey, J. M. (1983), "Integrating Optimization Methods with Information Systems for Decision Support," in *Building Decision Support Systems,* J. L. Bennett Ed., Addison-Wesley, Reading, MA, pp. 89–109.

Harker, P. T. (1989), "The Art and Science of Decision Making: The Analytic Hierarchy Process," in *The Analytic Hierarchy Process: Applications and Studies,* B. L. Golden, E. A. Wasil, and P. T. Harker Eds., pp. 3–36.

Hill, T. (1994), *Manufacturing Strategy: Text and Cases*, Richard D. Irwin, Burr Ridge, IL.

Hwang, C., and Masud, A. (1979), *Multiple Objective Decision Making-Methods Applications*, Springer, New York.

Khumawala, B. M. (1972), "An Efficient Branch and Bound Algorithm for the Warehouse Location Problem," *Management Science,* Vol. 18, No. 12, pp. B718–B733.

Krarup, J., and Pruzan, P. M. (1983), ‘‘The Simple Plant Location Problem: Survey and Synthesis,’’ *European Journal of Operational Research,* Vol. 12, pp. 36–81.

LINGO Systems Inc. (1995), *LINGO: The Modeling Language and Optimizer*, Chicago, IL.

Maltz, A. (1998), *The Changing Role of Warehousing*, Warehouse Education Research Council, Oak Brook, IL.

Massie, C. (2000), ‘‘Webhousing,’’ *WERC Sheet*, April, pp. 1–5.

Meidan, A. (1978), ‘‘The Use of Quantitative Techniques in Warehouse Location,’’ *International Journal of Physical Distribution and Materials Management,* Vol. 8, No. 6, pp. 347–358.

Min, H. (1989), ‘‘A Model-Based Decision Support System for Locating Banks,’’ *Information and Management,* Vol. 17, pp. 207–215.

Napolitano, M. (1998), *Using Modeling to Solve Warehousing Problems: A Collection of Decision-Making Tools for Warehouse Planning and Design*, Warehousing Education and Research Council, Oak Brook, IL.

Pagh, J. D., and Cooper, M. C. (1998), ‘‘Postponement and Speculation Strategies: How to Choose the Right Strategy,’’ *Journal of Business Logistics*, Vol. 19, No. 2, pp. 13–33.

Saaty, T. L. (1980), *The Analytic Hierarchy Process*, McGraw-Hill, New York.

Speh, T. W. (1990), *A Model for Determining Total Warehousing Costs for Private, Public and Contract Warehouses*, Warehousing Education and Research Council, Oak Brook, IL.

Speh, T. W. (1999), *Warehouse Inventory Turnover*, Warehousing Education and Research Council, Oak Brook, IL.

Sprague, R. H., and Carlson, E. D. (1982), *Building Effective Decision Support Systems*, Prentice Hall, Englewood Cliffs, NJ.

Steuer, R. (1986), *Multiple Criteria Optimization: Theory, Computation, and Application*, John Wiley & Sons, New York.

Vargas, L. G. (1990), ‘‘An Overview of the Analytic Hierarchy Process and Its Applications,’’ *European Journal of Operational Research,* Vol. 48, pp. 2–8.

Venable, T. (1996), ‘‘The New Business Location Process: Who’s Driving, and What’s Steering?’’ *Site Selection,* Vol. 41, No. 2, pp. 436–437.

CHAPTER 81

Warehouse Management

Gunter P. Sharp
Georgia Institute of Technology

1. INTRODUCTION **2083**
2. FUNCTIONAL DESCRIPTION OF WAREHOUSE OPERATIONS **2084**
  2.1. Functional Structure of Warehouse Operations 2084
  2.2. Classification of Warehouses Based on Mission 2085
    2.2.1. Factory Warehouse 2085
    2.2.2. Retail Distribution Warehouse 2085
    2.2.3. Catalog Retailer 2086
    2.2.4. Support for Manufacturing Assembly 2086
    2.2.5. Some Terminology 2087
3. STRATEGIC AND TACTICAL FACTORS IN WAREHOUSE OPERATION **2087**
  3.1. Classification by Implementation Time 2087
  3.2. Operational Planning 2088
    3.2.1. Space-Planning and Space-Adjustment Factors 2088
    3.2.2. Individual Product Assignment to Storage Positions 2090
    3.2.3. Forward–Reserve Allocation 2093
    3.2.4. Zoning and Batching, Order Retrieval 2093
    3.2.5. Pick Wave Planning 2095
4. DATABASE CONSIDERATIONS **2095**
  4.1. Products and Orders 2097
  4.2. Flow Control 2097
  4.3. Building, Equipment, and Personnel Assets 2097
  4.4. Operating Rules 2102
  4.5. Links to Hardware Controllers 2103
  4.6. Data Backups 2103
5. DAILY OPERATIONAL FACTORS IN WAREHOUSE OPERATION **2103**
  5.1. Receiving Operations 2103
  5.2. Storage and Inventory Control 2104
  5.3. Order Processing 2104
  5.4. Order Picking 2104
  5.5. Order Consolidation 2106
  5.6. Additional Factors 2107
6. CONCLUSION **2108**
REFERENCES **2108**

1. INTRODUCTION

Modern warehouses operate with sophisticated equipment for storage, handling, data capture, and communication. A competitive environment requires warehouse managers to have tight control over inventories and orders shipped, with rapid response being the norm. At the same time, efficient strategies for storage and retrieval offer significant operating savings. To satisfy these objectives, an effective warehouse management system (WMS) is essential. A WMS may be a combination of paper-based systems together with computerized inventory records, but the trend clearly is toward an integrated, computer-based system that handles management and automatic identification. The major benefits of a WMS are typically the following:

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

- Improved inventory accuracy, which allows for higher order-fill rates and faster response times
- Increased efficiency as a result of eliminated operations, reduced deadhead travel of workers and vehicles, and sharing of workers across departments
- More timely replenishments of forward pick areas, which result in higher order fill rates
- Better pick wave planning and execution in zone pick systems, which result in better worker utilization and earlier completion of the picking activity

This chapter presents the main concepts of warehouse management. First, there is a functional description of a typical warehouse operation, with emphasis on order picking because that is where most of the labor costs in a warehouse are incurred. This is followed by a discussion of strategic and tactical factors for warehouse operation, and then a discussion of database considerations for WMS. The last part of the chapter describes how users interact with a WMS and what functions they should expect it to perform. The purpose here is not to describe how the WMS is structured, since that varies with the software vendors, but rather to present a user's viewpoint of the major aspects of the system.

A distinction should be made among three different aspects of use of a WMS: planning, executing, and reporting (Rouwenhorst et al. 2000). Once a system is installed, effective reporting feeds back to planning. Although most users focus on the execution aspect, the planning aspect is just as important since that involves strategic (more than six months in the future) and tactical decisions (one to six months in the future) that affect operational effectiveness just as much as short-term planning and execution do.

2. FUNCTIONAL DESCRIPTION OF WAREHOUSE OPERATIONS

2.1. Functional Structure of Warehouse Operations

A functional structure of a warehouse with order-picking activities, as shown in Figure 1, consists of at least 8 functional areas (departments) and more than 15 material flows (represented by arrows) with various load types (Yoon and Sharp 1996). This functional structure is a road map, and not all facilities will contain all departments and flows. The major departments can be categorized as breakdown or consolidation area, based on major functions, and as storage or transfer area, based on the duration of product stay:

- *Receiving area:* incoming shipments are unloaded, and inspected if necessary (transfer area).
- *Pallet reserve area:* an area where products are stored and retrieved in whole pallet quantities, without pallet breakdown (storage area).
- *Carton pick area:* an area from which products are retrieved in carton quantities. Incoming loads and storage units are usually pallets, but may also be cartons and mixed unit loads (storage area).
- *Item pick area:* an area from which products are retrieved in item (less-than-carton) quantities. Incoming loads and storage units are often cartons, but may also be totes and items (storage area).
- *Sorting area:* an area where different items of an order are consolidated, if this function is needed because of orders being split into suborders for picking efficiency (transfer area).
- *Consolidation area:* an area where the different items, cartons, and totes belonging to an order are unitized, such as into a shrink-wrapped pallet (transfer area).
- *Shipping area:* an area where outgoing items are checked and loaded onto vehicles (transfer area). Consolidation and shipping areas are often combined.
- *Auxiliary areas:* may include labeling, repackaging, and processing of items returned from customers.

Material flows can be classified as order and replenishment flows. For replenishment the relevant flows are from receiving to pallet reserve, carton pick, and item pick areas; from pallet reserve to carton pick area; from carton pick to item pick area. The other flows, for customer orders, are relatively more frequent. Consider the flow of a typical high-activity product, which is received in pallets and stored in the pallet reserve area (e.g., the upper levels of pallet rack). It is then moved to the carton pick area (e.g., the lower levels of pallet rack). Individual cartons are removed from the carton pick area and placed in the item pick area (e.g., a gravity flow rack holding cartons). The order pickers selectively retrieve items from the flow racks and place them on conveyors, which carry the items to the sorting area. There the items of each order are sorted into totes, which are then sent to the consolidation area and then to shipping. Some customers may order this product in carton quantities, which may be retrieved directly from the carton pick area. If an order requests a quantity equivalent to 4.2 cartons, then 4 cartons may be retrieved from the carton pick area and the remaining

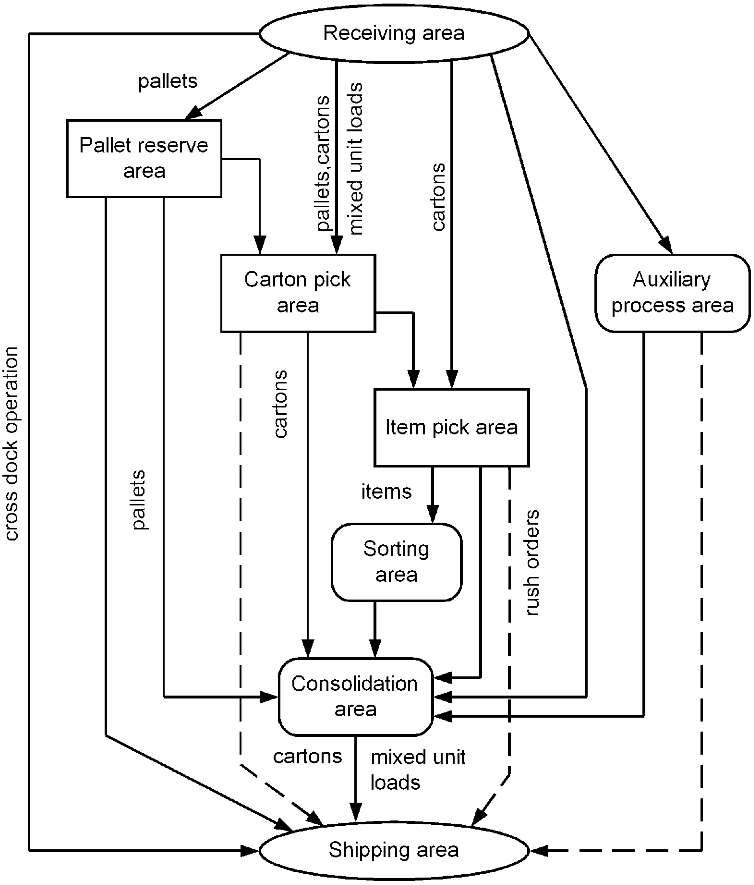

Receiving area
pallets
Pallet reserve area
pallets,cartons
mixed unit loads
cartons
Carton pick area
Auxiliary process area
cross dock operation
cartons
Item pick area
pallets
items
rush orders
Sorting area
Consolidation area
cartons
mixed unit loads
Shipping area

Figure 1 Functional Structure of a Warehouse with Order Picking.

item quantity from the item pick area. Similar logic applies to full pallet quantities of a product on an order.

2.2. Classification of Warehouses Based on Mission

Warehouses differ greatly in purpose and size, even within the same organization, so it is useful to examine some popular types. The purpose and size of a warehouse influence the factors that are important to the user of a WMS in the facility.

2.2.1. Factory Warehouse

A factory warehouse supplying wholesalers might process a small number of large orders daily with advance information about the composition of orders. The definition of "small number of orders" is based on the number of consolidation/shipping lanes: if during a process cycle, such as a shift, a packing/shipping lane can be dedicated to an order, we say the number of orders is "small." The definition of "large order" refers to the number of different products on an order: if there are 10 or more different products, we say the order is "large." The operating criteria in such a facility are usually cost and accuracy. Response time is often dependent on production scheduling in the factory and thus isusually beyond the control of the warehouse operator. The type of picking is usually pallet retrieval and carton picking, with emphasis on pallets.

2.2.2. Retail Distribution Warehouse

A warehouse serving captive retail units usually has advance information about the composition of orders. Typically, there are more orders per shift than consolidation/shipping lanes. Picking is usually

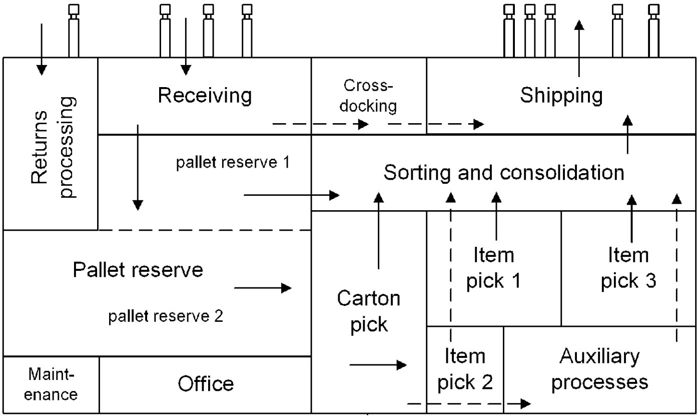


Figure 2 Example Layout of a Retail Distribution Warehouse.

carton and item picking, and usually there is a forward pick area (see Figure 2). Criteria are usually cost, accuracy, and fill rate (fraction of items requested that are actually shipped with the main part of the order). Response times are usually fixed and heavily dependent on truck routing schedules. If the retail units are not captive (not constrained to order from the particular warehouse), then response time becomes an important criterion.

2.2.3. Catalog Retailer

A catalog retailer typically processes a large number of small orders. While the definition of "small order" refers to fewer than 10 products, in the catalog retail business it is often closer to 2 or even 1 product per order. The number of orders per day may in the hundreds or even thousands. During the early part of the shift the composition of orders is usually known. Orders that arrive late during the shift may also be rushed through the system. For planning labor and replenishment of the forward pick area there is only statistical information available on these late-arriving orders. The primary criterion is often cost, but response time is also important. The picking here is usually item picking, with some carton picking.

2.2.4. Support for Manufacturing Assembly

A stockroom serving a manufacturing facility might process many small order with perhaps only statistical information about the composition of orders. If the response time is on the order of 30 minutes, then in effect any daily planning must be based on statistical information. The primary criterion is often response time, with cost a secondary criterion. The major types of picking may include item picking and carton picking.

Precise definitions of the descriptors "small" and "large" are not possible. For example, order size can be determined by cubic volume and/or number of different products on the order (number of line items). It is suggested that a small order be defined as one that contains fewer than 10 line items. If the quantity per line item results in a cubic volume less than 0.01 m^3 (0.35 ft^3), then a small order is also limited roughly to 0.1 m^3 (3.5 ft^3). A large order, containing 10 or more line items, would often have a volume greater than 0.1 m^3. In some situations, a large order might fill a truck. These definitions allow for some awkward, in-between situations. The demarcation based on the number of line items may be more useful because it relates more to operating strategies than a demarcation based on cubic volume.

The difference between advance and statistical information relates to the ability to process the order data for more efficient operation. In the example of a stockroom serving a manufacturing facility, the requirement for fast response (such as 30 minutes) probably would preclude the types of batching strategies used by a catalog retailer, who usually has several hours at night for data processing and a late-afternoon shipping deadline. If there is adequate time for preprocessing, we say there is advance information. Otherwise, the information is statistical. The information availability is directly related to the control rules in the operation of the OPS, since planning and execution based on statistical information may be different from that based on advance information.

2.2.5. Some Terminology

The language used by warehouse designers and operators contains some commonly used terms that need to be defined:

SKU: stock-keeping unit, the type identification of a product for purposes of distribution; for example, Coca-Cola Classic may be sold in 2-liter bottles packed 6 to a carton, 1-liter bottles packed 12 to a carton, 12-ounce cans packed 24 to a carton, 12-ounce cans arranged as 4 6-can inner packs in a carton, and 12-ounce cans packed 24 to a flat; these would all be different SKUs.

Item: the smallest unit of a product sold by the distribution center if many items are contained in a carton; sometimes called piece or each; examples are a 1-liter bottle of soft drink, a 12-pack of inexpensive ball-point pens, a 4-pack of spark plugs, a carton of cigarettes (containing 200 cigarettes in 10 packs of 20 each), a gift pack of stationery and envelopes.

Carton: a paperboard container holding identical product; usually of a size and weight allowing manual handling; example dimensions are 14 × 10 × 20 in. and 300 × 200 × 400 mm; recent trends in some industry sectors are towards a flimsier container, often called a flat, where the upper portion is plastic wrap.

Inner pack: several units of a product secured together (usually by paper or cellophane wrap) and sold by the distribution center as a unit if many items are contained in a carton and purchase quantities per item are large; a carton contains several inner packs.

Pallet: a set of cartons or totes of identical product arranged in a cubical pattern and usually supported by a base that may be of wood or plastic; example dimensions are 40 × 48 × 54 in. and 800 × 1200 × 1000 mm.

Mixed unit load: a set of cartons or totes of different products arranged in a cubical pattern similar to a pallet, often wrapped or strapped for stability.

Overpack: a large carton or tote containing different products; smaller than a pallet but larger than a carton so that manual handling may be difficult.

Tote: a container usually made of plastic and often used for storing and handling different products; usually similar in size to a carton; may be nestable or collapsible, with or without lid.

Order: a document from a customer requesting specific SKUs in specific quantities.

Line item: a “line” on an order document relating to a specific SKU and quantity.

Zone: a part of a distribution center to which an order picker is restricted; an example is a 40-aisle system divided into four zones of 10 aisles each, or an aisle-captive person-aboard system with 6 aisles and thus six zones.

Suborder: the portion of a customer order relating to a zone in the distribution center.

Batch: a set of suborders in a zone assigned to an order picker.

Time window: a portion of a day during which an order is processed; for example, an eight-hour day may be split into four two-hour time windows such that an order is processed completely during one of the time windows.

Pick wave: the set of orders processed during a time window; waves are often associated with zone picking.

Order class: a subset of the daily orders with some common characteristic, such as the same customer type, same size of order, or same shipping method.

Product group: a subset of the products in a distribution center with some common characteristic, such as vendor, product type, activity, or brand.

I/O point: the input/output point of a system.

S/R: storage/retrieval.

ASN: advance shipping notice, issued by a supplier to inform the warehouse operator of a shipment to arrive in the near future, usually transmitted electronically.

3. STRATEGIC AND TACTICAL FACTORS IN WAREHOUSE OPERATION

3.1. Classification by Implementation Time

We define strategic factors as those with implementation six or more months in the future. Usually these involve design changes in the physical or operating system where procurement and installation take that much time. Examples include:

- Changing the holding capacity and processing capacity of a functional area
- Changes in hardware for storage, vehicles, conveyors, control system

Tactical factors are defined as those with implementation one to six months in the future. Examples include:

- Modifications to existing storage and handling equipment and major movements of products to new storage or pick positions that cannot be performed with in-house labor on nights and weekends
- Changes in operational rules for storage, retrieval, sorting, and merging that require software modifications
- Negotiating changes in labor mix, that is, the balance between permanent and temporary employees, the balance between regular and overtime

3.2. Operational Planning

In addition to these strategic and tactical factors above, operational planning must be performed on a monthly or more frequent basis. This includes the examples below:

- Reclassification of products based on activity and inventory requirements, including the phase-in of new products and deletion of obsolete ones
- Restructuring of pick zones
- Changing operational settings in the WMS to accommodate a change in volume or change in the mix of orders
- Volume forecasting and labor planning
- Retraining of employees based on performance measures, such as productivity and quality

3.2.1. Space-Planning and Space-Adjustment Factors

The determination of storage space requirements is one of the most fundamental functions in warehouse management. Space utilization in a storage system is affected by several factors:

- *Volumetric efficiency,* F_a, in placing a product load in a storage compartment. This may result from undersized loads and from obstructions in the rack area, such as columns and sprinkler pipes.
- *Utilization,* F_b, of a set of storage compartments by a product lot as withdrawals are made. Deep-lane storage systems are particularly affected by this phenomenon, although it also affects single-deep pallet rack and shelving systems when withdrawals are in small quantities.
- Number of *storage compartments not assigned* for operational flexibility, such as performing dual-command S/R operations. It is usually impossible to occupy every storage position and maintain any type of operational discipline. Some fraction of the spaces, say 5–10%, need to be empty to provide opportunities for activity-based storage when incoming loads arrive. The resulting factor, F_c, will be 90–95% when applied to maximum product storage requirements. For average product storage requirements the factor usually does not apply and can be set to 100%.
- *Cyclical fluctuations* in storage requirements, resulting in factor F_d. Most businesses experience inventory peaks at a particular time(s) of the year. If one designs for the peak (or 90% of the peak) requirement, then at other times of the year the system will be underutilized. One can minimize this effect by renting space from third-party providers during peak inventory periods, by smoothing out deliveries and shipments with incentives, and by more frequent inventory replenishments during peak inventory periods.

The volumetric efficiency, F_a, depends on the matching of load sizes with storage compartment sizes. The storage compartment size must allow for clearances and some load deformation; thus, 100% utilization of the storage compartment would reflect the minimum clearances.

In one case study, pallets in a manufacturer's warehouse were limited in height to 127 cm (50 in.) based on rack spacing, but the average pallet height was 112 cm (44 in.) (Sharp et al. 1994). This gave a factor value of

$$F_a = \frac{112}{127} = 0.88$$

In addition, 50% of the pallets in storage required a height less than 102 cm (40 in.), so for these the factor was

$$F_a = \frac{102}{127} = 0.80$$

Partial pallets that are received and stored without consolidation contribute to low factor values. Columns and other obstructions may cause a loss of 5% or more of the available space. This factor applies to the maximum product storage requirement.

The utilization of a set of storage compartments by a product lot depends on the number of compartments assigned and the withdrawal pattern. For example, if an incoming lot consists of three pallets of identical product, the product is stored in a single-deep rack system, and carton withdrawals occur uniformly, then the average compartment utilization is calculated as shown in Table 1. The calculations assume the ability to reassign empty spaces immediately to other products. The result here is a utilization of 75%, based on the average product storage requirement. If we were to calculate the utilization factor based on maximum product storage requirement, it would be 1.5/3.0 = 50%.

For deep-lane systems the calculations are similar but more involved. Often there is a mismatch between an incoming product lot and the available slots, despite the best efforts of the warehouse designer. One first needs to calculate the optimum lane depth, assuming that empty lanes can be reassigned to other products immediately. Table 2 shows a typical situation in floor stacking, where an incoming lot of 15 pallets may be assigned to a lane depth up to 5, with a stacking height of 3. The best situation requires a floor space of 6.3 m^2, which corresponds to floor space occupied by product (not counting aisle space) of either 3.6 or 4.2 m^2. The incoming lot of 15 pallets requires 5 floor spots of 1 m^2 each, or 5 m^2. The average inventory is 8 pallets (the time for zero inventory is assumed to be negligible), and these require an average of 8/3 = 2.67 m^2. Thus, the utilization based on average product storage requirement is 2.67/3.6 = 74% (for the 2-deep lanes), and 2.67/5 = 53% based on maximum product storage requirement (Tompkins et al. 1996). This underutilization of deep-lane systems is called honeycombing.

The cyclical fluctuations work together with the method (or lack) of shared storage.

Consider the forecasts of inventory requirements (in pallets) over eight time periods for 10 products shown in Table 3. The 10 products have been classified into two groups based on chemical compatibility, and within each group any pallet may occupy any storage position (shared storage). If there is no reorganization of the storage system over the eight time periods, then the positions needed for the groups are 32 and 20, respectively, for a total of 52 pallet positions, as shown in column 15 of Table 4. If the spaces are reorganized every two time periods, then the requirements decrease to a total of 44, for savings of 15%. In this example the reorganization might require some special cleaning to avoid chemical incompatibilities with residues left in the storage compartments. For the case of reorganizing every 8 periods, the sum, over products, of the average storage requirement is 36.75. So the utilization, based on average product storage requirement, is 36.75/52 = 71%; for reorganization every 2 periods, the factor is 36.75/44 = 84%. This dependence of utilization based on reorganization interval is examined in Sharp et al. (1999).

TABLE 1 Method of Calculating Average Storage Compartment Utilization

Col. 1 Number of Spaces Occupied	Col. 2 Percent of Time Number Occupied	Col. 3 (col.1)(col.2)	Col. 4 Average Product Quantity in Storage	Col. 5 (col.4)(col.2)
3	0.333	1.000	2.500	0.833
2	0.333	0.667	1.500	0.500
1	0.333	0.333	0.500	0.167
	Total	2.000	Total	1.500

Average utilization, F_b = 1.500/2.000 = 0.75

TABLE 2 Lane Depth Analysis, Incoming Lot Size = 15 Pallets
Stacking height = 3, pallet dimensions are 1 × 1m, including horizontal clearances

Lane Depth		Modular Width		M^2 per	Lanes Needed		Average m^2	Floor Space
Pallets	Meters	Pallets	Meters	Lane	Maximum	Average	Occupied	Product
1	2.5	1	1	2.5	5	3	7.5	3
2	3.5	1	1	3.5	3	1.8	6.3	3.6
3	4.5	1	1	4.5	2	1.4	6.3	4.2
4	5.5	1	1	5.5	2	1.2	6.6	4.8
5	6.5	1	1	6.5	1	1	6.5	5

Explanation of averages, assuming smooth reduction of inventory:
For 1-deep, each of the 5 situations has equal probability, so average = 3.
For 2-deep, 3 lanes occurs 20% of the time, 2 lanes occurs 20 + 20 = 40% of time.
and 1 lane occurs 20 + 20 = 40% of time. So average is (3)(0.2) + 2(0.4) + 1(0.4) = 1.8.
For 3-deep, 2 lanes occurs 20 + 20 = 40% of time, 1 lane 60%, so average = 1.4.
For 4-deep, 2 lanes occurs 20% of time, 1 lane 80%, so average = 1.2.
For 5-deep, we always need the entire lane.
The time when no lanes are needed is assumed to be negligible.

The overall effect of these four factors can be quite shocking. Using values similar to those in the examples, we might have an overall utilization of storage space, based on average product storage requirements, equal to

$$F_a * F_b * F_c * F_d \;=\; 85\% * 75\% * 100\% * 80\% \;=\; 51\%$$

Based on maximum product storage requirements, the overall value would be much lower, depending on the method of shared storage. The above examples point to the need for the WMS to report information that enables the warehouse manager to estimate space requirements in a manner consistent with the ability to reallocate that space. Although these types of decisions are often thought of as being part of warehouse design, the ever-changing business conditions in the economy imply that they are also part of warehouse management.

3.2.2. Individual Product Assignment to Storage Positions

A storage rule is defined as a rule for assigning each SKU to storage locations. A primary objective of warehouse design and operation is to minimize the average process time per storage/retrieval (S/R) activity or per order. Even if the total cost of order picking may be the ultimate concern, a significant portion of the total cost is usually proportional to S/R time. Storage rules can be grouped into two broad classifications:

TABLE 3 Example Forecast, 10 Products in Two Groups

		Period							
Product	Group	1	2	3	4	5	6	7	8
1	1	4	5	6	7	9	0	3	4
2	1	7	9	11	0	1	3	5	6
3	1	9	10	0	1	2	5	7	8
4	1	12	0	2	3	4	7	8	10
5	1	0	6	6	4	3	3	2	1
6	2	4	0	0	1	1	2	3	4
7	2	0	1	1	2	3	4	7	9
8	2	0	1	2	2	3	4	5	0
9	2	2	3	4	5	6	7	0	0
10	2	1	2	2	3	3	3	0	1

TABLE 4 Analysis of Data in Example

col.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
			Period		Max.	Period		Max.	Period		Max.	Period		Max.	Max.		Average
Row	Product	Group	1	2	1, 2	3	4	3, 4	5	6	5, 6	7	8	7, 8	1–8		1–8
1	1	1	4	5	5	6	7	7	9	0	9	3	4	4	9	Total	4.75
2	2	1	7	9	9	11	0	11	1	3	3	5	6	6	11	of max.	5.25
3	3	1	9	10	10	0	1	1	2	5	5	7	8	8	10	values	5.25
4	4	1	12	0	12	2	3	3	4	7	7	8	10	10	12		5.75
5	5	1	0	6	6	6	4	6	3	3	3	2	1	2	6	48	3.13
			Subtotals		Max.	Subtotals		Max.	Subtotals		Max.	Subtotals		Max.	Max. 1–8		Subtotal
6	sum of rows 1–5:		32	30	32	25	15	25	19	18	19	25	29	29	32		24.13
7	6	2	4	0	4	0	1	1	1	2	2	3	4	4	4	Total	1.88
8	7	2	0	1	1	1	2	2	3	4	4	7	9	9	9	of max.	3.38
9	8	2	0	1	1	2	2	2	3	4	4	5	0	5	5	values	2.13
10	9	2	2	3	3	4	5	5	6	7	7	0	0	0	7		3.38
11	10	2	1	2	2	2	3	3	3	3	3	0	1	1	3	28	1.88
			Subtotals		Max.	Subtotals		Max.	Subtotals		Max.	Subtotals		Max.	Max. 1–8		Subtotal
12	sum of rows 7–11:		7	7	7	9	13	13	16	20	20	15	14	15	20		12.63
			Subtotals		Total	Totals		Total	Totals		Total	Totals		Total	Total	Total	Total
13	sum of rows 6 + 12:		39	37	39	34	28	38	35	38	39	40	43	44	52	76	36.75

1. *Dedicated storage:* Each product is assigned a set of storage compartments that no other product is allowed to occupy.
2. *Shared storage:* Each product group is assigned a set of storage compartments that no product in another group is allowed to occupy.

For dedicated storage, the most popular assignment rule is to assign the SKUs with the highest activity (fast movers) to the locations near the input/output (I/O) point(s). The motivation is to reduce average S/R time. This assignment of individual products to storage positions, called slotting, is usually supported by the WMS. The major principles were outlined in the 1960s (Heskett 1963); a more accessible reference is Tompkins et al. (1996). The activity level of a product is measured by its cube-per-order-index (COI), defined as:

$$\text{Cube per order index} = \frac{\text{access trips per period}}{\text{maximum storage space needed}} \tag{1}$$

A shared storage system where any product in any group may occupy a storage compartment is characterized as pure random storage. However, products are usually grouped by size, environmental requirements (temperature, humidity, hazard level, etc.), value, and chemical compatibility, so pure random storage systems are rare. More popular are class-based systems, where products are grouped by activity level in addition to the factors of size, environmental requirements, etc. A typical class-based system allocates products into three classes, usually called A, B, and C, following the principles of Pareto analysis (Tompkins et al. 1996). The left portion of Figure 3 shows a typical result of such ABC analysis for a pallet S/R system, where the 20% most active storage spaces account for 60% of movements, the next 20% of spaces account for 20% of movements, and the remaining 60% of least active spaces account for 20% of movements. The right portion of Figure 3 shows a corresponding grouping of storage compartments in a single-deep pallet system and the resulting assignment of product groups.

A major benefit of shared storage is reduced space needs. Although textbook formulas suggest savings on the order of 45% (Tompkins et al. 1996), in reality the savings are diminished by common seasonality among products, inbound transport consolidation, inherent variability of product inventory, and inability of the warehouse manager to reallocate space. Typical savings for a three-class system with monthly reallocation of space are 20-30% (Sharp et al. 1999).

Referring to Table 4, if the five products in group 1 were stored in a dedicated system, with space needs based on the forecast for eight periods, then a total of 48 spaces would be required (sum of the individual values in column 15). With shared storage for group 1, only 32 spaces are needed, for a savings of (48 − 32)/48 = 33% just within group 1.

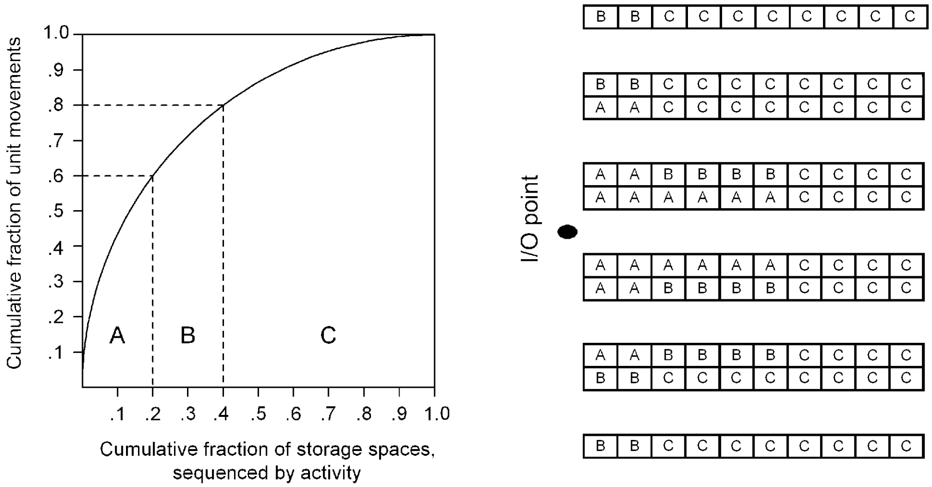


Figure 3 Pareto Analysis and Assignment of Product Groups to Storage Spaces.

A more refined version for pallet storage systems is based on the expected time in storage for each individual pallet (Goetschalckx and Ratliff 1990). For item pick areas a more refined measure is the viscosity of a product (Hackman and Rosenblatt 1990), defined as:

$$\text{Viscosity index} = \frac{\text{retrieval visits per period}}{(\text{cubic volume of product retrieved per period})\wedge 0.5} \tag{2}$$

The concept of family-based storage is sometimes applied in an effort to reduce picker travel time and to simplify order consolidation. The idea is based on identifying items that are likely to be ordered together and then to locate these items near each other. The method has been applied successfully in at least two situations: 1) selecting items based on bills-of-material for manufacturing, and 2) selecting clothing based on size in a catalog retailer operation (Frazelle 1989; Sadiq et al.1996; Amirhosseini and Sharp 1996).

3.2.3. Forward–Reserve Allocation

Referring to Figure 2, one of the crucial decisions in operational planning is deciding which products should occupy positions in the forward pick area and how space should be allocated to each. Depending on the storage technology used in the forward pick area, the total space allocated to it may be a strategic, design decision or a tactical, operational decision. The most commonly used methods use product activity to select products for the forward area and allocate an equal time value of inventory per product. This method, however, is inferior to the recommended method for making these decisions as described in Bartholdi and Hackman (1998). The concept of the viscosity index presented above is used in a recursive algorithm, using the following formula:

$$\text{Fraction of space for product } i = \frac{\sqrt{\text{cubic demand of product } i}}{\sum_j \sqrt{\text{cubic demand of product } j}}\,(V) \tag{3}$$

Dynamic reallocation of products to the forward pick area is practiced by some operators. Using this method, when a pick slot in the forward area becomes empty, it may be reassigned to another product that was previously being picked from the reserve area. Other operators periodically move slow-moving items from the forward area back to reserve.

3.2.4. Zoning and Batching, Order Retrieval

Order retrieval in large and/or busy systems is often characterized by zone picking, where the pick area is divided into smaller areas and a picker is restricted to a zone. An order that requests items from more than one zone is split into suborders, and the picker(s) in each zone selects the respective items. In a grocery distribution center an order may consist of a truckload, and this would be divided into perhaps 20 suborders. In this situation the zone picking would be simultaneous; otherwise it would take too long to complete the order. In a pharmaceutical distribution center the total cubic volume would be less, and the items might be picked into totes. In that situation the zone picking might be progressive, where the picker in the first zone starts the order and passes the partially filled tote to the picker in the next zone, and so forth. Figure 4 illustrates these concepts.

Zone picking is motivated by necessity in the case of large orders and efficiency in terms of operator training with equipment and familiarity with the products. When part-to-picker equipment is used, such as carousel, vertical storage column, and miniload, it is preferable to keep an operator at the retrieval station for that equipment to reduce operator idle time, and thus zones are usually defined by the equipment units.

When the orders or suborders are small relative to the zone size, batch picking is often used: each picker is assigned a set of suborders in the zone, and the items are selected in the most effective manner to reduce picker travel (in picker-to-part systems) or machine travel (in part-to-picker systems). If the pick vehicle has compartments that allow the picker to keep separate the items of the orders (maintain order integrity), then the method is sort-while-pick. Otherwise, downstream sorting is used to reestablish order integrity. The travel time savings from batch picking are discussed in (Armstrong et al. 1979; Choe et al. 1993).

Considering all these options, these are the following logical ways to process orders:

- *Single-order-pick:* One picker works on one order at a time until the order is filled.
- *Sort-while-pick, no zoning:* One picker works on several orders at time with a vehicle that has compartments for maintaining order integrity.

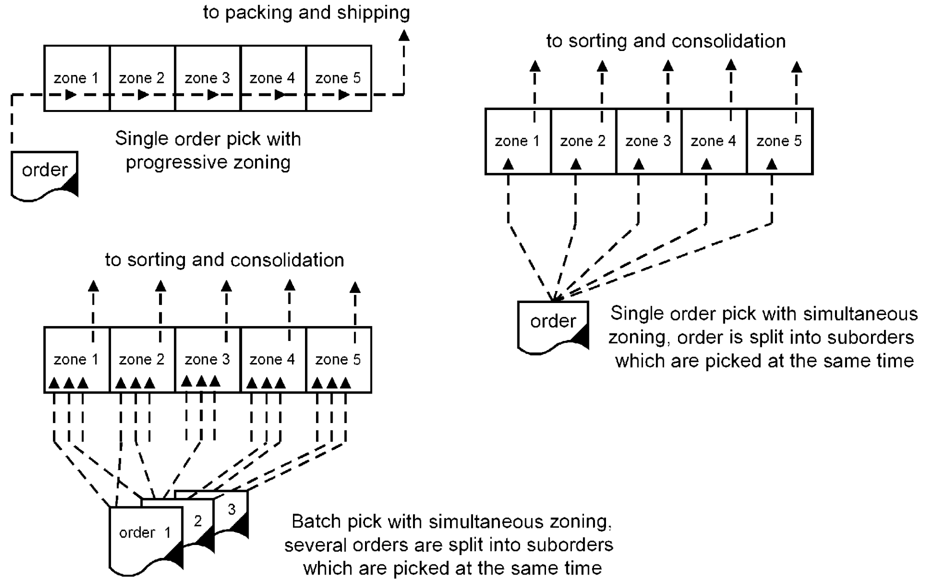

to packing and shipping
zone 1
zone 2
zone 3
zone 4
zone 5
order
Single order pick with progressive zoning
to sorting and consolidation
zone 1
zone 2
zone 3
zone 4
zone 5
order
Single order pick with simultaneous zoning, order is split into suborders which are picked at the same time
to sorting and consolidation
zone 1
zone 2
zone 3
zone 4
zone 5
order 1
2
3
Batch pick with simultaneous zoning, several orders are split into suborders which are picked at the same time

Figure 4 Zoning and Batch Picking.

- *Batch picking with downstream sorting, no zoning:* Several orders are picked by one person completely, often applied with conveyor transport of items to the sorting area.
- *Single-order-pick with zoning:* An order is split into suborders by zone and a picker in each zone fills the corresponding suborder; may be applied with progressive or simultaneous zoning.

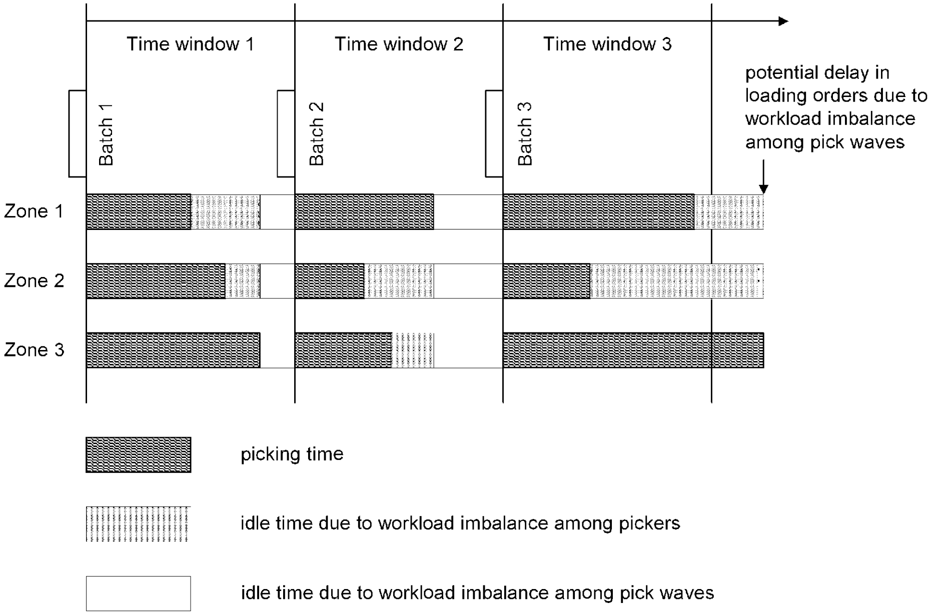


Figure 5 Workload Imbalances among Picker Zones.

TABLE 5 Picker Productivity as a Function of Zone Size and Time Window

Line Items per Consolidated Pick List	Number of Pick Waves	Time Window Duration, hr	Average Number of Orders in Process*	Productivity, Line Items per hr	Relative Productivity
50	12	0.5	25	64.6	65.3%
100	12	0.5	25	67.7	68.5%
50	6	1	50	72.7	73.5%
100	6	1	50	76.7	77.6%
200	6	1	50	79.7	80.6%
50	4	1.5	75	75.9	76.7%
100	4	1.5	75	80.2	81.1%
50	3	2	100	77.6	78.5%
100	3	2	100	82.2	83.1%
200	3	2	100	85.7	86.7%
50	2	3	150	79.4	80.3%
100	2	3	150	84.2	85.1%
200	2	3	150	87.9	88.9%
100	1	6	300	98.8	99.9%
200	1	6	300	98.9	100.0%

*Line items per order distributed Poisson with mean of 4.0.

- *Sort-while-pick with zoning:* An order is split into suborders by zone and a picker in each zone fills the corresponding suborders using a set of containers or a vehicle that has compartments for maintaining order integrity.
- *Batch picking with downstream sorting and zoning:* Several orders are split into suborders and the suborders for each zone are filled by the picker(s) operating in that zone, usually applied with simultaneous zoning.

3.2.5. Pick Wave Planning

When there are many orders to be processed daily, it may be desirable to split them into groups, called pick waves. The time for processing a pick wave in the pick area is called a time window. Pick waves may be necessary for one or more of the following reasons:

- Downstream sorting is used, and the design of the sorting system limits the number of orders that may be in process at one time. Note that the number of orders in process can often exceed the number of output chutes of a conveyor sorter (Bozer et al. 1988).
- The capacity of the forward pick area is insufficient to allow the daily orders to be filled without replenishment, and it is desirable to stop the pick process and replenish, for reasons of safety and/or efficiency.

There are some fundamental trade-offs among the size of a pick wave, picking efficiency, and size of the downstream sorting operation. The smaller the time window, the greater the workload imbalances among zones and across time windows. Depending upon the WMS and conveyor accumulation capacity, a picker who finishes early may or may not have to wait until all pickers are finished with the wave. Figure 5 illustrates this phenomenon of induced idle time. Also, smaller pick waves imply more travel per item selected. On the other hand, a small pick wave allows for a smaller and less expensive sorting system and allows for earlier completion of some orders. In the situation where items are sorted according to loading dock doors, the number of doors may limit the size of the wave. For a more complete discussion of these issues see (Choe 1991; Amirhosseini 1999). Table 5 shows some results for picker productivity and number of orders in process at a time for the situation where pickers must work in synchronized mode (pickers who finish early must wait for all to finish the wave).

4. DATABASE CONSIDERATIONS

Although the detailed structure of a WMS varies according to vendor, the WMS will typically be represented by a number of databases and protocols (methods) that reflect operating rules. In this section an overview of the more important databases is given, as well as some of the linkages among them. Further details may be found in Huffman (1985) and Lin (2000).

TABLE 6 Example Item Master Table for Products

Item SKU	Parent SKU	Vendor	Date Introduced	Vendor Part Number	Lot Control Logic	Perish Date, Months	Alternate Vendors	Substitute Items	Pricing Schedule
4539056512-1	none	153	99.10.14	23331175	none	24	none	4539056513	153-427
4539056512-2	4539056512-1	153	00.02.08	23331175	none	24	none	4539056513	153-427
4540076423	none	28	00.02.15	8004550034	none	24	29	4540076431	29-080

Description	Major Product Group	Minor Product Group	Status	Current Reserve Zones	Current Reserve Slots	Current Forward Zones	Current Forward Slots	Hazard Class	Item Weight, kg	Item Cube, liters
Speaker set, for PC	12	4	active	P1, P3	0958, 0129	C2	2583	none	5.680	3.00
Speaker set, for PC	12	4	active	P2	0125	C2	2583	none	5.680	3.00
Voltage adapter, general	12	7	active	P1	1018	C2	2674	none	0.6	0.60

Item L, cm	Item W, cm	Item H, cm	Items Per Carton	Carton L, cm	Carton W, cm	Carton H, cm	Cartons per Tier	Items per Tier	Tiers per Pallet	Cartons per Pallet	Items per Pallet	Pallet L, cm	Pallet W, cm	Pallet H, cm
20.0	15.0	10.0	8	40.0	30.0	20.0	8	64	5	40	320	120	80	100
20.0	15.0	10.0	8	40.0	30.0	20.0	8	64	7	56	448	120	80	140
12.0	10.0	5.0	16	24.0	20.0	20.0	20	320	5	100	1600	120	80	100

Total Sales Last 6 Months	Pallet Requests Last 6 Months	Sales, Pallet Requests, Last 6 Months	Carton Requests Last 6 Months	Sales, Carton Requests, Last 6 Months	Item Requests, Last 6 Months	Sales, Item Requests, Last 6 Months	Retail Items per WH Item	Retail Item L, cm	Retail Item W, cm	Retail Item H, cm
317	0	0	17	204	49	113	1	20.00	15.00	10.00
							1	20.00	15.00	10.0
2893	0	0	87	2352	78	541	1	12.0	10.0	5.0

4.1. Products and Orders

Two important databases relate to products and orders, respectively. The item master defines the products that are handled by the warehouse. Included in this database are product identification, physical characteristics, vendor sources, replenishment and safety stock levels, service level, and so forth (see Table 6). The order master contains the records of customer orders (see Table 7). In addition, closely related tables include a vendor master, a customer master, and an inventory master. The last table should reflect not only quantities and locations of products on hand, but also back-ordered quantities and quantities reserved for customers or other reasons.

4.2. Flow Control

Another series of database tables is used to track the flow of material: advance shipping notice (ANS) table for those receipts that are expected soon, putaway tables for items received that are to be stored, replenish tables for forward pick areas, pick tables for pick waves in each area, load consolidation tables, truck manifest tables, and additional move tables for other moves not specified above. There may be more than one version of each table, one as planned and another as executed, or these characteristics may be captured in one table. These tables, which are often subdivided by priority, are particularly important when automatic identification, such as bar code, is used to control flow. In the situation of batch picking with zoning the pick table needs to contain details on zone, picker, pick wave, sorting area, packing method, and so on. (see Table 8).

4.3. Building, Equipment, and Personnel Assets

The equipment master contains information on the physical characteristics of storage and transport equipment, speeds, compatibilities with respect to products and containers by size, and access times

TABLE 7 Example Order Master Table

Date	Order	Customer	Address	PO Number	Ship Date
00.10.25	392457	ABC Mfg. Co.	12 Main Street, Anytown, State	00100479	00.10.27

Purch Agent	Sales Rep	Tel Ref	Fax Ref	E Ref
Mary Doe	Jane Dowling	201.456.3333	201.456.3339	mdoe@abc-mfg.com

Ship Address	Ship Method	Preferred Carrier	Freight Charges	Ship Short	Split Shipment	Billing	Credit Status
14 Main Street, Anytown, State	LTL	Fast Truck Co.	customer	yes	no	30 2% 10	1

SKU	Quantity	Unit Price	Discount Logic
3092874101	4	34.00	2
3092874217	7	122.00	2
3092921004	1	345.00	3

Date	Order	Customer	Address	PO Number	Ship Date
00.10.25	392458	XYZ Dist. Co.	1 Little Lane, Smalltown, State	04005682	00.10.28

Purch Agent	Sales Rep	Tel Ref	Fax Ref	E Ref
Ron Howard	Linda Cabot	715.233.8700	715.233.8710	ronnie@xyzdistn.com

Ship Address	Ship Method	Preferred Carrier	Freight Charges	Ship Short	Split Shipment	Billing	Credit Status
1 Little Lane, Smalltown, State	package	Zip Express	customer	no	no	30 net	1

SKU	Quantity	Unit Price	Discount Logic
5200346721	8	55.00	1

TABLE 8 Example Pick Table for Batch Picking with Zoning

Date 8	01.03.0	Wave	2-08	Zones C1		Release Time 1030		
Empl	31	Method 2	Order Picker	Packing Lane 7		Standard Time 30		
Stop	Slot	SKU	Description	Package	Total	Quantity	Customer	Notes
1	1005	6540049423	PCV valve, GM 98 type 5	carton	3	3	Right tune #47	
2	1008	6840052109	tune-up kit, 6G, platinum	carton	7	2	J Discount #2	label 2 in from bottom
						1	AAFES TX	
						4	AAFES SF	export label, US mfg
3	1017	7134458302	oil filter, N30	carton	9	5	AAFES SF	export label, US mfg
						2	Right tune #47	
						2	J Discount #2	label 2 in from bottom
4	1024	6424083023	transmission filter, F28	carton	4	3	L&M Service	
						1	Right tune #47	

TABLE 9 Example Equipment Master Table

Type	Class	Parent	Date	Mfg.	Desc.	Qty.	W. cm	D, cm	H, cm	Aisle cm	Max. Lift, cm	Level	Levels	Deep	Wide	Flow
storage	010	none	96.05.08	ABC	pallet rack	450	320	280	530	300	0	1	1	2	3	B-B
	011	10	96.05.09	ABC	pallet rack	450	320	280	530	300	420	2–4	3	2	3	B-B
	012	10	96.05.10	ABC	pallet rack	350	320	280	530	300	540	5	2	2	3	B-B, Top thru
	015	none	96.05.11	ABC	p flow rack	55	110	500	300	300	180	1–2	2	4	1	
	021	none	96.05.12	DEF	shelf	120	120	65	200	150		1–6	6	1	1	B-B

Surface	Sprinkler	Std Load W, cm	Std Load D, cm	Std Load H, cm	Min. Load W, cm	Min. Load D, cm	Min. Load H, cm	Max. Load W, cm	Max. Load D, cm	Max. Load H, cm	Comments
floor	2,4	80	120	110	60	120	20	300	130	110	
none	2,5	80	120	110	60	120	20	300	130	110	
mesh	2,6	80	120	110	60	120	20	300	130	220	
roller	2,7	80	120	110	80	120	50	80	120	110	pallet QC
metal		30	30	27	10	10	10	110	35	30	

Type	Class	Mfg.	Date	Desc.	Power	Duration, hr	W, cm	L, cm	H, cm	Pick Aisle	Cross-Aisle	Max. Lift. cm	Max. wt, kg	Velocity H, m/sec
Vehicle	100	GHI	98.02.10	pallet jack	none	NA	80	110	130	200	200	8	800	0.6
	102	JKL	98.03.02	pallet jack	battery	8	85	120	115	220	220	10	1000	1.2
	201	GHI	98.05.04	fork lift, CB	propane	14	105	170	255	300	320	610	1500	2.8
	202	GHI	98.06.10	fork lift, CB	propane	14	105	170	255	300	320	640	1800	3.0
	251	JKL	99.03.02	order picker 1	battery	6	120			180	380	640	1400	2.0

Velocity V, m/sec	Std Load W, cm	Std Load D, cm	Std Load H, cm	Min. Load W, cm	Min. Load D, cm	Min. Load H, cm	Max. Load W, cm	Max. Load D, cm	Max Load H, cm	Permitted Accessories	IDs
0	80	120	110	60	120	10	100	140	140	container 1	100-01, 100-02
0	80	120	110	60	120	10	105	160	140	container 1	102-09, 102-11
1.3	80	120	110	60	120	10	120	120	220	container 1	201-04
1.3	80	120	110	60	120	10	120	120	220	container 1	202-07
1.2	80	120	130	40	80	10	80	120	160	container 1 container 3	251-01 251-02

TABLE 10 Example Personnel Master Table

ID	Name	Taxpayer ID	Date Hired	Phone	Pager	Pay Classification	Level
22	Cyril Bezukhov	123 45 6789	86.11.02	201.055.6000	201.055.6000	Manager 2	120
26	Ilya Rostov	087 65 4321	89.03.15	201.055.7000	201.055.7000	Manager 1	110
29	Mary Bolkonskaya	010 79 2468	84.05.22	201.055.7575	201.055.7575	Regular 4	130
31	Helene Kuragina	021 54 9876	93.01.08	201.055.8888	201.055.8888	Regular 2	115
36	Anna Drubetskaya	064 29 7531	98.07.14	201.055.9999	201.055.9999	Regular 1	100
40	Valentin Berg	004 02 3856	00.06.27	201.055.4444	201.055.4444	Clerical 1	100

Assignment Priorities											Other Shift Schedules		
A1	A2	R1	R2	P1	P2	I1	I2	L1	L2	Primary Shift			
2	1			3						1	2	3	4
1	2			3						1	2	3	
		3		2	1					1	2		
				1	2			3		2	1	4	
		1						2		3	none		
	1					2	5	3	4	1	2		

TABLE 11 Example Zone Master Table

Zone ID	Parent	Desc.	Primary Function	Other Functions	Storage Equip. ID	Quantity	Description	Vehicle Class	Description	ConvID	Description	Employee Type	Start Location	End Location
0	none	system												
1	0	Rec-1	Receiv	PalStor	010	30	p rack floor	100	pallet jack	301	skatewheel, extendible	R1		
					011	30	p rack	201	fork lift, CB	310	belt, 8m	R2		
2	0	Rec-2	Receiv	QuarInv	010	20	p rack floor	100	pallet jack	302	skatewheel, extendible	R2		
					011	20	p rack	201	fork lift, CB					
10	0	Pal-1	PalStor	QuarInv	010	120	p rack floor	201	fork lift, CB			R2		
					011	120	p rack	202	fork lift, CB			R2		
11	10	Pal-1A	PalStor	Cpick	010	120	p rack floor	102	pallet jack			P1		
								202	forklift, CB			R2		
12	10	Pal-1B	PalStor	QuarInv	011	120	p rack	202	forklift, CB			R2		
20	0	Pal-2	PalStor	QuarInv	010	150	p rack floor	201	fork lift, CB			R2		
					011	150	p rack	202	fork lift, CB			R2		
					012	150	p rack top	202	fork lift, CB			R2		
21	20	Pal-1A	PalStor	CPick	010	150	p rack floor	103	pallet jack			P1		
								202	fork lift, CB			R1		
22	20	Pal-1B	PalStor		011	150	p rack	202	forklift, CB			R2		
23	20	Pal-1B	CPick		011	150		251	order picker			P3		
24	20	Pal-1C	PalStor	QuarInv	012	150	p rack top	251	order picker			P3		
211	21	Pal-1A-A	CPick	Transient	010	30	p rack floor	103	pallet jack			P1	1131	1160
212	21	Pal-1A-B	CPick	Transient	010	50	p rack floor	103	pallet jack				1161	1210
213	21	Pal-1A-C	CPick	Transient	010	70	p rack floor	103	pallet jack				1001	1060
													1211	1220
30	0	Pal-3	CPick		015	55	p flow rack	201	fork lift, CB				1901	1950
40	0	Item-3	ItPick		070	1	carousel					P2	3201	3800
50	0	Item-4	ItPick		80	22	carton flow	401	belt+2			P3	4201	4222

(see Table 9). This table is usually subdivided into storage media (e.g., racks, shelving), automated storage/retrieval systems, vehicles, and conveyors. This type of information is accessed when decisions on product storage are made. Racks and conveyors are usually divided into zones, and this information is reflected in the tables. Compatibility of storage media with vehicles is also recorded.

The personnel master contains information on the skills of employees, availability by shift, and sometimes their performance (productivity and quality) ratings (see Table 10).

The most important table for building and equipment is the zone master. This table relates building space and equipment units in a logical manner, and it provides the links for inventory, product storage assignment, product retrieval, and flow control. In some applications the table is hierarchical, with overlapping zones. Zone overlap may occur when some areas of a building are used for one purpose, such as single-order-pick during one part of the shift, and another purpose, such as batch-picking-with-zoning, during another part. Table 11 shows some information in an example zone master table.

4.4. Operating Rules

A series of protocols is used to drive the functions in a warehouse according to specified rules. There is an umbrella protocol, called the function flow map, that defines the functions that are allowed to occur in the different zones (areas) of the warehouse. This protocol may be in the form of a table that can be controlled by the warehouse operator, or it may be coded directly into the WMS software. Table 12 is an example where the allowable functions are prioritized. In addition, there are tables that prescribe putaway rules and inventory allocation rules. The rules for picking flow path must reflect the zone structure: one zone, progressive zone picking, or simultaneous zone picking. Rules for picking method reflect single-order-pick and batch-picking. Picking control may be by paper labels, container license plate (e.g., bar code), or radio frequency (RF). Another set of rules applies

TABLE 12 Example Function Flow Map

Function number	Description	Origin Zone	Destinaion Zone, First Choice	Destination Zones, Other
101	receiving, general	none	1	2
102	receiving, R1 only	none	1	
103	receiving, R2 only	none	2	
201	putaway to pallet storage	1	10	20, 30
202	putaway to pallet storage	2	20	10, 30
203	putaway to pallet storage	1	10	
204	putaway to pallet storage	1	20	
205	putaway to pallet storage	1	30	
206	putaway to pallet storage	2	10	
207	putaway to pallet storage	2	20	
208	putaway to pallet storage	2	30	
221	move to cross dock operation	1	90	100
299	putaway to quarantine storage	1	10	20
301	putaway to item pick	1	40	50
302	putaway to item pick	1	50	40
303	putaway to item pick	2	50	40
401	pallet pick	10	100	90
402	pallet pick	20	100	90
501	replenish carton forward area	22	211	
502	replenish carton forward area	22	50	
601	carton picking, one zone only	211	100	
602	carton picking, multiple zones	211	80	
701	item pick, single-line orders	40	95	
702	item pick, single-line orders	50	95	
703	item pick, multi-line orders	40	85	90
704	item pick, multi-line orders	50	85	90
801	carton sorting	80	100	
901	item sorting	85	100	
950	packing	95	100	
960	outbound hold	95	98	
990	loading	100	none	

to consolidation, packing, and loading of orders. If value added functions are performed, such as labeling, pricing, or repackaging for retail display, then rules must be established for them.

4.5. Links to Hardware Controllers

Mechanized and automated equipment units will have their own hardware controllers, and these must be linked to the WMS. The hardware controllers are typically coded in C language in a UNIX operating system. Besides flow-control logic, information on system integrity, battery voltage, conveyor utilization, equipment malfunction, and so on is reported. Automated storage/retrieval systems, high-speed sorting conveyors, and merge conveyors, if present, must be integrated into the overall software system for the warehouse. Conveyor systems have numerous photo-eye detectors, which can send thousands of signals per hour; such high transaction rates must be handled by separate controllers that are linked to the WMS. The interfaces typically involve the transfer of batch instructions between the WMS and the hardware controllers. The main computing platform for the WMS may be a mainframe system, an IBM AS-400 system, or a Windows NT PC-based system. The major WMS vendors support multiple platforms (Seidl 1997; Hahn-Woernle 2000; IT logistiek 2000).

4.6. Data Backups

Systems that attempt to operate in real time require a hot backup or mirror computer operation, augmented by an emergency power generator. In essence, two computer operations run simultaneously, and either one can control the WMS in case the other fails. Otherwise, or in addition, a system of backup data tables should be maintained on an independent computer, so that they may be printed and used for operation in case the WMS computer fails. These backup tables should be updated once every shift or pick wave.

5. DAILY OPERATIONAL FACTORS IN WAREHOUSE OPERATION

5.1. Receiving Operations

It is logical to think of receiving as the first operation in warehousing. The physical receipt of goods, however, is typically preceded by an ASN. This allows for planned cross-dock operations, which are increasing in popularity. Out-of-stock conditions are flagged, and the corresponding goods go directly to packing and shipping for back-orders. Some goods may go to replenishment of forward pick areas directly from receiving (subject to stock rotation policies). Upon receipt of goods from qualified vendors, the inventory values in destination areas may be adjusted, or the goods may be logged into a transient state. Goods from non-qualified vendors may undergo a quality check and/or count. If this is time-consuming, the goods are held in a quarantine inventory status. If the quality check is based on a sample, it is usually possible to move the bulk of the product to another location and control the quarantine status through the WMS. Figure 6 shows the possible logical flows for receiving operations.

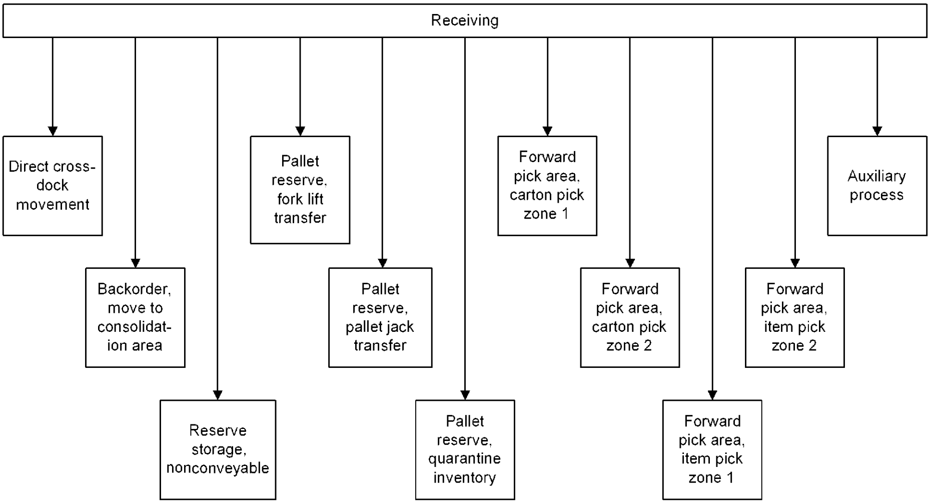

Receiving
Direct cross-dock movement
Backorder, move to consolidation area
Reserve storage, nonconveyable
Pallet reserve, fork lift transfer
Pallet reserve, pallet jack transfer
Pallet reserve, quarantine inventory
Forward pick area, carton pick zone 1
Forward pick area, carton pick zone 2
Forward pick area, item pick zone 1
Forward pick area, item pick zone 2
Auxiliary process

Figure 6 Receiving Operations Material Flows.

There should be the capability to process partial receipts, generate labels for pallets and cartons, and assign storage locations. The assignment of storage locations may follow one of several disciplines:

- *User is completely directed by WMS:* The WMS identifies the location based on product characteristics and available locations. The operator is not allowed to store the item in a different location, except by going through exception routines.
- *User is partially directed by WMS:* An example would be when the WMS directs the user to a zone or a portion of a zone (e.g., an aisle or portion of an aisle), but the user has flexibility within the zone of where to place the item.
- *User has complete flexibility:* The WMS allows the user to place the item in any location, and reports an error only when there is some physical incompatibility, such as compartment too small, compartment contains too much already, etc.

Returns processing is increasingly being assigned to outside operators, but most warehouses still accept some return goods from customers. This is often a labor-intensive operation, involving unpacking, inspecting, repackaging, and labeling. After the last step the goods can be ‘‘received’’ as if they came from a vendor. Manual date entry and error correction are also functions that must be included in receiving.

5.2. Storage and Inventory Control

In addition to the concepts of dedicated vs. shared storage discussed earlier, the WMS should have capability for lot control (typically by date or production run), fractional unit loads, and storage in more than one location. When the incoming quantity of a product is less than a unit load, there is an opportunity for consolidation when the products are stored. Various rules may govern this consolidation, such as same SKU, same SKU and same lot, allow different SKUs in storage compartment, compute remaining compartment capacity by cubic volume, compute remaining compartment capacity by dimensions of items, and so on. When goods that were in a transient state upon receipt are stored, the inventory values are adjusted. In addition, the actual storage location is recorded, or acknowledged in the situation of the user being completely directed by the WMS. Some goods may be reserved for specific customers when they are stored; such stock reservation may also occur later. Cycle counting may be performed when the items are stored; such opportunistic counting helps reduce extra travel for inventory control. Other functions that occur on a regular basis are stock activity reports for fast, medium, slow, and dead stock, and empty location reports. Manual data entry and error correction are also needed.

5.3. Order Processing

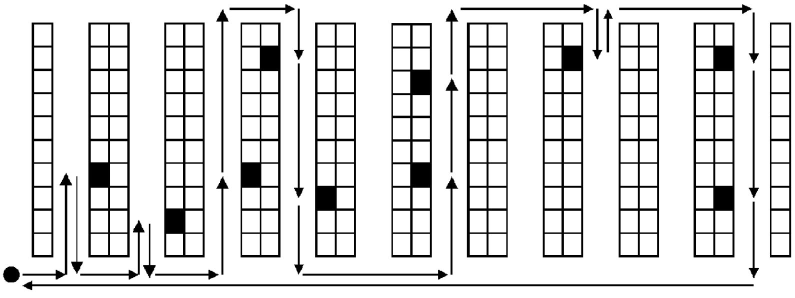
The first step in order processing, which is usually performed by the customer service department, is verification of item availability. This is preferably done online, or electronically in the situation of computer-based orders. On-line verification of customer credit status is next, followed by inventory reservation, if appropriate. The pricing structure may reflect quantity restrictions, such as full carton or full pallet, and these are applied at this time. The customer service department should suggest the closest or next quantity multiple, with price breaks if any, to simplify the work in the warehouse. The software used by the customer service department, which may be linked to the WMS, should have flexibility in pricing by customer and order type, flexibility for partial shipments, flexibility in picking and packing orders according to customer needs (including priority class), and flexibility in handling shipping charges (customer pays, price includes transportation, etc.) After the order has been selected, the WMS (or linked software) generates an invoice and bill of lading. Manual date entry and error correction are also included as functions.

5.4. Order Picking

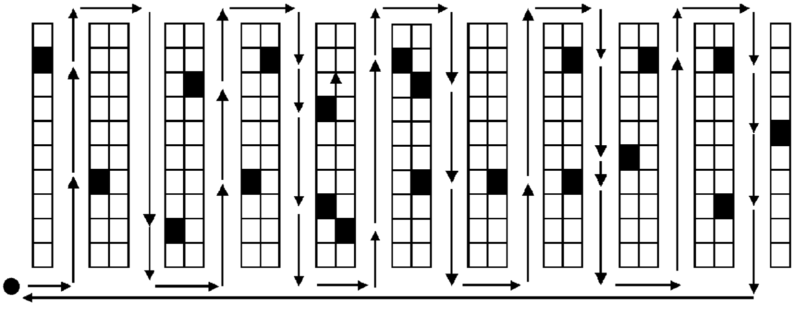
There are several planning functions precede the actual retrieval of products for customer orders. The first of these is to check current inventory levels in the forward pick areas and generate replenishment reports. Most warehouse operators prefer to replenish at the beginning of the shift, for reasons of safety and efficiency. Some replenishments may occur during the pick process, especially if the information about orders is incomplete, or if operators select full cartons from the item pick area when they should be selected from the carton pick area. The WMS should support workload balancing in the pick operation: reflect different picker capabilities according to data in the personnel master (Table 10), and reflect different number of operators according to pick wave and shift. The ability to balance workload over more than one day is desirable, but it is usually not available in the typical WMS.

The WMS must be capable of supporting the different pick methods described in Section 3.2.4, along with estimates of completion times for the major steps, such as the pick waves and sorting and loading operations for waves. For pick wave formation a variety of criteria may apply: group rush orders first, group orders by type (single-order-pick, batch-pick, etc.), group orders by packing method, group orders by loading dock door in reverse sequence based on delivery routes, and group orders by proximity of retrieval locations for sort-while-pick operation (Gibson and Sharp 1992). In zone systems it is desirable to have flexibility for changing the zone configuration, either by pick wave or dynamically. Dynamic adjustment of zones in item-pick operations has proven very effective (Bartholdi and Eisenstein 1996). The routing of pickers in a multi-aisle system is usually accomplished by software according to the number of retrieval stops and the travel metric:

- *Single-command operation:* The operator makes only one stop in the storage system before returning to the I/O point.
- *Dual-command operation:* The operator makes two stops in the storage system before returning to the I/O point. If one stop is for storage and the other for retrieval, as in pallet operations, this method is called interleaving. The rules in the WMS determine the travel savings to be gained by such methods (Graves et al. 1977).
- *Multicommand operation in a sparse system:* The operator makes several stops (more than two) in a system but considerably fewer than the number of pick aisles. For a single block of aisles, called a ladder structure (see Figure 7), an adaptation of the traveling salesman problem may be applied to the routing (Ratliff and Rosenthal 1983). When there are cross-aisles or multiple blocks of aisle, heuristic algorithms are available (Kees 2000).
- *Multicommand operation in a busy system:* When the number of retrieval stops is more than the number of pick aisles, serpentine routing is applied. Here the picker enters each aisle that contains product on the pick document and continues through to the end of the aisle, and then proceeds to the next aisle that contains product to be retrieved. In busy systems, or systems where U-turns or reverse travel in an aisle are not desirable, a one-way flow in each aisle may also be imposed. This may induce unnecessary travel in aisles with no products for the current



Multistop picker routing in a sparse picking application



Multistop, serpentine picker routing in a busy picking application

Figure 7 Multistop Picker Routing.

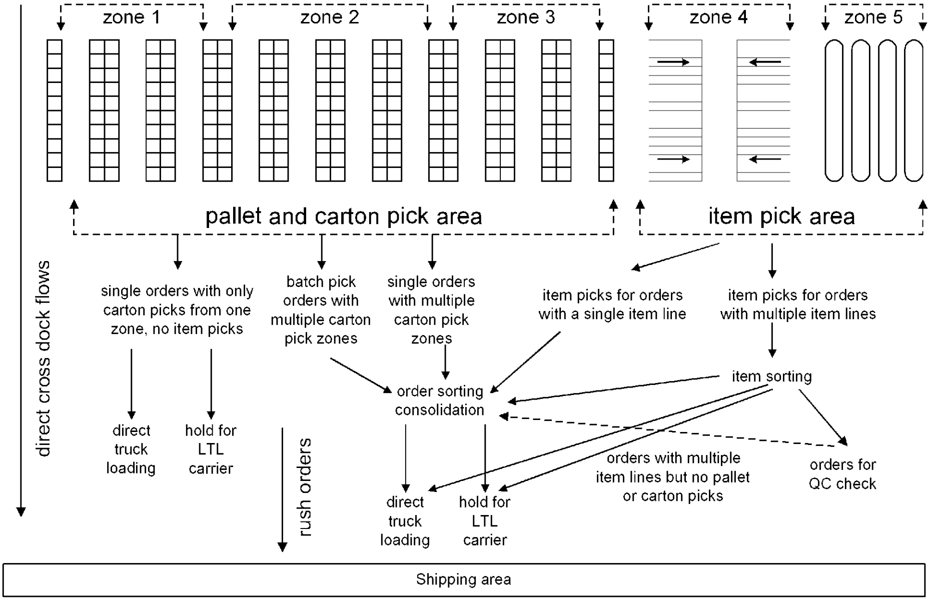
pick document. The expected picker travel in such systems may be obtained using the results in Choe (1991).

- *Multicommand operation with order-picker trucks:* When the vertical travel dimension is added, the picker routing becomes more complicated. Generally, it has been found that the best storage-assignment method is to place the more active items at the lower levels of the rack system. Placing more active items at the end of the aisle doesn't help much, since there is a good chance the picker will travel through the aisle anyway (Krueger 1999).
- *Multicommand operation with automated S/R systems:* In an automated S/R system, the characteristics of the S/R machine usually imply simultaneous movement in the horizontal and vertical dimensions. Most of the technology applications have one S/R machine per aisle. The picker routing then becomes a traveling salesman problem in the Chebyshev metric (Bozer et al. 1990).

The WMS will generate consolidated pick documents for each operator (see Table 8), sequenced by retrieval location. These documents may be paper, electronic, labels or electronic tags, or a combination. They must contain the picker or picker team identification, SKU location, quantity, inner-pack quantity if applicable, zone, wave, packing lane, packing method, and shipping method. Exceptions and interruptions occur frequently in order picking, and a series of protocols must be available to deal with them. The simplest exception decision to execute is to ship an incomplete order: the packing list and invoice are adjusted to reflect actual quantities, and the missing items are placed on back order for the customer if that is the applicable policy. More involved are decisions to substitute items, or to place nearly complete orders temporarily on hold to wait for a missing item. These actions usually require that the products be placed in a temporary storage area, which complicates the sequence of data-processing operations in the flow control tables. The discovery of wrong or damaged items at the loading operation may tie up a loading dock door and disrupt the release of the next pick wave.

5.5. Order Consolidation

After the multiple parts of an order are selected, they must be brought together in one location for packing and consolidation. This may be as simple as consolidating items in a staging lane near the loading dock. In other situations, such as the use of high-speed sorters, accumulation lanes feeding packing lanes are needed. The WMS or linked sorter software controls the flow of goods from picking through sorting and order consolidation and on to packing and loading. Items that differ physically



zone 1
zone 2
zone 3
zone 4
zone 5
pallet and carton pick area
item pick area
direct cross dock flows
single orders with only carton picks from one zone, no item picks
batch pick orders with multiple carton pick zones
single orders with multiple carton pick zones
item picks for orders with a single item line
item picks for orders with multiple item lines
direct truck loading
hold for LTL carrier
rush orders
order sorting consolidation
item sorting
orders with multiple item lines but no pallet or carton picks
orders for QC check
direct truck loading
hold for LTL carrier
Shipping area

Figure 8 Example of Order-Consolidation Material Flows.

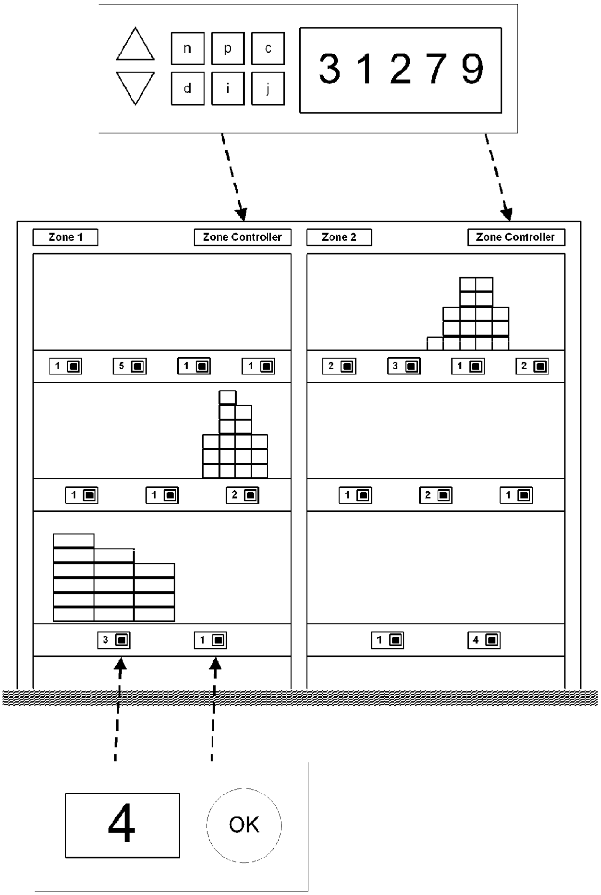
according to shape, such as pallets, nonconveyable items, conveyable items, totes containing multiple items, and so on pose the greatest challenge in this situation. Value-added operations such as labeling, pricing, and repackaging for retail display impose additional tracking and coordination needs. A strict control system, usually involving bar code, and sufficient floor space are necessary ingredients of an effective solution to such problems.

Figure 8 shows some possible flows for order consolidation where pallet and carton picks are sorted manually in a consolidation area and item picks are sorted in a separate area, perhaps with a conveyor.

5.6. Additional Factors

A number of human factor and hardware enhancements are available to increase the productivity and accuracy of order picking, including pick-to-light, voice-guided picking, and voice-recognition systems. Pick-to-light systems (see Figure 9) offer productivity increases of 30–100%, with corresponding error reductions of up to 80% (Sharp et al. 1997). Voice-guided picking allows the operator to use both hands to retrieve products without the need to refer to a paper pick document. Voice-recognition systems enable the picker to interact with the WMS without cumbersome key-entry

Pick-to-light zone controller display:
up-down buttons, function buttons, order number



Pick-to-light pick slot display:
order quantity, acknowledgment button

Figure 9 Pick-to-Light System.

devices. Radio-frequency (RF) systems are available to allow bar code readers to interact directly with the WMS from any point in the warehouse. Optical character recognition is available to eliminate the need for operators reading labels (Auto ID 2000). These systems must be integrated with the WMS and tested to ensure functionality.

Implementation should be done in gradual steps. One should not put in more functions and reports initially than can be managed. It may take one or two years for an organization to implement completely a new WMS. Employees should not receive more information than they need to perform their tasks. At the same time, managers should receive summary information on system performance, with details reported only for exceptions. Whenever possible, management reports should include utilization of labor hours by activity type and capacity utilization for equipment.

Linkages to supply chain management systems will occur if the parent organization owns other storage facilities or if the warehouse participates with other firms in a coordinated supply chain. In such cases it is recommended that the WMS vendor provide a unified software package for all relevant facilities. Each separate facility can be represented by one or more zones, and transportation links among them can be represented by transient states. The presence of an enterprise management system will also require linking.

6. CONCLUSION

Warehouse management consists of strategic, tactical, and operational factors. All of these require good information for making decisions: the selection of technologies for storage and retrieval, the assignment of incoming products to storage locations, the retrieval of products for customer orders, the assembly, packing, and loading of orders, the specification of a labor mix consisting of permanent and temporary employees, and the scheduling of labor and equipment. Modern equipment for storage, handling, data capture, and communication enable warehouse managers to have tight control over inventories and orders shipped while achieving short response times to customer orders. At the same time, there are available efficient strategies for storage, retrieval, and order assembly. To take advantage of such opportunities, the facility operator must have an effective, flexible warehouse management system. The trend today is toward an integrated, computer-based system that controls the flow of material and the actions of employees.

REFERENCES

Amirhosseini, M. M. (1999), “Effect of Time Windows and Zoning,” in *Advanced Order Picking Short Course,* Georgia Institute of Technology, Atlanta, pp. 1–10.

Amirhosseini, M. M., and Sharp, G. P. (1996), “Simultaneous Analysis of Products and Orders in Storage Assignment,” *Manufacturing Science and Engineering—1996,* MED-Vol. 4, pp. 803–811.

Armstrong, R. D., Cook, W. D., and A. L. Saipe, A. L. (1979), “Optimal Batching in a Semi-Automated Order Picking System,” *Journal of the Operational Research Society,* Vol. 30, pp. 711–720.

Bartholdi, J. J., and Eisenstein, D. (1996), “A Production Line That Balances Itself,” *Operations Research,* Vol. 44, No.1, pp. 21–34.

Bartholdi, J. B., and Hackman, S. (1998), “Warehousing and Distribution Science,” manuscript.

Bozer, Y. A., Quiroz, M., and Sharp, G. P. (1988), “An Evaluation of Alternative Control Strategies and Design Issues for Automated Order Accumulation and Sortation,” *Material Flow,* Vol. 4, No. 4, pp. 265–282.

Bozer, Y. A., Schorn, E. C., and Sharp, G. P. (1990), “Geometric Approaches to Solve the Chebyshev Traveling Salesman Problem,” *IIE Transactions,* Vol. 22, No. 3, pp. 238–254.

Choe, K.-I. (1991), “Aisle-Based Order Pick Systems with Batching, Zoning and Sorting,” Ph.D. thesis, Georgia Institute of Technology, Atlanta.

Choe, K.-I., Sharp, G. P., and Serfozo, R. F. (1993), “Aisle-Based Order Pick Systems with Batching, Zoning, and Sorting,” in *Progress in Material Handling Research: 1992,* R. J. Graves, L. M. McGinnis, R.E. Ward, and M.R. Wilhelm, Eds., Material Handling Institute, Charlotte, NC.

Frazelle, E. H. (1989), “Stock Location Assignment and Order Picking Productivity,” Ph.D. thesis, Georgia Institute of Technology, Atlanta.

Goetschalckx, M., and Ratliff, H. D. (1990), “Shared Storage Policies Based on the Duration Stay of Unit Loads,” *Management Science,* Vol. 36, No. 9, pp. 53–62.

Graves, S. C., Hausman, W. H., and Schwarz, L. B. (1977), “Storage–Retrieval Interleaving in Automatic Warehousing Systems,” *Management Science,* Vol. 23, No. 9, pp. 935–945.

Hackman, S. T., and Rosenblatt, M. J. (1990), “Allocating Items to an Automated Storage and Retrieval System,” *IIE Transactions,* Vol. 22, pp. 7–14.

Hahn-Woernle, C. (2000), ‘‘Trends in Warehouse Management,’’ in *Dortmunder Gespräche 2000,* Dortmund University, Dortmund, Germany, CD-ROM.

Heskett, J. L. (1963), ‘‘Cube-per-Order Index—A Key to Warehouse Stock Location,’’ *Transportation and Distribution Management,* Vol. 3, pp. 27–31.

Huffman, J. R. (1985), ‘‘Computers in the Warehouse,’’ in *Materials Handling Handbook,* R. Kulwiec, Ed., John Wiley & Sons, New York, pp. 669–703.

IT logistiek, Logistiek Krant, and Berenschot (2000), *Selection Methods for Warehouse Management Systems,* Elsevier, Amsterdam, CD-ROM.

Kees, J. R. (2000), ‘‘Optimal Configuration of Aisle-Based Pick Systems with Cross Aisles,’’ working paper, Erasmus University, Amsterdam.

Krueger, K. W. (1999), ‘‘Simulation Software Tool for Order Picking in a Person-Aboard Storage/Retrieval System,’’ Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta.

Lin, L.-C. (2000), ‘‘A Modularized Operations System Approach for the Distribution Center Design,’’ working paper, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan.

Ratliff, H. D., and Rosenthal, A. S. (1983), ‘‘Order Picking in a Rectangular Warehouse: A Solvable Case of the Travelling Salesman Problem,’’ *Operations Research,* Vol. 31, No. 3, pp. 507–521.

Rouwenhorst, B., Reuter, B., Stockram, V., van Houtum, G. J., Mantel, R. J., and Zijm, W. H. M. (2000), Warehouse Design and Control: Framework and Literature Review,’’ *European Journal of Operational Research,* Vol. 122, pp. 515–533.

Sadiq, M., Landers, T. L., and Taylor, G. D. (1996), ‘‘An Assignment Algorithm for Dynamic Picking Systems,’’ *IIE Transactions,* Vol. 28, pp. 607–616.

Seidl, J. (1997), *Warehouse Management Systems: Market Overview and Project Life Cycle,* Deloitte & Touche Consulting Group/Garr, Atlanta.

Sharp, G. G., Amirhosseini, M. M., and Shamanna, S. K. (1994), ‘‘Analysis of a Company’s Order Picking Rack System,’’ MHRC-OP-94-01, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta.

Sharp, G. P., Handelsmann, R., Light, D., and Yeremeyev, A. (1997), ‘‘Productivity and Quality Impacts of Pick-to-Light Systems,’’ in *Progress in Material Handling Research: 1996,* R. J. Graves, L. F. McGinnis, D. J. Medeiros, R. E. Ward, and M. R. Wilhelm, Eds., Material Handling Institute, Charlotte, NC.

Sharp, G. P., Amirhosseini, M. M., and Schwarz, F. (1999), ‘‘New Approaches and Results for Product Storage Assignment: Consideration of Demand Variability, Demand Correlation, Storage Compartment Size, and Degree of Order Completion,’’ in *Progress in Material Handling Research: 1998,* R. J. Graves, L. F. McGinnis, D. J. Medeiros, R. E. Ward, and M. R. Wilhelm, Eds., Material Handling Institute, Charlotte, NC.

Tompkins, J., White, J., Bozer, Y., Frazelle, E., Tanchoco, J., and Trevino, J. (1996), *Facilities Planning,* 2nd Ed., John Wiley & Sons, New York.

Yoon, C. S., and Sharp, G. P. (1996), ‘‘A Structured Procedure for Order Pick System Analysis and Design,’’ *IIE Transactions,* Vol. 28, pp. 379–389.

CHAPTER 82

Supply Chain Planning and Management*

Douglas M. Lambert
The Ohio State University

Edward A. Siecienski
JMA Supply Chain Management

1. INTRODUCTION 2111

2. SUPPLY CHAIN MANAGEMENT VS. LOGISTICS 2111
  2.1. Definition of Supply Chain Management 2111
  2.2. Difference between Logistics and Supply Chain Management 2112

3. CHANNEL STRUCTURE 2115
  3.1. Outsourcing Pieces of the Supply Chain 2115
  3.2. Postponement and Speculation 2115
  3.3. Time-to-Market Pressures 2116
  3.4. Other Issues Affecting Channel Structure 2116

4. SUPPLY CHAIN NETWORK STRUCTURE 2116
  4.1. Identifying Supply Chain Members 2117
  4.2. The Structural Dimensions of the Network 2117
  4.3. Types of Business Process Links 2118
  4.4 Supply Chain Mapping Considerations 2120

5. SUPPLY CHAIN BUSINESS PROCESSES 2120
  5.1. Customer Relationship Management Process 2121
  5.2. Customer Service Management Process 2121
  5.3. Demand Management Process 2121
  5.4. Customer Order-Fulfillment Process 2121
  5.5. Manufacturing Flow Management Process 2122
  5.6. Procurement Process 2122
  5.7. Product Development and Commercialization 2122
  5.8. Returns Process 2122
  5.9. SCM Process Summary 2122

6. BUSINESS PROCESS CHAINS 2123
  6.1. Linking Members of the Supply Chain 2123
  6.2. Information Flow Enablers 2124
  6.3. Software Packages 2125

7. THE MANAGEMENT COMPONENTS OF SUPPLY CHAIN MANAGEMENT 2125
  7.1. Planning and Control Methods 2125
  7.2. Work Flow/Activity Structure 2125
  7.3. Organization Structure 2125
  7.4. Communication and Information Flow Facility Structure 2125
  7.5. Product Flow Facility Structure 2125

---

*This chapter is adapted from D. M. Lambert and J. R. Stock, "Supply Chain Management," in *Strategic Logistics Management,* 4th Ed., Irwin/McGraw-Hill, Burr Ridge, IL, 2001. Used with permission. All rights reserved.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

7.6. Management Methods 2126
7.7. Power and Leadership Structure 2126
7.8. Risk and Reward Sharing 2126
7.9. Culture and Attitude 2126
8. SUPPLY CHAIN DESIGN 2127
8.1. The Manufacturer's Perspective 2127
8.2. The Wholesaler's Perspective 2128
8.3. The Retailer's Perspective 2128
9. SUPPLY CHAIN DESIGN CONSIDERATIONS 2128
9.1. Market Coverage Objectives 2128
9.1.1. Customer Buying Behavior 2128
9.1.2. Type of Distribution 2129
9.1.3. Channel Structure 2129
9.1.4. Control 2129
9.2. Product Characteristics 2129
9.2.1. Value 2129
9.2.2. Technicality 2129
9.2.3. Market Acceptance 2130
9.2.4. Substitutability 2130
9.2.5. Bulk 2130
9.2.6. Perishability 2130
9.2.7. Market Concentration 2130
9.2.8. Seasonality 2130
9.2.9. Width and Depth 2130
9.3. Customer Service Objectives 2130
9.3.1. Availability 2131
9.3.2. Order Cycle 2131
9.3.3. Communication 2131
10. SUPPLY CHAIN PERFORMANCE MEASUREMENT 2131
11. REENGINEERING IMPROVEMENT INTO THE SUPPLY CHAIN 2132
12. IMPLEMENTING INTEGRATED SUPPLY CHAIN MANAGEMENT 2133
13. MANAGING SUPPLIER RELATIONSHIPS 2134
13.1. Types of Partnerships 2135
13.1.1. The Partnership Model 2135
14. SUMMARY 2138
REFERENCES 2138
ADDITIONAL READING 2140

1.0 INTRODUCTION

In any society, industrialized or nonindustrialized, goods must be physically moved or transported between the place they are produced and the place they are consumed. Except in very primitive cultures, where each family meets its own household needs, the exchange process has become the cornerstone of economic activity. Exchange takes place when there is a discrepancy between the amount, type, and timing of goods available and the goods needed. If a number of individuals or organizations within the society have a surplus of goods that someone else needs, there is a basis for exchange. When many exchanges take place between producers and consumers, the alignment of firms that bring products or services to market has been called the supply chain, the demand chain, or the value chain. In this chapter we will use the term *supply chain* to represent this alignment of firms.

2. SUPPLY CHAIN MANAGEMENT VS. LOGISTICS

Supply chain management is a term that has grown significantly in use and popularity since the late 1980s, although considerable confusion exists about what it actually means. Many people use the term as a substitute or synonym for *logistics*. However, the definition of supply chain management used in this chapter is much broader than logistics.

2.1. Definition of Supply Chain Management

"Supply chain management is the integration of key business processes from end user through original suppliers that provides products, services, and information that add value for customers and other stakeholders" (Lambert et al. 1998). There are a number of important differences between this definition of supply chain management and the Council of Logistics Management* definition of logistics.

---

* The Council of Logistics Management is the leading-edge professional logistics organization, with a current membership of over 15,000.

First and foremost, supply chain management is the management of all key business processes, including customer relationship management, customer service management, demand management, order fulfillment, manufacturing flow management, procurement, product development and commercialization, and returns. Key requirements for successful implementation of supply chain management are executive support, leadership, commitment to change, and empowerment. These requirements will be described along with the key processes later in the chapter.

Thus, supply chain management (SCM) is a systems approach that is highly interactive and complex and requires simultaneous consideration of many trade-offs. As shown in Figure 1, SCM spans organizational boundaries, considering trade-offs both within and among organizations regarding where inventory should be held and where activities should be performed.

In addition to the processes involved in supply chain management, this figure illustrates the product flows and information linkages that must take place in a supply chain. Remember that product flows take place only after information flows are initiated.

Due to the dynamic nature of the business environment, management must monitor and evaluate the performance of the supply chain regularly and frequently. When performance goals are not met, management must evaluate possible supply chain alternatives and implement changes. SCM is particularly important in mature and declining markets, during periods of economic slowdown when market growth cannot conceal inefficient practices, and when product life cycles are extraordinarily short. It is also critical in new product/market development, when the organization is making decisions related to supply chain configuration.

2.2. Difference between Logistics and Supply Chain Management*

The term *supply chain management* (SCM) was originally introduced by consultants in the early 1980s (Oliver and Webber 1982) and has subsequently gained tremendous attention (La Londe 1998). Since 1989, academics have attempted to give structure to SCM (Stevens 1989; Towill et al. 1992; Ellram and Cooper 1993; Bechtel and Jayaram 1997).

Until recently, most practitioners (Davis 1993; Arntzen et al. 1995; Lee and Billington 1995; Camp and Colbert 1997), consultants (Scharlacken 1998; Tyndall et al. 1998; Copacino 1997), and

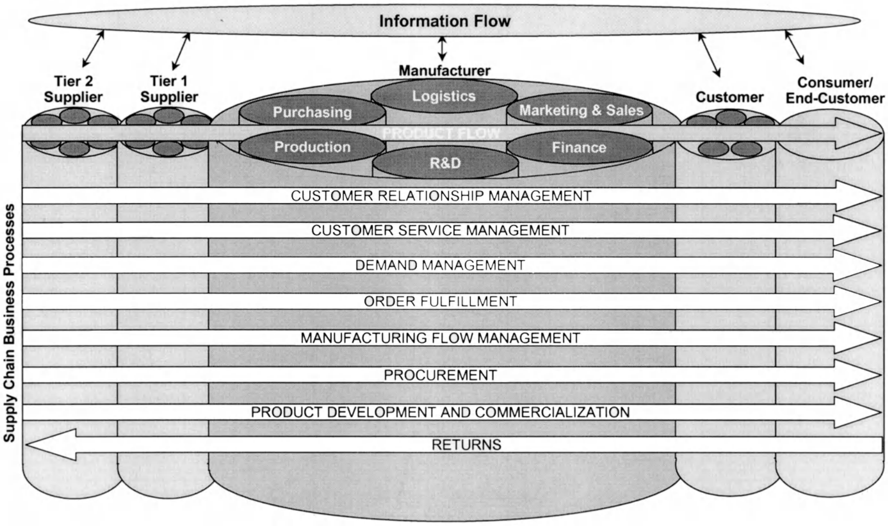

Information Flow
Tier 2 Supplier
Tier 1 Supplier
Manufacturer
Logistics
Purchasing
Marketing & Sales
Production
Finance
R&D
PRODUCT FLOW
Customer
Consumer/ End-Customer
Supply Chain Business Processes
CUSTOMER RELATIONSHIP MANAGEMENT
CUSTOMER SERVICE MANAGEMENT
DEMAND MANAGEMENT
ORDER FULFILLMENT
MANUFACTURING FLOW MANAGEMENT
PROCUREMENT
PRODUCT DEVELOPMENT AND COMMERCIALIZATION
RETURNS

Figure 1 Supply Chain Management: Integrating and Managing Business Processes across the Supply Chain. (From D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, 1998, pp. 1–19. Reprinted with permission)

---

*This section is taken from D. M. Lambert, M. C. Cooper and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, 1998, pp. 2–5.

academics (Fisher 1997; Lee and Billington 1992; Handfield and Nichols 1999; Bowersox and Closs 1996) viewed SCM as not appreciably different from the contemporary understanding of logistics management. That is, SCM was viewed as logistics outside the firm to include customers and suppliers. However, in 1986, the Council of Logistics Management (CLM) defined logistics management as ''The process of planning, implementing, and controlling the efficient, cost-effective flow and storage of raw materials, in-process inventory, finished goods, and related information flow from point-of-origin to point-of-consumption for the purpose of conforming to customer requirements.''

Logistics as defined by the Council of Logistics Management always represented a supply chain orientation, ''from point of origin to point of consumption.'' Then why the confusion? Probably because logistics is a functional silo within companies and is also a bigger concept that deals with the management of material and information flows across the supply chain. This is similar to the confusion over marketing as a concept and marketing as a functional area. Thus the quote from the CEO: ''Marketing is too important to be left to the marketing department.'' Everybody in the company should have a customer focus. The marketing concept does not apply just to the marketing department. It is everybody's responsibility to focus on serving the customer's needs.

Executives in companies leading the drive to implement SCM visualize the necessity of integrating all key business operations across the supply chain (Giunipero and Brand 1996; Bowersox 1997a). This broader understanding of SCM is likewise the core message in the following statement by James E. Morehouse, Vice President of A.T. Kearney, management consultants. ''For companies to survive and prosper, they will need to operate their supply chains as extended enterprises with relationships which embrace business processes, from materials extraction to consumption.'' Thus, the understanding of SCM has been reconceptualized from integrating logistics across the supply chain to integrating and managing key business processes across the supply chain (Cooper et al. 1997b). Based on this emerging distinction between SCM and logistics, in 1998, CLM announced a modified definition of logistics. The modified definition explicitly declares CLM's position that logistics management is only a part of SCM: ''Logistics is that part of the supply chain process that plans, implements, and controls the efficient, effective flow and storage of goods, services, and related information from the point-of-origin to the point-of-consumption in order to meet customers' requirements'' (Council of Logistics Management 1998).

Managing the supply chain is a complicated task and even managing logistics from point of origin to point of consumption is a lot easier to write on a piece of paper than actually to do. Imagine the degree of complexity if you are actually going to manage all suppliers back to the point of origin and all products/services out to the point of consumption. It is probably easier to understand why executives would want to manage their supply chains to the point of consumption, because whoever has the relationship with the end user has the power in the supply chain. Intel has created a relationship with the end user by having computer manufacturers place an ''Intel chip inside'' label on their computers. This affects the computer manufacturer's ability to switch chip suppliers. But managing all tier 1 suppliers' networks to the point of origin is an enormous undertaking. Managing the entire supply chain is a very difficult and challenging task, as illustrated in Figure 2.

The early marketing channel researchers, such as Wroe Alderson and Louis P. Bucklin, conceptualized the key factors for why and how channels are created and structured (Alderson 1950; Cox and Alderson 1950; Bucklin 1966). From a supply chain standpoint, these researchers were basically on the right track, particularly in the areas of identifying who should be a member of the marketing channel, describing the need for channel coordination, and drawing actual marketing channels. However, for the last 30 years the channels researchers studied power and conflict with questionable results and ignored two critical issues. First, they did not build on the early contributions by including suppliers to the manufacturer, and thus neglected the importance of a total supply chain perspective. Second, they focused on marketing activities and flows across the channel and overlooked the need to integrate and manage multiple key processes across companies.

Unlike the marketing channels literature, a major weakness of the SCM literature to date is that the authors appear to assume that everyone knows who is a member of the supply chain. Little effort has been given to identifying specific supply chain members, key processes that require integration, or what management must do to successfully manage the supply chain. The SCM framework presented here encompasses the combination of three closely interrelated elements: the structure of the supply chain, the supply chain business processes, and the supply chain management components (see Figure 3). We believe that the combination of these three elements captures the essence of SCM.

The supply chain structure is the network of members and the links between members of the supply chain. Business processes are the activities that produce a specific output of value to the customer. The management components are the managerial variables by which the business processes are integrated and managed across the supply chain. In combination, the SCM definition and this new framework move the SCM philosophy to its next evolutionary stage.

The implementation of SCM involves identifying the supply chain members, with whom it is critical to link, what processes need to be linked with each of these key members, and what type/level of integration applies to each process link. The objective of SCM is to maximize competitiveness

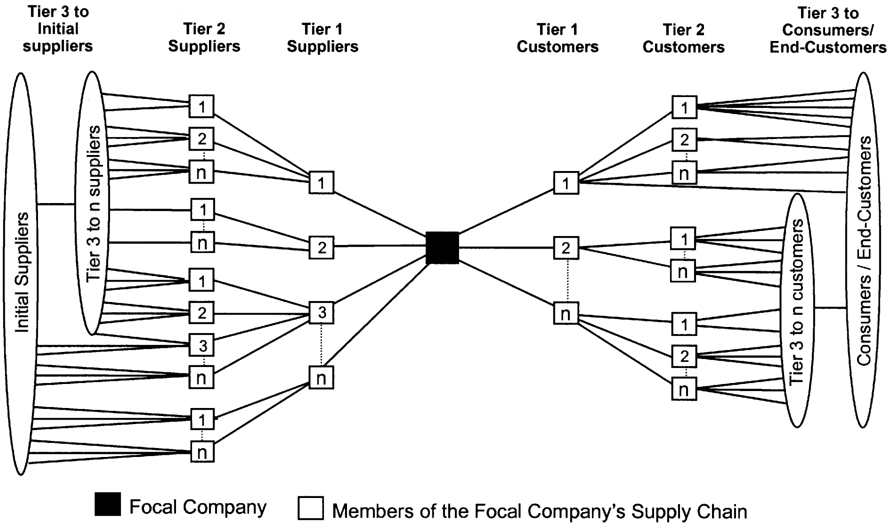

Tier 3 to Initial suppliers
Tier 2 Suppliers
Tier 1 Suppliers
Tier 1 Customers
Tier 2 Customers
Tier 3 to Consumers/ End-Customers
Initial Suppliers
Tier 3 to n suppliers
Tier 3 to n customers
Consumers / End-Customers
1
2
3
n
Focal Company
Members of the Focal Company's Supply Chain

Figure 2 Supply Chain Network Structure. (From D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, 1998, pp. 1–19. Reprinted with permission)

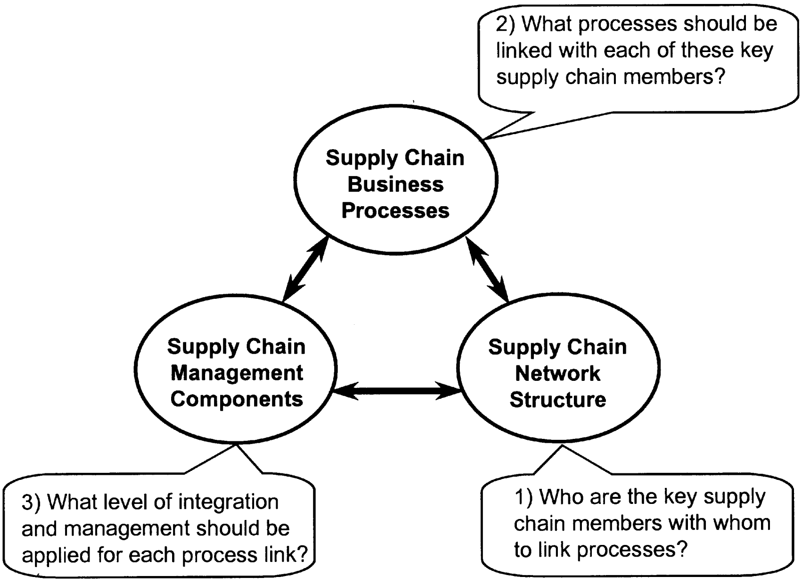

2) What processes should be linked with each of these key supply chain members?
Supply Chain Business Processes
Supply Chain Management Components
Supply Chain Network Structure
3) What level of integration and management should be applied for each process link?
1) Who are the key supply chain members with whom to link processes?

Figure 3 Supply Chain Management Framework: Elements and Key Decisions. (From D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, 1998, pp. 1–19. Reprinted with permission)

and profitability for the company as well as the whole supply chain network, including the end customer. Consequently, supply chain process integration and reengineering initiatives should be aimed at boosting total process efficiency and effectiveness across members of the supply chain.

3. CHANNEL STRUCTURE

The theory on channel structure described in the marketing literature provides a useful foundation for studying supply chain structure. According to this literature, channel structure may be viewed as a function of product life cycle, logistics systems, effective communication networks (Michman 1971; Ellram and Cooper 1990), product characteristics (Aspinwall 1958), and/or firm size (Weigand 1963). The most detailed theory of channel structure was developed by Bucklin (1966). He stated that the purpose of the channel is to provide consumers with the desired combination of its outputs (lot size, delivery time, and market decentralization) at minimal cost. Consumers determine channel structure by purchasing combinations of service outputs. The best channel has formed when no other group of institutions generates more profits or more consumer satisfaction per dollar of product cost. Bucklin concluded that functions will be shifted from one channel member to another in order to achieve the most efficient and effective channel structure.

Given a desired level of output by the consumer and competitive conditions, channel institutions will arrange their functional tasks in such a way as to minimize total channel costs. This shifting of specific functions may lead to the addition or deletion of channel members.

In deciding when and where to use channel intermediaries, the firm is really considering the make/buy or "outsourcing" decision. Does the organization need to develop the required skills and capabilities internally, or can it be done faster and more efficiently by using a third party?

3.1. Outsourcing Pieces of the Supply Chain

Approximately $40 billion of logistics services in the United States are being outsourced. (Piper Jaffray Equity Research 1999). And there are significant opportunities to outsource additional logistics services. Some examples of outsourcing services that are available include:

- A large pharmaceutical company outsources its worldwide distribution, providing on-site pharmacists at some centers to dispense high-value products.
- A third party handles the entire finished goods inventory for a large women's clothing company. When garments are purchased by a retailer, the distributor attaches the store's private label, refreshes the garment, packs it in the store's packaging, and ships to the retailer.
- A mail-order retailer is having FedEx handle not only their shipments, but also storage and management of the inventory and all aspects of distribution.
- In addition to handling store replenishment and delivery of product to consumers for a tool manufacturer, UPS now is going to handle the warehouse. If the retail store needs product, an order that reaches the distribution center by 9:00 p.m. will be at the store by the next morning (Richardson 1994).

Thus, outsourcing represents an opportunity that should be considered in supply chain design and evaluation of existing supply chains. In addition, the role and utility of the distributor are changing. In some cases, consolidation of suppliers and customers has reduced the value and functionality of distributors.

For example, Wal-Mart's large stores, which use direct distribution, replace small stores that may have used distributors. Similarly, advanced technology such as EDI trades information for inventory, reducing the need to hold inventory at distributors as well as at retailers (see Section 6.2). Better information technology and increased service offerings by carriers (such as cross-docking) also reduce the need for distributor's services (Copacino 1994).

3.2. Postponement and Speculation

Bucklin's theory of channel structure is based on the concepts of postponement and speculation (Bucklin 1965). Costs can be reduced by (1) postponing changes in the form and identity of a product to the last possible point in the marketing process and (2) postponing inventory location to the last possible point in time, since risk and uncertainty costs increase as the product becomes more differentiated. Postponement results in savings because it moves differentiation nearer to the time of purchase, when demand is more easily forecast. This reduces risk and uncertainty costs. Logistics costs are reduced by sorting products in large lots in relatively undifferentiated states. Third-party service providers can support postponement by mixing pallets for individual customers as orders are received, repackaging product to fit specific customer or country requirements, and performing final assembly or customization in the field.

Companies can use postponement to shift the risk of owning goods from one channel member to another. That is, a manufacturer may refuse to produce until it receives firm orders; a middleman may postpone owning inventories by purchasing from sellers who offer faster delivery, purchasing on consignment, or purchasing only when a sale has been made; and consumers may postpone ownership by buying from retail outlets where the products are in stock.

An excellent example of postponement is the mixing of paint colors at the retail store. Rather than having to forecast the exact colors that consumers will want to buy, the retailer mixes paint in any color the consumer wishes to acquire at the time of purchase. Other examples include color panels in the front of built-in kitchen appliances that enable the same unit to be any one of a number of colors; the centralization of slow-selling products in one warehouse location; and the assembly of slow-moving items only after orders have been received.

Speculation is the opposite of postponement: that is, a channel institution assumes risk rather than shifting it. Speculation can reduce marketing costs through (1) the economies of large-scale production; (2) the placement of large orders, which reduces the costs of order processing and transportation; (3) the reduction of stockouts and their associated cost; and (4) the reduction of uncertainty. To reduce the need for speculative inventories, managers in many firms are exploring strategies of time-based competition (Handfield 1991; Rafuse 1995; Christopher and Peck 1997). By using time-based competition, management can reduce significantly the firm's time to manufacture products while reducing inventory, improving inventory turns, reducing cost of ownership, and improving customer satisfaction.

3.3. Time-to-Market Pressures

Speed can be used as a source of competitive advantage. This is true in virtually all market sectors: services, manufacturing, and retailing. Retailers have been leaders in the area of time-based competition, relying heavily on advanced computer systems involving bar coding and EDI to support quick response (this will be described further in Section 6.2). The use of such systems is growing among carriers. But computer systems are not enough to create speed-to-market; fundamental changes in operational relationships are required. This includes information sharing among suppliers, manufacturers, and retailers, including lead times, forecasts of sales, production, and purchase needs, shipping, new product plans, and payment information.

Some of the benefits of effective time-based management include:

- Enhanced customer value through better responsiveness
- Reduced inventory requirements due to shorter lead times
- Reduced cost-added/duplicate functions
- Improved quality/product freshness through reduced handling and lower inventories
- Improved competitive position
- Increased responsiveness to changing market needs
- Improved productivity

3.4. Other Issues Affecting Channel Structure

Additional factors that can influence channel structure include:

- Technological, cultural, physical, social, and political factors
- Physical factors such as geography, size of market, location of production centers, and concentration of population
- Local, state, and federal laws
- Social and behavioral variables

For example, social, cultural, political, and economic variables may support channels that are not necessarily as efficient or effective as they should be.

4. SUPPLY CHAIN NETWORK STRUCTURE*

One key element of managing the supply chain is to have an explicit knowledge and understanding of how the supply chain network structure is configured. The three primary structural aspects of a

---

*This section is taken from D. M. Lambert, M. C. Cooper and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, 1998, pp. 5–9.

company's network structure are the members of the supply chain, the structural dimensions of the network, and the different types of process links across the supply chain. These three issues are all related to the first element: supply chain network structure, shown in Figure 3. Now each issue will be addressed.

4.1. Identifying Supply Chain Members

When determining the network structure, it is necessary to identify who the members of the supply chain are. Including all types of members may cause the total network to become highly complex, since it may explode in the number of members added from tier level to tier level (Cooper et al. 1997a). To integrate and manage all process links with all members across the supply chain would, in most cases, be counterproductive, if not impossible. The key is to sort out some basis for determining which members are critical to the success of the company and the supply chain and thus should be allocated managerial attention and resources.

Marketing channels researchers identified members of the channel based on who takes part in the various marketing flows, including product, title, payment, information, and promotion flows (Stern et al. 1996). Each flow included relevant members, such as banks for the payment flow and advertising agencies for the promotion flow. The channels researchers sought to include all members taking part in the marketing flows, regardless of how much impact each member had on the value provided to the end customer or other stakeholders.

The members of a supply chain include all companies/organizations with whom the focal company interacts directly or indirectly through its suppliers or customers, from point of origin to point of consumption. However, to make a very complex network more manageable, it seems appropriate to distinguish between primary and supporting members. Primary members of a supply chain are defined to be all those autonomous companies or strategic business units who actually perform operational and/or managerial activities in the business processes designed to produce a specific output for a particular customer or market.

In contrast, the supporting members of a supply chain are companies that simply provide resources, knowledge, utilities, or assets for the primary members of the supply chain.

For example, supporting companies include those that lease trucks to the manufacturer, banks that lend money to a retailer, the owner of the building that provides warehouse space, or companies that supply production equipment or print marketing brochures or provide temporary secretarial assistance. These supply chain members support the primary members now and in the future. Resource, knowledge, utility, or asset providers are important, if not vital, contributors to a company and the supply chain, but they do not directly participate in or perform activities in the value-adding processes of transforming inputs to outputs for the end customer.

The same company can perform both primary and supportive activities. Likewise, the same company can perform primary activities related to one process and supportive activities related to another process. An example from one of the case studies is an OEM that buys some critical and complex production equipment from a supplier. When the OEM develops new products, it works very closely with the equipment supplier, and thus the supplier is a primary member of the OEM's product-development process. However, when looking at the manufacturing flow management process, the supplier is a supportive and not a primary member, since supplying the equipment does not in itself add value to the output of the processes, even though the equipment does add value.

It should be noted that the distinction between primary and supporting chain members is not obvious in all cases. Nevertheless, this distinction provides a reasonable managerial simplification and yet captures the essential aspects of who should be considered key members of the supply chain. The approach for differentiating between types of members is to some extent similar to how Porter distinguished between value-adding and support activities in his value chain framework (Porter 1984).

The definitions of primary and supporting members make it possible to define the point of origin and the point of consumption of the supply chain. The point of origin of the supply chain occurs where no primary suppliers exist. All suppliers to the point of origin members are solely supporting members. The point of consumption is where no further value is added and the product and/or service is consumed.

4.2. The Structural Dimensions of the Network

Three structural dimensions of the network are essential when describing, analyzing, and managing the supply chain. These dimensions are the horizontal structure, the vertical structure, and the horizontal position of the focal company within the end points of the supply chain.

The horizontal structure refers to the number of tiers across the supply chain. The supply chain may be long, with numerous tiers, or short, with few tiers. The vertical structure refers to the number of suppliers/customers represented within each tier. A company can have a narrow vertical structure, with few companies at each tier level, or a wide vertical structure with many suppliers and/or customers at each tier level. The third structural dimension is the company's horizontal position within the supply chain. A company can be positioned at or near the initial source of supply, at or near to the ultimate customer, or somewhere between these end points of the supply chain.

Different combinations of these structural variables are possible. For example, a narrow and long network structure on the supplier side can be combined with a wide and short structure on the customer side. Increasing or reducing the number of suppliers and/or customers will affect the structure of the supply chain. As companies move from multiple- to single-source suppliers, the supply chain will become narrower. Outsourcing logistics, manufacturing, marketing, or product-development activities is another example of decision making that likely will change the supply chain structure. It may increase the length and width of the supply chain and likewise influence the horizontal position of the focal company in the supply chain network.

Supply chains that burst to many tier 1 customers/suppliers will strain corporate resources and limit the number of process links the focal company can integrate and closely manage beyond tier 1. In general, managers in companies with immediately wide vertical structures actively manage only a few tier 2 customers or suppliers. Some companies have transferred servicing small customers to distributors, thus moving the small customers farther down in the supply chain from the focal company. This principle, known as functional spin-off, is described in the channels literature (Stern et al. 1996) and can be applied to the focal company's network of suppliers as well as to its customers.

Supply chains look different from each company's perspective because management of each company sees its firm as the focal company and views membership and network structure differently. Thus, the perceived supply chain network structure is arbitrary. However, because each firm is a member of the other's supply chain, it is important for management of each firm to understand its interrelated roles and perspectives. This is because the integration and management of business processes across company boundaries will be successful only if it makes sense from each company's perspective (Cooper et al. 1997b).

4.3. Types of Business Process Links

Integrating and managing all business process links throughout the entire supply chain is likely not appropriate. Since the drivers for integration are situational and different from process link to process link, the levels of integration will/should likewise vary from link to link and over time. Thus, some links are more critical than others (Håkansson and Snehota 1995). As a consequence, the task of allocating scarce resources among the different business process links across the supply chain becomes crucial. Four fundamentally different types of business process links can be identified between members of a supply chain (Lambert et al. 1998): managed business process links, monitored business process links, not-managed business process links, and nonmember business process links.

Managed process links. Managed process links are links that the focal company finds important to integrate and manage. This might be in collaboration with other member companies of the supply chain. In the supply chain drawn in Figure 4, the managed process links are indicated by the thickest solid lines. The focal company will integrate and manage process links with tier 1 customers and suppliers. As indicated by the remaining thick solid lines in Figure 4, the focal company is actively involved in the management of a number of other process links beyond tier 1.

Monitored process links. Monitored process links are not as critical to the focal company. However, it is important to the focal company that these process links be integrated and managed appropriately between the other member companies. Thus, the focal company, as frequently as necessary, simply monitors or audits how the process link is integrated and managed. The thick dashed lines in Figure 4 indicate the monitored process links.

Not-managed process links. Not-managed process links are links that the focal company is not actively involved in and are not critical enough to use resources for monitoring. In other words, the focal company fully trusts the other members to manage the process links appropriately, or because of limited resources leaves it up to them. The thin solid lines in Figure 4 indicate the not-managed process links. For example, a manufacturer has a number of suppliers for cardboard shipping cartons. Usually the manufacturer will not choose to integrate and manage the links beyond the cardboard supplier all the way back to the growing of the trees. The manufacturer wants certainty of supply, but it may not be necessary to integrate and manage the links beyond the cardboard supplier.

The three alternatives for integrating and managing links are illustrated in Figure 5. Company A may chose to integrate with and actively manage link 2 (alternative 1). Or company A could chose not to integrate but only to monitor the procedures of companies B and C for integrating and managing link 2 (alternative 2). Both alternatives 1 and 2 necessitate some level of resource allocation from company A. Finally, company A can choose not to be involved and leave the integration and management of link 2 up to companies B and C (alternative 3).

Nonmember process links. Managers understand that their supply chains are influenced by decisions made in other connected supply chains. For example, a supplier to the focal company is also a supplier to the chief competitor. Such a supply chain structure may have implications for the

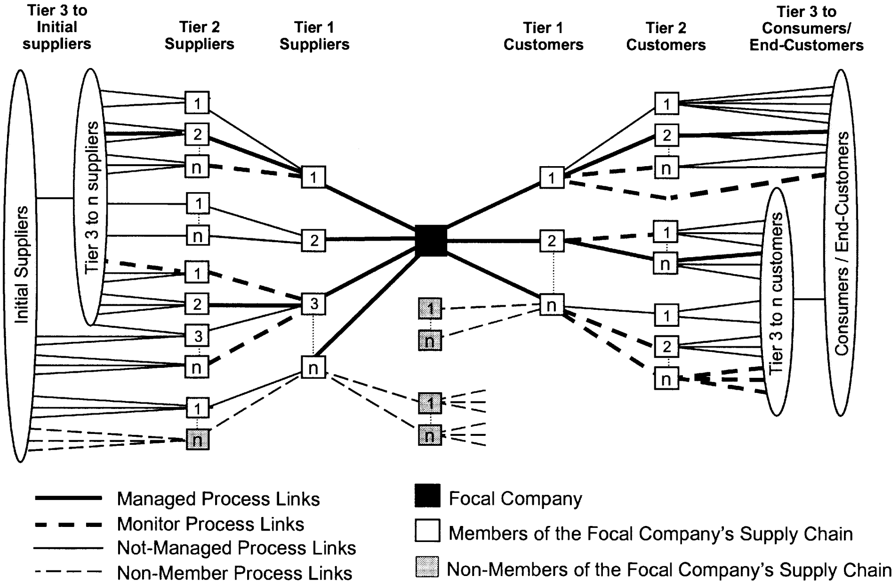

Tier 3 to Initial suppliers
Tier 2 Suppliers
Tier 1 Suppliers
Tier 1 Customers
Tier 2 Customers
Tier 3 to Consumers/ End-Customers
Initial Suppliers
Tier 3 to n suppliers
Tier 3 to n customers
Consumers / End-Customers
1
2
n
1
n
1
2
3
n
1
n
1
2
3
n
1
n
1
n
1
2
n
1
2
n
1
n
1
2
n
Managed Process Links
Monitor Process Links
Not-Managed Process Links
Non-Member Process Links
Focal Company
Members of the Focal Company's Supply Chain
Non-Members of the Focal Company's Supply Chain

Figure 4 Types of Inter-company Business Process Links. (From D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, 1998, pp. 1–19. Reprinted with permission)

supplier's allocation of manpower to the focal company's product-development process, availability of products in times of shortage, and/or protection of confidentiality of information. This leads us to identify a fourth type of business link, nonmember process links. Nonmember process links are process links between members of the focal company's supply chain and nonmembers of the supply chain. Nonmember links are not considered as links of the focal company's supply chain structure, but they can and often will affect the performance of the focal company and its supply chain. The thin dashed lines in Figure 4 illustrate examples of nonmember process links.

Based on the process links just described, there is variation in how closely companies integrate and manage links farther away from the first tier. In some cases, companies work through or around

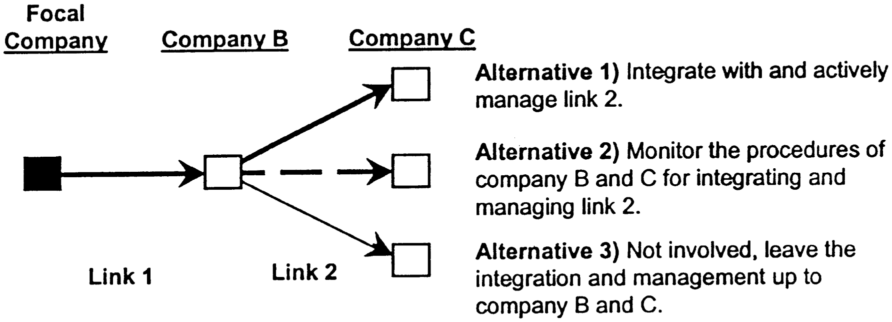

Focal Company
Company B
Company C
Link 1
Link 2
Alternative 1) Integrate with and actively manage link 2.
Alternative 2) Monitor the procedures of company B and C for integrating and managing link 2.
Alternative 3) Not involved, leave the integration and management up to company B and C.

Figure 5 The Focal Company's Alternatives for Involvement with Link 2. (From D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, 1998, pp. 1–19. Reprinted with permission)

other members/links in order to achieve specific supply chain objectives, such as product availability, improved quality, and reduced overall supply chain costs. For example, a tomato ketchup manufacturer in New Zealand conducts research on tomatoes in order to develop plants that provide larger tomatoes with fewer seeds. Their contracted growers are provided with young plants in order to ensure the quality of the output. Since the growers tend to be small, the manufacturer negotiates contracts with suppliers of equipment and supplies such as fertilizer and chemicals. The farmers are encouraged to purchase their raw materials and machinery using the contract rates. This results in higher-quality raw materials and lower prices without sacrificing the margins and financial strength of the growers.

There are several examples of companies who, in times of shortage, discovered that it was important to manage beyond tier 1 suppliers for critical times. One example involves a material used in the manufacture of semiconductors. The manufacturer had six tier 1 suppliers from which to purchase. However, when shortages occurred, it became apparent that all six tier 1 suppliers purchased from the same tier 2 supplier. It turned out that the most critical relationship was with the tier 2 supplier.

4.4. Supply Chain Mapping Considerations

The mapping of the supply chain for a focal company is an important step in understanding the interrelationships that impact the success of the business. Several important points need to be kept in mind before proceeding with the mapping of the supply chain network:

- Supply chain mapping should begin from the right side (customer side) of the network, as shown in Figure 2. This will bring the correct focus to why the map is being constructed.
- Once the customer segments are identified, each segment should be evaluated in terms of its importance to the success of the focal company. Factors such as sales, profit, contribution margin, and competitive threat can be used to select the priority that each customer or customer segment has in the supply chain.
- Rather than proceeding with comprehensive mapping from right to left (customers to tier *n* Suppliers), it is suggested that the mapping proceed for one or two of the highest-priority customer segments. This process will limit the complexity that can quickly develop by building a complete map that may become weighed down by trivial nodes and a need for vast amounts of information.
- At this point, existing processes and performance can be included in the supply chain map.
- Lastly, performance goals should be established for each customer segment. With performance goals in place, the development of revised intermediate performance metrics and adjusted business processes that support the new goals can proceed.

5. SUPPLY CHAIN BUSINESS PROCESSES*

Successful supply chain management requires a change from managing individual functions to integrating activities into key supply chain processes. Traditionally, both upstream and downstream portions of the supply chain have interacted as disconnected entitles receiving sporadic flows of information over time.

The purchasing department placed orders as requirements became necessary, and marketing, responding to customer demand, interfaced with various distributors and retailers and attempted to satisfy this demand. Orders were periodically given to suppliers and their suppliers, who had no visibility of demand at the point of sale or use. Satisfying the customer often required expedited operations throughout the supply chain as member firms reacted to unexpected changes in demand.

Operating an integrated supply chain requires continuous information flows, which in turn help to create the best product flows. The customer remains the primary focus of the process. However, improved linkages with suppliers are necessary because controlling uncertainty in customer demand, manufacturing processes, and supplier performance is critical to effective supply chain management (SCM). Achieving a good customer-focused system requires processing information both accurately and in a timely fashion because quick-response systems require frequent changes in response to fluctuations in customer demand.

In many major corporations, such as 3M, management has reached the conclusion that optimizing the product flows cannot be accomplished without implementing a process approach to the business. The key supply chain processes are:

---

*This material is adapted from D. M. Lambert, L. C. Guinipero, and G. Ridenhower, "Supply Chain Management: A Key to Achieving Business Excellence in the 21st Century," manuscript. All rights reserved.

- Customer relationship management
- Customer service management
- Demand management
- Order fulfillment
- Manufacturing flow management
- Procurement
- Product development and commercialization
- Returns

These processes were identified in Figure 1. While the specific processes identified by individual firms may vary somewhat from those above, supply chain management must consider five fundamental processes: selling, customer order fulfillment, manufacturing flow, procurement, and product development.

Of course, performance metrics must be changed to reflect process performance across the supply chain, and rewards and incentives must be aligned to these metrics in order to effect change. Each of the eight processes will now be described.

5.1. Customer Relationship Management Process

The first step towards integrated SCM is to identify key customer or customer groups that the organization targets as critical to its business mission. The corporate business plan is the starting point for this analysis. Customer service teams develop and implement partnering programs with key customers. Product and/or service agreements specifying the levels of performance are established with these key customer groups. In many cases, the product/service agreements will be tailored to meet the needs of key individual customers.

New customer interfaces lead to improved communications and better predictions of customer demand, which leads to improved service for customers. Customer service teams work with customers to identify further and eliminate sources of demand variability. Performance evaluations are undertaken to analyze the levels of service provided to customers as well as customer profitability.

5.2. Customer Service Management Process

Customer service provides the single source of customer information. It becomes the key point of contact for administering the product/service agreement. Customer service provides the customer with real-time information on promised shipping dates and product availability through interface with the organizations' production and distribution operations.

Managing customer service in an SCM environment requires an online, real-time system to provide product and pricing information to support customer inquiries and facilitate order placement. After-sales service is also a requirement. Finally, the technical customer service group must be able to assist the customer efficiently with product applications and recommendations.

5.3. Demand Management Process

Hewlett-Packard's experience with SCM indicates that inventory is either essential or variability driven (Davis 1993). Essential inventory includes work-in-process in factories and products in the pipeline moving from location to location. Time-based and periodic-review systems lead to certain amounts of incoming inventory stock. Variability stock is present due to variance in process, supply, and demand. Customer demand is by far the largest source of variability and stems from irregular order patterns. Given this variability in customer ordering, demand management is a key to an effective SCM process.

The demand management process must balance the customer's requirements with the firm's supply capabilities. Part of managing demand involves attempting to determine what and when customers will purchase. A good demand management system uses point-of-sale and key customer data to reduce uncertainty and provide efficient flows throughout the supply chain.

Marketing requirements and production plans should be coordinated on an enterprise-wide basis. Thus, multiple sourcing and routing options are considered at the time of order receipt, which allows market requirements and production plans to be coordinated on an organization-wide basis. In very advanced SCM systems, customer demand and production rates are synchronized to manage inventories globally.

5.4. Customer Order-Fulfillment Process

The key to effective SCM is meeting or exceeding customer need dates. It is important to achieve high order-fill rates either on a line item or order basis. Performing the order-fulfillment process effectively requires integration of the firm's manufacturing, distribution, and transportation plans. As previously discussed, partnerships should be developed with key supply chain members and carriers to meet customer requirements and reduce total delivered cost to customer. The objective is to develop a seamless process from the supplier to the organization and then on to its various customer segments.

With the growth of Internet-based businesses that are almost virtual in nature (i.e., no manufacturing or fulfillment assets), the process is carried out by contracted specialists. This implies that information integration becomes the key interaction between the supply chain entities.

5.5. Manufacturing Flow Management Process

The manufacturing process in make-to-stock firms traditionally produced and supplied product to the distribution channel based on historical forecasts. Products were pushed through the plant to meet a schedule. Often the wrong mix of products was produced, resulting in unneeded inventories, excessive inventory carrying costs, mark-downs, and transshipments of product.

With SCM, product is pulled through the plant based on customer needs. Manufacturing processes must be flexible to respond to market changes. This requires the flexibility to perform rapid changeover to accommodate mass customization. Orders are processed on a just-in-time basis in minimum lot sizes. Production priorities are driven by required delivery dates.

At 3M, manufacturing planners work with customer planners to develop strategies for each customer segment. Changes in the manufacturing flow process lead to shorter cycle times, meaning improved responsiveness to customers.

5.6. Procurement Process

Strategic plans are developed with suppliers to support the manufacturing flow management process and the development of new products. Suppliers are strategically categorized based on several dimensions, such as their contribution and criticality to the organization. In companies where operations extend world-wide, sourcing should be managed from corporate on a global basis.

Long-term partnerships are developed with a small core group of suppliers. The desired outcome is a win–win relationship where both parties benefit. This is a change from the traditional bid-and-buy system to involving a key supplier early in the design cycle, which can lead to dramatic reduction in product-development cycle times. Having early supplier input reduces time by getting the required coordination between engineering, purchasing, and the supplier prior to design finalization.

The purchasing function develops rapid communication mechanisms such as EDI and Internet linkages to transfer requirements quickly. These rapid communication tools provide a means to reduce time and cost spent on the transaction portion of the purchase. Purchasers can focus their efforts on managing suppliers as opposed to placing orders and expediting.

5.7. Product Development and Commercialization

If new products are the lifeblood of a corporation, then product development is the lifeblood of a company's new products. Customers and suppliers must be integrated into the product-development process in order to reduce time to market. As product life cycles shorten, the right products must be developed and successfully launched in ever-shorter time frames in order to remain competitive.

Managers of the product-development and commercialization process must:

- Coordinate with customer relationship management to identify customer articulated and unarticulated needs
- Select materials and suppliers in conjunction with procurement
- Develop production technology in manufacturing flow to assess manufacturability and integration into the best supply chain flow for the product/market combination

5.8. Returns Process

Managing the returns channel as a business process offers the same opportunity to achieve a sustainable competitive advantage as managing the supply chain from an outbound perspective (Clendenin 1997). Effective process management of the returns channel enables identification of productivity-improvement opportunities and breakthrough projects.

At Xerox, returns are managed in four categories: equipment, parts, supplies, and competitive trade-ins. "Return to available" is a velocity measure of the cycle time required to return an asset to a useful status. This metric is particularly important for those products where customers are given an immediate replacement in the case of product failure. Also, equipment destined for scrap and waste from manufacturing plants is measured in terms of the time until cash is received.

5.9. SCM Process Summary

Focusing efforts on these key business processes, which extend from the end users to original suppliers, provides the foundation for a supply chain management philosophy. The goals or outcomes of these processes are to:

- Develop customer-focused teams that provide mutually beneficial product and service agreements to strategically significant customers
- Provide a point of contact for all customers that efficiently handles their inquiries
- Continuously gather, compile, and update customer demand to match requirements with supply
- Develop flexible manufacturing systems that respond quickly to changing market conditions
- Manage supplier partnerships that allow for quick response and continuous improvement
- Fill 100% of customer orders accurately and on time
- Minimize the return to available cycle time

A responsive, flexible integrated supply chain can accomplish these objectives. Because as previously mentioned, these processes cut across business functions, it is important to examine or reengineer each key process using a systematic approach.

6. BUSINESS PROCESS CHAINS*

Thousands of activities are performed and coordinated within a company, and every company is by nature in some way involved in supply chain relationships with other companies (Bowersox 1997b; Stigler 1951; Coase 1937). When two companies build a relationship, certain of their internal activities will be linked and managed between the two companies (Håkansson and Snehota 1995). Since both companies have linked some internal activities with other members of their supply chain, a link between two companies is thus a link in what might be conceived as a supply chain network. For example, the internal activities of a manufacturer are linked with and can affect the internal activities of a distributor, which in turn are linked with and can have an effect on the internal activities of a retailer. Ultimately, the internal activities of the retailer are linked with and can affect the activities of the end customer.

6.1. Linking Members of the Supply Chain

Håkansson and Snehota (1995) stress that ‘‘the structure of activities within and between companies is a critical cornerstone of creating unique and superior supply chain performance.’’ Executives in leading companies believe that competitiveness and profitability can increase if internal key activities and business processes are linked and managed across multiple companies. Thus, ‘‘successful supply chain management requires a change from managing individual functions to integrating activities into key supply chain business processes’’ (Lambert et al. 1997).

Companies in the same supply chain may have different activity structures. Some companies emphasize a functional structure, some a process structure, and others a combined structure of processes and functions. Companies with processes often have different numbers of processes consisting of different activities and links between activities. Further, different names are used for similar processes and similar names for different processes. This lack of intercompany consistency is a cause of significant friction and inefficiencies in supply chains. At least with functional silos, there is generally an understanding of what functions like marketing, manufacturing, and accounting/finance represent. If each firm identifies its own set of processes, how do we communicate and link these processes across firms? A simplified illustration of such a disconnected supply chain is shown in Figure 6.

A process can be viewed as a structure of activities designed for action with a focus on end customers and on the dynamic management of flows involving products, information, cash, knowledge, and/or ideas.

In an exploratory study involving 30 successful supply chain redesign practitioners, Hewitt found that companies identified between 9 and 24 internal business processes. The two most commonly identifiable processes were order fulfillment and product development (Hewitt 1994).

A prerequisite for successful SCM is to coordinate activities within the firm. One way to do this is to identify the key business processes and manage them using cross-functional teams. In some cases the internal business processes have been extended to suppliers and managed to some extent between the two firms involved. This may imply that when a leadership role is taken, a firm’s internal business processes can become the supply chain business processes. The obvious advantage when this is possible is that each member of the band is playing the same tune.

The number of business processes that it is critical and/or beneficial to integrate and manage between companies will likely vary. In some cases it may be appropriate to link just one key process, and in other cases it may be appropriate to link multiple or all key business processes. However, in

---

*This section is taken from D. M. Lambert, M. C. Cooper and J. D. Pagh, ‘‘Supply Chain Management: Implementation Issues and Research Opportunities,’’ *International Journal of Logistics Management,* Vol. 9, 1998, pp. 9–11.

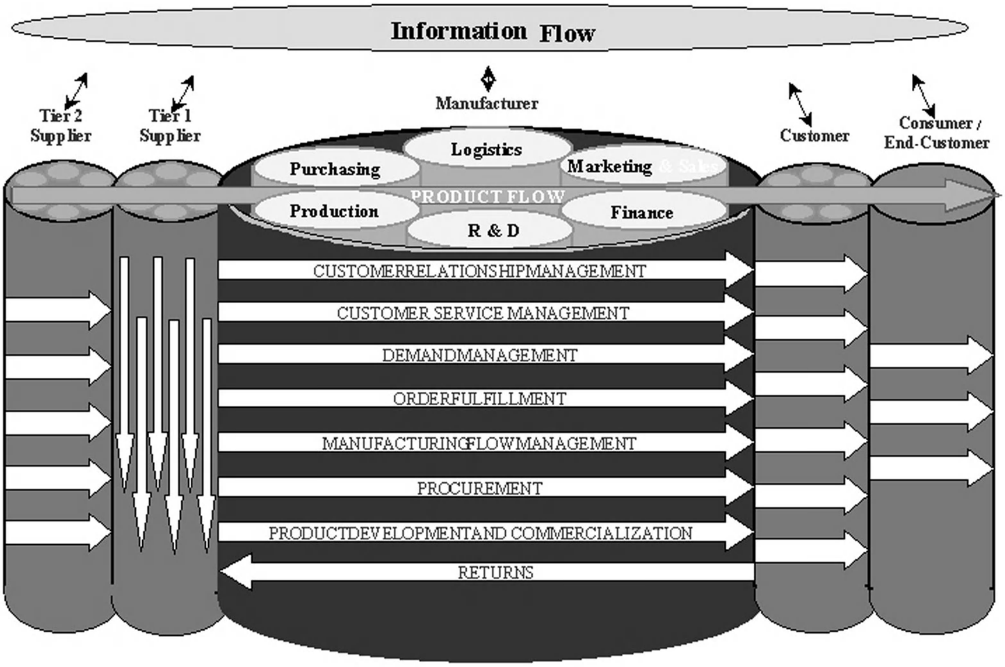

Information Flow
Tier 2 Supplier
Tier 1 Supplier
Manufacturer
Customer
Consumer / End-Customer
Logistics
Purchasing
Marketing & Sales
PRODUCT FLOW
Production
Finance
R & D
CUSTOMERRELATIONSHIPMANAGEMENT
CUSTOMER SERVICE MANAGEMENT
DEMANDMANAGEMENT
ORDERFULFILLMENT
MANUFACTURINGFLOWMANAGEMENT
PROCUREMENT
PRODUCTDEVELOPMENTAND COMMERCIALIZATION
RETURNS

Figure 6 Supply Chain Management: The Disconnects. (From D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, 1998, pp. 1–19. Reprinted with permission)

each specific case it is important that executives thoroughly analyze and discuss which key business processes to integrate and manage. The major components for integrating and managing a supply chain network are addressed next.

6.2. Information Flow Enablers

Process integration between members of the supply chain is dependent on the timely flow of the information that is identified as critical in the process mapping. Information can be exchanged in many forms, such as a telephone call or a paper report. In today's world of time-based competition, information exchange is normally based on computer-to-computer interaction. For many years, electronic data interchange (EDI), was the most common vehicle for exchanging information between computer systems, primarily between different companies. EDI is simply an agreed-to standardized format for a specific type of business transaction. Several organizations, such as the American National Standards Institute (ANSI), were instrumental in developing these standards. The standards are flexible, which is both a strength and a weakness. Specific details need to be worked out between the supply chain members for each type of transaction (e.g., purchase order or transportation information). This specificity adds cost for systems development as the number of entities is increased. The transmission of EDI information is usually facilitated by a value-added network (VAN) company that acts as an intermediary between the firms.

EDI is still used in many industries. The Internet has also become a transmission medium for transactions; however, newer methods are quickly becoming available for internet data communication. The most promising of these is extensible markup language (XML). The Internet is rapidly becoming the medium of choice for business-to-business e-commerce, and XML is becoming the language of this communication. The flexible structure of the language will allow for more trading partners to establish Internet links economically. At this time, the development of XML is moving so quickly that a good single reference for the reader is available on the Internet at www.xml.com. This source will provide the reader with references that will address the most timely status of XML specification and application.

6.3. Software Packages

Computer software is available for analysis, planning, and operation of various aspects of the supply chain. Since this Handbook will serve as a reference for several years before the next edition, it is not reasonable to mention specific software packages. However, the Council of Logistics Management provides an annual review of commercially available software. This catalog can serve as an important resource for finding potential software tools (Anderson Consulting 1999). Additional information regarding the Council of Logistics Management can be found at its website, www.clm1.org.

7. THE MANAGEMENT COMPONENTS OF SUPPLY CHAIN MANAGEMENT

The SCM management components are the third element of the SCM framework (see Figure 3). An essential underlying premise of the SCM framework is that certain management components are common across all business processes and members of the supply chain (Cooper et al. 1997b). We believe these common management components to be critical and fundamental for successful SCM because they essentially represent and determine how each process link is integrated and managed. The level of integration and management of a business process link is a function of the number and level, ranging from low to high, of components added to the link (Lambert et al. 1996a, b; Cooper et al. 1997a). Consequently, adding more management components or increasing the level of each component can increase the level of integration of the business process link.

The literature on SCM, business process reengineering, and buyer–supplier relationships suggests numerous possible components that must receive managerial attention when managing supply relationships (Cooper et al. 1997b; Lambert, et al 1996a, b; Olsen and Ellram 1997; Turnbull 1990). Each component can have several subcomponents, whose importance can vary depending on the process being managed. But the primary components are planning and control methods; work flow/activity structure; organization structure; communication and information flow facility structure; product flow facility structure; management methods; power and leadership structure; risk and reward structure; and culture and attitude. Each component is briefly described next.

7.1. Planning and Control Methods

Planning and control of operations are keys to moving an organization or supply chain in a desired direction. The extent of joint planning is expected to bear heavily on the success of the supply chain. Different components may be emphasized at different times during the life of the supply chain, but planning transcends the phases (Cooper and Ellram 1993). The control aspects can be operationalized as the best performance metrics for measuring supply chain success.

7.2. Work Flow/Activity Structure

The work flow/activity indicates how the firm performs its tasks and activities. The level of integration of processes across the supply chain would be a measure of organizational structure. All but one of the literature sources examined cites work structure as an important component.

7.3. Organization Structure

Organizational structure can refer to the individual firm and the supply chain. The use of cross-functional teams would suggest more of a process approach. When these teams cross-organizational boundaries, such as in-plant supplier personnel, the supply chain should be more integrated.

7.4. Communication and Information Flow Facility Structure

The information flow facility structure is key. The kind of information passed among supply chain members and the frequency of information updating has a strong influence on the efficiency of the supply chain. This may well be the first component integrated across part or all of the supply chain.

7.5. Product Flow Facility Structure

Product flow facility structure refers to the network structure for sourcing, manufacturing, and distribution across the supply chain. With reductions in inventory, fewer warehouses would be needed. Inventory is necessary in the system, but some supply chain members may keep a disproportionate amount of inventory. Since it is less expensive to have unfinished or semifinished goods in inventory than finished goods, upstream members may bear more of this burden. Rationalizing the supply chain network has implications for all members.

Product structure issues include how coordinated new product development is across the supply chain and the product portfolio. Lack of coordination in new product development can lead to inefficiencies of production, but there is also the risk of giving away corporate competence. The complexity of the product will likely affect the number of suppliers for the different components and the challenge of integrating the supply chain.

7.6. Management Methods

Management methods include the corporate philosophy and management techniques. It is very difficult to integrate a top-down organization structure with a bottom-up structure. The level of management involvement in day-to-day operations can differ across supply chain members.

7.7. Power and Leadership Structure

The power and leadership structure across the supply chain will affect its form. One strong leader will drive the direction of the supply chain. In most supply chains studied to date, there are one or two strong leaders among the firms. The exercise of power, or lack of it, can affect the level of commitment of other supply chain members. Forced participation will encourage exit behavior, given the opportunity (Macneil 1980; Williamson 1975).

7.8. Risk and Reward Sharing

The anticipation of sharing of risks and rewards across the supply chain affects the long-term commitment of its members. The recent fire at a Toyota supplier demonstrated Toyota's commitment to its suppliers and the assistance from other members of the chain.

7.9. Culture and Attitude

The importance of corporate culture and its compatibility across supply chain members cannot be underestimated. Meshing cultures and individuals' attitudes is time consuming but is necessary at some level for the supply chain to perform as a coordinated network. Aspects of culture include how employees are valued and incorporated into the management of the firm.

Figure 7 illustrates how the management components can be divided into two groups, to point out some basic differences. The first group is the physical and technical group, which includes the most visible, tangible, measurable, and easy-to-change components. Much of the literature on change management (Jaffe and Scott 1998; Andrews and Stalick 1994; Hammer 1990; Hammer and Champy

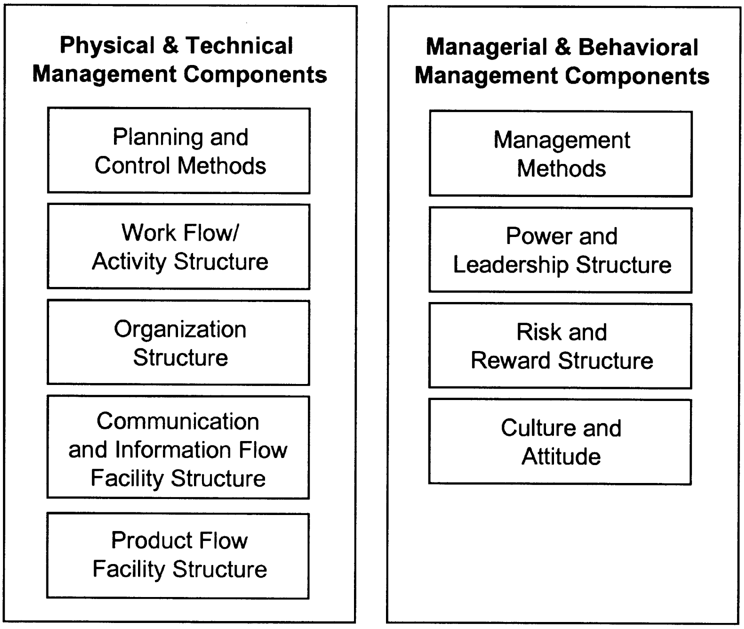

Physical & Technical Management Components
Planning and Control Methods
Work Flow/ Activity Structure
Organization Structure
Communication and Information Flow Facility Structure
Product Flow Facility Structure
Managerial & Behavioral Management Components
Management Methods
Power and Leadership Structure
Risk and Reward Structure
Culture and Attitude

Figure 7 Supply Chain Management: Fundamental Management Components. (From D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, 1998, pp. 1–19. Reprinted with permission)

1993; Towers 1994) shows that if this group of management components is the only focus of managerial attention, managing the supply chain will most likely be doomed to fail.

The second group is composed of the managerial and behavioral components. These components are less tangible and visible and are often difficult to assess and alter. The managerial and behavioral components define the organizational behavior and influence how the physical and technical management components can be implemented. If the managerial and behavioral components are not aligned to drive and reinforce an organizational behavior supportive of the supply chain objectives and operations, the supply chain will likely be less competitive and profitable. If one or more components in the physical and technical group are changed, management components in the managerial and behavioral group likewise may have to be readjusted. Consequently, the groundwork for successful SCM is established by understanding each of these management components and their interdependence. Hewitt states that true intra- and intercompany business process management, or redesign, is likely to be successful only if it is recognized as a multicomponent change process, simultaneously and explicitly addressing all SCM components (Hewitt 1994).

The physical and technical components are best understood and applied/managed the farthest up and down the supply chain. For example, in one case, the focal company had integrated its demand management process across four links by applying the following components: planning and control methods, work flow/activity structure, communication and information flow facility structure, and product flow facility structure. The managerial and behavioral management components are, in general, less well understood, and more difficulties are encountered in their implementation.

8. SUPPLY CHAIN DESIGN

Even though leading-edge firms are doing more planning of their supply chains, evidence suggests that the majority of supply chains were not designed but developed over time. For example, companies like Hewlett-Packard (Lee and Billington 1992; Davis 1993) and Digital Equipment (Arntzen et al. 1995) plan new channels/supply chains and use supply chain management strategies to modify existing networks. However, these examples appear to be the exception rather than the rule.

Current practice reveals a lack of planning by most firms. Better management of supply chains can create many benefits. For example, in many cases not all supply chain alternatives are known when structural arrangements are initially negotiated; these decisions may later prove to be less than optimal. Identifying suboptimal supply chain arrangements and making structural changes will lead to increased profitability.

In addition, unanticipated changes in the environment may make it necessary to reconsider the supply chain and reevaluate partnership arrangements. Environmental factors may include changes in end-consumer needs, markets, products and product lines, the competitive situation, the economic environment, and government regulation and incentives.

Supply chain strategy must be aligned with overall corporate strategy. Supply chain performance goals must be stated in operational terms, such as projected market coverage, sales and service support, sales volume, profitability, inventory turns, cash-to-cash cycle times, and return on investment. The supply chain strategy includes decisions regarding intensity of distribution, use of direct or indirect channels, the services of intermediaries in each geographic area, and implementation plans.

A firm must become involved in the supply chain design process when it is considering entering the market with a new product or when existing supply chains are falling short of performance objectives. The supply chain design process consists of the following steps (Lambert 1978):

1. Establish supply chain objectives.
2. Formulate a supply chain strategy.
3. Determine supply chain structure alternatives.
4. Evaluate supply chain structure alternatives.
5. Select supply chain structure.
6. Determine alternatives for individual supply chain members.
7. Evaluate and select individual supply chain members.
8. Measure and evaluate supply chain performance.
9. Evaluate supply chain alternatives when performance objectives are not met or attractive new options become available.

The manufacturer, wholesaler, or retailer may lead the design process, depending on the relative market power, financial strength, and availability of desired supply chain members.

8.1. The Manufacturer's Perspective

A manufacturer has market power when customers demand its product. When consumers demand a manufacturer's brand, retailers and consequently wholesalers are anxious to market its existing and new products because such products will draw customers. Increasingly the consolidation of manufacturers, wholesalers, and retailers on a national and global basis has resulted in a power shift to retailers since they have access to consumers. The consolidation of manufacturers results in a reduced set of global suppliers that produce brands which are increasingly viewed as substitutable by consumers. The store brands of retailers such as Wal-Mart become national and some cases global brands themselves, which has further contributed to the weakening of traditionally strong manufacturer brands.

A small manufacturer of a little-known brand may find it difficult to attract supply chain members for its existing or new product offerings. Such a manufacturer lacks market power when entering supply chain negotiations. Also, since financial resources determine a manufacturer's ability to perform marketing functions internally, small manufacturers usually cannot afford to distribute directly to retailers or geographically dispersed industrial customers and must rely on wholesalers. Furthermore, in some locations acceptable middlemen may not be available in every line of trade. Firms in this situation include some manufacturers of electrical supplies and small hand tools.

Even the manufacturer of a full line of products who has geographically concentrated customers may find direct channels less profitable than indirect channels for some of the products and customers. For example, many pharmaceutical companies have increased their use of wholesalers, even in concentrated market areas, because of the high customer service levels required.

8.2. The Wholesaler's Perspective

Wholesalers make it possible to provide possession, time, and place utility efficiently. Wholesalers are economically justified because they improve distribution efficiency by breaking bulk, building assortments of goods, and providing financing for retailers or industrial customers.

Wholesalers' market power is greatest when retailers order a small amount of each manufacturer's products or when the manufacturers involved have limited financial resources. For some products, such as Whirlpool appliances and some lines of jewelry and fashion apparel, per-unit prices and margins may be large enough to enable the manufacturer to sell directly to retailers, even when the number of items sold to each retailer is small. But manufacturers of low-value or low-margin items such as cigarettes and some food items may find it profitable to sell only through wholesalers, even though each retailer may order in relatively large quantities.

Wholesalers' and distributors' financial strength determines the number of services they can perform. Each service represents a profit opportunity as well as an associated risk and cost. The presence or absence of other firms offering comparable services influences the market power of individual wholesalers. Traditionally, wholesalers have been regional in scope. In some industries, such as pharmaceutical, wholesaler mergers have occurred. Cardinal Heath, McKesson, and Bergen Brunswig are large pharmaceutical wholesalers that have become national in scope. Together they control over one-half of the drug store wholesale business in the United States.

8.3. The Retailer's Perspective

Retailers exist when they provide convenient product assortment, availability, price, and image within the geographic market served. The degree of customer preference (loyalty due to customer service and price/value performance) that a retailer enjoys in a specific area directly affects its ability to negotiate supply chain relationships. The retailer's financial capability and size also determine its degree of influence over other supply chain members.

9. SUPPLY CHAIN DESIGN CONSIDERATIONS

Among the factors management must consider when establishing a supply chain are market coverage objectives, product characteristics, customer service objectives, and profitability (Bowersox et al. 1980).

9.1. Market Coverage Objectives

In order to establish market coverage objectives, management must consider consumer/customer buying behavior, the type of distribution required, supply chain structure, and the degree of control necessary for success.

9.1.1. Customer Buying Behavior

The buying motives of potential customer segments must be determined in order to design a supply chain that can perform most efficiently and effectively. This analysis enables the designer to determine the retail segment or segments most capable of reaching the target market or markets. Industrial marketers also must identify potential users and determine how these customers will make the purchase decision. The industrial purchaser's decision-making process depends on whether the firm is a user, an OEM, or a distributor.

9.1.2. Type of Distribution

There are basically three types of distribution that can be used to make product available to consumers: intensive distribution, selective distribution, and exclusive distribution. In intensive distribution, the product is sold to as many appropriate retailers or wholesalers as possible. Intensive distribution is appropriate for products such as chewing gum, candy bars, soft drinks, bread, film, and cigarettes, where the primary factor influencing the purchase decision is convenience. Industrial products that may require intensive distribution include pencils, paper clips, transparent tape, file folders, typing paper, transparency masters, and screws and nails.

In selective distribution, the number of outlets that may carry a product is limited, but not to the extent of exclusive dealing. By carefully selecting wholesalers and/or retailers, the manufacturer can concentrate on potentially profitable accounts and develop solid working relationships to ensure that the product is properly merchandised. The producer may also restrict the number of retail outlets if the product requires specialized servicing or sales support. Selective distribution may be used for product categories such as clothing, appliances, televisions, stereo equipment, home furnishings, and sports equipment.

When a single outlet is given an exclusive franchise to sell the product in a geographic area, the arrangement is referred to as exclusive distribution. Products such as specialty automobiles, some major appliances, some brands of furniture, and certain lines of clothing that enjoy a high degree of brand loyalty are likely to be distributed on an exclusive basis. This is particularly true if the consumer is willing to overcome the inconvenience of traveling some distance to obtain the product. Usually, exclusive distribution is undertaken when the manufacturer desires more aggressive selling on the part of the wholesaler or retailer or when channel control is important. Exclusive distribution may enhance the product's image and enable the firm to charge higher retail prices.

Sometimes manufacturers use multiple brands in order to offer exclusive distribution to more than one retailer or distributor. Exclusive distribution occurs more frequently at the wholesale level than at the retail level. Anheuser-Busch, for example, offers exclusive rights to distributors, who in turn use intensive distribution at the retail level (in states such as Florida where this is allowed). In general, exclusive distribution lends itself to direct channels (manufacturer to retailer). Intensive distribution is more likely to involve indirect channels with two or more intermediaries.

9.1.3. Channel Structure

With customer requirements and the type of distribution determined, management must select supply chain institutions for both inbound and outbound portions of the supply chain. Factors to consider when selecting supply chain members include financial strength, capabilities; ability to link up processes, ability to grow with the business, and competing supply chains.

9.1.4. Control

In many cases a firm may have to exercise some control over other members of the supply chain to ensure product quality and/or post-purchase services. The need for control stems from management's desire to protect the firm's long-term profitability.

9.2. Product Characteristics

Product characteristics are a major consideration in supply chain design. Nine product characteristics should be analyzed by the designer: the product's value, the technicality of the product, the degree of market acceptance, the degree of substitutability, the product's bulk, the product's perishability, the degree of market concentration, seasonality, and the width and depth of the product line.

9.2.1. Value

Products with a high per-unit cost require a large inventory investment. Consequently, high-value products typically will require shorter supply chains (fewer members) in order to minimize total inventory investment. But supply chains tend to be longer when the unit value is low, unless sales volume is high. In general, intensive distribution is used for low-value products.

The product's value also influences its inventory carrying cost and the desirability of premium transportation. Low-value, low-margin grocery products may be shipped by rail car and stored in field warehouses. High-value component parts and products such as high-fashion merchandise may be shipped by air freight to minimize in-transit inventories and reduce inventory carrying costs and markdowns.

9.2.2. Technicality

Highly technical products usually require demonstration by a salesperson as well as prepurchase and postpurchase service that often requires that repair parts be stocked. Technical products include such items as computers, high-priced stereo components, expensive cameras and video equipment, imported sports cars, and a multitude of industrial products. Generally, direct channels and selective or exclusive distribution policies are used for these kinds of products.

9.2.3. *Market Acceptance*

The degree of market acceptance determines the amount of selling effort required. If a leading manufacturer offers a new product and plans significant introductory advertising, customer acceptance will be high and intermediaries will want to carry the product. But new products with little market acceptance and low brand identification require aggressive selling.

9.2.4. *Substitutability*

Product substitutability is closely related to brand loyalty. When brand loyalty is low, product substitution is likely and intensive distribution is required. Firms place a premium on point-of-purchase displays in high-traffic areas. To gain support from wholesalers and/or retailers, the producer may offer higher-than-normal margins. Selective or exclusive distribution makes product support easier.

9.2.5. *Bulk*

Generally, low-value, high-weight products are restricted to markets close to the point of production. These products often require special materials-handling skills. With low weight and small cubes, more units can be shipped in a truck, rail car, or container, thereby reducing the per-unit cost of transportation. Tank truck shipment of orange juice concentrate from Florida to northern markets for packaging is an example of moving a product closer to the point of consumption to overcome value and bulk restrictions.

9.2.6. *Perishability*

Perishability refers to physical deterioration or to product obsolescence caused by changing customer buying patterns or technological change. Perishable products are usually sold on a direct basis in order to move product through the supply chain more quickly and reduce the potential for inventory loss.

9.2.7. *Market Concentration*

When the market is concentrated in a geographic area, short supply chains may be the most effective and efficient method. When markets are widely dispersed, however, specialized intermediaries are necessary; they can capitalize on the efficiencies associated with moving larger quantities. Because of widely dispersed markets, many food-processing companies use brokers to market their products. This factor also explains the existence of pooling agencies, such as freight forwarders and local cartage firms, that aggregate small shipments into truckload or carload units for movement to distant points.

9.2.8. *Seasonality*

Seasonality must be considered when applicable. For some products, sales volumes peak at certain times of the year (such as toy sales at Christmas); in other cases, raw materials, such as fresh fruits and vegetables, may only be available at specific times. Both cases require out of-season storage. Manufacturers must invest in warehouses, use third parties, or provide incentives to intermediaries so they perform the storage function. For example, manufacturers might offer a seasonal discount or consignment inventories to wholesalers or retailers who agree to take early delivery.

9.2.9. *Width and Depth*

The width and depth of a supplier's product line influence supply chain design. A manufacturer of products with low per-unit values may use intensive distribution with direct sales if the product line is broad enough to result in a relatively large average sales volume. Grocery manufacturers such as Kellogg's and General Foods are examples. Usually, a manufacturer of a limited line of products will use wholesalers to achieve adequate market coverage at a reasonable cost.

9.3. Customer Service Objectives

Customer service represents the place component of the marketing mix. Customer service can be used to differentiate the product or influence the market price—if customers are willing to pay more for better service. In addition, the supply chain structure will determine the costs of providing a specified level of customer service.

Customer service is a complex subject. However, it is usually measured in terms of the level of product availability, speed and consistency of the customer's order cycle, and communication that takes place between seller and customer. Management should establish customer service levels only after carefully studying customer needs.

9.3.1. Availability

The most important measure of customer service is inventory availability within a specified order cycle time. A common measure of availability is the number of orders shipped complete within a specified time period as a percentage of total orders received. The measure(s) selected should reflect the customer's view of customer service. The best measure of customer service reflects the product's importance to the customer and the customer's importance to the company.

9.3.2. Order Cycle

The order cycle is the time that elapses between the customer's order placement and the time the product is received. The ability to achieve the targeted order cycle time consistently influences the amount of inventory held throughout the supply chain. Consequently, the speed and consistency of the order cycle are prime factors in supply chain design. Most customers prefer consistent service to fast service because the former allows them to plan inventory levels to a greater extent than is possible with a fast but highly variable order cycle.

9.3.3. Communication

Communication refers to the firm's ability to supply timely information to the customer regarding such factors as order status, order tracking, back-order status, order confirmation, product substitution, product shortages, and product information requests. The use of automated information systems usually results in fewer errors in shipping, picking, packing, labeling, and documentation. The ability of supply chain members to provide good communications systems is a major factor in supply chain design.

10. SUPPLY CHAIN PERFORMANCE MEASUREMENT

The literature rarely focuses on measuring supply chain performance, for a number of reasons:

1. Measuring supply chain performance is difficult.
2. Some aspects of supply chain performance are difficult to quantify, making it difficult to establish a common performance standard.
3. Differences in supply chains make it difficult to establish standards for comparison.

One measure of supply chain performance is the extent to which the company's target market(s) are being satisfied, given the firm's goals and objectives. This would include measures of product availability, adequacy of customer service, and strength of brand image.

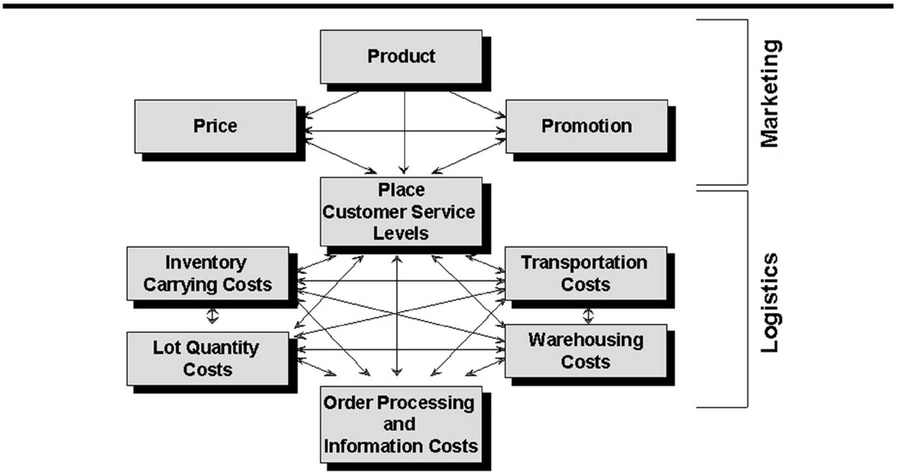
Next, management must analyze supply chain structure to determine whether the corporate strategy has been successfully implemented. Measures of structure efficiency include member turnover, competitive strength, and related issues. When management evaluates supply chain structure, it must compare the firm's ability to perform the functions/activities internally with another member's ability to perform these functions/activities.

Some potential quantitative measures of supply chain performance include logistics cost per unit, errors in order filling, and percent of damaged merchandise. Qualitative measures that managers may use when reevaluating the supply chain and specific members include degree of coordination, degree of conflict, and availability of information as needed. Management should set objectives for the supply chain and individual members and measure actual performance against planned performance. Also, evaluation measures should be developed over time and used to isolate potential problem areas. Perhaps the best measures of performance are the value created for customers and the profitability of the supply chain and its members.

For the individual firm, the goal is to find the most efficient way to offer the desired level of customer service (see Figure 8). For the supply chain, the goal is to improve overall efficiency by reallocating functions, and therefore costs, among its members. The level of customer service offered by the individual member firms, for example, will have a significant impact on other members and total supply chain performance.

For example, a manufacturer whose product availability is poor and order cycle times inconsistent may force wholesalers to carry more inventory as safety stock in order to offer an acceptable level of service to the retailers. In this case, lower logistics costs for the manufacturer were achieved at the expense of other members of the supply chain, and the entire supply chain may be less efficient.

However, if management concentrates on systems changes that improve logistics efficiency or effectiveness, it may be possible to satisfy all of the firm's objectives. For example, by linking members of the supply chain, using advanced information technology and sharing key data, a firm may be able to achieve some or all of the following: increased customer service levels, lower inventories, speedier collections, decreased transportation costs, lower warehousing costs, improvement in



Marketing Objective: Allocate resources to the marketing mix in such a manner as to maximize the long-term profitability of the firm.

Logistics Objective: Minimize Total Costs given the customer service objective where total costs = Transportation Costs + Warehousing Costs + Order Processing and Information Costs + Lot Quantity Costs + Inventory Carrying Costs

Figure 8 Cost Trade-offs Required in Marketing and Logistics. (From D. M. Lambert, *The Development of an Inventorry Costing Methodology: A Study of the Cost Associated with Holding Inventory,* National Council of Physical Distribution Management, Chicago, 1976. Reprinted with permission)

cash flow, and high return on assets. Thus, all supply chain decisions are best viewed from a systems perspective, as an integrated whole.

A manufacturer has minimal additional cash invested in inventory held by the customer rather than in the manufacturer's warehouse. Furthermore, the non-cost-of money components of inventory carrying cost are shifted to the next level of the supply chain. However, this may not be most efficient for the supply chain as a whole, as the value of inventory increases as it gets closer to the consumer because of mark-ups by each subsequent member and/or value added at various stages in the supply chain. The supply chain would be better off as a whole to have inventory held in the least valuable forms. In addition, the less differentiated the inventory has become, the more likely, in general, that it can be used in a different application.

In addition to rethinking traditional strategies for improving supply chain cash flow and return on assets, supply chain leaders may wish to consider automating and integrating the information systems within the supply chain. This can reduce lead-time variability and create time for planning. If communications flows throughout the supply chain are improved, all members will be able to reduce inventories while improving customer service.

In addition, the extra planning time that results due to increased communication speed will allow freight consolidations, warehousing cost savings, and lower lot quantity costs. Customer service levels can be improved and total operating costs reduced—truly a unique opportunity.

11. REENGINEERING IMPROVEMENT INTO THE SUPPLY CHAIN

A critical part of streamlining supply chains involves reengineering the key processes to meet customer needs. Reengineering is a process aimed at producing dramatic changes quickly. Hammer and Champy (1993) define it as the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical contemporary measures of performance such as cost, quality service, and speed. Improvement through reengineering cannot be accomplished in a haphazard manner. These changes must be supported at the top and driven through an overall management plan.

A typical reengineering process proceeds through three stages: fact finding, identifying areas for improvement to business process redesign, and creative improvements. The fact-finding stage is a very detailed examination of the current systems, procedures, and work flows. Key focus is placed on separating facts from opinions.

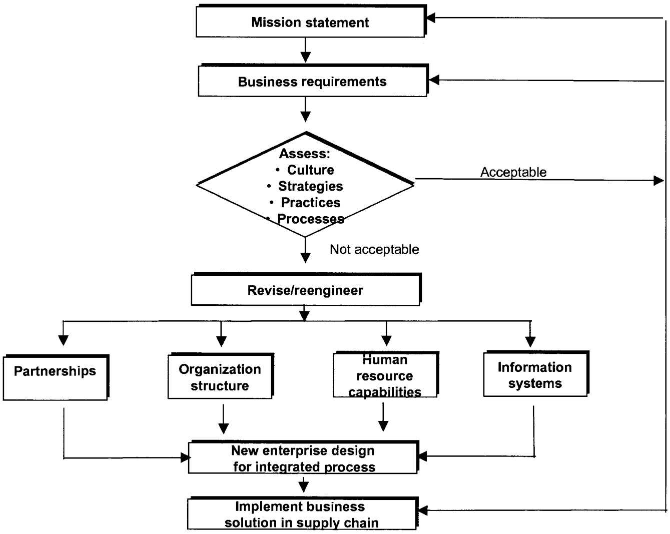


Figure 9 Reengineering SCM Process Flow Chart. (From D. M. Lambert, L. C. Guinipero, and G. J. Ridenhower, "Supply Chain Management: A Key to Achieving Business Excellence in the 21st Century," manuscript. All rights reserved)

Armed with the facts collected in the first stage, reengineering teams identify areas for improvement. They analyze where value was added for the final customer, with particular emphasis on customer contact points and product information transfers that are currently ineffective or inefficient. After identification of improvement points the creative phase of redesigning business process and information flow begins. The outcomes of the creative phase will fundamentally change both the nature of the work and how it is performed.

Figure 9 illustrates a general plan when undertaking a process reengineering approach. Organizational energy needs to focus on the firm's mission statement. The mission statement drives the business requirements in the organization. A complete assessment is made of the firm's culture, strategies, business practices, and processes.

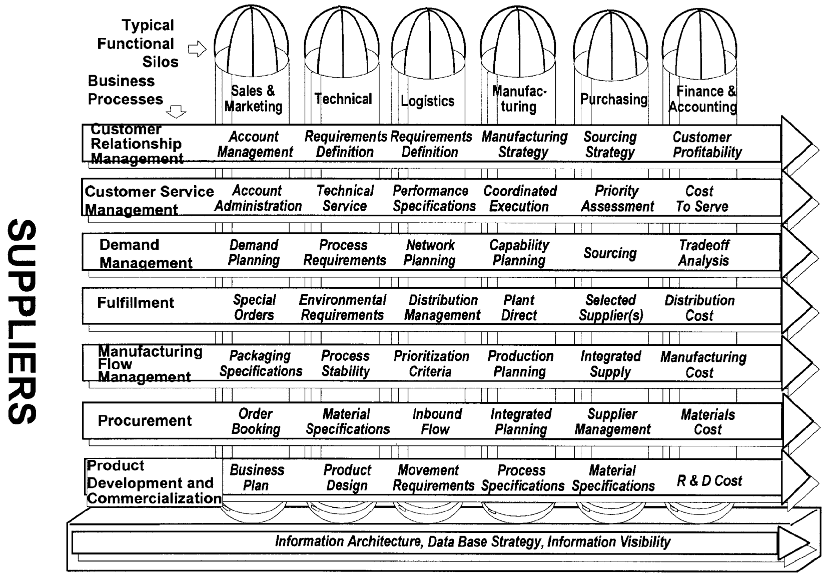
If this analysis proves acceptable, management implements its business solution across the supply chain. Typically, improvements are required in one of the areas to enhance supply chain performance. An example of this reengineering is the new Mercedes-Benz microcar, which is based on the principle of systems supply (Coleman et al. 1995). This reengineering of the process results in delegating more design activities to suppliers, reducing the amount of engineering and labor at the primary manufacturer. The result is passing the savings of these efficiencies along to the customer in the form of increased value.

12. IMPLEMENTING INTEGRATED SUPPLY CHAIN MANAGEMENT

Implementing SCM requires making the transition from a functional organization to a focus on process. Figure 10 illustrates how each function in the organization maps with the seven key processes.

In the customer relationship management process, sales and marketing provides the account management expertise, engineering provides the specifications that define the requirements, logistics provides knowledge of customer service requirements, manufacturing provides the manufacturing strategy, purchasing provides the sourcing strategy, and finance and accounting provides customer profitability reports. The customer service requirements must be used as input to manufacturing, sourcing, and logistics strategies.

If the proper coordination mechanisms are not in place across the various functions, the process will be neither effective or efficient. Taking a process focus means that all functions that touch the product or provide information must work together. For example, purchasing depends on sales/marketing data fed through a production schedule that are used to assess specific order levels and timing of requirements. These orders drive production requirements, which in turn are transmitted upstream to suppliers.


Typical Functional Silos
Business Processes
Sales & Marketing
Technical
Logistics
Manufacturing
Purchasing
Finance & Accounting
SUPPLIERS
CUSTOMERS
Customer Relationship Management
Account Management
Requirements Definition
Requirements Definition
Manufacturing Strategy
Sourcing Strategy
Customer Profitability
Customer Service Management
Account Administration
Technical Service
Performance Specifications
Coordinated Execution
Priority Assessment
Cost To Serve
Demand Management
Demand Planning
Process Requirements
Network Planning
Capability Planning
Sourcing
Tradeoff Analysis
Fulfillment
Special Orders
Environmental Requirements
Distribution Management
Plant Direct
Selected Supplier(s)
Distribution Cost
Manufacturing Flow Management
Packaging Specifications
Process Stability
Prioritization Criteria
Production Planning
Integrated Supply
Manufacturing Cost
Procurement
Order Booking
Material Specifications
Inbound Flow
Integrated Planning
Supplier Management
Materials Cost
Product Development and Commercialization
Business Plan
Product Design
Movement Requirements
Process Specifications
Material Specifications
R & D Cost
Information Architecture, Data Base Strategy, Information Visibility

Note: Process sponsorship and ownership must be established to drive the attainment of the supply chain vision and eliminate the functional barriers that artificially separate the process flows.

Figure 10 Implementation of Supply Chain Management. (From D. M. Lambert, L. C. Guinipero, and G. J. Ridenhower, "Supply Chain Management: A Key to Achieving Business Excellence in the 21st Century," manuscript. All rights reserved)

The increasing use of outsourcing has accelerated the need to coordinate supply chain processes because the organization becomes more dependent on outside contractors' suppliers. Consequently, coordination mechanisms must be in place within the organization. Where to place these coordination mechanisms and which team and functions are responsible become critical decisions.

Several process redesign and reengineering techniques can be applied to the seven key processes. Chrysler Corporation's development of Neon was accomplished through the efforts of 150 internal employees. This core group leveraged its efforts to 600 engineers, 289 suppliers, and line employees. Concurrent engineering techniques required the involvement of personnel from all key functional areas working with suppliers to develop the vehicle in 42 months. The use of concurrent engineering resulted in the avoidance of later disagreements, misunderstandings, and delays.

All firms within the supply chain will have their own functional silos that must be overcome and a process approach that must be accepted in order to successfully implement SCM. The requirements for successful implementation of SCM include:

- Executive support, leadership, and commitment to change
- An understanding of the degree of change that is necessary
- Agreement on the SCM vision and the key processes
- The necessary commitment of resources and empowerment to achieve the stated goals

13. MANAGING SUPPLIER RELATIONSHIPS

Supplier partnerships have become one of the hottest topics in interfirm relationships. Business pressures such as shortened product life cycles and global competition are making business too complex and expensive for one firm to go it alone. Despite all the interest in partnerships, a great deal of confusion still exists about what constitutes a partnership and when it makes the most sense to have one. This section will present a model (Lambert et al. 1996a, b) that can be used to identify when a partnership is appropriate as well as the type of partnership that should be implemented.

While there are countless definitions for partnerships in use today, we prefer the following definition: "A partnership is a tailored business relationship based on mutual trust, openness, shared risk and shared rewards that yields a competitive advantage, resulting in business performance greater than would be achieved by the firms individually" (Lambert et al. 1996a, b).

13.1. Types of Partnerships

Relationships between organizations can range from arm's-length relationships (consisting of either one-time exchanges or multiple transactions) to vertical integration of the two organizations, as shown in Figure 11. Most relationships between organizations have been at arm's length where the two organizations conduct business with each other, often over a long period of time and involving multiple exchanges. However, there is no sense of joint commitment or joint operations between the two companies. In arm's-length relationships, a seller typically offers standard products/services to a wide range of customers, who receive standard terms and conditions. When the exchanges end, the relationship ends. While arm's length represents an appropriate option in many situations, there are times when a closer, more integrated relationship, called a partnership, would provide significant benefits to both firms.

A partnership is not the same as a joint venture, which normally entails some degree of shared ownership across the two parties. Nor is it the same as vertical integration. Yet a well-managed partnership can provide benefits similar to those found in joint ventures or vertical integration. For instance, Pepsi, by acquiring restaurants such as Taco Bell, Pizza Hut, and KFC, ensured that Coca-Cola would never be served in these outlets. Coca-Cola has achieved a similar result without the cost of vertical integration through its partnership with McDonald's.

While most partnerships share some common elements and characteristics, there is no one ideal or benchmark relationship that is appropriate in all situations. Because each relationship has its own set of motivating factors as well as its own unique operating environment, the duration, breadth, strength, and closeness of the partnership will vary from case to case and over time. Research (Lambert et al. 1996) has indicated that three types of partnerships exist.

- *Type 1:* The organizations involved recognize each other as partners and, on a limited basis, coordinate activities and planning. The partnership usually has a short-term focus and involves only one division or functional area within each organization.
- *Type II:* The organizations involved progress beyond coordination of activities to integration of activities. Although not expected to last forever, the partnership has a long-term horizon. Multiple divisions and functions within the firm are involved in the partnership.
- *Type III:* The organizations share a significant level of integration. Each party views the other as an extension of their own firm. Typically no end date for the partnership exists.

Normally, a firm will have a wide range of relationships spanning the entire spectrum, the majority of which will not be partnerships but arm's-length associations. Of the relationships that are partnerships, the largest percentage will be type I, and only a limited number will be type III partnerships. Type III partnerships should be reserved for those suppliers or customers who are critical to an organization's long-term success. The previously described relationship between Coke and McDonald's has been evaluated as a type III partnership.

13.1.1. The Partnership Model

The partnership model shown in Figure 12 has three major elements that lead to outcomes: drivers, facilitators, and components. Drivers are compelling reasons to partner. Facilitators are supportive corporate environmental factors that enhance partnership growth and development. Components are joint activities and processes used to build and sustain the partnership. Outcomes reflect the performance of the partnership.

13.1.1.1. Drivers Both parties must believe that they will receive significant benefits in one or more areas and that these benefits would not be possible without a partnership. The primary potential benefits that drive the desire to partner include asset/cost efficiencies, customer service improvements, marketing advantage, and profit stability/growth (see Table 1 for examples). While the presence of strong drivers is necessary for successful partnerships, the drivers by themselves do not ensure success. The benefits derived from the drivers must be sustainable over the long term. If, for instance,

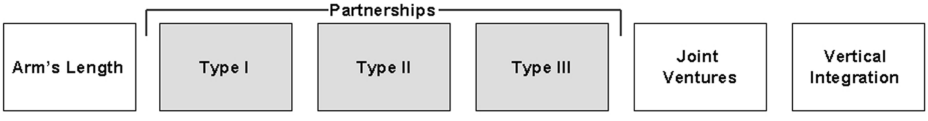


Figure 11 Types of Relationships. (From D. M. Lambert, M. A. Emmelhainz, and J. T. Gardner, "Developing and Implementing Supply Chain Partnerships," *International Journal of Logistics Management,* Vol. 7, No. 2, 1996, pp. 1–7. Reprinted by permission)

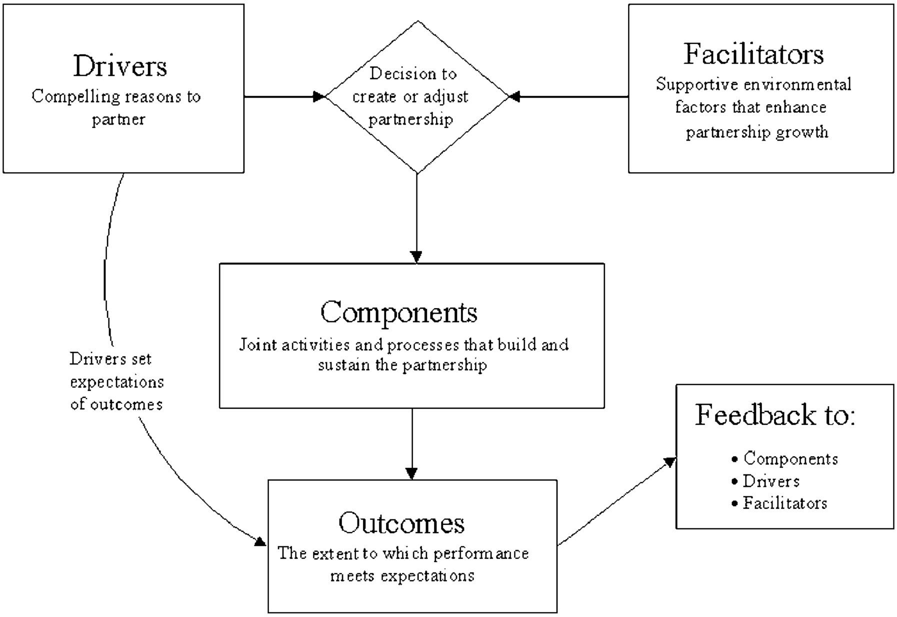

Drivers
Compelling reasons to partner
Decision to create or adjust partnership
Facilitators
Supportive environmental factors that enhance partnership growth
Components
Joint activities and processes that build and sustain the partnership
Drivers set expectations of outcomes
Feedback to:
• Components
• Drivers
• Facilitators
Outcomes
The extent to which performance meets expectations

Figure 12 The Partnering Process. (From D. M. Lambert, M. A. Emmelhainz, and J. T. Gardner, "Developing and Implementing Supply Chain Partnerships," *International Journal of Logistics Management,* Vol. 7, No. 2, 1996, pp. 1–7. Reprinted by permission)

the marketing advantage or cost efficiencies resulting from the relationship can be easily matched by a competitor, the probability of long-term partnership success is reduced.

In evaluating a relationship, how does a manager know if there are enough drivers to pursue a partnership? First, drivers must exist for each party. It is unlikely that the drivers will be the same for both parties, but they need to be strong for both. Second, the drivers must be strong enough to provide each party with a realistic expectation of significant benefits through a strengthening of the relationship. Each party should independently assess the strength of its specific drivers and present its drivers to the other party. One of the reasons that partnerships fail is that one or both of the parties have unrealistic expectations. The process of evaluation of the drivers helps the parties set realistic expectations.

13.1.1.2. Facilitators Drivers provide the motivation to partner. But even with a strong desire for building a partnership, the probability of success is reduced if the corporate environments are not supportive of a close relationship. Just as the relationship of a young couple with a strong desire to marry can be derailed by unsupportive in-laws, different communication styles, and dissimilar values, so can a corporate relationship be sidetracked by a hostile environment. On the other hand, a supportive environment that enhances integration of the two parties will improve the success of the partnership. Facilitators are elements of a corporate environment that allow a partnership to grow and strengthen. They serve as a foundation for a good relationship. In the short run, facilitators cannot be developed; they either exist or they don't. And the degree to which they exist often determines whether a partnership succeeds or fails. Facilitators include corporate compatibility, similar managerial philosophy and techniques, mutuality, and symmetry (see Table 1 for details).

Facilitators apply to the combined environment of the two potential partners. Therefore, unlike drivers, which are assessed by managers in each firm independently, facilitators should be assessed jointly. The discussion of corporate values, philosophies, and objectives often leads to an improved relationship even if no further steps toward building a partnership are taken. The more positive the facilitators, the better the chance of partnership success.

If both parties realistically expect benefits from a partnership and if the corporate environments appear supportive, then a partnership is warranted. The appropriateness of any one type of partnership is a function of the combined strength of the drivers and facilitators. A combination of strong drivers and strong facilitators would suggest a type III partnership, while low drivers and low facilitators

TABLE 1 Partnership Drivers, Facilitators, and Components

Partnership Drivers
- Asset/cost efficiency: What is the probability that this relationship will substantially reduce channel costs or improve asset utilization, for example, product costs, distribution costs savings, handling costs savings, packing costs savings, information handling costs savings, managerial efficiencies, and assets and devoted to the relationship?
- Customer service: What is the probability that this relationship will substantially improve the customer service level as measured by the customer, for example, improved on-time delivery, better taking of movement, paperless order processing, accurate order deliveries, improved cycle times, improved fill rates, customer survey results, and process improvements?
- Marketing advantage: What is the probability that Ns relationship will lead to substantial marketing advantages, for example, new market entry, promotion joint advertising, sales promotion), price (reduced price advantage), product jointly developed product innovation, branding opportunities), place (expanded geographic coverage, market saturation), access to technology, and innovation potential?
- Profit stability/growth: What is the probability that this relationship will result in profit growth or reduced variability in profit, for example, growth, cyclical leveling, seasonal leveling, market share stability, sales volume, and assurance of supply?

Partnership Facilitators
- Corporate compatibility: What is the probability that the two organizations will mesh smoothly in terms of (1) culture, for example, both firms place a value on keeping commitments, constancy of purpose, employees valued as long-term assets, and external stakeholders considered important, and (2) business, for example, strategic plans and objectives consistent, commitment to partnership ideas, and willingness to change?
- Management philosophy and techniques: What is the probability that the management philosophy and techniques of the two companies will match smoothly, for example, organizational structure, use of TQM, degree of top management support, types of motivation used, importance of teamwork, attitudes toward "personnel churning," and degree of employee empowerment?
- Mutuality: What is the probability both parties have the skills and predisposition needed for mutual relationship building? Is management skilled at two-sided thinking and action, taking the perspective of the other company, expressing goals and sharing expectations, and taking a longer-term view, for example, or is management willing to share financial information and integrate systems?
- Symmetry: What is the probability that the parties are similar on the following important factors that will affect the success of the relationship: relative size in terms of sales, relative market sham in their respective industries, financial strength, productivity, brand image/reputation, and technological sophistication?

Partnership Components
- Planning (style, level, and content)
- Joint operating controls (measurement and ability to make changes)
- Communications (non-routine and day-to-day: organization, balanced flow, and electronic)
- Risk/reward sharing (loss tolerance, gain commitment, and commitment to fairness)
- Trust and commitment to each other's success)
- Contract style (time frame and coverage)
- Scope (share of partner's business, value added, and critical activities) Investment (financial, technology, and people)

Partnership Outcomes
- Global performance outcomes (enhancement of profits, leveling of profits over time)
- Process outcomes (improved service, reduced costs)
- Competitive advantage (market positioning, market share, access to knowledge)

Source: D. M. Lambert, M. A. Emmelhainz, and T. Gardner, "Developing and Implementing Supply Chain Partnerships," *International Journal of Logistics Management,* Vol. 7, No. 2, 1996, pp. 4–13.

suggest an arm's-length relationship. While it might seem, from all of the press on the importance of integrated relationships and alliances, that managers should attempt to turn all of their corporate relationships into type III partnerships, this is not the case. In partnering, more is not always better. The objective in establishing a partnership should not be to have a type III partnership; rather, it should be to have the most appropriate type of partnership given the specific drivers and facilitators.

In fact, in situations with low drivers and/or facilitators, trying to achieve a type III partnership is likely to be counterproductive. The necessary foundation is just not there. Once it has been determined that a partnership of a specific type is warranted and should be pursued, the next step is actually to put the partnership into place. This is done through the components.

An assessment of drivers and facilitators is used to determine the potential for a partnership, but the components describe the type of relationship that has actually been implemented.

13.1.1.3. Components Components are the activities and processes that management establishes and controls throughout the life of the partnership. Components make the relationship operational and help managers create the benefits of partnering. Every partnership has the same basic components, but the way in which the components are implemented and managed varies. Components include planning, joint operating controls, communications, risk/reward sharing, trust and commitment, contract style, scope, and financial investment. Table 1 summarizes the drives, facilitators, and components of partnership.

13.1.1.4. Outcomes and Feedback Whatever type of supplier partnership is implemented, the effectiveness of the relationship must be evaluated and possibly adjusted. The key to effective measurement and feedback is how well the drivers of partnership were developed at the outset. At this beginning point, the measurement and metrics of relating to each driver should have been made explicit. These explicit measures then become the standard in evaluation of the partnership outcomes. Feedback can loop back to any step in the model. Feedback can take the form of periodic updating of the status of the drivers, facilitators, and components.

Additional information on the partnership model can be found in Gardner et al. (1999). Information on this book and other current developments in the partnership research can be found at the website of the International Journal of Logistics Management at www.logisticssupplychain.org.

14. SUMMARY

In this chapter we saw that:

1. Supply chain management is different from managing logistics in the supply chain.
2. Various supply chain structures are used.
3. Supply chain management is a process-oriented approach to manage relationships in the supply chain, and leading-edge firms such as 3M are implementing SCM.
4. Communications can improve the efficiency and effectiveness of the supply chain.
5. A number of factors influence supply chain design, evolution, and performance.
6. The implementation of integrated supply chain management requires a process management team structure.
7. Partnerships with key suppliers and customers are an important part of supply chain management.

REFERENCES

Alderson, W. (1950), “Marketing Efficiency and the Principle of Postponement,” *Cost and Profit Outlook,* Vol. 3, September.

Anderson Consulting (1999), Logistics Software, 1999 Edition, CD-ROM (available from Council of Logistics Management, Oak Brook, IL).

Andrews, D. C. and Stalick, S. K. (1994), *Business Reengineering: The Survival Guide,* NJ. Yourdon Press, Englewood Cliffs.

Arntzen, B. C., Brown, G. G., and Traffton, L. L. (1995), “Global Supply Chain Management at Digital Equipment Corporation,” *Interfaces,* Vol. 25, No. 1, pp. 69–93.

Aspinwall, L. (1958), “The Characteristics of Goods and Parallel Systems Theories,” in *Marketing Management,* E. Kelley and W. Lazer, Richard D. Irwin, Homewood, IL, pp. 434–450.

Bechtel, C., and Jayaram, J. (1997), “Supply Chain Management: A Strategic Perspective,” *International Journal of Logistics Management,* Vol. 8, No. 1, pp. 15–34.

Bowersox, D. J. (1997a), “Lessons Learned from World Class Leaders,” *Supply Chain Management Review,* Vol. 1, No. 1, pp. 61–67.

Bowersox, D. J. (1997b), “Integrated Supply Chain Management: A Strategic Perspective,” *Annual Conference Proceedings,* Council of Logistics Management, Chicago, pp. 181–189.

Bowersox, D. J., and Closs, D. J. (1996), *Logistical Management—The Integrated Supply Chain Process,* McGraw-Hill, New York.

Bucklin, L. P. (1965), "Postponement, Speculation and the Structure of Distribution Channels," *Journal of Marketing Research,* Vol. 2, No. 1, pp. 26–31.

Bucklin, L. P. (1966), *A Theory of Distribution Channel Structure,* Institute of Business and Economic Research, University of California, Berkeley.

Camp, R. C., and Colbert, D. N. (1997), "The Xerox Quest for Supply Chain Excellence," *Supply Chain Management Review,* Bol. 1, No. 1, pp. 82–91.

Christopher, M., and Peck, H. (1997), "Managing Logistics in Fashion Markets," *International Journal of Logistics Management,* Vol. 8, No. 2, pp. 63–74.

Clendenin, J. A. (1997), "Closing the Supply Chain Loop: Reengineering the Returns Channel Process," *International Journal of Logistics Management,* Vol. 8, No. 1, pp. 75–85.

Coase, R. H. (1937), "The Nature of the Firm," *Economica,* Vol. 4, pp. 386–405.

Coleman, J. L., Bhattacharya, A. K., and Brace, G. (1995), "Supply Chain Reengineering: A Supplier's Perspective," *International Journal of Logistics Management,* Vol. 6, No. 1, pp. 85–92.

Cooper, M. C., and Ellram, L. M. (1993), "Characteristics of Supply Chain Management and the Implications for Purchasing and Logistics Strategy," *International Journal of Logistics Management,* Vol 4, No. 2, pp. 13–22.

Cooper, M. C., Ellram, L. M., and Gardner, J. T. (1997), "Meshing Multiple Alliances," *Journal of Business Logistics,* Vol. 18, No. 1, pp. 67–89.

Cooper, M. C., Lambert, D. M., and Pagh, J. D. (1997), "Supply Chain Management: More Than a New Name for Logistics," *International Journal of Logistics Management,* Vol. 8, No. 1, pp. 1–13.

Copacino, W. C. (1994), "The Changing Role of the Distributor," *Traffic Management,* Vol. 33, No. 2, p. 31.

Copacino, W. C. (1997), *Supply Chain Management: The Basics and Beyond,* St. Lucie Press, Boca Raton, FL.

Cox, R., and Alderson, W. (1950), *Theory in Marketing,* Richard D. Irwin, Chicago.

Davis, T. (1993), "Effective Supply Chain Management," *Sloan Management Review,* Vol. 34, No. 4, pp. 35–46.

Ellram, L. M., and Cooper, M. C. (1990), "Supply Chain Management, Partnerships and the Shipper-Third Party Relationship," *International Journal of Logistics Management,* Vol. 1, No. 2, p. 2.

Ellram, L. M., and Cooper, M. C. (1993), "The Relationship between Supply Chain Management and Keiretsu," *International Journal of Logistics Management,* Vol. 4, No. 1, pp. 1–12.

Fisher, M. L. (1997), "What Is the Right Supply Chain for Your Product?" *Harvard Business Review,* Vol. 75, No. 2, pp. 105–116.

Gardner, J. T., Lambert, D. M., and Emmelhainz, M. A. (1999), *Partnership Facilitator's Guide: Developing and Implementing Successful Partnerships in the Supply Chain,"* Center for Competitive Excellence, Jacksonville, FL.

Giunipero, L. C., and Brand, R. R. (1996), "Purchasing's Role in Supply Chain Management," *International Journal of Logistics Management,* Vol. 7, No. 1, pp. 29–37.

Håkansson, H., and Snehota, I. (1995), *Developing Relationships in Business Networks,* Routledge, London.

Hammer, M. (1990), "Reengineering Work: Don't Automate, Obliterate," *Harvard Business Review,* Vol. 68, No. 4, pp.104–112.

Hammer, M., and Champy, J. (1993), *Reengineering the Corporation: A Manifesto for Business Revolution,* Harper Business, New York.

Handfield, R. (1991), "The Role of Materials Management in Developing Time-Based Competition," *International Journal of Purchasing and Materials Management,* Vol. 29, No. 4, pp. 2–10.

Handfield, R. B., and Nichols, E. L., Jr. (1999), *Introduction to Supply Chain Management,* Prentice Hall, Upper Saddle River, NJ.

Hewitt, F. (1994), "Supply Chain Redesign," *International Journal of Logistics Management,* Vol. 5, No. 2, pp. 1–9.

La Londe, B. J. (1998), "Supply Chain Evolution by the Numbers," *Supply Chain Management Review,* Vol. 2, No. 1, pp. 7–8.

Lambert, D. M. (1978), *The Distribution Channels Decision,* National Association of Accountants, New York, and Society of Management Accountants of Canada, Hamilton, ON, pp. 44–45.

Lambert, D. M. (1996), *The Development of an Inventory Costing Methodology: A Study of the Cost Associated with Holding Inventory,* National Council of Physical Distribution Management, Chicago.

Lambert, D. M., Emmelhainz, M. A., and Gardner, J. T. (1996a), "So You Think You Want a Partner?" *Marketing Management,* Vol. 5, No. 2, pp. 25–41.

Lambert, D. M., Emmelhainz, M. A., and Gardner, J. T. (1996b), "Developing and Implementing Supply Chain Partnerships," *International Journal of Logistics Management* Vol. 7, No. 2, pp. 1–17.

Lambert, D. M., Cooper, M. C., and Pagh, J. D. (1998), "Supply Chain Management: Implementation Issues and Research Opportunities," *International Journal of Logistics Management,* Vol. 9, No. 2, p. 1–19.

Lambert, D. M., Guinipero, L. C., and Ridenhower, G. J. (1997), "Supply Chain Management: A Key to Achieving Business Excellence in the 21st Century," manuscript.

Lee H. L., and Billington, C. (1992), "Managing Supply Chain Inventory: Pitfalls and Opportunities," *Sloan Management Review,* Vol. 33, No. 3, pp. 65–73.

Lee, H. L., and Billington, C. (1995), "The Evolution of Supply Chain Management Models and Practice at Hewlett-Packard," *Interfaces,* Vol. 25, No. 5, pp. 42–63.

Macneil, I. R. (1980), *The New Social Contract, An Inquiry into Modern Contractual Relations,* Yale University Press, New Haven, CT.

Michman, R. (1971), "Channel Development and Innovation," *Marquette Business Review,* pp. 45–49.

Oliver, R. K., and Webber, M. D. (1982), "Supply Chain Management: Logistics Catches up with Strategy," in *Logistics: The Strategic Issues,* M. G. Christopher, Ed., Chapman & Hall, London.

Olsen, R. F., and Ellram, L. M. (1997), "A Portfolio Approach to Supplier Relationships," *Industrial Marketing Management,* Vol. 26, pp. 101–113.

Piper Jaffray Equity Research (1999), *Third-Party Logistics,* January, p. 18.

Porter, M. E. (1984), *Competitive Advantage—Creating and Sustaining Superior Performance,* Free Press.

Rafuse, M. (1995), "Reducing the Need to Forecast," *International Journal of Logistics Management,* Vol. 6, No. 2, pp. 103–108.

Richardson, H. L. (1994), "How Much Should You Outsource?" *Transportation and Distribution,* Vol. 35, No. 9, p. 61.

Scharlacken, J. W. (1998), "The Seven Pillars of Global Supply Chain Planning," *Supply Chain Management Review,* Vol. 2, No. 1, pp. 32–40.

Stern L. W., El-Ansary, A., and Coughlan, A. (1996), *Marketing Channels,* 5th Ed., Prentice Hall, Englewood Cliffs, NJ.

Stevens, G. C. (1989), "Integration of the Supply Chain," *International Journal of Physical Distribution and Logistics Management,* Vol. 19, No. 8, pp. 3–8.

Stigler, G. E. (1951), "The Division of Labor Is Limited by the Extent of the Market," *Journal of Political Economy,* Vol. 59, No. 3, pp. 185–193.

Towers, S. (1994), *Business Process Re-engineering: A Practical Handbook for Executives,* Stanley Thorns, Cheltenham, UK.

Towill, D. R., Naim, M. M., and Wikner, J. (1992), "Industrial Dynamics Simulation Models in the Design of Supply Chains," *International Journal of Physical Distribution and Logistics Management,* Vol. 22, No. 5, pp. 3–13.

Turnbull, P. W., (1990), "A Review of Portfolio Planning Models for Industrial Marketing and Purchasing Management," *European Journal of Marketing,* Vol. 24, No. 3, pp. 7–22.

Tyndall, G., Partsch, W., and Kamauff, J. (1998), *Supercharging Supply Chains,* John Wiley & Sons, New York.

Weigand, R. E. (1963),"The Marketing Organization, Channels and Firm Size," *Journal of Business,* Vol. 36, pp. 228–36.

Williamson, O. E. (1975), *Markets and Hierarchies: Analysis and Antitrust Implications,* Free Press, New York.

ADDITIONAL READING

Bowersox, D. J., Bixby Cooper, M., Lambert, D. M., and Taylor, D. A. (1980), *Management in Marketing Channels* McGraw-Hill, New York, pp. 201–209.

Christopher, M. G., *Logistics: The Strategic Issues,* Chapman & Hall, London, 1982.

Council of Logistics Management (1998), Annual Meeting (Anaheim, CA, October), Council of Logistics Management, Oak Brook, IL.

Jaffe, D. T., and Scott, C. D., "Reengineering in Practice: Where Are the People? Where Is the Learning," *Journal of Applied Behavioral Science,* Vol. 34, No. 3, 1998, pp. 250–267.

SECTION V

METHODS FOR DECISION MAKING

A. Probabilistic Models and Statistics
B. Economic Evaluation
C. Computer Simulation
D. Optimization

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

V.A

Probabilistic Models and Statistics

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 83

Stochastic Models

Colm A. O'Cinneide
Purdue University

1. INTRODUCTION **2146**
- 1.1. Overview 2146
- 1.2. Simulation vs. Mathematical Analysis 2146
- 1.3. Preliminaries on Probability 2146

2. POINT PROCESSES **2149**
- 2.1. The Poisson Process 2149
- 2.2. Renewal Processes 2150

3. MARKOV CHAINS **2150**
- 3.1. The Markov Property 2150
- 3.2. Transition Matrices 2151
- 3.3. Irreducibility and Steady State 2152
- 3.4. Analysis of a Simple Discrete-Time Queue 2153
- 3.5. Markov Chains in Continuous Time 2154
- 3.6. Reversible Markov Chains and Birth-and-Death Models 2156

4. SIMPLE QUEUEING MODELS **2157**
- 4.1. Notation 2157
- 4.2. Simple Markovian Queueing Models 2158
  - 4.2.1. $M/M/1$ Queue 2158
  - 4.2.2. $M/M/2$ Queue 2158
  - 4.2.3. $M/M/\infty$ Queue 2158
  - 4.2.4. Erlang Loss System $M/M/s/s$ 2158
- 4.3. A Comparison of Systems 2159
- 4.4. Models Based on the $M/G/1$ Queue 2159
  - 4.4.1. $M/G/1$ Waiting Time Formula 2159
  - 4.4.2. $M/G/1$ under a Priority Discipline 2159
  - 4.4.3. The $M/G/1$ Queue with Batch Arrivals 2160
- 4.5. Examining the Assumptions of a Simple Queueing Model 2160

5. SOME GENERAL PRINCIPLES **2161**
- 5.1. Long-Run Behavior 2161
  - 5.1.1. Steady-State vs. Long-Run Averages 2161
  - 5.1.2. What Goes in Must Come Out 2161
  - 5.1.3. Little's Law and Other Conservation Laws 2162
- 5.2. Behavior of a Bottleneck Queue 2162
- 5.3. Deleterious Effects of Variability 2162
- 5.4. Rate of Convergence to Steady State 2162
- 5.5. ASTA and PASTA 2163

6. QUEUEING NETWORKS **2163**
- 6.1. Jackson Networks 2164
- 6.2. General Product-Form Networks 2165
- 6.3. General Networks: Stability and Instability 2166

7. TWO-MOMENT APPROXIMATIONS AND DECOMPOSITION METHODS **2167**
- 7.1. Introduction to Decomposition 2168
- 7.2. A Simple Decomposition Method 2168
  - 7.2.1. Approximation for Splitting 2169
  - 7.2.2. Approximation for Superposition 2169
  - 7.2.3. Approximation for Queueing 2169
- 7.3. Insights from the Decomposition Approach 2170

REFERENCES **2171**

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

1.1. Overview

Many systems of interest to industrial engineers involve randomness or unpredictability; for example, in the arrival of jobs requiring processing, or in machine breakdowns. In attempting to understand these systems for the purpose of design or control, a mathematical model is needed. Often, randomness and uncertainty must be captured explicitly in the model in order to represent the system reasonably faithfully. Such a model is called a *stochastic model*. The value of these models is that they enable us to predict the performance of a new system or the effect of a change in an existing system.

This chapter is a tutorial covering those aspects of the theory of stochastic models that are of greatest importance for applications in the manufacturing and service industries. After some preliminaries in this section, we cover four families of stochastic models. *Point processes,* treated in Section 2, are collections of random times, representing, for example, arrival times of jobs at a service facility. *Markov chains* are treated in Section 3. These offer both mathematical tractability and the flexibility to model a broad range of systems. *Queueing systems,* treated in Sections 4 and 5, capture the phenomenon of waiting. These are the central focus of this chapter. Indeed, the primary reason for treating point processes and Markov chains is to provide a foundation from which to build and analyze useful queueing models. Queueing networks, treated in Sections 6 and 7, are models of two or more interacting queues.

This chapter deals only with the simplest and most basic stochastic models and their mathematical analysis. Among the many excellent sources for the introductory material discussed here are the books by Ross (1997) and Wolff (1989), who treat the subject fairly mathematically, and Hall (1991), who focuses more on practical considerations. Ross (1996) is more advanced. For a comprehensive treatment of stochastic models of manufacturing systems, see Buzacott and Shanthikumar (1993). Larson and Odoni (1981) and Hall (1991) cover a variety of examples of stochastic modeling for service industries. Other chapters of the Handbook with a substantial stochastic modeling component are Chapters 60, 72, and 94.

1.2. Simulation vs. Mathematical Analysis

One of the primary applications of stochastic models is in discrete-event simulation of engineering systems that are subject to randomness. The first step in this methodology is to develop a stochastic model of the system in question. Simulation is used to analyze the model because models of real systems are usually too complex for direct mathematical analysis. Simulation is treated thoroughly in Chapters 93–96 of the Handbook.

Only relatively simple stochastic models may be examined in any detail using mathematical analysis, whereas almost any stochastic model may be analyzed using simulation. Because of this, it is necessary to state clearly what benefits are derived from mathematical analysis of stochastic models. These are as follows.

1. The mathematics yields insights and general principles, such as Little's law (see Section 5.1), which cannot be readily discovered from simulations.
2. The mathematics can be used to validate simulation results. This may be especially important in designing a new system, when one cannot validate the simulation by comparison with observations on an existing system. The goal of validating simulations motivates much of the interest in stochastic models; see Subsection 5.4 and Section 7 below.
3. The mathematics can sometimes show explicitly how performance is affected by system parameters and configuration choices. This is true for certain special families of models, including many of those treated in Sections 4 to 6 below. To get similar information through simulation may require a large amount of computation.
4. The mathematics may help us to understand a system that is fundamentally hard to simulate. Examples of this include models involving queues in heavy traffic (Subsection 5.4) or problems in which a very small probability must be estimated. Long simulation runs may be needed to get any accuracy in such situations, whereas various limit theorems in probability theory often give very accurate results.

1.3. Preliminaries on Probability

Here we briefly review some of the fundamental concepts in probability used in later sections. To define probabilities, we begin with an *experiment,* which may be thought of as some process leading to one of a collection of *outcomes*. This collection of outcomes is called the *sample space*. An *event* is a subcollection of outcomes. We say that an event *occurs* if the outcome that results from the experiment is in the event. Events are precisely the objects to which probabilities are assigned. The probability of an event E is denoted by $P(E)$.

The *conditional probability* of an event A given an event B is the probability that they both occur divided by the probability that B occurs:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{1}$$

This is also called simply the probability of A given B. One of the most important concepts in probability is the formal notion of independence or ‘‘unrelatedness’’—that one event has no bearing on another. Formally, two events A and B are said to be *independent* if either of the following holds.

$$P(A|B) = P(A) \text{ or } P(A \text{ and } B) = P(A)P(B) \tag{2}$$

The first of these may be expressed by saying that the knowledge that B will occur has no effect on the chance that A will occur.

A *random variable* X is a quantity whose value depends on the outcome of an experiment. The *distribution* of a random variable is any specification of all probabilities associated with that random variable. One such specification is the *distribution function* F of X, which is defined by

$$F(x) = P(X \leq x), \; -\infty < x < \infty \tag{3}$$

If X is *discrete*, which is to say its possible values may be written as a list $x_1, x_2, x_3, \ldots$, we may specify its distribution through its *probability mass function,* f, defined by

$$f(x) = P(X = x), \; -\infty < x < \infty \tag{4}$$

Note that $f(x) = 0$ unless x is one of the possible values x_i of X. To say that X is a *continuous random variable* is to say that there is a function f for which

$$P(X \leq x) = \int_{-\infty}^{x} f(u)\, du, \; -\infty < x < \infty \tag{5}$$

In this situation, f is called the *probability density function* of X, and f is the derivative F' of the distribution function F of X.

The *expected value* $E(X)$ of a random variable X is its weighted average value, in which each possible value of X is weighted by its probability. This has the interpretation of being a central value, sometimes also a typical value, of X. It may be written in the following ways, depending on whether the random variable is discrete or continuous.

$$\begin{aligned} E(X) &= \sum_x x f(x) \text{ in the discrete case, and} \\ E(X) &= \int_{-\infty}^{\infty} x f(x)\, dx \text{ in the continuous case} \end{aligned} \tag{6}$$

Sometimes it is convenient to use the Steiltjes integral notation to combine these into the single formula

$$E(X) = \int_{-\infty}^{\infty} x\, dF(x) \tag{7}$$

The *variance* of a random variable is its *expected squared deviation* from its mean, or

$$\text{Var}(X) = E[(X - E(X))^2] = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\, dx \tag{8}$$

the integral form being valid in the case of a continuous random variable only. The *standard deviation* SD(X) is the square root of the variance. This has the interpretation of being a ‘‘typical distance’’ between the value of the random variable and its mean. Because of this simple interpretation, the standard deviation may be considered a natural measure of variability of a random variable. In the context of stochastic models, another measure of variability is often more convenient to work with than the standard deviation. This is the *coefficient of variation,* or CV, defined as the ratio $\text{CV}(X) = \text{SD}(X)/E(X)$ of the standard deviation to the mean. The CV may be interpreted as a typical deviation of a random variable from its mean, expressed relative to that mean. The squared coefficient of variation, or SCV, is often used below also.

We briefly discuss three distributions of importance, the *Poisson distribution,* the *exponential distribution,* and the *normal distribution.* A random variable X has the Poisson distribution with mean m if

$$P(X = x) = e^{-m}\frac{m^x}{x!} \text{ for } x = 0, 1, 2, \ldots \tag{9}$$

This is a discrete distribution, over the non–negative integers. Its expected value and variance are equal:

$$E(X) = \mathrm{Var}(X) = m \tag{10}$$

See Subsection 2.1 for an explanation of how the Poisson distribution arises naturally.

A random variable Y is said to have the exponential distribution with *rate* λ if its probability density function is

$$f(y) = \lambda e^{-\lambda y} \text{ for } y \geq 0 \tag{11}$$

The mean and variance are

$$E(Y) = \frac{1}{\lambda} \text{ and } \mathrm{Var}(Y) = \frac{1}{\lambda^2} \tag{12}$$

It follows from this that the CV of the exponential distribution is 1. The exponential distribution has the remarkable *memoryless property,* which states that

$$P(Y > s + t | Y > s) = P(Y > t) \text{ for all } s, t \geq 0 \tag{13}$$

Thinking of Y as a random time, this says that the conditional distribution of the excess of Y over a fixed time s, given that Y does indeed exceed s, is the same as the unconditional distribution of Y. To express this in a more down-to-earth way, suppose that Y is the time at which a train will come and that Y has the exponential distribution. Then, no matter how long we have waited for the train, the distribution of the time we have left to wait is always the same. If this seems counterintuitive, then it may help to consider the memoryless property as it arises in a "discrete-time" setting. If you keep tossing a coin until you first get a head, then as long as you have not yet got a head, it is clear that the distribution of the number of tosses you have left to go is always the same. The number of tails before the first head here has the *geometric distribution*, which is a discrete-time analog of the exponential distribution. See Eq. (24) for its probability mass function.

A random variable Z is said to have the *standard* normal distribution if it has probability density function

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \text{ for } -\infty < z < \infty \tag{14}$$

The mean and variance are 0 and 1, respectively. A random variable X is said to have the normal distribution with mean m and variance σ^2 if $(X - m)/\sigma$ has the standard normal distribution.

To understand how the normal distribution arises, we must introduce another basic concept, that of *independent random variables*. Two random variables, X and Y, are said to be independent if

$$P(X \leq x \text{ and } Y \leq y) = P(X \leq x)P(Y \leq y), \; -\infty \leq x, y \leq \infty \tag{15}$$

This is equivalent to saying that any event defined in terms of X is independent of any event defined in terms of Y. For independent random variables X and Y, the variance of the sum is the sum of the variances:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \tag{16}$$

This simple formula often makes it easy to compute variances. So while it is the standard deviations that are most easily interpreted, typically the variances are the first quantities to be calculated. A significant analogy is that it is the *squared* length of the hypotenuse of a right-angled triangle that is got by adding the *squared* lengths of the other sides—throughout it is squared lengths, not the lengths themselves, that have this simple additive behavior. Often in stochastic models and statistics, collections X_1, X_2, X_3, . . . of independent and identically distributed random variables play a key role. See, for example, Subsection 2.2. For such random variables, with $m = E(X_1)$ and $\sigma^2 = \mathrm{Var}(X_1)$ denoting the common expected value and variance, and with $\overline{X}_n = (X_1 + X_2 + \ldots + X_n)/n$ denoting the mean of the first n of the X_i's, we have

$$E(\overline{X}_n) = m \text{ and } \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \tag{17}$$

Thus, $\mathrm{SD}(\overline{X}_n) = \sigma/\sqrt{n}$, and so, while an individual X_i is typically about σ in distance from m, the mean $\overline{X}_n$ is typically only about $\sigma/\sqrt{n}$ from m. This result quantifies the manner in which the individual deviations of the X_i's from their common expected value tend to cancel one another out as they are added together in the process of forming the "sample mean" $\overline{X}_n$. Furthermore, the *central limit theorem* states that the distribution of $\overline{X}_n$ is *asymptotically normal* with mean and variance (17). Roughly speaking, this means that $(\overline{X}_n - m)/(\sigma/\sqrt{n})$ is approximately standard normal. This is one of the most fundamental results in probability and one of the key consequences of independence.

2. POINT PROCESSES

In the applications we consider, a point process is a model for a collection of random times. Typical examples are the arrival times of parts at a machine in a manufacturing system or the times at which 911 emergency calls are made in a city. Point processes are also sometimes known by the more informal term *streams*. Mathematically, a point process may be described as a collection of increasing times $0 \leq T_1 < T_2 < T_3 < \ldots$ We often refer to these times generically as arrival times, even though in a given application they may represent departure times or times of some other type of occurrence. Another way to describe a point process is to specify the corresponding *counting process*, which is defined by

$$N(t) = \text{the number of arrivals in the interval } [0, t],\ t \geq 0$$

So N "counts" the number of arrivals up to time t. Yet another way to present a point process is through the *interarrival times* $X_1 = T_1$, $X_2 = T_2 - T_1$, $X_3 = T_3 - T_2, \ldots$ In summary, there are three natural ways to specify a point process: using the arrival times T_i themselves, using the counting process N, or using the interarrival times X_i. For any particular point process, one or another of these views may give a special insight.

2.1. The Poisson Process

The Poisson process is a process of completely random arrivals. It may be described informally as follows. Suppose we break up time into small intervals of length, say, h. Suppose that in each time interval we toss a coin with probability of heads p, again a small value. (Generally we think of coin tosses as fair in that each outcome has an equal 50% chance. Here we need an unfair coin, but we keep two other properties of coin tossing: that the outcomes of different tosses are independent and that on each toss there is the same probability p of heads.) For example, if we take h to be one second and $p = 1/3600$, then we get one head per hour on average. If a given toss results in heads, we place an arrival at the center of the corresponding interval. Otherwise, there is no arrival in the interval. That this process represents arrivals completely at random is because of two observations: the probability of having an arrival in any one of these one-second intervals is the same, and what happens in one interval is independent of what happens in all the other intervals. On average, this model produces arrivals at rate p/h per unit time, or p arrivals per second. Using the standard symbol λ for an arrival rate, we have $\lambda = p/h$. To define the true Poisson process of rate λ, we must take a limit as $h \to 0$ and $p \to 0$ in such a way that the overall arrival rate p/h is always λ.

The main properties of the Poisson process are:

1. $N(t)$, the number of arrivals by time t, has the Poisson distribution with mean λt, which is given by (9) with $m = \lambda t$.
2. The numbers of arrivals in several nonoverlapping time intervals are independent of one another.
3. The interarrival times $X_1, X_2, X_3, \ldots$ are independent with the exponential distribution, given by (11).

If, for a given point process N, the expected number of arrivals in any interval is proportional to the length of the interval, then we say that *the arrival rate of N is constant*. This is equivalent to the property that $E(N(t)) = \lambda t$ for all $t \geq 0$. It is a remarkable fact that if a point process has a constant arrival rate and satisfies property 2 above, then it is a Poisson process. The Poisson process arises naturally in many ways. For example, it holds quite generally that if a large number of independent point processes are combined, or superposed, each having a constant arrival rate and each making a small contribution to the total, then the superposition is approximately a Poisson process. The arrival times of telephone calls at a telephone exchange typically behave like a Poisson process, over any period where the overall arrival rate is fairly constant, because they are in fact the superposition of point processes generated by a great many people and computers acting somewhat independently. Similar effects are sometimes seen in manufacturing systems. For example, in a job shop with highly diverse routing (Buzacott and Shanthikumar 1993), arrival streams tend to behave like Poisson processes.

2.2. Renewal Processes

If the interarrival times $X_1, X_2, X_3, \ldots$ of a point process are independent and identically distributed, then it is called a *renewal process*. To motivate the use of renewal processes in modeling, consider the standard example of replacing light bulbs as they burn out. If the lifetimes X_i, $i = 1, 2, \ldots$ of the bulbs are independent and have the same distribution, then the times of replacement form a renewal process. By property 1 of Subsection 2.1, the Poisson process is a renewal process with exponentially distributed interarrival times.

Renewal processes, being a larger class than the Poisson processes, are more difficult to analyze. We do not have a simple description of the distribution of the number of arrivals by time t, $N(t)$, for a renewal process. However, we do have approximations for its mean and variance for large t. Let $m = E(X_1)$ and $\sigma^2 = \mathrm{Var}(X_1)$ denote the common mean and variance of the X_i's, and let c^2 denote the SCV $(\sigma/m)^2$ (see Subsection 1.3). Then the (long-run) *arrival rate* and *asymptotic variance* of the counting process N are defined in general and evaluated in the case of the renewal process in the following:

$$\lambda = \lim_{t\to\infty} \frac{E(N(t))}{t} = \frac{1}{m} \text{ and } \lim_{t\to\infty} \frac{\mathrm{Var}(N(t))}{t} = \lambda c^2 = \frac{\sigma^2}{m^3} \tag{18}$$

The first result here is known as the *elementary renewal theorem*. Simply put, it says that if the average time between arrivals is a half-hour, then on average there will be two arrivals per hour ($\lambda = 2$) in the long run. To enhance the results (2.18), we may invoke an extension of the central limit theorem of Subsection 1.3, stating that $N(t)$ is asymptotically normally distributed with mean λt and variance $\lambda c^2 t$ for large times t. This fact is useful in understanding queues in heavy traffic; see Subsection 5.2.

Example 2.1: Suppose that 911 calls arrive in a Poisson process of rate one per minute. The probability of more than 80 calls in a 60-minute period may be calculated approximately as follows. Let Z denote a standard normal random variable. By (12) and (18), the mean and variance of $N(60)$ are both approximately equal to 60, and so we have $P(N(60) > 80) \approx P(Z > (80 - 60)/\sqrt{60}) = P(Z > 2.582) = 0.005$. Of course, in this case we are dealing with a special renewal process, namely the Poisson process. Therefore, property 1 of Subsection 2.1 may be used as an alternative approach to computing this probability.

3. MARKOV CHAINS

A stochastic process is a collection of random variables X_t defined for a set of times t. The counting processes of Section 2 were our first examples of stochastic processes. The term *stochastic model* refers simply to a stochastic process used to model something. There is a large family of stochastic processes, known as the Markov chains, or simply chains, that have enough generality to be genuinely useful in applications and yet enough structure to be mathematically tractable. Markov chains may be thought of as the simplest processes that allow some level of dependence over time. We begin with a simple example to illustrate the Markov property in the context of building a stochastic model.

3.1. The Markov Property

Consider a machine that at any given time is either operational or not. We observe it in successive time periods and record a 1 if it is operational and a 0 if it is not. Over time, this produces a sequence of observations, perhaps in this example 1, 1, 1, 0, 0, 1, . . . , indicating that the machine ran for three periods, was then down for two, and so forth. If these observations cannot be predicted with certainty in advance, we may wish to model them as a sequence of random variables $X_1, X_2, \ldots, X_T$, each taking as its value either 0 or 1, with T denoting the number of observation periods. Then the collection of random variables $(X_1, X_2, \ldots, X_T)$ is an example of a stochastic model. It is convenient to denote the entire process simply by X. X_t is referred to as the state of the process X at time t. The set $S = \{0, 1\}$ of possible values of the X_t's is called the *state space* of the process. The observations 1, 1, 1, 0, 0, 1, . . . constitute a *realization* or *path* of X. If we observe the process for T time periods, then each path is a sequence of T zeroes and ones.

A specification of the probabilities associated with a stochastic model is called the *law* or *distribution* of the model. To define the law of X, we must assign a probability to each possible path of X. Doing this directly is often cumbersome, and the first step in developing stochastic models is to find natural ways of specifying the probabilities of the paths. The simplest specification is to assume that the states at different times are independent and that the probability that the machine is operational at any fixed time t is some constant p. Then $1 - p$ is the probability that the machine is inoperational at a time t. The probability of any particular path is then $p^r (1 - p)^{T-r}$, where r is the number of time periods for which the machine is operational on this path (and so $T - r$ is the number of time periods for which it is inoperational). This comes from independence of all the X_t's, using an extension of (2) to the effect that the probability of several independent events all occurring is the product of their probabilities.

In the model for machine failures just described, if we have observed the machine up to some time period t, these observations are irrelevant from the point of view of predicting the state of the machine in the next time period, period $t + 1$. This is because the model assumes that the successive states of the machine are independent. However, it is quite common in practice that the history of a process is informative regarding its future behavior. In the present example, it may be that when the machine is operational in one time period, it has a better than average chance of being operational in the next time period also. But to allow complete generality in the manner in which future states depend on present and past states would leave us with the problem of overchoice. We will never have enough data to fit such a detailed model. So we make an assumption that allows some dependence, but not too much. We suppose that, having observed the machine for periods 0, 1, 2, . . . , t, the distribution of the state of the machine in period $t + 1$ depends only on its state in period t, the immediately preceding period. In other words, if we know the condition of the machine at the "present" time t, the past conditions are uninformative as to future conditions. This is the *Markov property*. Mathematically, it is expressed by the condition

$$\begin{aligned} &P(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \ldots, X_1 = i_1) \\ &\quad = P(X_{t+1} = j | X_t = i) \\ &\quad \text{for all states } i_1, i_2, \ldots, i_{t-1}, i, j \in S \text{ and times } t \geq 0 \end{aligned} \tag{19}$$

This conditional probability is called the *transition probability* from state i to state j at time t. Once these quantities are specified, the entire law of the Markov process is specified. The Markov process is said to have *stationary transition probabilities* if $P(X_{t+1} = j | X_t = i)$ does not depend on t, in which case it is denoted simply by p_{ij}. We assume this property from now on, as this is the most important situation in applications, being the situation where the least number of parameters is needed to specify the model.

3.2. Transition Matrices

For a Markov chain $X = (X_0, X_1, X_2, \ldots)$ observed at all times $t = 0, 1, 2, \ldots$ and having state space $S = \{1, 2, \ldots, n\}$, the matrix P with entries p_{ij} is called the *transition matrix*. This is an $n \times n$ matrix whose rows add to 1. In the example of the machine above, where the state space $S = \{0, 1\}$ has only two states, the transition matrix is a 2×2 matrix

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{01} & p_{11} \end{pmatrix}$$

Here, for example, p_{01} is the conditional probability that the machine is operational in the next time period, given that it is now inoperational. The other entries have parallel interpretations.

With the introduction of P, it is natural to define the vectors

$$\pi(t) = [\pi_1(t), \pi_2(t), \ldots, \pi_n(t)] = [P(X_t = 1), P(X_t = 2), \ldots, P(X_t = n)]$$

which give the distribution of the chain at time t. These may all be computed, once the initial distribution $\pi(0)$ of the chain is known, using the following:

$$P(X_{t+1} = j) = \sum_{i=1}^{n} P(X_{t+1} = j | X_t = i) P(X_t = i) = \sum_{i=1}^{n} \pi_i(t) p_{ij} \tag{20}$$

In matrix form, this is the recurrence relation $\pi(t + 1) = \pi(t)P$, repeated application of which gives

$$\pi(t+s) = \pi(t)P^s. \tag{21}$$

It is the compactness of (21) as compared to (20) that motivates the use of vector and matrix notation in dealing with Markov chains. The matrix P^s, which is the sth power of the matrix P, is known as the *s-step transition matrix,* its entry $p_{ij}(s)$ being the probability that the chain, when initialized in state i, enters state j upon the sth transition.

3.3. Irreducibility and Steady State

A *steady-state distribution* of a Markov chain is a distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$, say, such that if X_0 has the distribution π, then X_1 also has that distribution, as do X_2, X_3, and so forth. This is also known as a *stationary distribution.* Using (21), the condition that π be a stationary distribution may be expressed in the form

$$\pi P = \pi \tag{22}$$

These are called the *steady-state equations.* A Markov chain is said to be irreducible if any state can be reached from any other through a sequence of transitions whose probabilities are positive. For irreducible chains, there is precisely one steady-state distribution π, and it is the only vector x that satisfies the steady-state equations $xP = x$ and also satisfies the condition $x_1 + x_2 + \ldots + x_n = 1$.

Example 3.1: Figure 1(a) represents a Markov chain on the state space $S = \{1, 2\}$ with transition matrix

$$P = \begin{pmatrix} 0.5 & 0.5 \\ 1 & 0 \end{pmatrix}$$

This chain is irreducible because each state may be reached from the other. Its steady-state distribution may be computed by solving (22) to give $\pi_1 = 2/3$ and $\pi_2 = 1/3$. Figure 1(b) represents a Markov chain that is not irreducible because, for example, state 1 cannot be reached from state 3.

Another important concept related to long-run behavior is that of a *limiting distribution* of a chain. This is a limit of the form

$$\pi_j(\infty) = \lim_{t\to\infty} P(X_t = j) \text{ or, using vectors and matrices,}$$

$$\pi(\infty) = \lim_{t\to\infty} \pi(t) = \lim_{t\to\infty} \pi(0)P^t \tag{23}$$

when we have convergence. We see that a steady-state distribution π is also a limiting distribution by taking $\pi(0) = \pi$ here and using (22). Conversely, any limiting distribution is also a steady-state distribution. To establish this, we have

$$\pi(\infty)P = (\lim_{t\to\infty} \pi(0)P^t)P = \lim_{t\to\infty} \pi(0)P^{t+1} = \pi(\infty)$$

showing that $\pi(\infty) = \pi(\infty)P$. Because of this, in later sections of this chapter we do not carefully distinguish between steady-state distributions and limiting distributions.

For some chains, the limit in (23) may fail to exist. Existence is guaranteed by irreducibility along with a new property, *aperiodicity,* which requires that, for sufficiently large s, every entry of P^s is positive. Under irreducibility and aperiodicity, the limits of (23) exist and are equal for any initial distribution. The common limit is the unique steady-state distribution π. In what follows, we always assume aperiodicity and irreducibility, unless otherwise stated.

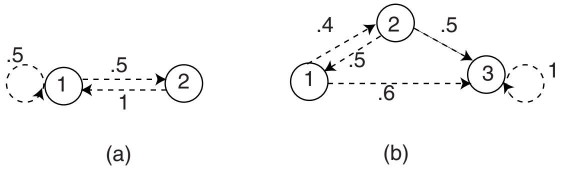


Figure 1 Two Simple Markov Chains.

Given a Markov chain, a natural question to ask is What is the long-run fraction of time that the chain spends in a given state $i \in S$? The idea of long-run behavior is quite different from that of steady-state/limiting distributions. In the latter, the motivating idea is to understand the behavior of the chain at some fixed time very far in the future. But the question of long-run behavior concerns what happens to the chain over a very long time interval. This distinction will be discussed further in the context of queues in Subsection 5. The answer to the question above is again the solution π to the steady-state equations (22), as long the chain is irreducible. To illustrate, the chain of Example 3.1 spends two-thirds of its time in state 1 and one-third in state 2, in the long run.

We summarize the discussion of steady-state behavior of Markov chains as follows. For "well-behaved" chains, such as irreducible, aperiodic, finite-state chains, there is a unique steady-state distribution that is also the limiting distribution. Furthermore, this distribution gives the long-run fraction of time spent in each state. It may be computed easily, by solving the steady-state equations (22). The steady-state behavior of the chain does not depend on the initial state.

Many interesting models require an infinite state space. The results just stated for the finite-state case extend to the infinite-state case once we adopt the condition of *positive recurrence* in addition to irreducibility and aperiodicity. A chain is positive recurrent if the expected time to return to any state is finite. We now turn to a simple infinite-state example.

3.4. Analysis of a Simple Discrete-Time Queue

In this subsection, we develop a very simple queueing model. This model is a Markov chain $L(t)$ representing the number of jobs present in a queueing system observed at regular discrete times $t = 0, 1, 2, \ldots$ The state space is $\{0, 1, 2, \ldots\}$. There are two types of transitions possible: arrivals and departures. We write p for the probability that a job arrives in the next time step. We write q for the probability that a job will complete service in the next time step, assuming that there is at least one job present ($L(t) > 0$). If we write r for $1 - p - q$, which is the probability of no state change when there is at least one job present, then the transition matrix of the chain is

$$p = \begin{pmatrix} 1-p & p & 0 & 0 & \ldots \\ q & r & p & 0 & \ldots \\ 0 & q & r & p & \ldots \\ 0 & 0 & q & r & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

If there are no jobs present, then there cannot be a departure, and this is why the first row of P does not follow the same pattern as the others. In the case of this chain, the steady-state equations (22) become

$$\pi_0(1 - p) + \pi_1 q = \pi_0 \text{ and, for } i > 1,\ \pi_{i-1} p + \pi_i(1 - p - q) + \pi_{i+1} q = \pi_i$$

These may be reorganized as

$$-p\pi_0 + q\pi_1 = 0, \text{ and, for } i > 1, -p\pi_{i-1} + q\pi_i = -p\pi_i + q\pi_{i+1}$$

Thus $p\pi_i = q\pi_{i+1}$ for all i, and, once we require that the π_i's sum to 1, we get the following unique solution:

$$\pi_i = (1 - \rho)\rho^i,\ i = 0, 1, 2, \ldots, \text{ when } \rho = \frac{p}{q} \tag{24}$$

From this, the steady-state probability that the system is empty is $\pi_0 = 1 - \rho$. This is also the long-run fraction of time that the system is empty (see Subsection 3.3). The probability that the system is not empty is $1 - \pi_0 = \rho$. Fleshing out the interpretation of the model, we imagine that the server is busy as long as there is at least one job in the system. Then ρ may be interpreted as the long-run proportion of time that the server is busy. This is known as the *server utilization*. The distribution (24) is called the *geometric distribution with parameter* ρ. Its mean is given by

$$\frac{\rho}{1 - \rho}, \tag{25}$$

and this may be described as the steady-state expected number of jobs in the system. As the utilization ρ approaches 1, the expected number of jobs in the system increases in approximate inverse proportion to $1 - \rho$. If the utilization is 99%, then the expected number of jobs waiting is 99. This quantifies the trade-off between utilization of the server, which we would like to keep high, and work-in-process, which we would like to keep low. This is a significant insight into the behavior of queues, gained from a very simple model.

3.5. Markov Chains in Continuous Time

Suppose now that we have a stochastic process $X(t)$, $t \geq 0$, observed for all times, not just integer times. We take the state space to be the finite set $\{1, 2, \ldots, n\}$ as before. The Markov property in this situation is just as described in Subsection 3.1 for the discrete-time case: knowing the state of X at a time t, its evolution after time t is independent of its history before time t. Briefly, the future is independent of the past, given the present.

Just as the rule whereby a discrete-time chain evolves is defined by a matrix P, the transition matrix, the rule whereby a continuous-time chain evolves is defined by a matrix Q, the generator matrix. The off-diagonal entry q_{ij}, $i \neq j$, of Q has the interpretation that $q_{ij}h$ is approximately the probability that the chain, starting from state i at a time t, makes a transition to state j by time $t + h$, where h is small:

$$P(X(t + h) = j \mid X(t) = i) \approx q_{ij}h \quad \text{for} \quad i \neq j \tag{26}$$

(The error in this approximation is small compared to h.) Because of this, q_{ij} is called the *transition rate* from state i to state j. It may be described more precisely as the rate of transitions into state j while the process is in state i. The diagonal entries q_{ii} of Q are determined by the condition that the rows of Q must add to 0, which is analogous to the property that the rows of a transition matrix add to 1. So q_{ii} is negative and we denote it by $-q_i$, where

$$q_i = \sum_{j \mid j \neq i} q_{ij} \tag{27}$$

the sum on the right being over all states except i. (A technical condition assumed throughout is that these quantities are finite.) The transition matrix of X over a time t is defined by

$$P(t) = (p_{ij}(t)), \quad \text{where } p_{ij}(t) = P(X(t) = j \mid X(0) = i) \tag{28}$$

(Here and below, the notation (a_{ij}) means the matrix of the a_{ij}'s.) Soon we shall develop a simple formula for $P(t)$ in terms of the basic data of the chain, its generator Q.

By the Markov property of X, the process $X(0)$, $X(h)$, $X(2h)$, . . . that results when we observe X at intervals of length h is a discrete-time Markov chain with transition matrix $P(h)$. If we think of h as a small time interval, then equation (26) shows that the transition probabilities of this discrete-time chain are given approximately by

$$\begin{gathered} p_{ij}(h) \approx q_{ij}h \quad \text{for} \quad i \neq j, \text{ and } p_{ii}(h) \approx 1 - q_i h, \text{ or,} \\ \text{in matrix form, } P(h) \approx I + hQ \end{gathered} \tag{29}$$

The rows of the matrix $I + hQ$ add to 1 because those of Q add to 0.

We have related the continuous-time chain to a discrete-time chain with a fast clock, whose time unit is the small quantity h but whose transition probabilities $p_{ij}(h)$ are proportionately small for $i \neq j$ by (29). This allows us to analyze the continuous-time chain using discrete-time results. All the basic calculations for continuous-time, finite-state Markov chains may be carried out by taking a limit as $h \to 0$ of the discrete-time approximation. For example, the transition matrix $P(t)$, defined in (28), may be derived as follows. We divide the time interval $[0, t]$ into a large number N of short intervals of length $h = t/N$, so that the transition matrix $P(t)$ is the N-step transition matrix corresponding to $P(h)$. It follows from (29) that $P(t)$ is approximately the N-step transition matrix corresponding to the transition matrix $I + hQ$. This approximation becomes exact as $h \to 0$, and we have

$$\begin{aligned} P(t) &= (p_{ij}(t)) = (P(X(t) = j \mid X(0) = i)) \\ &= \lim_{h \to 0} (I + hQ)^N = \lim_{N \to \infty} \left(I + \frac{tQ}{N}\right)^N = e^{tQ} \end{aligned} \tag{30}$$

In the final equality here we are using the well-known limit $\lim_{N \to \infty} (1 + x/N)^N = e^x$, which holds even when x is replaced by a matrix, here by tQ. The exponential of a matrix is defined by the Taylor series

$$e^A = I + A + \frac{A^2}{2} + \ldots = \sum_{n=0}^{\infty} \frac{A^n}{n!}$$

Now, some consequences of (30). If we differentiate with respect to t we get

$$\frac{d}{dt} e^{tQ} = Qe^{tQ} \quad \text{or} \quad P'(t) = QP(t), \text{ and similarly } P'(t) = P(t)Q \tag{31}$$

These are called *Kolmogorov's backward and forward equations,* respectively. The *Chapman–Kolmogorov equations* $P(s + t) = P(s)P(t)$ may be deduced from (30) as follows:

$$P(s + t) = e^{(s+t)Q} = e^{sQ+tQ} = e^{sQ}e^{tQ} = P(s)P(t) \tag{32}$$

The generator Q determines how a continuous-time Markov chain evolves via (26). There is another, more direct prescription for the evolution of the chain in terms of Q. If the chain is now in state i, then the time T until the next change of state has the exponential distribution with rate q_i:

$$P(T \leq t) = 1 - e^{-q_i t}, \; t \geq 0 \tag{33}$$

When the chain does leave state i, it chooses its next state $j \neq i$ according to the probabilities q_{ij}/q_i. Why is the time T exponentially distributed? Because the Markov property of X implies that T must have the memoryless property (13), and this in turn implies that T has the exponential distribution.

Example 3.2: Let us illustrate this using the machine-failure example introduced in Subsection 3.1, now reworked in continuous time. The state space $S = \{0, 1\}$ has two states, representing inoperational and operational, as before. The order-2 generator matrix is of the form

$$Q = \begin{pmatrix} -q_0 & q_0 \\ q_1 & -q_1 \end{pmatrix}$$

Here, q_0 may be interpreted as the rate of repair when the machine is inoperational, and q_1 as the rate of breakdown when the machine is operational. In other words, the chain evolves with alternating exponentially distributed inoperational and operational periods, the inoperational periods having rate q_0 and the operational periods having rate q_1.

The Poisson counting process of Section 2 is a continuous-time Markov chain N on the infinite state space $\{0, 1, 2, \ldots\}$, with generator

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \ldots \\ 0 & -\lambda & \lambda & 0 & \ldots \\ 0 & 0 & -\lambda & \lambda & \ldots \\ 0 & 0 & 0 & -\lambda & \ldots \\ \ldots & & \ldots & & \ldots \end{pmatrix} \tag{34}$$

Thus, the transitions are always from a state n to the state $n + 1$. The transitions are, of course, arrivals because they cause the count N to increase by 1. The probability of a transition in a short interval of time h is approximately λh for any n by (26). This observation corresponds precisely with the description of the Poisson process in terms of coin tossing in Section 2. Moreover, the fact that the time between arrivals in a Poisson process is exponential may be seen now as a consequence of the fact, expressed in (33), that the holding times in any continuous-time Markov chain are exponentially distributed.

Let us briefly describe the long-run behavior of continuous-time, finite-state Markov chains. We assume irreducibility as before, which in this case means simply that the entries of $P(t)$ are all positive for $t > 0$. Periodicity does not arise in continuous time. There is a unique distribution π satisfying

$$\pi P(t) = \pi \text{ for all } t \geq 0$$

This is the steady-state distribution of the chain. By differentiation at $t = 0$ using (31), we find that

$$\pi Q = 0, \text{ or } \pi_j q_j = \sum_{i|i \neq j} \pi_i q_{ij} \text{ for all states } j = 1, 2, \ldots, n \tag{35}$$

This relates π to the parameters of the chain—the entries of its generator Q. These are the steady-state equations in the continuous-time case. For finite-state irreducible chains, these equations have a unique solution whose components add to 1, and this solution is the steady-state distribution π. As in the discrete-time case, π is also the limiting distribution of the Markov chain and gives the long-run proportion of time spent in each state. These results extend to the infinite-state case, assuming positive recurrence, as in Subsection 3.3.

Example 3.3: The steady-state distribution of the machine-failure model of Example 3.3 is given by

$$\pi_0 = \frac{q_1}{q_0 + q_1} \text{ and } \pi_1 = \frac{q_0}{q_0 + q_1}$$

This may be seen by solving the equations (35) or by noting that the steady-state probabilities must be proportional to the mean times spent in the states, which are given by (12) as $1/q_0$ and $1/q_1$.

3.6. Reversible Markov Chains and Birth-and-Death Models

Now we consider some special continuous-time Markov chains that arise in discussing simple queueing models. We restrict attention to the continuous-time case, as this is the setting of most of the practical models treated later.

Let X be a continuous-time Markov chain with generator Q and steady-state distribution π. The quantity $\pi_i q_{ij}$, sometimes called the *flux from i to j*, is the rate at which transitions from i to j occur in steady state. Contrast this with the transition rate q_{ij} itself, which is the rate of transitions from i to j when the chain is in state i.

Suppose now that X has state space $S = \{0, 1, 2, \ldots\}$ and each transition is to a neighboring state: the chain goes either up by one or down by one on any transition. Such a process is known as a *birth-and-death process* and is characterized by the *birth rates* $\lambda_i = q_{i,i+1}$ and *death rates* $\mu_i = q_{i,i-1}$. The generator of a birth-and-death process takes the form

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \ldots \\ \mu_1 & -\mu_1 - \lambda_1 & \lambda_1 & 0 & \ldots \\ 0 & \mu_2 & -\mu_2 - \lambda_2 & \lambda_2 & \ldots \\ 0 & 0 & \mu_3 & -\mu_3 - \lambda_3 & \ldots \\ \ldots & & \ldots & & \ldots \end{pmatrix} \tag{36}$$

Now, in the long run, the rate of transitions from state i to $i + 1$ must equal the rate from $i + 1$ to i, because the number of the former type of transition must always be within 1 of the number of the latter type, and therefore

$$\pi_i q_{ij} = \pi_j q_{ji} \tag{37}$$

for $j = i + 1$. This equation states that the flux from i to $j = i + 1$ equals the flux from j to i. This is easily seen to be true not just for neighboring pairs i, j, but for all pairs of distinct states $i \neq j$. The equations (37) are called the *detailed balance equations*. It is easy to deduce the steady-state distribution from (37), the result being

$$\pi_n = C \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n}, \; n = 0, 1, 2, \ldots \tag{38}$$

where C is a constant chosen to ensure that the π_n's add to 1. (If no such constant exists, then the process is not positive recurrent.) This formula determines the steady-state distribution of a great many simple queueing models, several of which will be discussed in the next section.

A Markov chain satisfying (37) for all pairs of distinct states $i \neq j$ is said to be *reversible,* a somewhat unfortunate choice of term referring to the fact that, if viewed in reverse time while in steady state, it is probabilistically indistinguishable from its forward-time version. (Such a chain might be described more precisely by saying that it is unchanged by time reversal rather than reversible.) Many of the queueing models to be treated later are in fact Markov chains satisfying the property (37) or some related property. See Kelly (1979) for a thorough treatment of such chains.

4. SIMPLE QUEUEING MODELS

Of central interest in designing industrial and service systems is the problem of providing service to a stream of arriving customers or jobs. Because of irregularities in the arrival or service processes, there are times when too much work has arrived in too short a time and so jobs have to wait for service. Even though the system has the capacity to serve all arrivals, randomness generally results in some waiting. Waiting could be eliminated if the irregularity could be eliminated, without increasing overall service capacity or diminishing the overall flow of arriving work. Waiting is therefore a consequence of irregularity. This waiting is central: it occupies resources in the form of waiting space, contributes to work-in-process, causes due dates to be missed, and may cause machines and workers to be starved of work. Much system design focuses on minimizing waiting time. Therefore, many stochastic models also focus on waiting, and models of waiting are called queueing systems, or simply queues.

A queueing system may be divided loosely into three subsystems: the arrival process, the waiting area, and the service system. Each of these subsystems can operate in a variety of ways. The following are some of the more common possibilities. The arrival process specifies a sequence of jobs or batches of jobs and the times at which they enter the system. The arrival times may constitute a Poisson process or a renewal process, for example. The jobs may be indistinguishable from one another, or they may be of several classes, to be treated differently in the waiting or service area. The waiting area may be managed in various ways. Jobs may be ordered according to time of arrival, the most recent arrival being the first to be served (first-in-first-out, or FIFO). This is the most common discipline used in serving people because of the sense of fairness it engenders. To give some simple alternatives, jobs may be served in random order (SIRO) or last-in-first-out (LIFO). Under the processor sharing (PS) discipline, all jobs present share the server's attention equally. Jobs may also be served according to priorities determined based on job class or service needs. Preemption, in which an arriving high-priority job ejects a low-priority job from service, may or may not be allowed. The service system is where the jobs receive what they have waited for. There may be one or many servers, and the servers may be subject to breakdowns. Jobs may be served singly or in batches, and service times may depend on job class. See Hall (1991) for a discussion of modeling real systems as queues, with a careful treatment of the role of these three subsystems.

4.1. Notation

The queues of this section all belong to the class $G/G/s/K$. The four elements in this notation have the following meanings. The first element refers to the arrival process, G indicating generally distributed interarrival times. Independence of these times is assumed. Thus, the first G indicates that arrivals form a renewal process. The second element describes the service times, the G again indicating a general distribution. Independence is again assumed, and moreover, it is assumed that interarrival times are independent of service times. The third element, s, is the number of servers, and the last element, K, is the capacity of the system. $K - s$ is therefore the capacity of the waiting area. If the parameter K is not present, an unlimited capacity is assumed. If one of the G's is replaced by M, this indicates that the corresponding distribution is exponential. The M connotes the property of memorylessness (13). For example, the $M/M/1$ queue is the special case of the $G/G/1$ queue in which the arrivals are Poisson and the service distribution is exponential. D, standing for deterministic, is used to indicate constant service or interarrival times.

It is standard to denote the long-run arrival rate of a queue by λ and the long-run service rate (of one server) by μ. To avoid having to deal with unimportant special cases, we always assume that these rates are positive—that is, that they are not zero. In the case of renewal arrivals and independent, identically distributed service times, the expected interarrival time is then $1/\lambda$ and the expected service time is $1/\mu$. Throughout, ρ denotes the *traffic intensity* λ/μ of the arriving stream of jobs, which is the rate at which work arrives, work itself being measured in terms of the time it takes a server to perform it. Another quantity of importance is the server utilization, which is the proportion of time that a server is busy. For a single-server queue, the utilization and the traffic intensity are equal.

$L(t)$ will denote the load, or number of jobs in the system, waiting or being served, at time $t \geq 0$. For each model considered in this section, the distribution of $L(t)$ approaches a limit as $t \rightarrow \infty$, and the limit is referred to as the steady-state distribution of the number in the system. (As we said in Subsection 3.3, we do not distinguish carefully between steady-state and limiting distributions.) We denote by L a random variable whose distribution is this steady-state distribution. Let W_n denote the time-in-system or *flow time* of the nth arrival to the system. Then, for the models considered here, the distribution of W_n also has a limit as $n \rightarrow \infty$, which we call the *steady-state flow-time distribution*. We denote by W a random variable with this distribution. In the same way we can define a random variable W_Q whose distribution is the steady-state distribution of waiting time. Flow time and waiting time differ only in that the latter does not include service time. Most of the results given here concern the expected values of the random variables L, W, and W_Q for various queues. These expected values are called *steady-state expected values.* For some technicalities concerning steady-state and long-run averages, see Subsection 5.1. For more on these models, see the books referred to in Subsection 1.1 or any of the many introductory texts on queues.

4.2. Simple Markovian Queueing Models

In this subsection we treat several queues of the $M/M/s/K$ type. These queues have Poisson arrivals, exponential service times, s servers, and capacity K. For these queues, the number-in-system $L(t)$, $t \geq 0$, is a continuous-time Markov chain, in fact, a birth-and-death process (Subsection 3.6). The Markov property arises from the exponentiality of service and interarrival times—see the discussion following (33). The queueing discipline is taken to be FIFO in every case. The results presented here follow fairly directly from (38).

4.2.1. *M/M/1 Queue*

This is the single-server queue with Poisson arrivals and exponential service. It is the continuous-time analog of the simple queue treated in Subsection 3.4. It may be viewed as a birth-and-death process, with generator (36) where $\lambda_n = \lambda$ and $\mu_n = \mu$ for all $n \geq 0$. The fact that the birth rates are constant is what makes the arrival process Poisson. Suppose that the traffic intensity $\rho = \lambda/\mu$ is less than 1. This condition is an example of the general *capacity condition* (57) below, as it concerns the capacity of the server to handle the arriving work. It implies that the process is positive recurrent. The steady-state distribution of the number in the system L may be derived directly from (38) and is the same geometric (ρ) distribution that arose in (24) for the discrete-time queue of Subsection 3.4. The steady-state distribution of flow time W is exponential with rate $\mu(1 - \rho)$. The steady-state expected number-in-system, time-in-system, and waiting time are

$$E(L) = \frac{\rho}{1-\rho};\ E(W) = \frac{1}{\mu(1-\rho)};\ E(W_Q) = \frac{\rho}{\mu(1-\rho)} \tag{39}$$

4.2.2. *M/M/2 Queue*

Here we increase the number of servers to two. Assume now that $\rho < 2$, so that the two servers are sufficient to handle the arriving work. The $M/M/2$ queue is the birth-and-death process with $\lambda_n = \lambda$ for all $n \geq 0$, $\mu_1 = \mu$, and $\mu_n = 2\mu$ for $n \geq 2$. The new μ_n's express the idea that when there are two or more jobs in the system, both servers are busy and so the overall rate at which service is provided is doubled. The steady-state probabilities are given by

$$\pi_0 = \frac{2-\rho}{2+\rho};\ \pi_n = 2\pi_0\left(\frac{\rho}{2}\right)^n,\ n > 0 \tag{40}$$

For this queue we record the steady-state expected number-in-system and flow time:

$$E(L) = \frac{\rho}{1-\rho^2/4};\ E(W) = \frac{1}{\mu(1-\rho^2/4)} \tag{41}$$

4.2.3. *M/M/∞ Queue*

This is the birth-and-death process with $\lambda_n = \lambda$ and $\mu_n = n\mu$ for all $n \geq 0$. The service rate is proportional to the number of jobs in the system, and this captures the idea that each job is being served simultaneously at the same rate, μ. The steady-state distribution of the number in the system is Poisson (ρ), again by (38), where $\rho = \lambda/\mu$ is the traffic intensity as usual. Thus, the expected number in the system is simply $E(L) = \rho$. Since there is no waiting in this system, other quantities are easy to derive but perhaps not very informative. For example, the distribution of the steady-state time-in-system W is the same as the service distribution, namely, exponential with rate μ.

4.2.4. *Erlang Loss System M/M/s/s*

Although this model is associated with telephony rather than industrial engineering, its historical importance and simplicity justify mentioning it. Suppose we have a telephone exchange with s circuits, meaning that it can handle at most s simultaneous calls. The key measure of performance of this system is the fraction of calls that are "lost," i.e., that arrive to find all circuits busy. Erlang modeled this system as an $M/M/s/s$ queue in which calls arrive in a Poisson process of rate λ and last an exponential time with mean $1/\mu$. The servers here are the circuits. This may be viewed as the birth-and-death process with $\mu_n = n\mu$ for $n \geq 0$, $\lambda_n = \lambda$ for $n \leq s$, and $\lambda_n = 0$ for $n > s$. The steady-state distribution π may be found from (38), giving

$$\pi_n = \frac{\lambda^n/n!}{1 + \lambda + \lambda^2/2! + \ldots + \lambda^s/s!}, \; n = 0, 1, \ldots, s \tag{42}$$

With $n = s$, this is the proportion of time that all s circuits are busy. This is also the proportion of arriving calls that are lost, because of the PASTA property to be discussed below in Section 5.5. Equation (42) in the case $n = s$ is the celebrated *Erlang loss formula.*

4.3. A Comparison of Systems

To illustrate what can be gained from the mathematical results of the previous subsection, we present a comparison of three systems. The three systems are identical in the load that must be served and in the service capacity available, but differ slightly in the queueing discipline. Here are the systems.

- *System 1* consists of two separate single-server queues, each having arrivals at rate λ and service at rate μ. This system is modeled as two independent $M/M/1$ queues.
- *System 2* consists of a single queue served by a pair of servers. The arrival stream has rate 2λ, and each server serves at rate μ. This is modeled as an $M/M/2$ queue.
- *System 3* consists of a single queue with an arrival stream of rate 2λ and a single server serving at rate 2μ. This is modeled again as an $M/M/1$ queue.

Denoting λ/μ by ρ, the overall traffic intensity in each case is 2ρ (not ρ). Using formula (39) above for the first and third systems and (41) for the second, the steady-state mean numbers-in-system for the three cases may be found, leading to

$$E(L_1) = 2\frac{\rho}{1-\rho} \geq E(L_2) = \frac{2\rho}{1-\rho^2} = \left(\frac{2}{1+\rho}\right)\frac{\rho}{1-\rho} \geq E(L_3) = \frac{\rho}{1-\rho}$$

Thus, from the viewpoint of work-in-process, the third system is the most favorable and the first is the least favorable. This has a simple explanation. System 1 compares poorly to system 2 in that the former may have an idle server when there are two jobs in the system, whereas this cannot occur in the latter. Also, system 2 compares poorly to system 3 in that the former serves at half the rate of the latter when there is only one job in the system. Combining the queues almost halves average work-in-process for high traffic intensities. Using a single fast server to serve the combined queue halves average work-in-process for all traffic intensities. The comparison is intuitive. The process of modeling has led to quantification of the intuition that is satisfying and natural, even if the assumptions of the models, such as exponentiality, are somewhat arbitrary. The advantage of maintaining a single queue served by several servers over having a separate queue for each server, demonstrated here in the comparison of system 1 and system 2, is well known and is put to good use in many retail banks and at airline check-in counters.

4.4. Models Based on the $M/G/1$ Queue

The $M/G/1$ queue is a single-server queue with Poisson arrivals and independent, identically distributed service times having an arbitrary distribution. Here we analyze this queue and several of its variants. The results show that quite a broad range of phenomena can be modeled simply enough to allow basic performance measures to be given by explicit formulas. As always, λ and μ denote the arrival and service rates, respectively.

4.4.1. M/G/1 Waiting Time Formula

The *Pollaczek–Khintchine formula* gives the steady-state expected waiting time for this queue as

$$E(W_Q) = \frac{\lambda E(S^2)}{2(1-\rho)} = \frac{\rho(1 + c_s^2)}{2\mu(1-\rho)} \tag{43}$$

where S is a random variable whose distribution is the service distribution and c_s^2 is its SCV. The $M/M/1$ queue is a special case, for which $c_s^2 = 1$. See (39).

4.4.2. M/G/1 under a Priority Discipline

Suppose that, instead of a homogeneous stream of arriving jobs, the jobs are of different classes that are treated differently in the queue. The classes range from 1 to p, say, and jobs of class k, $1 \leq k \leq p$, are given priority k. This means that a job of class k in the queue would be allowed to enter service only if there were no higher priority jobs (i.e., jobs of class $k' < k$). The discipline is taken to be nonpreemptive, so that a low-priority job in service will complete service rather than being ejected by an arriving high-priority job. Jobs of class k are assumed to arrive in a Poisson process, with rate λ_k, and are assumed to have their own characteristic service distribution function G_k. We write λ for the overall arrival rate, ρ_k^+ for the traffic intensity of jobs of priority k or higher, and μ_k for the service rate of class k:

$$\lambda = \sum_{i=1}^{p} \lambda_k,\ \rho_k^+ = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} \text{ and } \frac{1}{\mu_k} = \int_0^\infty x dG_k(x)$$

We also introduce S, a random variable with the aggregated service distribution, whose distribution function is defined as $G = \Sigma\ \lambda_k G_k / \lambda$. With this notation, we have the following formulas for the steady-state expected waiting time of jobs of each class.

$$E(W_Q^1) = \frac{\lambda E(S^2)}{2(1 - \rho_1^+)} \text{ and, for } k > 1,\ E(W_Q^k) = \frac{\lambda E(S^2)}{2(1 - \rho_k^+)\,(1 - \rho_{k-1}^+)}$$

It may be shown from this that, in order to minimize the expected number of jobs in the system (work-in-process), the shorter jobs should receive higher priority. This is a manifestation of the general principle that short jobs should be served first. For details, see for example Buzacott and Shanthikumar (1993). This underlies the idea of having express checkout lines in supermarkets.

4.4.3. The M/G/1 Queue with Batch Arrivals

Suppose that, instead of jobs arriving singly at the Poisson arrival times, they arrive in batches. Suppose that the batch sizes are independent and identically distributed. This arrival process is known as a *compound Poisson process* and is fundamentally more variable than the Poisson process. We assume the FIFO discipline. We write m_B and c_B^2 for the mean and SCV of batch size. Then the traffic intensity is $\rho = \lambda m_B / \mu$, where λ is the arrival rate of batches and μ is the service rate of individual jobs. It is easy to derive the steady-state expected waiting time in this system. The nth *batch waiting time* is the time between the arrival of the nth batch and initiation of service for the first job in the batch. By treating an entire batch as a single job, the steady-state batch waiting time may be deduced directly from the Pollaczek–Khintchine formula (43) as

$$E(W_{Q(\text{Batch})}) = \frac{\rho(1 + c^2)}{2\mu(1 - \rho)} \tag{44}$$

where c^2 is the SCV of the service time of an entire batch, which may be expressed in terms of the SCV of service for a single job as $c^2 = c_B^2 + c_s^2 / m_B$. The expected waiting time for a job, rather than a batch, is found by adding to (44) the expected total service time of all jobs ahead of a typical job in a batch, and this gives

$$E(W_Q) = \frac{\rho(1 + c^2)}{2\mu(1 - \rho)} + \frac{1}{\mu} \frac{E(B(B - 1))}{2m_B} \tag{45}$$

where B is a random variable having the batch-size distribution.

4.5. Examining the Assumptions of a Simple Queueing Model

In building a stochastic model for a real system, we make various simplifying assumptions for reasons of tractablility or parsimony. To be effective in modeling, we must appreciate the consequences of these assumptions. In this subsection we illustrate what is involved in this aspect of model building by considering an idealized example. We consider the $M/M/1$ queue as a model for a single worker performing an operation on arriving workpieces. The primary reason for adopting such a model is the mathematical tractability that comes as a consequence of the Markov property. Now we consider the implications of this model.

As we saw in Section 2, Poisson arrivals may be thought of as completely random arrivals in a very exact sense. They are neither too regular, as for example if they came once every five minutes, nor too irregular, as for example if they came in batches of between 1 and 20 with highly variable interarrival times. The Poisson process implies memorylessness, in that observations of past arrivals give no hint as to future arrivals. To expand on this property, if the arrivals to the real system have been unusually heavy recently, this is no indication that they will continue to be heavy, nor is it an indication that they will become less heavy. If actual arrivals tend to come in clusters, or if the arrival rate fluctuates throughout the workday, the Poisson model may be inappropriate. The qualitative behavior of the Poisson process may or may not be consistent with the actual nature of the system. The success of the model may depend heavily on having consistency here.

The assumption that the service times are independent implies, for example, that knowledge of one service time tells us nothing about the next service time. If one service time is particularly long, then this gives us no reason to expect the next one to be long, or short, or unusual in any way. In reality, there may be serial correlations between the service times, caused for example by fluctuations in the attentiveness of the worker or in the quality of the arriving workpieces. This, of course, violates the assumption of independent service times. These effects may give rise to a tendency to have several long or short service times in a row, which would tend to increase waiting times.

Service times are also assumed to follow the exponential distribution in the $M/M/1$ model. One feature of the exponential distribution is that it is highly variable. For this distribution, 14% of service times are over twice the average service time and 39% of service times are less than half the average. This variability is inconsistent with a routine manual task performed on uniform parts. It may be consistent with a cognitive task, such as diagnosis of the cause of a malfunction. So whether this assumption is reasonable depends heavily on the nature of the task being performed.

Good modeling requires that we understand the qualitative consequences of modeling assumptions and that we use this understanding to make choices consistent with the reality we are trying to model.

5. SOME GENERAL PRINCIPLES

5.1. Long-Run Behavior

The first focus of the mathematics of stochastic models is to determine long-run or steady-state behavior. This is because long-run behavior requires less detail to describe than the full evolution of the system, and it often gives a good sense of the overall behavior of the system over the time frame for which it is to be in operation. In interpreting results on long-run behavior, one must bear in mind a sense of how long a time interval is needed before average performance may be approximated by long-run performance.

5.1.1. Steady-State vs. Long-Run Averages

As we saw in Subsection 3.3, there are several competing descriptors of long-run behavior, including steady-state expected values, which are averages with respect to the steady-state or stationary distribution, and long-run or time averages. Usually, the mathematics leads most easily to steady-state averages. This is exemplified by the fact that, in a Markov chain, the steady-state distribution π is easy to characterize and to compute. *Ergodic properties* are equalities between these two kinds of averages. To relate this to queues, as before let us denote by $L(t)$ the number-in-system at time t for a certain queueing system, and by W_n the time-in-system for the nth job to arrive. Define the long-run average number-in-system by

$$\overline{L} = \lim_{T\to\infty} \frac{\int_0^T L(t)\,dt}{T} \tag{46}$$

Similarly, define the long-run average time-in-system by

$$\overline{W} = \lim_{n\to\infty} \frac{W_1 + W_2 + \ldots + W_n}{n} \tag{47}$$

Let L and W be random variables whose distributions are the steady-state distributions of the number-in-system and time-in-system for the queue, respectively. Then, in great generality, we have the ergodic properties

$$\overline{L} = E(L) \text{ and } \overline{W} = E(W) \tag{48}$$

The difficulty in understanding these statements is not to see that they are true but to distinguish between the concepts represented by the two sides of the equalities.

5.1.2. What Goes In Must Come Out

This simple principle, more formally expressed, says that the long-run arrival rate to a system equals the long-run departure rate. The reason is clear: the difference between the number of arrivals and the number of departures over a time interval equals the change in the number of jobs actually in the system, and, as long as the number in the system is not allowed to grow relentlessly, the numbers of arrivals and departures cannot differ by very much. This simple fact leads quickly to the important *traffic equations* (53) for a queueing network.

5.1.3. Little's Law and Other Conservation Laws

Suppose jobs arrive at a system, spend some time there, perhaps waiting and being served, and then depart. The system may be a warehouse or other storage system, a queueing system, or perhaps something more complex, such as a queueing network. Using the notation of (46) and (47), it may be shown in great generality that

$$\overline{L} = \lambda \overline{W} \tag{49}$$

This is *Little's law*. Illustrations of it may be seen in (39) and (41), once the ergodic relations (48) are taken into account. To justify (49) heuristically, following Ross (1997), if jobs arrive at rate λ and on average spend time $\overline{W}$ in the system, then demand for occupancy of the system arrives at a rate of $\lambda\overline{W}$, and so occupancy must be supplied at a rate of $\lambda\overline{W}$ also, meaning that $\lambda\overline{W}$ jobs must be present on average.

This simple fact may be applied to various systems. For example, in the context of a queue, the "system" could be taken to mean the waiting area or the service area rather than the entire system. If applied to the service area of a single-server queue, Little's law yields

$$1 - P(L = 0) = \lambda \left(\frac{1}{\mu}\right), \text{ or } P(L = 0) = 1 - \rho \tag{50}$$

To see that this is a special case of (49), note that $1 - P(L = 0)$ is not only the steady-state probability that the server is busy, it is also the steady-state expected number of jobs in service, which in turn, by an ergodic property, is the time-average number of jobs in the system in question. Of course, $1/\mu$ is the expected time in service and plays the role of $\overline{W}$ in (49). When Little's law is applied to the waiting area, it tells us that the expected number of jobs waiting is the product of the arrival rate and the average waiting time.

5.2. Behavior of a Bottleneck Queue

A bottleneck station is a station with sufficiently high server utilization to cause substantial delays. These are also called stations in *heavy traffic*. In a manufacturing system, such stations are often the ones of greatest interest. They typically represent the critical resources and may be responsible for most of the work-in-process and waiting. Despite the complexity of the exact analysis of the $G/G/1$ queue, in heavy traffic the behavior is somewhat simple. It may be shown that the steady-state expected waiting time satisfies

$$E(W_Q) \approx \frac{\rho(c_a^2 + c_s^2)}{2\mu(1 - \rho)} \tag{51}$$

in the sense that the ratio of the two sides of the approximation converges to 1 as $\rho \rightarrow 1$. Here, c_a^2 and c_s^2 are the SCVs of interarrival and service time, respectively. For a more precise statement, see Asmussen (1987). This approximation is exact for the $M/G/1$ queue, in which case $c_a^2 = 1$. See (43).

A common feature of all the waiting-time formulas we have had, from the simple (25) to the general (51), is the presence of the factor $1 - \rho$ in the denominator. This is an important observation. As utilization at a station increases from 90% of capacity to 99%, waiting times [and so also loads, by (49)] typically increase by a factor of 10.

5.3. Deleterious Effects of Variability

In the Pollaczek–Khintchine formula, increasing the SCV of service causes the mean waiting time to increase also. This illustrates the general principle that increasing variability reduces performance. The same principle is illustrated more fully by the approximation (51) for the steady-state expected waiting time in a $G/G/1$ queue in heavy traffic. There we see that the SCVs of the interarrival and service times contribute equally to the approximate expected waiting time.

In the $G/G/1$ queue, it is not true in complete generality that increasing the *variance* of service times, say, will cause average waiting time to increase. To formalize this intuitive idea, we need a sufficiently stringent definition of increased variability. See Ross (1996) for a discussion of the stochastic ordering that makes the intuition precise.

5.4. Rate of Convergence to Steady State

An important example of the usefulness of the mathematics of stochastic models arises when simulating queues with fairly high server utilizations. It is not unusual to find that the mean waiting time, say, is highly variable. Here is a fairly common experience. We carry out a simulation run on a model of a manufacturing system, involving the processing of $n = 100{,}000$ jobs. We compute the mean waiting time $\overline{W}_n$ over the run. Now we repeat the experiment, with a different seed for the random number generator, and compute another mean waiting time $\overline{W}'_n$. $\overline{W}_n$ and $\overline{W}'_n$ may well differ by on the order of 20%. The source of this extreme variability may not be clear from the model being simulated, and yet such variability in a real system would be indicative of poor design. To explain this phenomenon mathematically, suppose we are simulating a very simple system, say a $G/G/1$ queue. It may be shown that the asymptotic distribution of the mean $\overline{W}_n$ of the first n waiting times, for n large and ρ near 100%, is normal with expected value given by (51) and CV

$$\frac{1}{(1-\rho)}\sqrt{\frac{2(c_a^2 + c_s^2)}{n}} \tag{52}$$

(See Whitt 1989 and Asmussen 1992 for more on these limit results.) Because we must choose the simulation run-length n large enough to make this small, a substantial multiple of $(1-\rho)^{-2}$ waiting times is needed to get a reliable estimate of $\overline{W}$ from simulation. If $\rho = 0.95$, then $(1-\rho)^{-2} = 400$, and for an $M/M/1$ queue, for which $c_a^2 = c_s^2 = 1$ (see Subsection 1.3), the indications are that about 160,000 waiting times are needed to estimate $\overline{W}$ to about 20% accuracy with 95% probability. One interpretation of this result is that the queue converges to steady state very slowly for high traffic intensities, resulting in persistent correlations between the successive waiting times and slow attenuation of the variance of the mean waiting time as run-length n increases. This capability that the mathematical analysis of stochastic models provides to explain qualitatively and thereby validate the behavior of simulations is perhaps one of its most important uses in engineering.

5.5. ASTA and PASTA

Consider an $M/G/1$ queue with batch arrivals, the average batch size being 10 jobs. Suppose that the traffic intensity is low, say $\rho = 0.1$. Then arriving jobs see the queue in atypical conditions because 90% of them see other jobs ahead of them in the queue, whereas, since the traffic intensity is 0.1, in fact the queue is empty 90% of the time by (50). So jobs may tend to arrive when the station is busy, as in the example just given, or they may tend to arrive when it is not busy. The latter occurs in the $D/D/1$ queue, in which arrival and service times are constant. Suppose that initially the queue is empty and each interarrival time is 1 unit while each service time is 0.9 units. Then, although the utilization is 90%, no job ever has to wait. Incidentally, this illustrates the remark made earlier that waiting can be eliminated without increasing server capacity if variability can be eliminated.

What of the $M/M/1$ queue? Again, it illustrates something interesting—a perfect balance between the two extreme examples of the previous paragraph. It illustrates the property that *Poisson arrivals see time averages* (PASTA). More informally, Poisson arrivals see typical conditions. This property means that if we were to make observations on the queue at the arrival times, the distribution of these observations would be the time-average distribution of the system. In particular, the average number of jobs in the system just before arrival times is the same as the long-run average number of jobs in the system. Similarly, the proportion of arrivals that find the queue empty is precisely $1 - \rho$, the long-run proportion of time that the queue is empty.

This property extends to some more complex situations. For example, in the *open product-form queueing networks* to be discussed in the next section, the arrivals to the individual queues see the entire network in time-average (or steady-state) conditions. These arrivals are not necessarily Poisson, and so the acronym ASTA (arrivals see time averages) has been coined to describe this situation.

6. QUEUEING NETWORKS

A queueing network is a collection of stations among which jobs move and compete for service. Each station is a queueing system in its own right. Queueing networks are commonly used as models for manufacturing, telecommunication, and transportation systems. These models are usually analyzed by simulation. The level of complexity of queueing network models, from the viewpoint of mathematical analysis, is far beyond that of the individual stations. This is because the process of going from a single queue to a queueing network is the process of increasing the dimensionality of the model—going from a single dimension to many. As in all areas of applied mathematics, this brings with it a qualitative increase in difficulty. This is Bellman's "curse of dimensionality." Because of this, queueing networks must be approached with modest expectations as to what may be accomplished.

To understand what is known about queueing networks, one must recognize first that some networks have truly complex behavior and that a general and detailed theoretical understanding of them is probably unattainable. On the other hand, there is a remarkable class of networks, the *product-form networks,* for which a great deal is known. To a certain extent, these networks behave as if they consisted of collections of independent stations, and to that extent the stations may be analyzed separately. Thus, a high-dimensional problem is replaced by several low-dimensional problems. In Subsections 6.1 and 6.2, we treat product-form networks. Then, to highlight the special nature of these networks, in Subsection 6.3 we discuss some networks whose behavior is complex and difficult to analyze. In particular, these networks may allow queues to grow without bound even when there is sufficient service capacity to process all arriving work. In between these extremes, there appear to be many queueing networks which, although not product form, are fairly amenable to simple *decomposition approximations.* This is the topic of Section 7.

It is difficult to describe the family of product-form networks qualitatively. One attempt is to say that they are a family of networks in which jobs behave in a highly random but nonconspiratorial manner. This is to say, the jobs do not interfere with one another in complex ways: they do not tend to batch together in their movements, and when a job arrives at a station it tends to see that station, and the larger network, in a typical condition, in the sense of the ASTA property discussed in Subsection 5.5. In particular, jobs are neither more nor less likely to arrive when a station is heavily loaded.

6.1. Jackson Networks

We describe the famous *Jackson networks* and discuss their remarkably simple steady-state distributions. Consider a network of M FIFO stations. Let s_m denote the number of servers at station m, $m = 1, 2, \ldots, M$. *Exogenous* arrivals, that is, arrivals from outside the network, enter station m in a Poisson process of rate λ_m^*. Jobs at station m require exponential service with rate μ_m. When a job completes service at a station, it either visits another station or leaves the network. The probability that a job that has just completed service at station m next goes to station n is written as p_{mn}, the choice of next station being independent of everything that has happened up to the time that choice is made. These routing probabilities are collected into an $M \times M$ routing matrix $P = (p_{mn})$. The technical condition that $I - P$ be invertible is needed to ensure that all jobs eventually leave the network.

This is quite a complicated process. Jackson's (1957) remarkable discovery was that the steady-state distribution is very simple. To identify it, we first find the overall arrival rate λ_n at each station n—this is the total of the exogenous arrival rate and the arrival rate of jobs from within the network. The exogenous arrival rate is λ_n^*. The exogenous arrival stream must be combined with the streams of jobs routed to station m after completing a service. A proportion p_{mn} of the jobs completing service at station m is routed to station n, for each $m = 1, 2, \ldots, M$. But the departure rate from station m is the same as its arrival rate, λ_m ("what goes in must come out"—see Section 5.1), and so we have

$$\lambda_n = \lambda_n^* + \sum_{m=1}^{M} \lambda_m p_{mn} \tag{53}$$

These equations are called the traffic equations. Invertibility of $I - P$ implies that these equations have a unique solution in the λ_ms. Suppose now that $\rho_m < s_m$, $\lambda_m > 0$ and $\mu_m > 0$ for each $m = 1, 2, \ldots, M$. These conditions ensure that the network is an irreducible, positive recurrent Markov chain. Jackson proved the following result: The steady-state distribution of the number of jobs at each station in the Jackson network is the same as that of M independent stations, the mth being an $M/M/s_m$ queue with arrival rate λ_m and service rate μ_m.

Here is a simple consequence. If we take each station to be a single-server queue, and denote by ρ_m the traffic intensity λ_m/μ_m of the mth station, then the steady-state expected flow time (of an exogenous arrival), written $E(T)$, of a job through the network is

$$E(T) = \frac{1}{\sum_{m=1}^{M} \lambda_m^*} \sum_{m=1}^{M} \frac{\rho_m}{1 - \rho_m} \tag{54}$$

To explain, the rightmost sum is the expected total number in the system by (39). This is divided by the overall exogenous arrival rate to get the average time-in-system, using Little's law (49).

Example 7.1: Consider the simple network depicted in Figure 2. There are two stations in tandem, processing an exogenous stream of rate $\lambda_1^* = 1$ arriving at the first station. There are no exogenous arrivals at the second station, so $\lambda_2^* = 0$. The service rates are $\mu_1 = 5$ and $\mu_2 = 4$. For any job completing service at the second station, there is a probability $p_{21} = 0.5$ that it will be sent back through the two stations again. Otherwise, it departs the network. By solving the traffic equations (53), we find the total arrival rates at the stations to be $\lambda_1 = \lambda_2 = 2$. Therefore the traffic intensities are $\rho_1 = \lambda_1/\mu_1 = 2/4 = .5$ and $\rho_2 = \lambda_2/\mu_2 = 2/5 = 0.4$, which are both less than 1 and so the

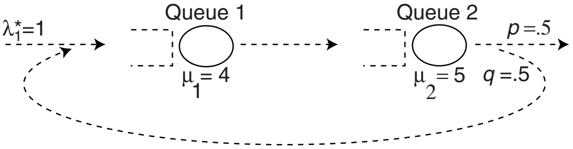


Figure 2 A Simple Queueing Network.

network is positive recurrent. Now from (54), the steady-state expected flow time of an exogenous arrival through the network is

$$E(T) = \frac{1}{\lambda_1^*}\left(\frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}\right) = 1 + \frac{2}{3} = 1.666$$

The mean waiting times per arrival (rather than per exogenous arrival) at the two stations are given by

$$E(W_Q^{-1}) = \frac{0.5}{(4)\,(1-0.5)} = 0.25 \text{ and } E(W_Q^2) = \frac{0.4}{(5)\,(1-0.4)} = 0.133$$

6.2. General Product-Form Networks

In the Jackson network, the jobs at a station are indistinguishable in that they all have the same service requirements and all follow the same routing rules. This is a feature possessed by few real-world systems. Jobs in a queue often differ in urgency, routing, and service requirements. In this subsection we discuss *multiclass product-form networks,* in which jobs are distinguishable into classes, each class having its own distinctive routing pattern. In the networks considered, jobs behave in the same highly random but nonconspiratorial manner we observed in Jackson networks. We also consider some variations involving nonexponential service distributions and priority disciplines.

The first step in defining multiclass networks is to introduce a framework for describing a variety of routing rules in a simple, unified way. In this framework, each job in the network has a class. This class characterizes which station the job is presently at and the rule whereby it chooses its next class. Each class is associated with one and only one station, and so, when a job chooses its next class, this also determines its next station. Suppose our network is composed of M stations serving a total of K classes of jobs. Let $P = (p_{k\ell})$ denote a $K \times K$ *routing matrix,* which is to say $p_{k\ell}$ is the probability that a class-k job becomes a class-ℓ job after completing service. We suppose that $I - P$ is invertible, which is to say each job eventually leaves the network. One important special case allowed under this model is that of several different *types* of job, each type making transitions from one station to another according to its own $M \times M$ routing matrix. In this special case, job class is determined by both current station and type, and the type of a job never changes.

Once we have define the routing rules, the traffic equations for our multiclass network may be written down, using the same logic as in (53). These are

$$\lambda_\ell = \lambda_\ell^* + \sum_{k=1}^{n} \lambda_k p_{k\ell},\ \ell = 1, 2, \ldots, K \tag{55}$$

where the exogenous arrival rates λ_ℓ^* and the total arrival rates λ_ℓ are now specified for each class $\ell = 1, 2, \ldots, K$ rather than for each station as in (53). Note that the service distributions do not enter into this equation. Once we specify the service rates μ_ℓ for each class, the class traffic intensities $\rho_\ell = \lambda_\ell/\mu_\ell$ are determined. The traffic intensity at a *station* is the sum of the traffic intensities of its classes:

$$\rho^{(m)} = \sum_{\substack{\text{all classes } \ell \text{ served} \\ \text{at station } m}} \rho_\ell,\ m = 1, 2, \ldots, M \tag{56}$$

To complete the description of a family of multiclass queueing networks, we suppose that all exogenous arrival processes are independent Poisson processes, that the queues are FIFO, and that the service distributions at each station are exponential. We write s_m for the number of servers at station m, $m = 1, 2, \ldots, M$. This network is not a product-form network, in general. However, if we further suppose that all classes ℓ served at a given station m have the same service rate $\mu^{(m)}$, so that $\mu_\ell = \mu^{(m)}$ for each such ℓ and m, then the network is product form. We refer to this network as the *standard multiclass product-form network*. We assume that the traffic intensity $\rho^{(m)}$ at each station m is less than the number of servers:

$$\rho^{(m)} < s_m, \; m = 1, 2, \ldots, M \tag{57}$$

This condition is known as the *capacity condition*. It ensures that sufficient service capacity is available to handle the arriving work. Special cases of this condition have arisen at several earlier points in this chapter. If the station service rates $\mu^{(m)}$ and the class total arrival rates λ_k are all positive, then (57) implies that the standard multiclass product-form network is an irreducible, positive recurrent Markov chain. Remarkably, an explicit formula for the steady-state distribution of this network is known (Kelly 1979). While the formula is complicated, it has a simple description in words.

1. The configurations (i.e., the classes and positions in queue of all the jobs) of the different stations at a fixed time in steady state are independent.
2. The number of jobs at station m has the same steady-state distribution as the $M/M/s_m$ queue with traffic intensity $\rho^{(m)}$.
3. Given that there are n jobs present at a particular station m, the number of jobs of each class has the multinomial distribution (Ross 1997) with n trials and with outcome probabilities $\rho_\ell/\rho^{(m)}$ for classes ℓ served at station m.

The distribution of the overall numbers of jobs at the stations described in (1) and (2) is exactly as in the Jackson network. Formulas for steady-state waiting and flow times in this multiclass network may be written down just as we did in the case of Jackson networks—see, for example, (54).

We can generalize the standard multiclass product-form network in several ways and still get the same simple steady-state distribution. Instead of the FIFO discipline, we can use any queueing discipline that depends only on the position of the jobs in the queue (and not, for example, on a job's class or service time) and the same steady-state distribution will apply. Under certain queue disciplines, nonexponential service distributions may be allowed, and the service distribution may be allowed to depend on class. These disciplines include SIRO, PS, and preemptive LIFO, which were discussed in the introduction to Section 4. The $M/M/s$ queue under any of these disciplines becomes what is known as a *symmetric queue* (Kelly 1979), and this property underlies the possibility of relaxing the exponentiality assumption while maintaining the product-form steady-state distribution.

All of these networks have very special structure, however, either having exponential service with a rate that does not depend on class or a somewhat exotic service discipline. And yet the product-form networks constitute a broad family having a simple steady-state behavior. It seems reasonable to view this steady-state behavior as in some sense typical of a well-behaved queueing network and to use the product-form solution as a first approximation to any reasonable network.

6.3. General Networks: Stability and Instability

A simple example of an unstable queue is an $M/M/1$ queue in which the arrival rate λ exceeds the service rate μ. For this queue, the number in the system grows without bound. In fact, the number of jobs in the system at time t, $L(t)$, satisfies

$$\lim_{t\to\infty} \frac{L(t)}{t} = \lambda - \mu > 0 \tag{58}$$

so the amount of work in the system grows roughly linearly over time. This kind of behavior cannot persist for very long in real systems, as waiting jobs will ultimately overwhelm any finite storage capacity. This is a strong form of instability. In contrast, if $\lambda < \mu$ then the limit of (58) is zero and, moreover, the expected queue size at time t converges to the value given in (39) as $t \to \infty$. This is typical of a stable queue. The critical case, in which $\lambda = \mu$, is more subtle; here the expected queue size at time t grows like $\sqrt{t}$. This we classify as unstable also.

For Markovian networks (with a discrete state space), there is a convenient definition of stability: a stable network is one that is a positive recurrent Markov chain. This definition may be extended to more general networks, but doing this formally would require a mathematical digression. For a detailed discussion of stability, see Meyn and Tweedie (1993).

The $M/M/1$ queue with $\lambda > \mu$ fails to be stable for a very simple reason: there is insufficient service capacity to handle the load. A feature of this situation is that the traffic equations (53) actually fail. The arrival and departure rates are *not* equal, the former being λ and the latter being the smaller value μ. In the case of a multiclass queueing network, we refer to the quantities λ_ℓ and ρ_ℓ determined by the traffic equations (53) as the nominal arrival rate and traffic intensity for class ℓ. Similarly, $\rho^{(m)}$

in (57) is the nominal traffic intensity for station m. A simple necessary condition for a queueing network to be stable, then, is that there be sufficient capacity to handle the load, which is condition (57) above. We have referred to this as the capacity condition for stability. This condition is sufficient to guarantee stability of product-form networks. However, as indicated earlier, the relative tractability of product-form networks conveys a false sense of how complex the behavior of general networks can be. Only in the late 1980s did researchers focus on the serious complications that may arise with non-product-form queueing networks. The Rybko–Stolyar (1992) network is an early example of a network exhibiting a certain key behavior, *subcritical instability*. This network is unstable in the sense that the amount of work in the system grows linearly with time, even though the capacity condition (57) is satisfied. The network is somewhat contrived in that it follows a discipline in which priority is given to slow jobs over fast jobs, contrary to the advice of Subsection 4.4 above. More recently, Bramson (1994) showed that the phenomenon of subcritical instability arises even under the FIFO discipline. While Bramson's network is again somewhat contrived, it clearly points to a need for a deeper understanding of instability. General necessary and sufficient conditions for stability of multiclass networks are not known at the present time and do not appear to be in the offing.

On the other hand, several results guaranteeing the stability of a queueing network are known. One approach to the question of stability is to initialize the network with a very large number of jobs in it and then to study its behavior over a proportionately long period of time. With this perspective, the jobs have so small an individual impact that they behave collectively like a fluid, and the network approaches a *fluid limit*, which is described by a *fluid model* reflecting the original network behavior. It has been proved in substantial generality that if a unit of fluid is emptied out of the fluid network within a fixed time, however it may be distributed among the stations, then the original network is stable. See Dai (1995) for details.

A *reentrant line* is a network in which all jobs follow the same route. The same station may be visited several times along the route—-hence the name "reentrant." These are natural models of certain semiconductor manufacturing systems. A feature of these networks is that the different classes served at a station have a natural ordering, from the earliest stage in the route ("first") to the latest stage in the route ("last"). Two common disciplines, first-buffer-first-served (FBFS) and last-buffer-first-served (LBFS), in which the buffers at a station receive priority according to how advanced they are along the common route, are known to be stable under subcritical nominal loads. See Dai and Weiss (1995) for details.

7. TWO-MOMENT APPROXIMATIONS AND DECOMPOSITION METHODS

Consider a queueing network, perhaps the Jackson network of Subsection 6.1 or the standard multiclass FIFO network described in Subsection 6.2. Suppose now we relax the assumption that service and external interarrival times are exponential. We may wish to do this, for example, because there are automated workstations in the system being modeled and service times at these stations are much less variable than the exponential distribution. Without exponential distributions, the explicit formulas for expected flow times and numbers-in-system of Section 6 fail, and, moreover, it is a formidable problem to devise a direct algorithm to compute these quantities. Simulation may be an effective answer to the problem. For this approach, see Subsection 1.2 and the chapters of the Handbook dealing with simulation referenced there. Here we address the question, What can be said about these more general networks, using a basic knowledge of stochastic models? In answer to this question, we present an approach to developing sensible and intuitive approximations. These approximations are built around two themes: (1) they are two-moment methods, requiring only means and variances (equivalently, SCVs) of service times and external interarrival times, and (2) they are decomposition methods, treating different stations as independent in steady state.

In support of this approach to queueing networks, there are the following points. In many situations, rough information on the mean and variability of service times and exogenous interarrival times is all we have to work with, even if the method of analysis is simulation. In a given context, it may also be intuitively clear that these quantities largely determine performance. Mathematics supports the two-moment approach to some degree. For example, the heavy-traffic mean of the single-server queue (51) is determined by the first two moments of the interarrival and service distributions. This phenomenon extends to single-class networks, whose heavy-traffic behavior depends only on two moments. See Reiman (1981) for an early result along these lines and Williams (1996) for a more recent survey. In support of decomposition as a theme for developing approximations, we note that a product-form network is a network for which decomposition is not merely an approximation but an exact mathematical property. It is natural in approximating more general networks to suppose that this decomposition property continues to hold.

On the other hand, there are many multiclass networks, for example the Rybko–Stolyar network discussed in Subsection 6.2, whose behavior is clearly not consistent with decomposition approximations, and so we must be modest in our expectations as to how far decomposition methods will take us in understanding queueing networks.

7.1. Introduction to Decomposition

Decomposition methods give quick, rough approximations for queueing networks using two-moment information. The effectiveness of these methods depends heavily on the nature of the network being approximated. It is not hard to find examples where any particular decomposition method will perform badly.

Decomposition is the process of approximating a system by breaking it up into parts, analyzing the parts, and then putting them back together. This general approximation strategy may be used in many diverse settings, offering fast approximate solutions to intractable computational problems. Typically, a decomposition method for queueing networks is based on two things: a method for approximating the behavior of individual stations and a method for approximating the flows of jobs between stations. The entire network is analyzed by characterizing the behavior of the individual stations in a manner that is consistent with the flows between the stations. Not only are decomposition methods fast, but they often also provide explicit approximate formulas that give insight into how variability affects performance in fairly complex systems. See Subsection 7.3 for an example. Decomposition methods are a central theme of two books on manufacturing systems, Buzacott and Shanthikumar (1993) and Gershwin (1994), and also feature in Hopp and Spearman (1996) and Compton (1997). They may be viewed as analogous to the analysis of variance (ANOVA) in statistics: a sensible elementary analysis, of great value as a benchmark, despite concerns about validity of approximations or assumptions. Just as ANOVA is valid under strict normality assumptions, the decomposition approach is exact for product-form networks.

In the next subsection we present a simple two-moment decomposition approximation scheme for queueing networks, based on the literature up to the time of publication of Whitt's 1983 paper describing the Queueing Network Analyzer (QNA). The underlying methodology is elegant and simple and makes small computational demands both in terms of CPU cycles and lines of computer code. The basic methodology is a part of many software packages for analyzing production systems, for example MPX (Suri and de Treville 1991).

7.2. A Simple Decomposition Method

We generally follow Whitt (1983), which contains both an elegant synthesis of early decomposition work and significant innovations. See also Nelson (1995, Subsection 8.10.2) for another simple analyzer. The network for which we develop the methodology is the *generalized Jackson network*, which is like the Jackson network of Subsection 6.1 with regard to routing, but its service times and external interarrival times are allowed to have general distributions. This network is stable under the capacity condition (57), and so we know that its behavior is not as erratic as, say, the Rybko–Stolyar network. This gives us some assurance that it is a reasonable network to attempt to approximate. The only change in the parameterization used in Subsection 6.1 is that we now must define an SCV of external interarrivals, c_{am}^{*2}, and of service times, c_{sm}^2, for each station m, $m = 1, 2, \ldots, M$.

We describe the behavior of a queueing network in terms of a collection of interacting point processes, or streams. These streams are the job flows of the network. All streams will be approximated by renewal processes. In dealing with these approximating renewal processes, we suppose that the only information we have about them is the mean m and variance σ^2 of the interarrival times. Knowing m and σ^2 is, of course, equivalent to knowing the rate $\lambda = 1/m$ and SCV c^2. All streams are treated as independent even if they are not.

In approximating a given stream N by a renewal process, we require the approximation to have the same rate λ as N. Here are the two main strategies for choosing the SCV $\tilde{c}^2$ of the approximating renewal process.

- *The stationary-interval method:* Here we choose the interarrival time SCV $\tilde{c}^2$ of the approximating renewal process to be the same as the SCV of the steady-state distribution of the interarrival times of N.
- *The asymptotic method:* Here we choose the interarrival-time SCV of the renewal process so that its asymptotic variance [defined in (18)] agrees with that of N. This means we choose $\tilde{c}^2$ so that

$$\lim_{t\to\infty} \frac{\mathrm{Var}(N(t))}{t} = \lambda\tilde{c}^2 \tag{59}$$

These two approaches typically lead to different approximations to the SCV, unless N itself is renewal, because the variance of $N(t)$ is affected by the correlations between the interarrival times.

Having decided how to deal with streams, we turn to the network. The dynamics of the network are described in terms of three key operations on streams: (1) splitting, (2) superposition, and (3) queueing. Splitting happens when a stream is divided into two or more streams that are directed to different destinations. Superposition happens when two or more streams combine to form a single stream. Queueing is the process of going through a queue—viewed as an operation that transforms the arrival stream into the departure stream. We treat each of these operations on streams through simple approximations, and this leads to a natural approximation for the entire network.

The arrival rates of all streams are determined exactly through the traffic equations (53), and so we need only explain how the SCVs are handled.

7.2.1. *Approximation for Splitting*

First consider a renewal process N with rate λ and interarrival time SCV c^2, and suppose a proportion p of this stream is to be routed to a certain station. The routing model described in Section 6 has the following effect. For each arrival of N we toss a coin with probability of heads p, independently of everything. We reject the points for which the coin comes up tails and keep those for which the coin comes up heads. Now the split stream is again a true renewal process—no approximation here—and its interarrival time SCV $\tilde{c}^2$ is given by

$$\tilde{c}^2 = pc^2 + 1 - p \tag{60}$$

This may be used as an approximation for the splitting operation in the case that N is not renewal. The approximation is linear (more precisely, affine) in the SCV of the input stream. Even when the original stream is not renewal, as $p \to 0$ the split stream approaches a Poisson process and so its SCV approaches 1. This behavior is mimicked by the approximate SCV $\tilde{c}^2$, which also approaches 1 as $p \to 0$.

7.2.2. *Approximation for Superposition*

Suppose we have two independent streams, N_1 and N_2, to be superposed. Denote their rates and SCVs by λ_1, c_1^2, λ_2, and c_2^2. Then, as the variance of the sum of independent things is the sum of the variances by (16), it follows that the asymptotic variance (59) of the superposition is the sum of the asymptotic variances of the individual streams. Using the asymptotic method, this leads to approximating the superposition by a renewal process with SCV $\tilde{c}^2$ determined by

$$\lambda\tilde{c}^2 = \lambda_1 c_1^2 + \lambda_2 c_2^2 \tag{61}$$

where $\lambda = \lambda_1 + \lambda_2$ is the rate of the superposition. Again, $\tilde{c}^2$ is linear in the input SCVs c_1^2 and c_2^2. This formula, extended to nonrenewal and nonindependent streams, is the approximation for superposition.

7.2.3. *Approximation for Queueing*

A queue is viewed as transforming its arrival stream into its departure stream. A natural approximation is

$$\tilde{c}^2 = \rho^2 c_s^2 + (1 - \rho^2)c_a^2 \tag{62}$$

relating the approximation for the SCV of the departure stream, $\tilde{c}^2$, to those of the arrival stream and service time, c_a^2 and c_s^2. See Whitt (1983) for justification. Again, the approximation is linear in the input SCVs. It has the intuitively natural behavior that, as $\rho \to 1$, the departure SCV approaches the service SCV, whereas as $\rho \to 0$, the departure SCV approaches the interarrival SCV.

Rather than giving general equations to complete the specification of the analyzer, we illustrate the process for a simple example.

Example 7.2: We choose the network of Figure 2, treated in Example 7.1, making a single modification. We suppose that the service distribution at station 1 has SCV $c_{s1}^2 = 2$. All other distributions and parameters are as before. Writing c_{am}^2 and c_{dm}^2 for the arrival and departure SCVs at stations $m = 1$ and 2, upon applying (62) to the first station we have

$$c_{d1}^2 = \rho_1^2 c_{s1}^2 + (1 - \rho_1^2)c_{a1}^2 = 0.5 + 0.75c_{a1}^2$$

Clearly $c_{a2}^2 = c_{d1}^2$, as the arrivals to the second station are the departures from the first. and so upon applying (62) to the second station we get

$$c_{d2}^2 = \rho_2^2 c_{s2}^2 + (1 - \rho_2^2)c_{a2}^2 = 0.16 + 0.84c_{a2}^2$$

Using (60) and writing c_{21}^2 for the SCV of the feedback stream from station 2 to station 1, we have

$$c_{21}^2 = p_{21}c_{d2}^2 + (1 - p_{21}) = 0.5c_{d2}^2 + 0.5$$

An equation for c_{a1}^2 may now be written down, using (61). The result is

$$2c_{a1}^2 = (1)c_{a1}^2 + (1)c_{21}^2 = 1 + c_{21}^2$$

These equations may be solved to give $c_{a1}^2 = 1.06$ and $c_{a2}^2 = 1.30$. Substituting these values into the approximation (5.6), we find the approximate waiting times per arrival at the stations to be

$$E(W_Q^{(1)}) = \frac{0.5(1.06 + 2)}{2(4)(1 - 0.5)} = 0.38 \text{ and } E(W_Q^{(2)}) = \frac{0.4(1.30 + 1)}{2(5)(1 - 0.4)} = 0.15$$

The first of these is substantially inflated over the result of Example 7.1, mostly because of the increase in c_{s1}^2 itself but also partly because of the increase in c_{a1}^2 due to the extra variability manifesting in the feedback stream.

7.3. Insights from the Decomposition Approach

An excellent use of the decomposition approach is in devising simple insights into practical questions. We follow Buzacott (1996) in treating the following question using decomposition approximations. *When several tasks must be performed on each of a stream of jobs, should these tasks be carried out by separate servers in a flow-line arrangement, or should they all be carried out by parallel single-server stations?* We invoke two models to draw a conclusion. The first model consists of a flow line of n buffered stations, that is, a tandem network. Each task is performed by a different server. In the second model we have N parallel single-server stations where each server can process a job from start to finish. The question is to quantify the trade-offs. A difference is to be expected in the work-in-process levels, and we focus on quantifying this. Since the capacity of the flow line is constrained by its slowest station, a feature that does not arise in the parallel system, we begin by assuming that the service times for the different tasks are exactly balanced in the sense that their distributions are all the same. We suppose that, in the parallel system, arrivals are allocated at random, each job being assigned independently to one of the stations with equal probability. We may now analyze this model in a variety of ways. We can consider heavy-traffic theory, product-form network theory, and approximations, to make the comparison. Using the heavy-traffic approximation (51) and the splitting approximation (60) in conjunction with the asymptotic method of Subsection 7.1 yields the following approximations for the steady-state expected work-in-process when ρ is close to 1. We have

$$\overline{L}_{\text{series}} \approx \frac{c_a^2 + (2m - 1)c_s^2}{2(1 - \rho)} \tag{63}$$

for the tandem system, whereas for the parallel system we have

$$\overline{L}_{\text{parallel}} \approx \frac{m - 1 + c_a^2 + mc_s^2}{2(1 - \rho)} \tag{64}$$

In the case of Poisson arrivals and exponential service, (7.5) is exact because in this case the network is product-form. In the case of Poisson arrivals but general service, (7.6) is exact as each station becomes an $M/G/1$ queue, for which the Pollaczek–Khintchine formula (43) holds. We see by comparing (63) and (64) that the flow line has a smaller approximate expected number-in-system if and only if $c_s^2 \leq 1/2$. This leads us to conclude that low variability favors the flow line. Thus, repetitive tasks are suited to the flow-line arrangement, whereas tasks demanding cognitive skills may be better grouped.

Had arrivals to the parallel system been assigned to stations *cyclically*, so that the first arrival goes to the first queue, the second to the second, and so forth, the conclusion would have been that the parallel system is always better. This is because of the reduced variability in the arrival processes. Buzacott (1996) gives much more on this and related issues.

Acknowledgements

The author thanks Sunkyo Kim and Graeme Warren for many helpful ideas and suggestions.

REFERENCES

Asmussen, S. (1987), *Applied Probability and Queues.* John Wiley & Sons, New York.

Asmussen, S. (1992), "Queueing Simulation in Heavy Traffic," *Mathematics of Operations Research,* Vol. 17, No. 1, pp. 84–111.

Bramson, M. (1994), "Instability of FIFO Queueing Networks," *Annals of Applied Probability,* Vol. 4, No. 2, pp. 414–431.

Buzacott, J. A. (1996), "Commonalities in Reengineered Business Processes: Models and Issues," *Management Science,* Vol. 42, No. 5, pp. 768–781.

Buzacott, J. A., and Shanthikumar, J. G. (1993), *Stochastic Models of Manufacturing Systems,* Prentice Hall, Englewood Cliffs, NJ.

Compton, W. D. (1997), *The Management of World-Class Manufacturing Enterprises.* Prentice Hall, Upper Saddle River, NJ.

Dai, J. G. (1995), "On Positive Harris Recurrence of Multiclass Queueing Networks: A Unified Approach Via Fluid Limit Models," *Annals of Applied Probability,* Vol. 5, pp. 49–77.

Dai, J. G., and Weiss, G. (1996), "Stability and Instability of Fluid Models for Re-entrant Lines," *Mathematics of Operations Research,* Vol. 21, pp. 115–134.

Gershwin, S. B. (1994), *Manufacturing Systems Engineering,* Prentice Hall, Englewood Cliffs, NJ.

Hall, R. W. (1991), *Queueing Methods for Services and Manufacturing,* Prentice Hall, Englewood Cliffs, NJ.

Hopp, W. J., and Spearman, M. L. (1996), *Factory Physics: The Foundations of Manufacturing Management,* Irwin, Chicago.

Jackson, J. R. (1957), "Networks of Waiting Lines," *Operations Research,* Vol. 5, pp. 518–521.

Kelly, F. P. (1979), *Reversibility and Stochastic Networks,* John Wiley & Sons, New York.

Larson, R. C., and Odoni, A. R. (1981), *Urban Operations Research,* Prentice Hall, Englewood Cliffs, NJ.

Meyn, S. P., and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability,* Springer, New York.

Nelson, B. L. (1995), *Stochastic Modeling, Analysis and Simulation,* McGraw Hill, New York.

Reiman, M. I. (1984), "Open Queueing Networks in Heavy Traffic," *Mathematics of Operations Research,* Vol. 9, pp. 441–458.

Ross, S. M. (1996), *Stochastic Processes,* 2nd Ed., John Wiley & Sons, New York.

Ross, S. M. (1997), *Introduction to Probability Models,* 6th Ed., Academic Press, Boston.

Rybko, A. N., and Stolyar, A. L. (1992), "Ergodicity of Stochastic Processes Describing the Operation of Open Queueing Networks," *Problems of Information Transmission,* Vol. 28, pp. 199–220.

Suri, R., and de Treville, S. (1991), "Full Speed Ahead: A Timely Look at Rapid Modeling Technology in Operations Management," *OR/MS Today,* June, pp. 34–42.

Whitt, W. (1989), "Planning Queueing Simulations," *Management Science,* Vol. 35, pp. 1341–1366.

Whitt, W. (1983), "The Queueing Network Analyzer," *Bell System Technical Journal,* Vol. 62, No. 9, pp. 2279–2815.

Williams, R. J. (1996), "On the Approximation of Queueing Networks in Heavy Traffic," in *Stochastic Networks: Theory and Applications,* F. P. Kelly, S. Zachary, and I. Zeidins, Eds, Oxford University Press, Oxford, pp. 35–56.

Wolff, R. W. (1989), *Stochastic Modeling and the Theory of Queues,* Prentice Hall, Englewood Cliffs, NJ.

CHAPTER 84

Decision-Making Models

Mark R. Lehto
Purdue University

1. INTRODUCTION 2173
  1.1. Role and Utility of Chapter 2173
  1.2. Elements of Decision Making 2174
  1.3. Integrative Model of Decision Making 2175
2. CLASSICAL DECISION THEORY 2178
  2.1. Choice Procedures 2178
    2.1.1. Axioms of Rational Choice 2178
    2.1.2. Dominance 2179
    2.1.3. Lexicographic Ordering and EBA 2179
    2.1.4. Minimum Aspiration Level and Satisficing 2180
    2.1.5. Minimax (Cost and Regret) and the Value of Information 2180
    2.1.6. Maximizing Expected Value 2181
    2.1.7. Subjective Expected Utility (SEU) Theory 2182
    2.1.8. Multiattribute Utility Theory 2183
    2.1.9. Holistic Comparison 2184
  2.2. Statistical Inference 2184
    2.2.1. Bayesian Inference 2184
    2.2.2. Signal-Detection Theory 2185
    2.2.3. Dempster-Schafer Method 2186
3. DECISION ANALYSIS 2187
  3.1. Structuring Decisions 2187
    3.1.1. Decision Matrices and Trees 2187
    3.1.2. Value Trees 2188
    3.1.3. Event Trees or Networks 2189
    3.1.4. Influence Diagrams and Cognitive Mapping 2190
  3.2. Probability Assessment 2191
    3.2.1. Direct Numerical Assessment 2191
    3.2.2. Fitting a Subjective Belief Form 2191
    3.2.3. Bisection Method 2191
    3.2.4. Conditioning Arguments 2192
    3.2.5. Reference Lotteries 2192
    3.2.6. Scaling Methods 2192
    3.2.7. Scoring Rules, Calibration, and Group Assessment 2192
  3.3. Utility Function Assessment 2193
  3.4. Preference Assessment 2194
    3.4.1. Indifference Methods 2194
    3.4.2. Direct-Assessment Methods 2195
    3.4.3. Indirect Measurement 2195
4. BEHAVIORAL DECISION THEORY 2195
  4.1. Statistical Estimation and Inference 2196
    4.1.1. Human Abilities/Limitations 2196
    4.1.2. Heuristics and Biases 2198
    4.1.3. Selective Processing of Information 2199
    4.1.4. Models of Human Judgment 2200
    4.1.5. Debiasing Human Judgments 2201
  4.2. Preference and Choice 2201
    4.2.1. Violation of Rationality Axioms 2202
    4.2.2. Framing of Decisions and Preference Reversals 2202
    4.2.3. Prospect Theory 2203

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

4.2.4. Labile Preferences 2204

5. DYNAMIC AND NATURALISTIC DECISION MAKING 2205
  5.1. Naturalistic Decision Making 2205
    5.1.1. Levels of Task Performance 2205
    5.1.2. Recognition-Primed Decision Making 2206
    5.1.3. Image Theory 2207
    5.1.4. Contingent Decision Making 2207
    5.1.5. Dominance Structuring 2207
    5.1.6. Explanation-Based Decision Making 2207
    5.1.7. Shared Mental Models and Awareness 2208
    5.1.8. Team Leadership 2208
  5.2. Time Pressure and Stress 2208

6. GROUP DECISION MAKING 2209
  6.1. Ethics and Social Norms 2209
  6.2. Group Processes 2210
    6.2.1. Conflict 2210
    6.2.2. Conflict Resolution 2211
  6.3. Group Performance and Biases 2212
  6.4. Prescriptive Approaches 2212
    6.4.1. Agendas and Rules of Order 2213
    6.4.2. Idea-Generation Techniques 2213
    6.4.3. Nominal Group and Delphi Technique 2213
    6.4.4. Structuring Group Decisions 2213
    6.4.5. Computer-Mediated Group Decision Making 2214

7. SUMMARY CONCLUSIONS 2214

REFERENCES 2215

1. INTRODUCTION

This chapter focuses on the broad topic of human decision making. Decision making is often viewed as a stage of human information processing because people must gather, organize, and combine information from different sources to make decisions. However, as decisions grow more complex, information processing actually becomes part of decision making and methods of decision support that help decision makers process information become of growing importance. Decision making also overlaps with problem solving. The point where decision making becomes problem solving is fuzzy, but many decisions require problem solving, and the opposite is true as well. Cognitive models of problem solving are consequently relevant for describing many aspects of human decision making. They become especially relevant for describing steps taken in the early stages of a decision where choices are formulated and alternatives are identified.

A complete treatment of human decision making is well beyond the scope of a single book chapter.* The topic has its roots in economics and is currently a focus of operations research and management science, psychology, sociology, and cognitive engineering. These fields have produced numerous models and a substantial body of research on human decision making. At least three objectives have motivated this work: to develop normative prescriptions that can guide decision makers, to describe how people make decisions and compare the results to normative prescriptions, and to determine how to help people apply their ‘‘natural’’ decision-making methods more successfully. The goals of this chapter are to synthesize the elements of this work into a single picture and to provide some depth of coverage in particularly important areas. The integrative model presented in Section 1.3 focuses on the first goal. The remaining sections address the second goal.

1.1. Role and Utility of Chapter

This chapter is intended to provide an overall perspective on human decision making to human factors practitioners, developers of decision tools (such as expert systems), product designers, researchers in related areas, and others who are interested in both how people make decisions and how decision making might be improved. The chapter consequently presents a broad set of prescriptive and descriptive approaches. Numerous applications are presented and strengths and weaknesses of particular approaches are noted. Emphasis is also placed on providing useful references containing additional information on topics the reader may find of special interest.

---

*No single book covers all of the topics addressed here. More detailed sources of information are references throughout the chapter. Sources such as von Neuman and Morgenstern (1947), Friedman (1990), Savage (1954), Luce and Raiffa (1957), and Shafer (1976), are useful texts for people desiring an introduction to normative decision theory. Raiffa (1968), Keeney and Raiffa (1976), Saaty (1988), Buck (1989), and Clemen (1996) provide applied texts on decision analysis. Kahneman et al. (1982), Winterfeldt and Edwards (1986), Sevenson and Maule (1993), Payne et al. (1993), Yates (1994), and Heath et al. (1994), among numerous others, provide recent texts addressing elements of behavioral decision theory. Klein et al. (1993) and Klein (1998) provide introductions to naturalistic decision making.

Section 2 addresses topics grouped under the somewhat arbitrary heading of classical decision theory. The presented material provides a normative and prescriptive framework for making decisions. Section 3 summarizes decision analysis, or the application of normative decision theory to improve decisions. The discussion considers the advantages of the various approaches, how they can be applied, and what problems might arise during their application. Section 4 addresses topics grouped under the heading of behavioral decision theory. The material in the latter section compares human decision making to the normative models discussed earlier. Several descriptive models of human judgment, preference, and choice are also discussed. Section 5 explores topics falling under the heading of dynamic and naturalistic decision theory. This material should be of interest to practitioners interested in the process followed when many real-world decisions are made, the quality of these decisions, and why people use particular methods to make decisions. The discussion provides insight into how people perform diagnostic tasks, make decisions to take risks when using products, and develop expertise. Section 6 introduces the topic of group decision making. The discussion addresses conflict resolution both within and between groups, group performance and biases, and methods of group decision making.

1.2. Elements of Decision Making

Decision making requires that the decision maker make a choice between two or more alternatives (note that doing nothing can be viewed as making a choice). The selected alternative then results in some real or imaginary consequences to the decision maker. Judgment is a closely related process where a person rates or assigns values to attributes of the considered alternatives. For example, a person might judge both the safety and attractiveness of a car being considered for purchase. Obtaining an attractive car is a desirable consequence of the decision, while obtaining an unsafe car is an undesirable consequence. A rational decision maker seeks desirable consequences and attempts to avoid undesirable consequences.

The nature of decision making can vary greatly, depending on the decision context. Certain decisions, such as deciding where and what to eat for lunch, are routine and repeated often. Other choices, such as purchasing a house, choosing a spouse, or selecting a form of medical treatment for a serious disease, occur seldom, may involve much deliberation, and take place over a longer time period. Decisions may also be required under severe time pressure and involve potentially catastrophic consequences, such as when a fire chief decides whether to send fire fighters into a burning building. Previous choices may constrain or otherwise influence subsequent choices (for example, a decision to enter graduate school might constrain a future employment-related decision to particular job types and locations). The outcomes of choices may be uncertain and in certain instances are determined by the actions of potentially adverse parties, such as competing manufacturers of a similar product. Decisions may be made by a single individual or by a group. Within a group, there may be conflicting opinions and differing degrees of power between individuals or factions. Decision makers may also vary greatly in their knowledge and degree of aversion to risk.

Conflict occurs when a single decision maker is not sure which choice should be selected or when there is lack of consensus within a group regarding the choice. Both for groups and single decision makers, conflict occurs, at the most fundamental level, because of uncertainty or conflicting objectives. Uncertainty can take many forms and is one of the primary reasons decisions can be difficult. In ill-structured decisions, decision makers may not have identified the current condition, alternatives to choose between, or their consequences. Decision makers also may be unsure what their aspirations or objectives are, or how to choose between alternatives. After a decision has been structured, at least four reasons for conflict may exist. First, when alternatives have both undesirable and desirable consequences, decision makers may experience conflict due to conflicting objectives. For example, a decision maker considering the purchase of an air bag-equipped car may experience conflict because an air bag increases cost as well as safety. Second, decision makers may be unsure of their reaction to a consequence. For example, people considering whether to enter a raffle where the prize is a sailboat may be unsure how much they want a sailboat. Third, decision makers may not know whether a consequence will happen for sure. Even worse, they may be unsure what the probability of the consequences is, or may not have enough time to evaluate the situation carefully. They also may be uncertain about the reliability of information they have. For example, it may be difficult to determine the truth of a sale person's claim regarding the probability of their product breaking down immediately after the warranty expires.

To resolve conflicts, decision makers must deal appropriately with uncertainty, conflicting objectives, or a lack of consensus. Conflict resolution, therefore, becomes a primary focus of decision theory. The following section presents an integrative model of decision making that relates conflict resolution to the above-discussed elements of decision making. This model specifically considers how decision making changes when different sources of conflict are present. It also matches methods of conflict resolution to particular sources of conflict and decision rules.

1.3. Integrative Model of Decision Making

Human decision making can be viewed as a stage of information processing that falls between perception and response execution (Welford 1976). The integrative model of human decision making, presented in Figure 1, shows how the elements of decision making discussed above fit into this perspective. From this view, decision making is the process followed when a response to a perceived

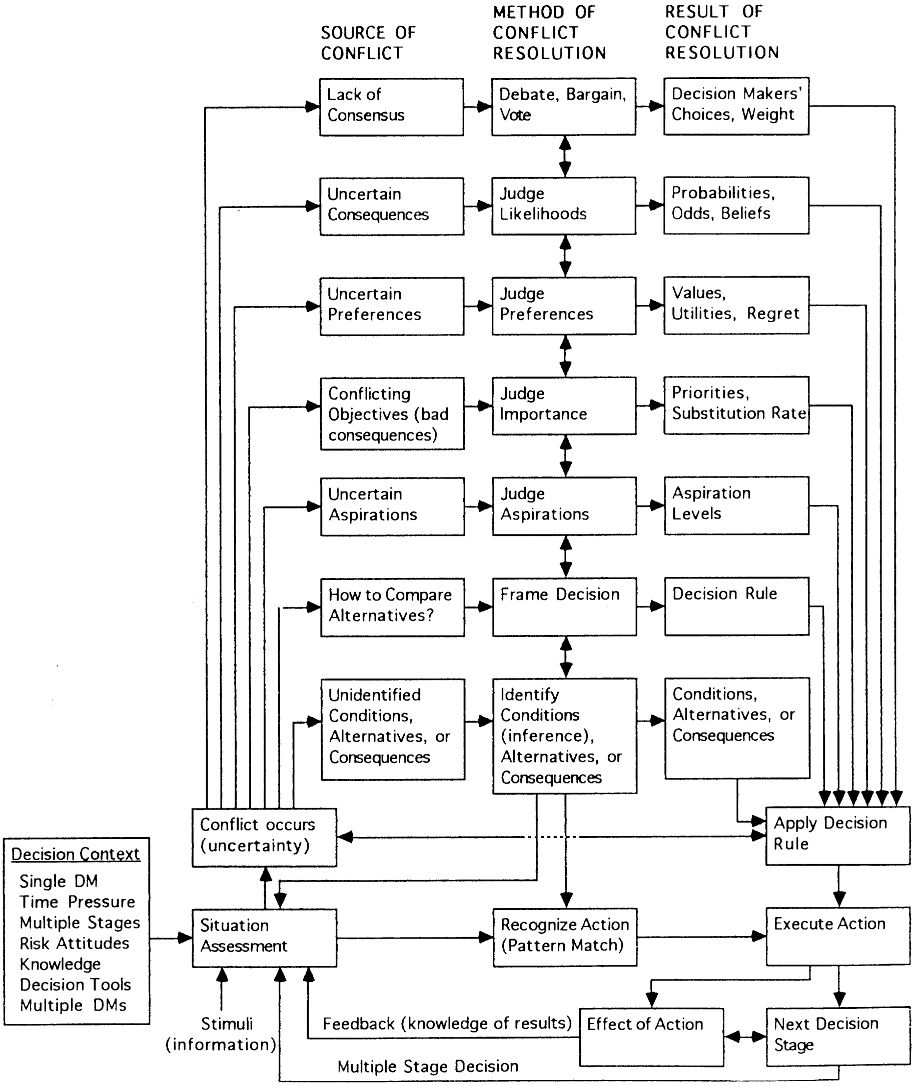

SOURCE OF CONFLICT
METHOD OF CONFLICT RESOLUTION
RESULT OF CONFLICT RESOLUTION
Lack of Consensus
Debate, Bargain, Vote
Decision Makers' Choices, Weight
Uncertain Consequences
Judge Likelihoods
Probabilities, Odds, Beliefs
Uncertain Preferences
Judge Preferences
Values, Utilities, Regret
Conflicting Objectives (bad consequences)
Judge Importance
Priorities, Substitution Rate
Uncertain Aspirations
Judge Aspirations
Aspiration Levels
How to Compare Alternatives?
Frame Decision
Decision Rule
Unidentified Conditions, Alternatives, or Consequences
Identify Conditions (inference), Alternatives, or Consequences
Conditions, Alternatives, or Consequences
Conflict occurs (uncertainty)
Apply Decision Rule
Decision Context
Single DM
Time Pressure
Multiple Stages
Risk Attitudes
Knowledge
Decision Tools
Multiple DMs
Situation Assessment
Recognize Action (Pattern Match)
Execute Action
Stimuli (information)
Feedback (knowledge of results)
Effect of Action
Next Decision Stage
Multiple Stage Decision

Figure 1 Integrative Model of Human Decision Making.

stimulus is chosen. The process followed depends on what decision strategy is applied and can vary greatly between decision contexts.* Decision strategies, in Figure 1, correspond to different paths between situation assessment and executing an action. The particular decision strategy followed depends upon both the decision context and whether or not the decision maker experiences conflict.†

At least four, sometimes overlapping, categories of decision making can be distinguished. *Group decision making* occurs when multiple decision makers interact and is represented at the highest level of the model as a source of conflict that might be resolved through debate, bargaining, or voting. For example, members of a university faculty committee might debate and bargain before voting between candidates for a job opening.

Dynamic decision making occurs in a changing environment, in which the results of earlier decisions impact future decisions. The decisions made in such settings often make use of feedback and are multistage in nature. For example, a decision to take a medical test almost always requires a subsequent decision regarding what to do after receiving the test results. Dynamic decision making is represented at the lowest level of the model by the presence of two feedback loops, which show how the action taken and its effects can feedforward to the assessment of a new decision or feedback to the reassessment of the current decision.

Routine decision making occurs when decision makers use knowledge and past experience to decide quickly what to do and is especially prevalent in dynamic decision making contexts. Routine decision making is represented in Figure 1 as a single pattern-matching step or associative leap between situation assessment and executing an action. For example, a driver after perceiving a stop sign decides to stop, or the user of a word-processing system after perceiving a misspelled word decides to activate the spell checker. Since routine decisions are often made in dynamic task environments, routine decision making is discussed in this chapter as a subtopic of dynamic decision making.

Conflict-driven decision making occurs when various forms of conflict must be resolved before an alternative action can chosen and often involves a complicated path between situation assessment and executing an action.‡ Before executing an action, the decision maker experiences conflict, somehow resolves it, and then either recognizes the best action (conflict resolution might transform the decision to a routine one) or applies a decision rule. Applying the decision rule ideally leads to a choice that is then executed. Attempting to apply the decision rule may, however, cause additional conflicts, leading to more conflict resolution. For example, decision makers may realize they need more information to apply a particular decision rule. In response, they might decide to use a different decision rule that requires less information. Along these lines, when choosing a home, a decision maker might decide to use a satisficing decision rule after seeing that hundreds of homes are listed in the classified ads of the local newspaper.

Potential sources of conflict, methods of conflict resolution, and the results of conflict resolution are listed at the top of Figure 1. Each source of conflict maps to a particular method of conflict resolution, which then provides a result necessary to apply a decision rule, as schematically illustrated in the figure.§ Table 1 presents a set of decision rules, briefly describes their procedural nature and their required inputs, and also lists the sections of this chapter where they are covered. The required inputs of particular decision rules can be easily mapped to sources of conflict. As shown in the table, each decision rule requires that alternatives and their consequences be identified. Other decision rules require measures of aspiration, importance, preference, and uncertainty for each consequence or consequence dimension. For example, to compare alternatives using expected value, the probability and value of each consequence must also be known. Certain decision rules also accept inputs describing the degree of consensus between decision makers.

Accordingly, conflict occurs at the most fundamental level when the current condition, alternative actions, or their consequences have not been identified. At the next most fundamental level, conflict occurs when the decision maker is unsure how to compare the alternatives. In other words, the

---

* The notion that the best decision strategy varies between decision contexts is a fundamental assumption of the theory of contigent decision making (Payne et al. 1993), cognitive continuum theory (Hammond 1980), and other approaches discussed later in this chapter.

† Conflict has been recognized as an important determinant of what people will do in risky decision-making contexts (Janis and Mann 1977). Janis and Mann focus on the stressful nature of conflict and on how affective reactions in stressful situations can impact the decision strategies followed.

‡ The distinction between routine and conflict-driven decision making made here is similar to Rasmussen's (1983) distinction between (a) routine skill or rule-based levels of control and (b) nonroutine knowledge-based levels of control in information-processing tasks.

§ Note that multiple sources of conflict are possible for a given decision context. An attempt to resolve one source of conflict may also make the decision maker aware of other conflicts that must first be resolved. For example, decision makers may realize they need to know what the alternatives are before they can determine their aspiration levels.

TABLE 1 Decision Rules, Required Inputs, and Procedure Applied by the Rules

Decision Rule	Required Inputs	Procedure Applied	Section Covered
Dominance	All alternatives, value of each consequence.	Select alternative best on all consequences.	2.1.2
EBA	All alternatives, value of each consequence.	Select first alternative found to be best on a consequence dimension. Random order of consequences.	2.1.3
Lexicographic	All alternatives, value of each consequence, priorities.	Order consequences by priority. Select first alternative found to be best on a consequence dimension.	2.1.3
Satisficing	At least one and up to all alternatives, aspiration level and value of each consequence	Sequentially evaluate each alternative. Stop if each consequence of an alternative equals or exceeds the aspiration level.	2.1.4
Minimax Cost	All alternatives, value of each consequence.	Compare the worst consequence values of each alternative	2.1.5
Minimax Regret	All alternatives, regret for each consequence.	Compare largest regrets of each alternative	2.1.5
EV	All alternatives, probability and value of each consequence.	Weight value of each consequence by its probability, for each alternative.	2.1.6
Laplace	All alternatives, value or utility of each consequence.	Weight value or utility of each consequence equally, for each alternative.	2.1.7
SEU	All alternatives, probability and utility of each consequence.	Weight utility of each consequence by its probability, for each alternative.	2.1.7
MAUT	All alternatives, value or utility of each consequence, priorities.	Weight value or utility of each consequence by priority, for each alternative.	2.1.8
Holistic	All alternatives and consequences.	Holistically compare the consequences of each alternative	2.1.9

decision maker has not yet selected a decision rule. Given that the decision maker has a decision rule, conflict can still occur if the needed inputs are not available. These sources of conflict and associated methods of conflict resolution are briefly addressed below in relation to the remainder of this chapter.

Identifying the current condition, alternative actions, and their consequences is an important part of decision making. This topic is emphasized in both naturalistic decision theory (Klein et al. 1993) and decision analysis* (Raiffa 1968; Clemen 1996). Decision trees, influence diagrams, and other tools for structuring decisions are covered in Section 3. Normative methods of identifying the current condition falling under the topic of inference (or diagnosis) are presented in Section 2.2. Section 4.1 describes several descriptive models of human inference and discusses their limitations. Section 6 includes discussion group decision-making methods that may be useful at this decision-making stage.

When decision makers are unsure how to compare alternatives, they must consider what information is available and then frame the decision appropriately. The way the decision is framed then determines (1) which decision rules are appropriate, (2) what information is needed to make the decision using the given rules (as discussed earlier in reference to Table 1), and (3) the choices selected. As discussed in Section 5.1, there is reason to believe that people apply different decision-making strategies in different decision contexts. Section 2.1 discusses the appropriateness of decision rules and how the particular rule used can impact choices. When the specific inputs needed by a decision rule are not available, the resulting conflict might be resolved by judging aspirations, importance, preference, or likelihood. It also might be resolved by choosing a different decision rule or strategy. As noted in Section 5.1, there is a prevalent tendency among decision makers in naturalistic settings to minimize analysis and its required cognitive effort. In group situations, conflict due to a lack of consensus between multiple decision makers might be resolved through debate, bargaining, or voting (Section 6).

---

*Clemen (1996) includes a chapter on creativity and decision structuring. Some practitioners claim that structuring the decision is the greatest contribution of the decision analysis process.

2. CLASSICAL DECISION THEORY

Classical decision theory began with the development of normative models in economics and statistics that specified optimal decisions (von Neumann and Morgenstern 1947; Savage 1954). Classical decision theory focuses heavily on the notion of rationality (Winterfeldt and Edwards 1986; Savage 1954). Emphasis is placed on the quality of the process followed when making a decision rather than on the ultimate outcome. Accordingly, a rational decision maker must think logically about the decision. To do this, the decision maker must first formally describe what is known about the decision. The decision is then made by applying principles of logic and Bayesian probability theory (Savage 1954). This approach is therefore quantitative, and also normative or prescriptive if the numerical inputs needed are available.

The classical approach has been applied to two related problems: (1) preference and choice, and (2) statistical inference.

2.1. Choice Procedures

Classical decision theory represents preference and choice problems in terms of four basic elements: (1) a set of potential actions (A_i) to choose between, (2) a set of events or world states (E_j), (3) a set of consequences (C_{ij}) obtained for each combination of action and event, and (4) a set of probabilities (P_{ij}) for each combination of action and event. For example, a decision maker might be deciding whether to wear a seatbelt when traveling in an automobile. Wearing or not wearing a seat belt corresponds to two actions A_1 and A_2. The expected consequence of either action depends upon whether an accident occurs. Having or not having an accident corresponds to two events E_1 and E_2. Wearing a seatbelt reduces the expected consequences C_{11} having an accident (E_1). As the probability of having an accident increases, use of a belt should therefore become more attractive.

Once a decision has been represented in terms of these basic elements, the choice is then made by applying decision rules. Numerous decision rules have been developed. Decision rules are based upon basic axioms (or what are felt to be self-evident assumptions) of rational choice. Not all rules, however, make use of the same axioms. Different rules make different assumptions and can provide different preference orderings for the same basic decision. The following discussion will first present some of the most basic axioms. Then several well-known decision rules will be briefly covered.

2.1.1. Axioms of Rational Choice

Numerous axioms have been proposed that are essential either for a particular model of choice or for the method of eliciting numbers used for a particular model (Winterfeldt and Edwards 1986). The best-known set of axioms (Table 2) establishes the normative principle of subjective expected utility (SEU) as a basis for making decisions (see Savage 1954; Luce and Raiffa 1957 for a more rigorous description of the axioms). On an individual basis, these axioms are intuitively appealing (Stukey and Zeckhauser 1978), but, as discussed in Section 4, people's preferences can deviate significantly from the SEU model in ways that conflict with certain axioms. Consequently, there has been a movement toward developing other, less restrictive standards of normative decision making (Frisch and Clemen 1994; Zey 1992).

Frisch and Clemen propose that "a good decision should (a) be based on the relevant consequences of the different options (*consequentialism*), (b) be based on an accurate assessment of the world and a consideration of all relevant consequences (*thorough structuring)*, and (c) make tradeoffs of some form (*compensatory decision rule*)." Consequentialism and the need for thorough structuring are both assumed by all normative decision rules. Most normative rules are also compensatory. However, when people make routine habitual decisions, they often don't consider the consequences of their choices, as discussed in Section 5. Also, because of cognitive limitations and the difficulty of obtaining information, it becomes unrealistic in many settings for the decision maker to consider all the options and possible consequences. To make a decision under such conditions, decision makers may limit the scope of the analysis by applying principles such as satisficing and other noncompensatory decision rules discussed below. They also may apply heuristics, based on their knowledge or experience, leading to performance that can approximate the results of applying compensatory decision rules (Section 4).

TABLE 2 Basic Axioms of Subjective Expected Utility Theory

A. Ordering/Quantification of Preference
Preferences of decision makers between alternatives can be quantified and ordered using the relations: $>$, where $A > B$ means that A is preferred to B $=$, where $A = B$ means that A and B are equivalent $\geq$, where $A \geq B$ means that B is not preferred to A
B. Transitivity of Preference
if $A_1 \geq A_2$ and $A_2 \geq A_3$, then $A_1 \geq A_3$
C. Quantification of Judgment
The relative likelihood of each possible consequence that might result from an alternative action can be specified.
D. Comparison of Alternatives
If two alternatives yield the same consequences, the alternative yielding the greater chance of the preferred consequence is preferred.
E. Substitution
If $A_1 > A_2 > A_3$, then the decision maker will be willing to accept a gamble $[p(A_1)$ and $(1 - p)(A_3)]$ as a substitute for A_2 for some value of $p \geq 0$.
F. Sure Thing Principle
If $A_1 \geq A_2$, then for all p, the gamble $[p(A_1)$ and $(1 - p)(A_3)] \geq [p(A_2)$ and $(1 - p)(A_3)]$.

2.1.2. Dominance

Dominance is perhaps the most fundamental normative decision rule. Dominance is said to occur between two alternative actions A_i and A_j when A_i is at least as good as A_j for all events E, and for at least one event E_k, A_i is preferred to A_j. For example, one investment might yield a better return than another regardless of whether the stock market goes up or down. Dominance can also be described for the case where the consequences are multidimensional. This occurs when for all events E, the kth consequence associated with action i (C_{ik}) and action j (C_{jk}), satisfies the relation $C_{ik} \geq C_{jk}$ for all k, and for at least one consequence $C_{ik} > C_{jk}$. For example, a physician choosing between alternative treatments has an easy decision if one treatment is *both* cheaper and more effective for all patients.

Dominance is obviously a normative decision rule, since a dominated alternative can never be better than the alternative that dominates it. Dominance is also conceptually simple, but it can be difficult to detect when there are many alternatives to consider or many possible consequences. The use of tests for dominance by decision makers in naturalistic settings in discussed further in Section 5.1.5.

2.1.3. Lexicographic Ordering and EBA

The lexicographic ordering principle (see Fishburn 1974) considers the case where alternatives have multiple consequences. For example, a purchasing decision might be based on both the cost and performance of the considered product. The different consequences are first ordered in terms of their importance. Returning to the above example, performance might be considered more important than cost. The decision maker then sequentially compares each alternative beginning with the most important consequence. If an alternative is found that is better than the others on the first consequence, it is immediately selected. If no alternative is best on the first dimension, the alternatives are compared for the next-most important consequence. This process continues until an alternative is selected or all the consequences have been considered without making a choice. The latter situation can happen only if the alternatives have the same consequences.

The elimination by aspects (EBA) rule (Tversky 1972) is similar to the lexicographic decision rule. It differs in that the consequences used to compare the alternatives are selected in random order, where the probability of selecting a consequence dimension is proportional to its importance. Both EBA and lexicographic ordering are noncompensatory decision rules, since the decision is made using a single consequence dimension. Returning to the above example, the lexicographic principle would result in selecting a product with slightly better performance, even if it costs much more. EBA would select either product depending on which of the consequences was first selected.

2.1.4. *Minimum Aspiration Level and Satisficing*

The minimum aspiration level or satisficing decision rule assumes that the decision maker sequentially screens the alternative actions until an action is found which is good enough. For example, a person considering the purchase of a car might stop looking once he or she found an attractive deal instead of comparing every model on the market. More formally, the comparison of alternatives stops once a choice is found that exceeds a minimum aspiration level S_{ik} for each of its consequences C_{ik} over the possible events E_k.

Satisficing can be a normative decision rule when (1) the expected benefit of exceeding the aspiration level is small, (2) the cost of evaluating alternatives is high, or (3) the cost of finding new alternatives is high. More often, however, it is viewed as an alternative to maximizing decision rules. From this view, people cope with incomplete or uncertain information and their limited rationality by satisficing in many settings instead of optimizing (Simon 1955, 1983).

2.1.5. *Minimax (Cost and Regret) and the Value of Information*

Minimax cost selects the best alternative (A_i) by first identifying the worst possible outcome for each alternative. The worst outcomes are then compared between alternatives. The alternative with the minimum worst-case cost is selected. Formally, the preferred action A_i is the action for which over the events k, $\text{MAX}_k(C_{ik}) = \text{MIN}_i[\text{MAX}_k(C_{ik})]$. For example, in Table 3, the maximum cost is 5 for alternative A_1, 7 for A_2, and 8 for A_3. A_1 would be chosen since it has the smallest maximum cost. Minimax cost corresponds to assuming the worst and therefore makes sense as a strategy where an adverse opponent is able to control the events (von Neumann and Morgenstern 1947). Along these lines, an airline executive considering whether to reduce fares might assume that a competitor will also cut prices, leading to a no-win situation.

Minimax regret involves a similar process, but the calculations are performed using regret instead of cost (Savage 1954). Regret is calculated by first identifying which alternative is best for each possible event. The regret R_{ik}, associated with each consequence (C_{ik}) for the combination of event E_k and alternative A_i then becomes: $R_{ik} = \text{MAX}_i(C_{ik}) - C_{ik}$. Returning to our earlier example, if E_1 occurs, alternative A_2 with a cost of 2 is best, resulting in a regret of 0 ($2 - 2$). A_1 has a cost of 5, resulting in a regret of 3 ($5 - 2$). A_3 has a cost of 6, resulting in a regret of 4 ($6 - 2$). These calculations are repeated for events E_2 and E_3, resulting in regret values for each combination of events and alternative actions. The preferred action A_i is the action for which over the events k, $\text{MAX}_k(R_{ik}) = \text{MIN}_i[\text{MAX}_k(R_{ik})]$. The maximum regrets for A_1 (a value of 3) and A_3 (a value of 4) are both found when event E_1 occurs. The maximum regret for A_2 (a value of 2) is found when event E_2 occurs. Alternative A_2 is then selected because it has the minimum maximum regret.

Note that the minimax cost and minimax regret principles do not always suggest the same choice (Table 3). Minimax cost is easily interpreted as a conservative strategy. Minimax regret is more difficult to judge from an objective or normative perspective (Savage 1954). As shown by the example, minimax regret can be less conservative than minimax cost. Alternatives that were not chosen can also impact choices made using minimax regret. For example, if alternative A_3 is removed from consideration, minimax regret and minimax cost will both select A_1. The interesting conclusion is that comparative and absolute measures of preference can result in different choices.

Bell (1982) argues persuasively that regret plays a very prominent role in decision making under uncertainty. For example, the purchaser of a new car might be happy, until finding out that a neighbor got the same car for $200 less from a different dealer. It is interesting to observe that regret is closely related to the value of information. This follows, since with hindsight, decision makers may regret their choice if they did not select the alternative giving the best result for the event (E_k) which actually took place. With perfect information, the decision maker would have chosen E_k. Conse-

TABLE 3 Example Comparison of Minimax Cost and Minimax Regret. Minimax cost selects A_1 and minimax regret selects A_2

	E_1	E_2	E_3	Max Cost	Max Regret
A_1	5	5	5	**5**	3
A_2	2	7	2	7	**2**
A_3	6	8	4	8	4

quently, the regret (R_{ik}) associated with having chosen alternative (A_i) is a measure of the value of having perfect information, or of knowing ahead of time that event E_k would occur. When each of the events (E_k) occur with probability P_k, it becomes possible to calculate the expected value of perfect information [EVPI(A_i)], given that the decision maker would chose action A_i before receiving this information with the following expression:

$$\text{EVPI}(A_i) = \sum_k P_k R_{ik} \tag{1}$$

The above approach can be extended to the case of imperfect information (Raiffa 1968) by replacing P_k in the above equation with the probability of event k (E_k) given the imperfect sample information (I). This results in an expression for the expected value of sample information [EVSI(A_i, I)], given that the decision maker would chose action A_i before receiving this information:

$$\text{EVSI}(A_i,I) = \sum_k (P_k|I) R_{ik} \tag{2}$$

The value of imperfect (or sample) information provides a normative rule for deciding whether to collect additional information. For example, a decision to perform a survey before introducing a product can be made by comparing the cost of the survey to the expected value of the information obtained. It is often assumed that decision makers are biased when they fail to seek out additional information. The above discussion shows that *not* obtaining information is justified when the information costs too much. From a practical perspective, the value of information can guide decisions to provide information to product users (Lehto and Papastavrou 1991).

2.1.6. Maximizing Expected Value

From elementary probability theory, return is maximized by selecting the alternative with the greatest expected value. The expected value of an action A_i is calculated by weighting its consequences C_{ik} over all events k, by the probability P_{ik} the event will occur. The expected value of a given action A_i is therefore:

$$\text{EV}[A_i] = \sum_k P_{ik} C_{ik} \tag{3}$$

More generally, the decision maker's preference for a given consequence C_{ik} might be defined by a value function $V(C_{ik})$, which transforms consequences into preference values. The preference values are then weighted using the same equation. The expected value of a given action A_i becomes:

$$\text{EV}[A_i] = \sum_k P_{ik} V(C_{ik}) \tag{4}$$

Monetary value is a common value function. For example, lives lost, units sold, or air quality might all be converted into monetary values. More generally, however, value reflects preference, as illustrated by ordinary concepts such as the value of money or the attractiveness of a work setting. Given that the decision maker has large resources and is given repeated opportunities to make the choice, choices made on the basis of expected monetary value are intuitively justifiable. A large company might make nearly all of its decisions on the basis of expected monetary value. Insurance buying and many other rational forms of behavior can not, however, be justified on the basis of expected monetary value. Many years ago, it was already recognized that rational decision makers made choices not easily explained by expected monetary value (Bernoulli 1738). Bernoulli cited the St. Petersburg paradox, in which the prize received in a lottery was 2^n and n was the number of times (n) a flipped coin turned up heads before a tails was observed. The probability of n flips before the first tail is observed is 0.5^n. The expected value of this lottery becomes:

$$\text{EV}[L] = \sum_k P_{ik} V(C_{ik}) = \sum_{n=0}^{\infty} 0.5^n 2^n = \sum_{n=1}^{\infty} 1 \Rightarrow \infty \tag{5}$$

The interesting twist is that the expected value of the above lottery is infinite. Bernoulli's conclusion was that preference cannot be a linear function of monetary value, since a rational decision maker would never pay more than a finite amount to play the lottery. Furthermore, the value of the lottery can vary between decision makers. According to utility theory, this variability reflects rational differences in preference between decision makers for uncertain consequences.

2.1.7. Subjective Expected Utility (SEU) Theory

Expected utility theory extended expected value theory to describe better how people make uncertain economic choices (von Neumann and Morgenstern 1947). In their approach, monetary values are first transformed into utilities, using a utility function $u(x)$. The utilities of each outcome are then weighted by their probability of occurrence to obtain an expected utility. Subjective utility theory (SEU) added the notion that uncertainty about outcomes could be represented with subjective probabilities (Savage 1954). It was postulated that these subjective estimates could be combined with evidence using Bayes' rule to infer the probabilities of outcomes* (see Section 2.2). This group of assumptions corresponds to the Bayesian approach to statistics. Following this approach, the SEU of an alternative (A_i), given subjective probabilities (S_{ik}) and consequences (C_{ik}) over the events E_k, becomes:

$$\mathrm{SEU}[A_i] = \sum_k S_{ik} U(C_{ik}) \tag{6}$$

Note the similarity between the above formulation for SEU and the earlier equation for expected value. EV and SEU are equivalent if the value function equals the utility function. Methods for eliciting value and utility functions differ in nature (Section 3). Preferences elicited for uncertain outcomes measure utility.† Preferences elicited for certain outcomes measure value. It accordingly has often been assumed that value functions differ from utility functions, but there are reasons to treat value and utility functions as equivalent (Winterfeldt and Edwards 1986). The latter authors claim that the differences between elicited value and utility functions are small and that "severe limitations constrain those relationships, and only a few possibilities exist, one of which is that they are the same."

When people are presented choices that have uncertain outcomes, they react in different ways. In some situations, people find gambling to be pleasurable. In others, people will pay money to reduce uncertainty; for example, when people buy insurance. SEU theory distinguishes between risk neutral, risk averse, risk seeking, and mixed forms of behavior. These different types of behavior are described by the shape of the utility function (Figure 2).

A risk-neutral decision maker will find the expected utility of a gamble to be the same as the utility of the gamble's expected value. That is, expected u(gamble) = u(gamble's expected value). For a risk-averse decision maker, expected u(gamble) < u(gamble's expected value); for a risk-

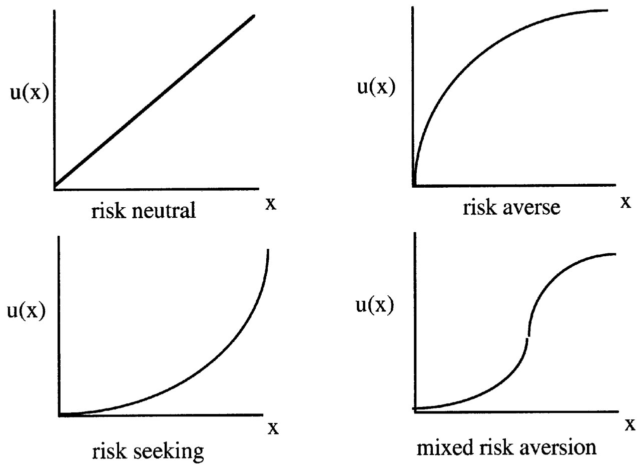


Figure 2 Utility Functions for Differing Risk Attitudes.

---

* When no evidence is available concerning the likelihood of different events, it was postulated that each consequence should be assumed to be equally likely. The Laplace decision rule makes this assumption and then compares alternatives on the basis of expected value or utility.

† Note that classical utility theory assumes that utilities are constant. Utilities may, of course, fluctuate. The random utility model (Bock and Jones 1968) allows such fluctuation.

seeking decision maker, expected u(gamble) $>$ u(gamble's expected value). On any given point of a utility function, attitudes towards risk are described formally by the coefficient of risk aversion:

$$C_{RA} = \frac{u''(x)}{u'(x)} \tag{7}$$

where $u'(x)$ and $u''(x)$ are respectively the first and second derivatives of $u(x)$ taken with respect to x. Note that when $u(x)$ is a linear function of x, that is, $u(x) = ax + b$, then $C_{RA} = 0$. For any point of the utility function, if $C_{RA} < 0$, the utility function depicts risk-averse behavior, and if $C_{RA} > 0$, the utility function depicts risk-seeking behavior. The coefficient of risk aversion therefore describes attitudes toward risk at each point of the utility function, given that the utility function is continuous. SEU theory consequently provides a powerful tool for describing how people might react to uncertain or risky outcomes. However, some commonly observed preferences between risky alternatives can not be explained by SEU. Section 4.2 focuses on experimental findings showing deviations from the predictions of SEU.

A major contribution of SEU is that it represents differing attitudes towards risk and provides a normative model of decision making under uncertainty. The prescriptions of SEU are also clear and testable. Consequently, SEU has played a major role in fields other than economics, both as a tool for improving human decision making and as a stepping stone for developing models that describe how people make decisions when outcomes are uncertain. As discussed further in Section 4, much of this work has been done in psychology.

2.1.8. Multiattribute Utility Theory

Multiattribute tility theory (Keeney and Raiffa 1976) extends SEU to the case where the decision maker has multiple objectives. The approach is equally applicable for describing utility and value functions. Following this approach, the utility (or value) of an alternative A, with multiple attributes x, is described with the multiattribute utility (or value) function $u(x_1 \ldots x_n)$, where $u(x_1 \ldots x_n)$ is some function $f(x_1 \ldots x_n)$ of the attributes x. In the simplest case, multiattribute utility theory (MAUT) describes the utility of an alternative as an additive function of the single attribute utility functions $u_n(x_n)$. That is,

$$u(x_1 \ldots x_n) = \sum_n k_n u_n(x_n) \tag{8}$$

where the constants k_n are used to weight each single attribute utility function (u_n) in terms of its importance. Assuming an alternative has three attributes, x, y, and z, an additive utility function is $u(x,y,z) = k_x u_x(x) + k_y u_y(y) + k_z u_z(z)$. Along these lines, a community considering building a bridge across a river vs. building a tunnel or continuing to use the existing ferry system might consider the attractiveness of each option in terms of the attributes of economic benefits, social benefits, and environmental benefits.*

More complex multiattribute utility functions, include multiplicative forms and functions that combine utility functions for subsets of two or more attributes (Keeney and Raiffa 1976). An example of a simple multiplicative function would be $u(x,y) = u_x(x)*u_y(y)$. A function that combines utility functions for subsets, would be $u(x,y,z) = k_{xy}u_{xy}(x,y) + k_z u_z(z)$. This latter type of function becomes useful when utility independence is violated. Utility independence is violated when the utility function for one attribute depends on the value of another attribute. Along these lines, when assessing $u_{xy}(x,y)$, it might be found that $u_x(x)$ depends on the value of y. For example, peoples' reaction to the level of crime in their own neighborhood might depend on the level of crime in a nearby suburb. In the latter case, it is probably better to directly measure $u_{xy}(x =$ crime in own neighborhood, $y =$ crime in nearby suburb) than to estimate it from the single-attribute functions. The assessment of utility and value functions is discussed later in Section 3.

MAUT has been applied to a wide variety of problems (Saaty 1988; Keeney and Raiffa 1976; Winterfeldt and Edwards 1986; Clemen 1996). An advantage of MAUT is that it helps structure complex decisions in a meaningful way. Alternative choices and their attributes often naturally divide into hierarchies. The MAUT approach encourages such divide-and-conquer strategies and, especially in its additive form, provides a straightforward means of recombining weights into a final ranking of alternatives. The MAUT approach is also a compensatory strategy that allows normative trade-offs between attributes in terms of their importance.

---

* To develop the multiattribute utility function, the single-attribute utility functions (u_n) and the importance weights (k_n) are determined by assessing preferences between alternatives. Methods of doing so are discussed in Section 3.4.

2.1.9. *Holistic Comparison*

Holistic comparison is a nonanalytical method of comparing alternatives. This process involves a holistic comparison of the consequences for each alternative instead of separately measuring and then recombining measures of probability, value, or utility (Sage 1981; Stanoulov 1994; Janis and Mann 1977). A preference ordering between alternatives is thus obtained. For example, the decision maker might rank in order of preference a set of automobiles that vary on objectively measureable attributes, such as color, size, and price. Mathematical tools can then be used to derive the relationship between observed ordering and attribute values and ultimately predict preferences for unevaluated alternatives, as discussed in Section 3.3.4.

One advantage of holistic comparison is that it requires no formal consideration of probability or utility. Consequently, decision makers unfamiliar with these concepts may find holistic comparison to be more intuitive, and potential violations of the axioms underlying SEU and MAUT, due to their lack of understanding, become of lesser concern. People seem to find the holistic approach helpful when they compare complex alternatives (Janis and Mann 1977). In fact, people may feel there is little additional benefit to be obtained from separately analyzing the probability and value attached to each attribute. This tendency becomes prevalent in naturalistic decision making, as addressed further in Section 5.

2.2. Statistical Inference

Inference is the procedure followed when a decision maker uses information to determine whether a hypothesis about the world is true. Hypotheses can specify past, present, or future states of the world, or causal relationships between variables. Diagnosis is concerned with determining past and present states of the world. Prediction is concerned with determining future states. Inference or diagnosis is required in many decision contexts. For example, before deciding on a treatment, a physician must first diagnose the illness.

From the classical perspective, the decision maker is concerned with determining the likelihood that a hypothesis (H_i) is true. Bayesian inference is the best-known technique, but signal detection theory, and fundamentally different approaches such as the Dempster–Schafer method, have seen application. Each of these approaches is discussed below.

2.2.1. *Bayesian Inference*

Bayesian inference is a well-defined procedure for inferring the probability (P_i) that a hypothesis (H_i) is true, from evidence (E_j) linking the hypothesis to other observed states of the world. The approach makes use of Bayes' rule to combine the various sources of evidence (Savage 1954). Bayes' rule states that the posterior probability of hypothesis H_i given that evidence E_j is present, or $P(H_i|E_j)$, is given by the equation:

$$P(H_i|E_j) = \frac{P(E_j|H_i)P(H_i)}{P(E_j)} \tag{9}$$

where $P(H_i)$ is the probability of the hypothesis being true prior to obtaining the evidence E_j and $P(E_j|H_i)$ is the probability of obtaining the evidence E_j given that the hypothesis H_i is true. For example, consider the case where a physician is attempting to determine whether a patient has a disease present in 10% of the general population. The physician has a test available that gives a positive result 90% of the time when administered to patients who actually have the disease. The test also gives a positive result 20% of the time when administered to patients who don't have the disease. If the test were to be administered to a member of the general population, Eq. (9) predicts that the probability of having the disease given a positive test result is:

$$P(\text{disease}|\text{positive test}) = \frac{P(\text{positive test}|\text{disease})P(\text{disease in general population})}{P(\text{positive test})}$$

Also,

$$\begin{aligned} P(\text{positive test}) = {} & P(\text{positive test}|\text{disease})P(\text{disease in general population}) \\ & + P(\text{positive test}|\text{no disease})P(\text{no disease in general population}) \end{aligned}$$

$$P(\text{disease}|\text{positive test}) = \frac{0.9*0.1}{0.9*0.1 + 0.2*0.9} = 0.33$$

As discussed further in Section 4.1, people often fail to combine evidence consistently with the above predictions of Bayes' rule. A common finding is that people fail to adequately consider the base rate of the hypothesis. In the above example, this would correspond to focusing on *P*(positive test|disease) = 0.9 and not considering P(disease in general population) = 0.1. As a consequence, many people might be surprised that P(disease|positive test) = 0.33 rather than a number close to 0.9.

When the evidence E_j consists of multiple states $E_1, \ldots, E_n$, each of which is conditionally independent, Bayes' rule can be expanded into the expression:

$$P(H_i|E_j) = \frac{\prod_{j=1}^{n} P(E_j|H_i)P(H_i)}{P(E_j)} \tag{10}$$

Calculating $P(E_j)$ can be somewhat difficult, due to the fact that each piece of evidence must be dependent,* or else it would not be related to the hypothesis. The odds forms of Bayes' rule provides a convenient way of looking at the evidence for and against a hypothesis that doesn't require $P(E_j)$ to be calculated. This results in the expression:

$$\Phi(H_i|E_j) = \frac{P(H_i|E_j)}{P(\sim H_i|E_j)} = \frac{\prod_{j=1}^{n} P(E_j|H_i)P(H_i)}{\prod_{j=1}^{n} P(E_j)|\sim H_i)P(\sim H_i)} \tag{11}$$

where $\Phi(H_i|E_j)$ refers to the posterior odds for hypothesis H_i, $P(\sim H_i)$ is the prior probability that hypothesis H_i is not true, and $P(\sim H_i|E_j)$ is the posterior probability that hypothesis H_i is not true.

The two latter forms of Bayes' rule provide an analytically simple way of combining multiple sources of evidence. Bayesian inference becomes much more difficult when the evidence is not certain or when the conditional independence assumption is not met. When evidence is not certain, complex multistage forms of Bayesian analysis are required that consider the probability of the evidence being true (Winterfeldt and Edwards 1986). When conditional independence is not true, the expanded form of Bayes' rule must be modified. For example, consider the case where the evidence consists of three events (E_1, E_2, E_3), where E_1 and E_2 are conditionally dependent and E_3 is conditionally independent of the two other events. The posterior probability, $P(H_i|E_1,E_2,E_3)$, then becomes:

$$P(H_i|E_1,E_2,E_3) = \frac{P(E_1,E_2|H_i)P(E_3|H_i)P(H_i)}{P(E_1,E_2)P(E_3|E_1,E_2)} \tag{12}$$

where $P(E_1,E_2|H_i)$ is the conditional probability of obtaining E_1 and E_2 given the hypothesis H_i, $P(E_3|H_i)$ is the conditional probability of obtaining E_3 given H_i, and $P(E_1,E_2)P(E_3|E_1,E_2)$ is the probability of obtaining the evidence (E_1, E_2, E_3).

2.2.2. Signal-Detection Theory

Bayesian inference combined with SEU leads to signal-detection theory (Tanner and Swets 1954), which has been applied in a large variety of contexts to model human performance (Wickens 1992). In signal-detection theory, the human operator is assumed to use Bayes' rule to estimate the probability that a signal actually is present from a noisy observation of the system. For example, an operator might estimate the probability a machine is going out of tolerance from a warning signal. The responses of the operator and the true state of the system together determine a set of four outcomes (Table 4).

TABLE 4 Potential Outcomes Considered by Signal Detection Theory

		State of the World	
		Noise (N)	Signal (S)
Response	yes	false alarm (fa)	hit (h)
	no	correct rejection (cr)	miss (m)

---

*Note that conditional independence between E_1 and E_2 implies that $P(E_1/H_i,E_2) = P(E_1/H_i)$ and that $P(E_2/H_i,E_1) = P(E_2/H_i)$. This is very different from simple independence, which implies that $P(E_1) = P(E_1/E_2)$ and that $P(E_2) = P(E_2/E_1)$.

The signal-detection model assumes an operator receives evidence from the environment regarding the true state of the world. The relationship between the signal (S) and the evidence (E) is measured by the conditional probability [$P(E|S)$] of obtaining the observed evidence given the signal is there. The decision maker is assumed to select a criterion value (x_c) that the evidence must exceed before saying yes. It is assumed that the value chosen will maximize utility. If the evidence is represented with a variable x, the expected utility of the operator can be described in terms of x, x_c and the four outcomes in Table 4. The expected utility for a given probability cutoff x_c, and utility function u, is given by the expression:

$$\begin{aligned}\mathrm{SEU}[x_c] &= P(x \geq x_c|S)P(S)u(h) + P(x \geq x_c|N)P(N)u(fa) \\ &\quad + P(x < x_c|S)P(S)u(m) + P(x < x_c|N)P(N)u(cr)\end{aligned} \tag{13}$$

where h is a hit, fa is a false alarm, m is a miss, and cr is a correct rejection. The above expression can be maximized by first substituting $1 - P(x \geq x_c|N)$ for $P(x < x_c|N)$ and also substituting $1 - P(x \geq x_c|S)$ for $P(x < x_c|S)$ into the equation for $\mathrm{SEU}[x_c]$, and then setting the derivative of $\mathrm{SEU}[x_c]$ with respect to x_c to zero. The result at the cutoff x_c is shown below:

$$\frac{P(x = x_c|S)}{P(x = x_c|N)} = \beta^* \geq \frac{P(N)(u(cr) - u(fa))}{P(S)(u(h) - u(m))} \tag{14}$$

where β^* is the optimal value of β. Substituting back the relation between $P(E|S)$ and the evidence x, the optimal decision rule is to say yes if

$$\frac{P(E|S)}{P(E|N)} \geq \beta^* \tag{15}$$

Equation (15) can be extended to multiple operators or multiple sources of evidence (Lehto and Papastavrou 1991). The resulting expression takes into account the probability of a false alarm and the probability of detection for the other source of information. Lehto and Papastavrou use this approach to analyze situations where the other source of information is a warning signal. The extent to which human judgments correspond to the predictions of Bayes' rule is further discussed in Section 4.1.

2.2.3. Dempster–Schafer Method

The Dempster–Schafer method (Schafer 1976; Fedrizzi et al. 1994) is an alternative to Bayesian inference for accumulating evidence for or against a hypothesis that has been proposed for use in decision analysis (Strat 1994). In this approach, the relation of hypotheses (H) to evidence (e) is described by a basic probability assignment (bpa) function, p. Given evidence (e), this function $p_e(n)$ assigns a value between 0 and 1 to each subset of H, such that the sum of the values assigned is 1. For example, consider the case where there are three hypotheses (A,B,C). When no evidence is available, the vacuous bpa assigns a value of 1 to the set of hypotheses $H = (A,B,C)$ and a 0 to all subsets. That is, the subsets (A), (B), (C), (A,B), and (A,C) are each assigned a value of 0. The Bayesian approach would instead assign a probability of 0.33 to A, B, and C respectively.

Also, given that evidence $p_e(A) = x$ supporting a specific hypothesis A is found, the Dempster–Schafer approach assigns $(1 - p_e(A))$ to H. The Bayesian approach, of course, assigns $(1 - p_e(A))$ to the complement of A. Returning to the above example, suppose the evidence supports hypothesis A to the degree $p_e(A) = 0.6$. Using the Dempster–Schafer approach, $p_e(A,B,C) = 0.4$. This, of course, is very different from the Bayes' interpretation, where $P(A) = 0.6$ and $P(\text{not } A) = 0.4$. The Dempster–Schafer method uses a belief function $B(n)$ to assign a total belief to n, where n is a subset of the set of possible hypotheses (H), as the sum of the beliefs assigned to m, where m is the set of possible subsets of n. In the above example, the belief in (A,B,C) after receiving evidence (e) is as given below:

$$\begin{aligned}B(A,B,C) &= p_e(A,B,C) + p_e(A,B) + \mathrm{p}_e(A,C) + p_e(B,C) + p_e(A) + p_e(B) + p_e(C) \\ &= 0.4 + 0 + 0 + 0 + 0.6 + 0 + 0 \\ &= 1.0\end{aligned} \tag{16}$$

Similarly, the belief in (A,B) after receiving the evidence (e) is:

TABLE 5 Tableau for Dempster–Schafer Method of Combining Evidence

X = [(A), (A,B,C)]	**Y = [(A,B), (A,B,C)]** $p_f(A,B) = 0.4$	$p_f(A,B,C) = 0.6$
$p_e(A) = 0.6$	A ; $p_e p_f = 0.24$	A ; $p_e p_f = 0.36$
$p_e(A,B,C) = 0.4$	A,B ; $p_e p_f = 0.16$	A,B,C ; $p_e p_f = 0.24$

$$\begin{aligned} B(A,B) &= p_e(A,B) + p_e(A) + p_e(B) \\ &= 0 + 0.6 + 0 \\ &= 0.6 \\ &= B(A) \end{aligned} \tag{17}$$

To combine evidence from multiple sources e and f, Dempster–Schafer theory uses the combining function $c(p_e(X),p_f(Y))$, where X and Y are both sets of subsets of H. For example, we might have $X = [(A), (A,B,C)]$ and $Y = [(A,B), (A,B,C)]$. The combining function then assigns a value to each subset n of H. The value assigned is determined by first describing the set of subsets n' within n defined by the intersection of subsets within X and subsets within Y. A value of 0 is assigned to all subsets of n not within n'. The products $p_e(X)*p_f(Y)$ are then summed and assigned to each subset within n'. Returning to the above example, we can calculate $c(n')$ using the values given in Table 5. First note that the set of subsets n' for the example is defined by the inner elements of the table. Specifically, $n' = [(A), (A,B), (A,B,C)]$. The values used by the combining function $c(n')$ are also shown. Using these numbers, the values of $c(n')$ become: $c(A) = 0.24 + 0.36 = 0.6$; $c(A,B) = 0.16$; $c(A,B,C) = 0.24$. All remaining subsets for this evidence are assigned a value of 0.

It has been argued that the Dempster–Schafer method of assigning evidence is better suited for diagnosing medical problems than the Bayesian method (Gordon and Shortliffe 1984). The latter researchers particularly criticize the Bayesian assumption that evidence partially supporting a hypothesis should also support its negation. Gordon and Shortliffe note that the Dempster–Schafer method shows promise as a means of accumulating belief in expert diagnostic systems used in medicine.

3. DECISION ANALYSIS

The application of classical decision theory to improve human decision making is the goal of decision analysis (Raiffa 1968; Howard 1968; 1988; Keeney and Raiffa 1976). Decision analysis requires inputs from decision makers, such as goals, preference and importance measures, and subjective probabilities. Elicitation techniques have consequently been developed that help decision makers provide these inputs. Particular focus has been placed on methods of quantifying preferences, tradeoffs between conflicting objectives, and uncertainty (Keeney and Raiffa 1976; Raiffa 1968). As a first step in decision analysis, it is necessary to do some preliminary structuring of the decision, which then guides the elicitation process. The following discussion first presents methods of structuring decisions and then covers techniques for assessing subjective probabilities, utility functions, and preferences.

3.1. Structuring Decisions

The field of decision analysis has developed many useful frameworks for representing what is known about a decision (Howard 1968; Winterfelt and Edwards 1986; Clemen 1996). In fact, the above authors and others have stated that the process of structuring decisions is often the greatest contribution of going through the process of decision analysis. Among the many tools used, decision matrices and trees provide a convenient framework for comparing decisions on the basis of expected value or utility. Value trees provide a helpful method of structuring the sometimes complex relationships among objectives, attributes, goals, and values and are used extensively in multiattribute decision-making problems. Event trees, fault trees, inference trees, and influence diagrams are useful for describing probabilistic relationships between events and decisions. Each of these approaches is briefly discussed below.

3.1.1. Decision Matrices and Trees

Decision matrices are often used to represent single-stage decisions (Figure 3). The simplicity of decision matrices is their primary advantage. They also provide a very convenient format for applying

$$\begin{array}{c|cc|}
 & E_1 & E_2 \\ \hline
A_1 & C_{11} & C_{12} \\
A_2 & C_{21} & C_{22} \\ \hline
 & P & (1-P)
\end{array}$$

Figure 3 Decision Matrix Representation of a Single-Stage Decision.

the decision rules discussed in the previous section. Decision trees are also commonly used to represent single-stage decisions (Figure 4) and are particularly useful for describing multistage decisions (Raiffa 1968). Note that in a multistage decision tree, the probabilities of later events are conditioned on the result of earlier events. This leads to the important insight that the results of earlier events provide information regarding future events.* Following this approach, decisions may be stated in conditional form. An optimal decision, for example, might be to first do a market survey, and then market the product only if the survey to is positive.

Analysis of a single or multistage decision tree involves two basic steps referred to as averaging out and folding back (Raiffa 1968). These steps, respectively, occur at chance and decision nodes.† Averaging out occurs when the expected value (or utility) at each chance node is calculated. In Figure 4, this corresponds to calculating the expected value of A_1 and A_2, respectively. Folding back refers to choosing the action with the greatest expected value at each decision node.

Decision trees consequently provide a straightforward way of comparing alternatives in terms of expected value or SEU. However, their development requires significant simplification of most decisions and the provision of numbers, such as measures of preference and subjective probabilities, that decision makers may have difficulty determining. In certain contexts, decision makers struggling with this issue may find it helpful to develop value trees, event trees, or influence diagrams, as expanded upon below.

3.1.2. Value Trees

Value trees hierarchically organize objectives, attributes, goals, and values (Figure 5). From this perspective, an objective corresponds to satisficing or maximizing a goal or set of goals. When there is more than one goal, the decision maker will have multiple objectives, which may differ in importance. Objectives and goals are both measured on a set of attributes. Attributes may provide (1) objective measures of an goal, such as when fatalities and injuries are used as a measure of highway safety, (2) subjective measures of an goal, such as when people are asked to rate the quality of life in the suburbs vs. the city, or (3) proxy or indirect measures of a goal, such as when the quality of ambulance service is measured in terms of response time.

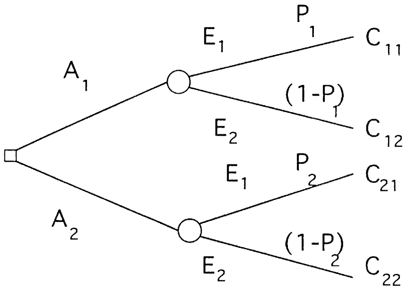


Figure 4 Decision Tree Representation of a Single-Stage Decision.

---

* For example, the first event in a decision tree might be the result of a test. The test result then provides information useful in making the final decision.

† Note that standard convention uses circles to denote chance nodes and squares to denote decision nodes (Raiffa 1968).

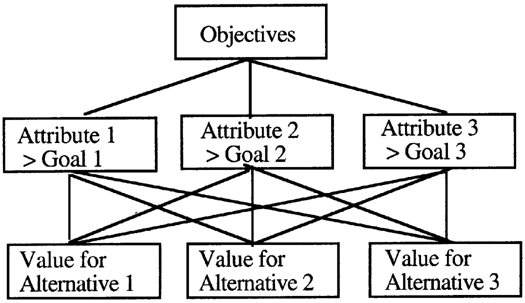

Objectives
Attribute 1 > Goal 1
Attribute 2 > Goal 2
Attribute 3 > Goal 3
Value for Alternative 1
Value for Alternative 2
Value for Alternative 3

Figure 5 Generic Value Tree.

In generating objectives and attributes, it becomes important to consider their relevance, completeness, and independence. Desirable properties of attributes (Keeney and Raiffa 1976) include:

1. *Completeness:* The extent to which the attributes measure whether an objective is met.
2. *Operationality:* The degree to which the attributes are meaningful and feasible to measure.
3. *Decomposability:* Whether the whole is described by its parts.
4. *Nonredundancy:* Correlated attributes give misleading results.
5. *Minimum size:* Considering irrelevant attributes is expensive and may be misleading.

Once a value tree has been generated, various methods can be used to assess preferences directly between the alternatives.

3.1.3. Event Trees or Networks

Event trees or networks show how a sequence of events can lead from primary events to one or more outcomes. Human reliability analysis (HRA) event trees are a classic example of this approach (Figure 6). If probabilities are attached to the primary events, it becomes possible to calculate the probability of outcomes, as illustrated in Section 3.2.4. This approach has been used in the field of risk assessment to estimate the reliability of human operators and other elements of complex systems (Gertman and Blackman 1994). Chapter 32 provides additional information on human reliability analysis and other methods of risk assessment.

Fault trees work backwards from a single undesired event to its causes (Figure 7). Fault trees are commonly used in risk assessment to help infer the chance of an accident occurring (Hammer 1993; Gertman and Blackman 1994). Inference trees relate a set of hypotheses at the top level of the tree to evidence depicted at the lower levels. This latter approach has been used by expert systems, such as PROSPECTOR (Duda et al. 1979). PROSPECTOR applies a Bayesian approach to infer the presence of a mineral deposit from uncertain evidence.

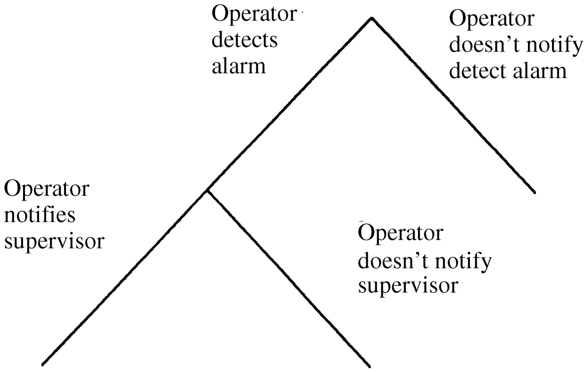

Operator detects alarm
Operator doesn't notify detect alarm
Operator notifies supervisor
Operator doesn't notify supervisor

Figure 6 HRA Event Tree. (Adapted from Gertman and Blackman 1994)

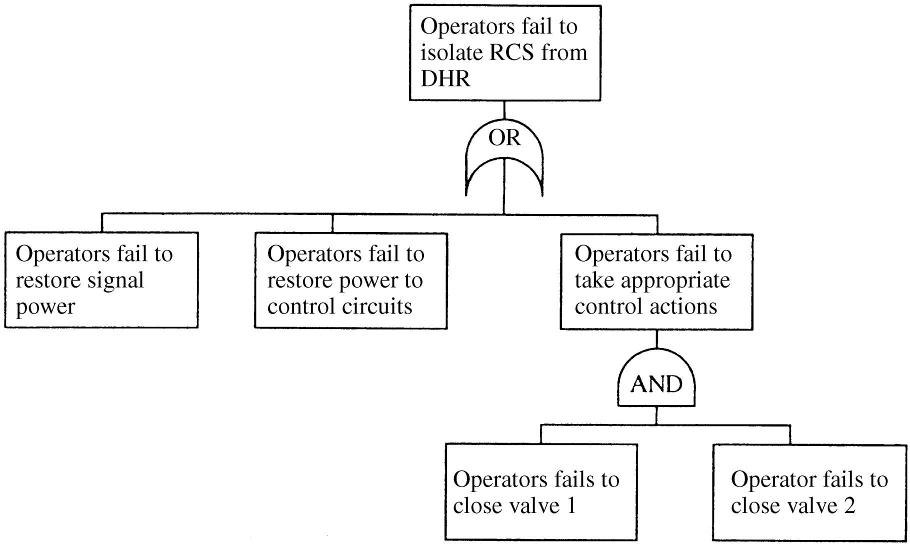

Operators fail to isolate RCS from DHR
OR
Operators fail to restore signal power
Operators fail to restore power to control circuits
Operators fail to take appropriate control actions
AND
Operators fails to close valve 1
Operator fails to close valve 2

Figure 7 Fault Tree for Operators. (Adapted from Gertman and Blackman 1994).

3.1.4. Influence Diagrams and Cognitive Mapping

Influence diagrams are often used in the early stages of a decision to show how events and actions are related. Their use in the early stages of a decision is referred to as knowledge (or cognitive) mapping (Howard 1988). Links in an inference diagram depict causal and temporal relations between events and decision stages.* A link leading from an event *A* to an event *B* implies that the probability of obtaining event *B* depends on whether event *A* has occurred. A link leading from a decision to a event implies that the probability of the event depends on the choice made at that decision stage. A link leading from an event to a decision implies that the decision maker knows the outcome of the event at the time the decision is made.

One advantage of influence diagrams in comparison to decision trees is that influence diagrams show the relationships between events more explicitly. Consequently, influence diagrams are often used to represent complicated decisions where events interactively influence the outcomes. For example, the influence diagram in Figure 8 shows that the true state of the machine affects both the

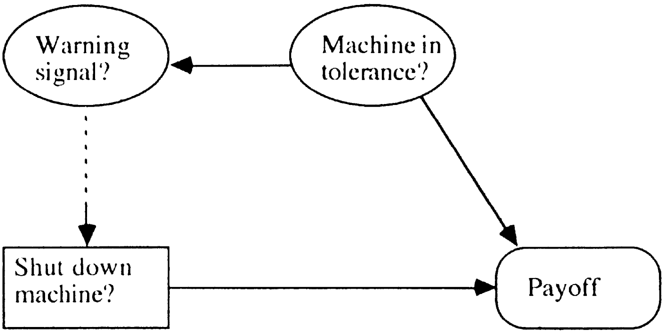

Warning signal?
Machine in tolerance?
Shut down machine?
Payoff

Figure 8 Influence Diagram Representation of a Single-Stage Decision.

---

*As for decision trees, the convention for influence diagrams is to depict events with circles and decisions with squares.

probability of the warning signal and the consequence of the operator's decision. This linkage would be hidden within a decision tree.* Influence diagrams have been used to structure medical decision-making problems (Holtzman 1989) and are emphasized in modern texts on decision analysis (Clemen 1996). Howard (1988) states that influence diagrams are the greatest advance he has seen in the communication, elicitation, and detailed representation of human knowledge. Part of the issue is that influence diagrams allow people who do not have deep knowledge of probability to describe complex conditional relationships with simple linkages between events. Once these linkages are defined, the decision becomes well defined and can be formally analyzed.

3.2. Probability Assessment

Several approaches have been used in decision analysis to assess subjective probabilities. In this section several of the more well-known techniques will be summarized. These techniques include: (1) direct numerical assessment, (2) fitting subjective belief forms, (3) the bisection method, (4) conditioning arguments, (5) preferences between reference gambles, and (6) scaling methods. Techniques proposed for improving the accuracy of assessed probabilities, including scoring rules, calibration, and group assessment, will then be presented.

3.2.1. Direct Numerical Assessment

In direct numerical estimation, decision makers are asked to give a numerical estimate of how likely they think the event is to happen. These estimates can be probabilities, odds, log odds, or words (Winterfeldt and Edwards 1986). Winterfeldt and Edwards argue that log odds have certain advantages over the other measures. Gertman and Blackman (1994) note that log odds are normally used in risk assessment for nuclear power applications because human error probabilities (HEPs) vary greatly in value. HEPs between 1 and 0.00001 are typical.

3.2.2. Fitting a Subjective Belief Form

Fitting a subjective belief form requires that the questions be posed in terms of statistical parameters. That is, decision makers could be asked to first consider their uncertainty regarding the true value of a given probability and then estimate their mean, mode, or median belief. This approach can be further extended by asking decision makers to describe how certain they are of their estimate. For example, a worker might subjectively estimate the mean and variance of the proportion of defective circuit boards before inspecting a small sample of circuit boards. If the best estimate corresponds to a mean, mode, or median, and the estimate of certainty to a confidence interval or standard deviation, a functional form such as the Beta-1 probability density function (pdf) can then be used to fit a subjective probability distribution (Clemen 1996; Buck 1989).

In other words, a distribution is specified that describes the subject's belief that the true probability equals particular values. This type of distribution can be said to express uncertainty about uncertainty (Raiffa 1968). Given that the subject's belief can be described with a Beta-1 pdf, Bayesian methods can be used to combine binomially distributed evidence easily with the subject's prior belief (Clemen 1996; Buck 1989). Returning to the above example, the worker's prior subjective belief can be combined with the results of inspecting the small sample of circuit boards, using Bayes' rule. As more evidence is collected, the weight given to the subject's initial belief becomes smaller compared to the evidence collected. The use of prior belief forms also reduces the amount of sample information that must be collected to show that a proportion, such as the percentage of defective items, has changed (Buck 1989).

3.2.3. Bisection Method

The bisection method (Raiffa 1968) is another direct technique for attempting to estimate a subjective probability density function (pdf). This technique is somewhat more general than fitting the subject's belief with a functional form, such as the beta-1, since it makes no parametric assumptions. The bisection method involves two steps which are repeated until the subject's belief is adequately described. Following this approach, the first step is to determine the median ($P_{0.5}$) of the subjective pdf. This question is posed to the decision maker in a form such as "For what value of p do you feel it is equally likely the true value p† is greater than or less than p?" This step is then repeated for subintervals to obtain the desired level of detail.

---

* The conditional probabilities in a decision tree would reflect this linkage, but the structure of the tree itself does not show the linkage directly. Also, the decision tree would use the flipped probability tree using P(warning) at the first stage and P(machine down|warning) at the second stage. It seems more natural for operators to think about the problem in terms of P(machine down) and P(warning|machine down), which is the way the influence diagram in Figure 7 depicts the relationship.

† Note that it has been shown that people viewing fault trees can be insensitive to missing information (Fischhoff et al. 1978).

3.2.4. *Conditioning Arguments*

Statistical conditioning arguments are based on the idea that the probability of a complicated event, such as the chance of having an accident, can be determined by estimating the probability of simpler events (or subsets). From a more formal perspective, a conditioning argument determines the probability of an event A by considering the possible conditions (C_i) under which A might happen, the associated conditional probabilities [$P(A|C_i)$], and the probability of each condition [$P(C_i)$]. The probability of A can then be represented as:

$$P(A) = \sum_i P(A|C_i)P(C_i) \tag{18}$$

This approach is illustrated by the development of event trees and fault tree analysis. In fault tree analysis, the probability of an accident is estimated by considering the probability of human errors, component failures, and other events. This approach has been extensively applied in the field of risk analysis (Gertman and Blackman 1994).* THERP (Swain and Guttman 1983) extends the conditioning approach to the evaluation of human reliability in complex systems.

SLIM-MAUD (Embrey 1984) implements a related approach in which expert ratings are used to estimate human error probabilities (HEPs) in various environments. The experts first rate a set of tasks in terms of performance-shaping factors (PSFs) that are present. Tasks with known HEPs are used as upper and lower anchor values. The experts also judge the importance of individual PSFs. A subjective likelihood index (SLI) is then calculated for each task in terms of the PSFs. A logarithmic relationship is assumed between the HEP and SLI, allowing calculation of the human error probability for task j (HEP_j) from the subjective likelihood index assigned to task j (SLI_j). More specifically:

$$\mathrm{Log}(1 - \mathrm{HEP}_j) = a\mathrm{SLI}_j + b \tag{19}$$

$$\text{where } \mathrm{SLI}_j = \sum_i \mathrm{PSF}_{ij} * I(\mathrm{PSF}_i) \tag{20}$$

$I(\mathrm{PSF}_i)$ is the importance of PSF_i, and PSF_{ij} is the rating given to PSF_i for task j. Gertman and Blackman (1994) provide guidelines regarding the use of this method and have generally positive conclusions. SLIM-MAUD is interesting in that it uses multiattribute utility theory as a basis for generating probability estimates.

3.2.5. *Reference Lotteries*

Reference lottery methods take a less direct approach to obtaining point estimates of the decision maker's subjective probabilities. When the objective is to measure how likely event A is to occur, the approach asks decision makers to consider a lottery where they will receive a prize x if event A occurs, and a prize y if it does not. They are then asked how much they would be willing to pay for the lottery. The amount they are willing to pay z is then equated to the lottery, using the relation $z = P(A)x + [1 - P(A)]y$. From this expression it becomes possible to estimate the decision maker's subjective estimate of $P(A)$. Specifically, $P(A) = (z - y)/(x - y)$. A variant of this approach that asks decision makers to compare two lotteries over the same range of preferences might be preferable because it removes the potential effect of risk aversion (Winterfelt and Edwards 1986).

3.2.6. *Scaling Methods*

Scaling methods ask subjects to rate or rank the probabilities to be assessed. Likert scales with verbal anchors have been used to obtain estimates of how likely people feel certain risks are (Kraus and Slovic 1988). Another approach has been to ask subjects to do pair-wise comparisons of the likelihoods of alternative events (Saaty 1988). Pairwise comparisons of probabilities on a ratio scale correspond to relative odds and consequently have high construct validity. In fact, much of the risk assessment focuses on determining order of magnitude differences in probability. Saaty (1988), however, argues that the psychometric literature indicates that people's ability to distinguish items on the same scale is limited to 7 ± 2 categories. He consequently proposes use of a relative scale to measure differences in importance, preference, and probability that uses verbal anchors corresponding to equal, weak, strong, very strong, and absolute differences between rated items. In perhaps the most controversial aspect of his approach, these five verbal anchors are assigned the numbers 1, 3, 5, 7, and 9. Using these numbers, subjective probabilities can then be calculated from pair-wise ratings on his verbal scale.

3.2.7. *Scoring Rules, Calibration, and Group Assessment*

A number of approaches have been developed for improving the accuracy of assessed probabilities (Winterfelt and Edwards 1986; Lichtenstein et al. 1982). Two desirable properties of elicited prob-

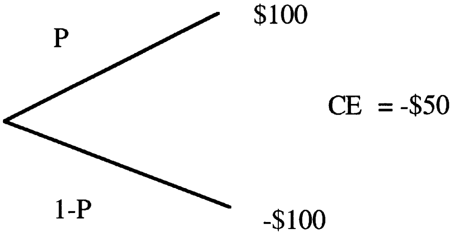


Figure 9 The Standard Gamble Used in the Variable Probability Method of Eliciting Utility Functions.

abilities include extremeness and calibration. More extreme probabilities (for example, P(good sales) = 0.9 vs. P(good sales) = 0.5) make decisions easier since the decision maker can be more sure of what is really going to happen. Well-calibrated probability estimates match the actual frequencies of observed events. Scoring rules provide a means of evaluating assessed probabilities in terms of both extremeness and calibration. If decision makers assess probabilities on a routine basis, feedback can be provided using scoring rules. Such feedback seems to be associated with the highly calibrated subjective probabilities provided by weather forecasters (Murphy and Winkler 1974).

Group assessment of subjective probabilities is another often-followed approach, as alluded to earlier in reference to SLIM-MAUD. There is evidence that group judgments are usually more accurate than individual judgments and that groups tend to be more confident in their estimates (Sniezek and Henry 1989; Sniezek 1992; also see Section 6.2). Assuming that individuals within a group independently provide estimates, which are then averaged, the benefit of group judgment is easily shown to have a mathematical basis. Simply put, a mean should be more reliable than an individual observation. Group dynamics, however, can lead to a tendency towards conformity (Janis 1972). Winterfeldt and Edwards (1986) therefore recommend that members of a group be polled independently.

3.3. Utility Function Assessment

Standard methods for assessing utility functions (Raiffa 1968) include (1) the variable probability method and (2) the certainty equivalent method. In the variable probability method, the decision maker is asked to give the value for the probability of winning at which they are indifferent between a gamble and a certain outcome (Figure 9). A utility function is then mapped out when the value of the certainty equivalent (CE) is changed over the range of outcomes. Returning to Figure 9, the value of P at which the decision maker is indifferent between the gamble and the certain loss of $50 gives the value for $u(-\$50)$. In the utility function in Figure 10, the decision maker gave a value of about 0.5 in response to this question.

The certainty equivalent method uses lotteries in a similar way. The major change is that the probability of winning or losing the lottery is held constant, while the amount won or lost is changed. In most cases, the lottery provides an equal chance of winning and losing. The method begins by

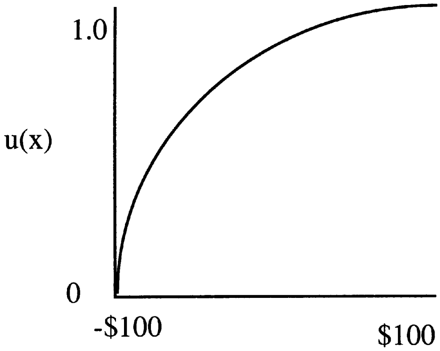


Figure 10 A Typical Utility Function.

asking the decision maker to give a certainty equivalent for the original lottery (CE_1). The value chosen has a utility of 0.5. This follows since the utility of the best outcome is assigned a value of 1 and the worst is given a utility of 0. The utility of the original gamble is therefore:

$$u(CE_1) = pu(\text{best}) + (1 - p)u(\text{worst}) = p(1) + (1 - p)(0) = p = 0.5 \tag{21}$$

The decision maker is then asked to give certainty equivalents for two new lotteries. Each uses the CE from the previous lottery as one of the potential prizes. The other prizes used in the two lotteries are the best and worst outcomes from the original lottery, respectively. The utility of the certainty equivalent (CE_2) for the lottery using the best outcome and CE_1 is given by the expression below:

$$u(CE_2) = pu(\text{best}) + (1 - p)u(CE_1) = p(1) + (1 - p)(0.5) = 0.75 \tag{22}$$

The utility of the certainty equivalent (CE_3) given for the lottery using the worst outcome and CE_1 is given by:

$$u(CE_3) = pu(CE_1) + (1 - p)u(\text{worst}) = p(0.5) + (1 - p)(0) = 0.25 \tag{23}$$

This process is continued until the utility function is specified in sufficient detail. A problem with the certainty equivalent method is that errors are compounded as the analysis proceeds. This follows since the utility assigned in the first preference assessment (i.e., $u(CE_1)$) is used throughout the subsequent preference assessments. A second issue is that the CE method uses different ranges in the indifference lotteries, meaning that the CEs are compared against different reference values. This might create inconsistencies since, as discussed later in Section 4, attitudes toward risk usually change depending upon whether outcomes are viewed as gains or losses. The use of different reference points may, of course, cause the same outcome to be viewed as either a loss or a gain. Utilities may also vary over time. Section 4.2 discusses some of these issues further.

3.4. Preference Assessment

Methods for measuring strength of preference include indifference methods, direct assessment, and indirect measurement (Keeney and Raiffa 1976; Winterfeldt and Edwards 1986). Indifference methods modify one of two sets of stimuli until subjects feel they are indifferent between the two. Direct-assessment methods ask subjects to rate or otherwise assign numerical values to attributes, which are then used to obtain preferences for alternatives. Indirect-measurement techniques avoid decomposition and simply ask for preference orderings between alternatives. Recently there has been some movement towards evaluating the effectiveness of particular methods for measuring preferences (Huber et al. 1993; Birnbaum et al. 1992).

3.4.1. Indifference Methods

Indifference methods are illustrated by the variable probability and certainty equivalent methods of eliciting utility functions presented in the previous section. There, indifference points were obtained by varying either probabilities or values of outcomes. Similar approaches have been applied to develop multiattribute utility or value functions. This approach involves four steps: (1) develop the single attribute utility or value functions, (2) assume a functional form for the multiattribute function, (3) assess the indifference point between various multiattribute alternatives, and (4) calculate the substitution rate or relative importance of one attribute compared to the other. The single-attribute functions might be developed by indifference methods (i.e., the variable probability or certainty equivalent methods) or direct-assessment methods, as discussed later. Indifference points between multiattribute outcomes are obtained through an interactive process in which the values of attributes are systematically increased or decreased. Substitution rates are then obtained from the indifference points.

For example, consider the case for two alternative traffic safety policies, A_1 and A_2. Each policy has two attributes, x = lives lost and y = money spent. Assume the decision maker is indifferent between A_1 and A_2, meaning the decision maker feels that $v(x_1,y_1) = v(20{,}000 \text{ deaths}; \$1 \text{ trillion})$ is equivalent to $v(x_2,y_2) = v(10{,}000 \text{ deaths}; \$1.5 \text{ T})$. For the sake of simplicity, assume an additive value function, where $v(x,y) = (1 - k)v_x(x) + kv_y(y)$. Given this functional form, the indifference point $A_1 = A_2$ is used to derive the relation:

$$(1 - k)v_x(20{,}000 \text{ deaths}) + kv_y(\$1 \text{ T}) = (1 - k)v_x(10{,}000 \text{ deaths}) + kv_y(\$1.5 \text{ T}) \tag{24}$$

This results in the substitution rate as shown below:

$$\frac{k}{1 - k} = \frac{v_x(20{,}000 \text{ deaths}) - v_x(10{,}000 \text{ deaths})}{v_y(\$1.5 \text{ T}) - v_y(\$1 \text{ T})} \tag{25}$$

If $v_x = -x$, and $v_y = -y$, a value of approximately 2^{-5} is obtained for k. The procedure becomes somewhat more complex when nonadditive forms are assumed for the multiattribute function (Keeney and Raiffa 1976).

3.4.2. Direct-Assessment Methods

Direct-assessment methods include curve fitting and various numerical rating methods (Winterfeldt and Edwards 1986). Curve fitting is perhaps the simplest approach. Here, the decision maker first orders the various attributes and then simply draws a curve assigning values to them. For example, an expert might draw a curve relating levels of traffic noise (measured in decibels) to their level of annoyance (on a scale of 0 to 1). Rating methods, as discussed earlier in reference to subjective probability assessment, include direct numerical measures on rating scales and relative ratings.

The analytic hierarchy process (AHP) provides one of the more implementable methods of this type (Saaty 1988). In this approach, the decision is first structured as a value tree (Figure 5). Then each of the attributes is compared in terms of importance in a pair-wise rating process. When entering the ratings, decision makers can enter numerical ratios (for example, an attribute might be twice as important as another) or use the subjective verbal anchors mentioned earlier in reference to subjective probability assessment. The AHP program uses the ratings to calculate a normalized eigenvector assigning importance or preference weights to each attribute. Each alternative is then compared on the separate attributes. For example, two houses might first be compared in terms of cost and then be compared in terms of attractiveness. This results in another eigenvector describing how well each alternative satisfies each attribute. These two sets of eigenvectors are then combined into a single vector that orders alternatives in terms of preference. The subjective multiattribute rating technique (SMART) developed by Edwards (see Winterfeldt and Edwards 1986) provides a similar, easily implemented approach. Both techniques are computerized, making the assessment process relatively painless.

3.4.3. Indirect Measurement

Indirect-measurement techniques avoid asking people to rate or rank directly the importance of factors that impact their preferences. Instead, subjects simply state or order their preferences for different alternatives. A variety of approaches can then be used to determine how individual factors influence preference. Conjoint measurement theory provides one such approach for separating the effects of multiple factors when only their joint effects are known. Application of the approach entails asking subjects to develop an ordered set of preferences for a set of alternatives that systematically vary attributes felt to be related to preference. The relationship between preferences and values of the attributes is then assumed to follow some functional form. The most common functional form assumed is a simple additive-weighting model. Preference orderings obtained using the model are then compared to the original rankings. Example applications of conjoint measurement theory to describe preferences between multiattribute alternatives are discussed in Winterfeldt and Edwards (1986). Related applications include the dichotomy-cut method, used to obtain decision rules for individuals and groups from ordinal rankings of multiattribute alternatives (Stanoulov 1994).

The policy capturing approach used in social judgment theory (Hammond et al. 1975; Hammond 1993) is another indirect approach for describing human judgments of both preferences and probability. The policy-capturing approach uses multivariate regression or other similar techniques to relate preferences to attributes for one or more decision makers. The obtained equations correspond to policies followed by particular decision makers. An example equation might relate medical symptoms to a physician's diagnosis. It has been argued that the policy-capturing approach measures the influence of factors on human judgments more accurately than decomposition methods. Captured weights might be more accurate because decision makers may have little insight into the factors that impact their judgments (Valenzi and Andrews 1973). People may also weigh certain factors in ways that reflect social desirability rather than influence on their judgments (Brookhouse et al. 1986). For example, people comparing jobs might rate pay as being lower in importance than intellectual challenge, while their preferences between jobs might be predicted entirely by pay. Caution must also be taken when interpreting regression weights as indicating importance, since regression coefficients are influenced by correlations between factors, their variability, and their validity (Stevenson et al. 1993).

4. BEHAVIORAL DECISION THEORY

As a normative ideal, classical decision theory has influenced the study of decision making in a major way. Much of the earlier work in behavioral decision theory compared human behavior to the prescriptions of classical decision theory (Edwards 1954; Slovic et al. 1977; Einhorn and Hogarth 1981). Numerous departures were found, including the influential finding that people use heuristics during judgment tasks (Tversky and Kahneman 1974). On the basis of such research, pyschologists have concluded that other approaches are needed to describe the process of human decision making. Descriptive models that relax assumptions of the normative models, but still retain much of their essence, are now being evaluated in the field of judgment and decision theory (Stevenson et al. 1993).

The following discussion summarizes findings from this broad body of literature. The discussion begins by considering research on statistical estimation and inference. Attention then shifts to the topic of decision making under uncertainty and risk.

4.1. Statistical Estimation and Inference

The ability of people to perceive, learn, and draw inferences accurately from uncertain sources of information has been a topic of much research. The following discussion first briefly considers human abilities and limitations on such tasks. The next section introduces several heuristics people seem to use to cope with their limitations and considers how their use can cause certain biases. Attention then shifts to probabilistic information-processing models and policy capturing models. These modeling approaches provide a mathematically oriented view of how people judge probabilities, the biases that might occur, and how people learn to perform probability judgment tasks. The final section briefly summarizes findings on debiasing human judgments.

4.1.1. Human Abilities/Limitations

Research conducted in the early 1960s tested the notion that people behave as ''intuitive statisticians'' who gather evidence and apply it in accordance with the Bayesian model of inference (Peterson and Beach 1967). Several studies evaluated how good people are at estimating statistical parameters, such as means, variances, and proportions. Other studies have compared human inferences obtained from probabilistic evidence to the prescriptions of Bayes' rule.

A number of interesting results were obtained (Table 6). The research first shows that people can be fairly good at estimating means, variances, or proportions from sample data. However, this ability drops greatly when the judged events occur either rarely or very often. In particular, when people are asked to estimate the risk associated with the use of consumer products (Dorris and Tabrizi 1978; Rethans 1980) or various technologies (Lichtenstein et al. 1978), estimates can be weakly related to accident data. Weather forecasters are one of the few groups of people that have been documented as being able to estimate high and low probabilities accurately (Winkler and Murphy 1973).

Part of the issue is that risk estimates are related to factors other than likelihood, such as catastrophic potential, degree of control, or familiarity (Lichtenstein et al. 1978; Slovic 1978; 1987; Lehto et al. 1994). Weber (1994) provides additional evidence that subjective probabilities are related to factors other than uncertainty and argues that people will overestimate the chance of highly positive outcome because of their desire to obtain it. Weber also argues that people will overestimate the chance of a highly undesirable outcome because of their fear of receiving it. Traditional methods of decision analysis separately elicit and then recombine subjective probabilities with utilities, as discussed earlier, and assume that subjective probabilities are independent of consequences. A finding of dependency therefore casts serious doubt upon the normative validity of this commonly accepted approach.

When studies of human inference are considered, several other trends become apparent (Table 6). In particular, several significant deviations from the Bayesian model have been found. These include:

1. Decision makers tend to be conservative in that they don't give as much weight to probabilistic evidence as Bayes' rule (Edwards 1968).
2. They don't consider base rates or prior probabilities adequately (Tversky and Kahneman 1974).
3. They tend to ignore the reliability of the evidence (Tversky and Kahneman 1974).
4. They tend to overestimate the probability of conjunctive events and underestimate the probability of disjunctive events (Bar-Hillel 1973).
5. They tend to seek out confirming evidence rather than disconfirming evidence and place more emphasis on confirming evidence when it is available (Einhorn and Hogarth 1978; Baron 1985).
6. They are overconfident in their predictions (Fischhoff et al. 1977), especially in hindsight (Fischhoff 1982; Christensen-Szalanski 1991).
7. They show a tendency to infer illusionary causal relations (Tversky and Kahneman 1973).

A lively literature has developed regarding these deviations and their significance (Evans 1989; Wickens 1992; Caverni et al. 1990; Klein et al. 1993). From one perspective, these deviations demonstrate inadequacies of human reason and are a source of societal problems (Hammond 1974). From the opposite perspective, it has been held that the above findings are more or less experimental

TABLE 6 Sample Findings on the Ability of People to Estimate and Infer Statistical Quantities

Statistical Estimation	
Accurate estimation of sample means	Peterson and Beach 1968
Variance estimates correlated with mean	Lathrop 1967
Variance biases not found	Levin 1975
Variance estimates based on range	Pitz 1980
Accurate estimates of sample proportions between 0.75 and 0.25	Edwards 1954
Severe overestimates of high probabilities; severe underestimates of low proportions	Fischhoff et al. 1977; Lichtenstein et al. 1982
Reluctance to report extreme events	
Weather forecasters provided accurate probabilities	Winkler and Murphy 1973
Poor estimates of expected severity	Dorris and Tabrizi 1977
Correlation of 0.72 between subjective and objective measures of injury frequency	Rethans 1980
Risk estimates lower for self than for others	Weinstein 1980, 1987
Risk estimates related to catastrophic potential, degree of control, familiarity	Lichtenstein et al. 1978
Evaluations of outcomes and probabilities are dependent	Weber 1994
Statistical Inference	
Conservative aggregation of evidence	Edwards 1966
Nearly optimal aggregation of evidence in naturalistic setting	Lehto et al. 2000
Failure to consider base rates	Tversky and Kahneman 1974
Base rates considered	Birnbaum and Mellers 1983 Koehler 1996
Overestimation of conjunctive events	Bar-Hillel 1973
Underestimation of disjunctive events	
Tendency to seek confirming evidence	Einhorn and Hogarth 1978; Baron 1985
Tendency to discount disconfirming evidence	
Tendency to ignore reliability of the evidence	Kahneman and Tversky 1973
Subjects considered variability of data when judging probabilities	Evans and Pollard 1985
People insensitive to information missing from fault trees	Fischhoff et al. 1978
Overconfidence in estimates	Fischhoff et al. 1977
Hindsight bias	Fishhoff 1982 Christensen-Szalanski and Willham 1991
Illusionary correlations	Tversky and Kahneman 1974
Gambler's fallacy	
Misestimation of covariance between items	Arkes 1981
Misinterpretation of regression to the mean	Tversky and Kahneman 1974

artifacts that do not reflect the true complexity of the world (Cohen 1993). From one such perspective, people deviate from Bayes' rule because it makes unrealistic assumptions about what is known or knowable. Simon (1955, 1983) makes a particularly compelling argument for the latter point of view. It also has been noted that researchers overreport findings of bias (Evans 1989; Cohen 1993).

There is an emerging body of literature that, on one hand, shows that deviations from Bayes' rule can in fact be justified in certain cases from a normative view and, on the other hand, shows that these deviations may disappear when people are provided richer information or problems in more natural contexts. For example, drivers performing a simulated passing task combined their own observations of the driving environment with imperfect information provided by a collision-warning system, as predicted by a distributed signal detection theoretic model of optimal team decision making (Lehto et al. 2000). Other researchers have pointed out that:

1. A tendency towards conservatism can be justified when evidence is not conditionally independence (Navon 1979).
2. Subjects do use base rate information and consider the reliability of evidence, in slightly modified experimental settings (Birnbaum and Mellers 1983; Koehler 1996).
3. A tendency to seek out confirming evidence can offer practical advantages (Cohen 1993) and may reflect cognitive failures, due to a lack of understanding of how to falsify hypotheses, rather than entirely a motivational basis (Klayman and Ha 1987; Evans 1989).
4. Subjects prefer stating subjective probabilities with vague verbal expressions rather than precise numerical values (Wallsten et al. 1993), demonstrating that they are not necessarily overconfident in their predictions.*
5. There is evidence that the hindsight bias can be moderated by familiarity with both the task and the type of outcome information provided (Christensen-Szalanski and Willham 1991).

Koehler (1996) provides a particularly compelling reexamination of the base rate fallacy. He concludes that the literature does not support the conventional wisdom that people routinely ignore base rates. To the contrary, he states that base rates are almost always used and that their degree of use depends on task structure and representation as well as their reliability compared to other sources of information. Because such conflicting conclusions can be obtained, depending upon the setting in which human decision making is observed, Koehler and researchers in the field of naturalistic decision making (Klein 1998; Klein et al. 1993) strongly emphasize the need to conduct ecologically valid research in rich realistic decision environments.

4.1.2. Heuristics and Biases

Tversky and Kahneman (1973, 1974) made a key contribution to the field when they showed that many of the above-mentioned discrepancies between human estimates of probability and Bayes' rule could be explained by the use of three heuristics.† The three heuristics they proposed were those of representativeness, availability, and anchoring and adjustment.

The *representativeness* heuristic holds that the probability of an item *A* belonging to some category *B* is judged by considering how representative *A* is of *B*. For example, a person is typically judged more likely to be a librarian than a farmer when described as "A meek and tidy soul, who has a desire for order and structure and a passion for detail." Application of this heuristic will often lead to good probability estimates but can lead to systematic biases. Tversky and Kahneman (1974) give several examples of such biases. In each case, representativeness influenced estimates more than other, more statistically oriented information. In the first study, subjects ignored base rate information (given by the experimenter) about how likely a person was to be either a lawyer or an engineer. Their judgments seemed to be based entirely on how representative the description seemed to be of either occupation. Tversky and Kahneman (1983) found people overestimated conjunctive probabilities in a similar experiment. Here, after being told that "Linda is 31 years old, single, outspoken, and very bright," most subjects said it was more likely she was both a bank teller and active as a feminist than simply a bank teller. In a third study, most subjects felt that the probability of more than 60% male births on a given day was about the same for both large and small hospitals (Tversky and Kahneman 1974). Apparently, the subjects felt large and small hospitals were equally representative of the population.

Other behaviors explained in terms of representativeness by Tversky and Kahneman included gambler's fallacy, insensitivity to predictability, illusions of validity, and misconceptions of statistical regression to the mean. With regard to gambler's fallacy, they note that people may feel long sequences of heads or tails when flipping coins are unrepresentative of normal behavior. After a sequence of heads, a tail therefore seems more representative. Insensitivity to predictability refers to a tendency for people to predict future performance without considering the reliability of the information they base the prediction upon. For example, a person might expect an investment to be profitable solely on the basis of a favorable description without considering whether the description has any predictive value. In other words, a good description is believed to be representative of high profits, even if it states nothing about profitability. The illusion of validity occurs when people use highly correlated evidence to make a conclusion. Despite the fact that the evidence is redundant, the presence of many representative pieces of evidence increases confidence greatly. Misconception of regression to the mean occurs when people react to unusual events and then infer a causal linkage when the process returns to normality on its own. For example, a manager might incorrectly conclude that punishment works after seeing that unusually poor performance improves to normal levels following punishment. The same manager might also conclude that rewards don't work after seeing that unusually good performance drops after receiving a reward.

---

* It is interesting to note that Dawes and Mulford (1996) claim that the empirical support for the overconfidence effect is inadequate and logically flawed.

† It is important to point out that heuristic reasoning can lead to excellent results.

The availability heuristic holds that the probability of an event is determined by how easy it is to remember the event happening. Tversky and Kahneman state that perceived probabilities will, therefore, depend on familiarity, salience, effectiveness of memory search, and imaginability. The implication is that people will judge events as more likely when the events are familiar, highly salient (such as an airplane crash), or easily imaginable. Events also will be judged more likely if there is a simple way to search memory. For example, it is much easier to search for words in memory by the first letter rather than the third letter. It is easy to see how each of the above items impacting the availability of information can influence judgments. Biases should increase when people lack experience or when their experiences are too focused.

Anchoring and adjustment holds that people start from some initial estimate and then adjust it to reach some final value. The point initially chosen has a major impact on the final value selected when adjustments are insufficient. Tversky and Kahneman refer to this source of bias as an anchoring effect. They show how this effect can explain under- and overestimates of disjunctive and conjunctive events. This happens if the subject starts with a probability estimate of a single event. The probability of a single event is of course less than that for the disjunctive event and greater than that for the conjunctive event. If adjustment is too small, then under- and overestimates respectively occur for the disjunctive and conjunctive events. Tversky and Kahneman also discuss how anchoring and adjustment may cause biases in subjective probability distributions. Hogarth and Einhorn (1992) present an anchoring and adjustment model of how people update beliefs that explains a number of ordering effects, such as the primacy and recency effects. This latter model holds that the degree of belief in a hypothesis after collecting k pieces of evidence can be described as follows:

$$S_k = S_{k-1} + w_k[s(x_k) - R] \tag{26}$$

where S_k is the degree of belief after collecting k pieces of evidence, S_{k-1} is the anchor or prior belief, w_k is the adjustment weight for the kth piece of evidence, $s(x_k)$ is the subjective evaluation of the kth piece of evidence, and R is the reference point against which the kth piece of evidence is compared. In evaluation tasks, $R = 0$. This corresponds to the case where evidence is either for or against a hypothesis.* For estimation tasks, $R \neq 0$. The different values of R result in an additive model for evaluation tasks and an averaging model for estimation tasks. Also, if the quantity, $s(x_k) - R$, is evaluated for several pieces of evidence at a time, the model predicts primacy effects. If single pieces of evidence are individually evaluated in a step-by-step sequence, recency effects become more likely.

The notion of heuristics and biases has had a particularly formative influence on decision theory. A substantial recent body of work has emerged that focuses on applying research on heuristics and biases (Kahneman et al. 1982; Heath et al. 1994). Applications include medical judgment and decision making, affirmative action, education, personality assessment, legal decision making, mediation, and policy making. It seems clear that this approach is excellent for describing many general aspects of decision making in the real world. However, research on heuristics and biases has been criticized as being pretheoretical (Slovic et al. 1977). Koehler (1996) also points out that efforts to confirm the representativeness heuristic has contributed to overselling of the ''base rate'' bias. Other views of human judgment are expanded upon below.

4.1.3. Selective Processing of Information

Evans (1989) argues that factors which cause people to process information in a selective manner or attend to irrelevant information are the major cause of biases in human judgment. Factors assumed to influence selective processing include the availability, vividness, and relevance of information, and working memory limitations. The notion of availability refers to the information actually attended to by a person while performing a task. Evans's model assumes that relevant information elements are determined during a heuristic, preattentional stage. This stage is assumed to involve unconscious processes and is influenced by stimulus salience (or vividness) and the effects of prior knowledge.

In the next stage of his model, inferences are drawn from the selected information. This is done using rules for reasoning and action developed for particular types of problems. Working memory influences performance at this stage by limiting the amount of information that can be consciously attended to while performing a task. The knowledge used during the inference process might be organized in schemas that are retrieved from memory and fit to specific problems (Cheng and Holyoak 1985). Support for this latter conclusion is provided by studies showing that people are able to develop skills in inference tasks but may fail to transfer these skills (inference related) from one setting to another. Evans also provides evidence that prior knowledge can cause biases when it is inconsistent with provided information and that improving knowledge can reduce or eliminate biases.

---

* It is easy to see that Eq. (26) approximates the log-odds form of Bayes' rule where evidence for or against the hypothesis is additively combined.

Evans's model of selective processing of information is consistent with other explanations of biases. Among such explanations, information overload has been cited as a reason for impaired decision making by consumers (Jacoby 1977). The tendency of highly salient stimuli to capture attention during inference tasks has also been noted by several researchers (Nisbett and Ross 1980; Payne 1980). Nisbett and Ross suggest that vividness of information is determined by its emotional content, concreteness and imagability, and temporal and spatial proximity. As noted by Evans, these factors have also been shown to affect the memorability of information. This provides a plausible explanation of both the availability heuristic and the experimental results mentioned earlier regarding biases in risk perceptions.

4.1.4. Models of Human Judgment

A number of approaches have been developed for mathematically describing human judgments. These approaches include social judgment theory, policy capturing, multiple-cue probability learning models, information integration theory, and conjoint measurement approaches.

Social judgment theory (SJT) implements an ecological approach for explaining how environmental cues are related to psychological responses (Hammond et al. 1975; Hammond 1993; Brehmer and Joyce 1988). The approach can be traced back to the Brunswick lens model (Brunwick 1952), which describes human judgments in terms of perceived environmental cues. Emphasis is placed on performing experiments where information cues reflect the statistical characteristics of the real world. Policy-capturing models are also derived from the lens model and have been applied to a wide number of real-world applications to describe expert judgments (Brehmer and Joyce 1988). For example, policy-capturing models have been applied to describe software selection by management information system managers (Martocchio et al. 1993), medical decisions (Brehmer and Joyce 1988), and highway safety (Hammond 1993). As mentioned earlier with regard to preference assessment, linear or nonlinear forms of regression are used in this approach to relate judgments to environment cues. These equations provide surprisingly good fits to expert judgments. In fact, there is evidence, and consequently much debate, over whether the models can actually do better than experts on many judgment tasks (Slovic et al. 1977; Brehmer 1981; Kleinmuntz 1984).

Cognitive continuum theory (Hammond 1980) builds upon Brunswick's earlier work by distinguishing judgments on a cognitive continuum varying from highly intuitive decisions to highly analytical decisions. Hammond (1993) summarizes earlier research showing that task characteristics cause decision makers to vary on this continuum. A tendency towards analysis increases, and reliance on intuition decreases, when (1) the number of cues increases, (2) cues are measured objectively instead of subjectively, (3) cues are of low redundancy, (4) decomposition of the task is high, (5) certainty is high, (6) cues are weighted unequally in the environmental model, (7) relations are nonlinear, (8) an organizing principle is available, (9) cues are displayed sequentially instead of simultaneously, and (10) the time period for evaluation is long. Intuitive methods can be better than analytical methods in some situations (Hammond et al. 1987).

Multiple cue probability learning models extend the lens model to the psychology of learning (Brehmer and Joyce 1988). Research on multiple-cue probability learning has provided valuable insight into factors affecting learning of inference tasks. One major finding is that providing cognitive feedback about cues and their relationship to the inferred effects leads to quicker learning than feedback about outcomes (Balzer et al. 1989). Stevenson et al. (1993) summarize a number of other findings, including that (1) subjects can learn to use valid cues, even when they are unreliable, (2) subjects are better able to learn linear relationships than nonlinear or inverse relationships, (3) subjects do not consider redundancy when using multiple cues, (4) source credibility and cue validity are considered, and (5) the relative effectiveness of cognitive and outcome feedback depends on the formal, substantive, and contextual characteristics of the task.

Information integration theory (Anderson 1981) takes a somewhat different approach than SJT or the lens model to describe how cue information is used when making judgments. A major deviation is that information-integration theory emphasizes the use of factorial experimental designs where cues are systematically manipulated. The goal of this approach is to determine first how people scale cues when determining their subjective values, and second how these scaled values are combined to form overall judgments. Various functional forms of how information is integrated are considered, including additive and averaging functions. A substantial body of research follows this approach to test various ways people might combine probabilistic information. A primary conclusion is that people tend to integrate information using simple averaging, adding, subtracting, and multiplying models. Conjoint measurement approaches (Wallsten 1972, 1976), in particular, provide a convenient way of both scaling subjective values assigned to cues and testing different functional forms describing how these values are combined to develop global judgments. By applying this approach, Wallsten (1976) was able to model primacy and recency effects.

4.1.5. Debiasing Human Judgments

The notion that many biases (or deviations from normative models) in statistical estimation and inference can be explained has led researchers to consider the possibility of debiasing human judgments (Keren 1990). Part of the issue is that heuristics often work very well. It seems logical that biases based on both the availability and representativeness heuristics might be reduced if people were provided more information. As discussed earlier in Section 4.1.1, there is evidence that biases can be moderated by familiarity with both the task and the type of outcome information provided. However, debiasing research has provided mixed results. Many biases, such as optimistic beliefs regarding health risks, have been difficult to modify (Weinstein and Klein 1995). People show a tendency to seek out information that supports their personal views (Weinstein 1979) and are quite resistant to information that contradicts strongly held beliefs (Nisbett and Ross 1980; McGuire 1966). Evans (1989) concludes that "pre-conceived notions are likely to prejudice the construction and evaluation of arguments."

Other evidence shows that experts may have difficulty providing accurate estimates of subjective probabilities even when they receive feedback. For example, many efforts to reduce both overconfidence in probability estimates and the hindsight bias have been unsuccessful (Fischhoff 1982). One problem is that people may not pay attention to feedback (Fischhoff and MacGregor 1982). They also may only attend to feedback that supports their hypothesis, leading to poorer performance and at the same time greater confidence (Einhorn and Hogarth 1978). Efforts to reduce confirmation biases through training have also in general been unsuccessful (Evans 1989).

On the positive side, there is evidence that providing feedback on the accuracy of weather forecasts may help weather forecasters (Winkler and Murphy 1973). There is also some evidence that people can learn to perform statistical reasoning more accurately after training in statistics (Fong et al. 1986). Failure to consider sample size was significantly reduced after training. Another study showed that asking people to write down reasons for and against their estimates of probabilities improved calibration and reduced overconfidence (Koriat et al. 1980). There is evidence that overconfidence is reduced when decision makers represent subjective probabilities verbally (Zimmer 1983; Wallsten et al. 1993). Conservatism, or the failure to modify probabilities adequately after obtaining evidence, was also reduced in Zimmer's study.

The conclusion is that debiasing human judgments is difficult but not impossible. Some perspective can be obtained by considering that most studies showing biases have focused on statistical inference and generally involved people not particularly knowledgeable about statistics, who are not using decision aids such as computers or calculators. It naturally may be expected that people will perform poorly on such tasks, given their lack of training and forced reliance on mental calculations (Winterfeldt and Edwards 1986). The finding that people can improve their abilities on such tasks after training in statistics is particularly telling, but also encouraging. Another encouraging finding is that biases are occasionally reduced when people process information verbally instead of numerically. This result might be expected, given that most people are more comfortable with words than numbers.

4.2. Preference and Choice

Much of the research on human preference and choice has focused on comparing observed preferences to the predictions of subjective utility theory (SEU) (Goldstein and Hogarth 1997). Early work, examining SEU as a descriptive theory, drew generally positive conclusions. However, it soon became apparent that people's preferences for risky or uncertain alternatives often violated basic axioms of SEU theory. The finding that people's preferences change when the outcomes are framed in terms of costs, as opposed to benefits, has been particularly influential. Several other common deviations from SEU have been observed. One potentially serious deviation is that preferences can be influenced by sunk costs or prior commitment to a particular alternative. Preferences change over time and may depend upon which alternatives are being compared, or even the order in which they are compared. The regret associated with making the "wrong" choice seems to play a major role when people compare alternatives. Accordingly, the satisfaction people derive from obtaining particular outcomes after making a decision is influenced by positive and negative expectations prior to making the decision. Other research on human preference and choice has shown that people choose between and apply different decision strategies, depending upon the cognitive effort required to apply a decision strategy successfully, the needed level of accuracy, and time pressure. Certain strategies are more likely than others to lead to choices consistent with those prescribed by SEU theory.

Alternative models, such as prospect theory and random utility theory, were consequently developed in order to explain human preferences under risk or uncertainty.* The following discussion will first summarize some common violations of the axioms underlying SEU theory before moving on to framing effects and preference reversals. Attention will then shift to models of choice and preference. The latter discussion will begin with prospect theory before addressing other models of labile or conditional preferences. Decision-making strategies, and how people choose between them, will be covered in Section 6.

4.2.1. Violation of Rationality Axioms

Several studies have shown that people's preferences between uncertain alternatives can be inconsistent with the axioms underlying subjective expected utility (SEU) theory. One fundamental violation of the assumptions is that preferences can be intransitive (Tversky 1969; Budescu and Weiss 1987). Also, as mentioned in the previous section, subjective probabilities may depend upon the values of consequences (violating the independence axiom) and, as discussed in the next section, the framing of a choice can impact preference. Another violation is given by the Myers effect (Myers et al. 1965), where preference reversals between high (H) and low (L) variance gambles can occur when the gambles are compared to a certain outcome, depending upon whether the certain outcome is positive (H preferred to L) or negative (L preferred to H). This latter effect violates the assumption of independence because the ordering of the two gambles depends on the certain outcome.

Another commonly cited violation of SEU theory is that people show a tendency towards uncertainty avoidance which can lead to behavior inconsistent with the ''sure-thing'' axiom. The Ellsburg and Allais paradoxes (Ellsburg 1961; Allais 1953) both involve violations of the sure-thing axiom (see Table 2) and seem to be caused by people's desire to avoid uncertainty. The Allais paradox is illustrated by the following set of gambles. In the first gamble, a person is asked to choose between gambles $A1$ and $B1$, where:

> Gamble $A1$ results in \$1 million for sure. Gamble $B1$ results in \$2.5 million with a probability of 0.1, \$1 million with a probability of 0.89, and \$0 with a probability of 0.01.
>
> In the second gamble, the person is asked to choose between gambles *A2 and* $B2$, where: $A2$ results in \$1 million with a probability of 0.11 and \$0 with a probability of 0.89. Gamble $B2$ results in \$2.5 million with a probability of 0.1 and \$0 with a probability of 0.9.

Most people prefer gamble $A1$ to $B1$ and gamble $B2$ to $A2$. It is easy to see that this set of preferences violates expected utility theory. First, if $A1 > B1$, then $u(A1) > u(B1)$, meaning that: $u(\$1\text{ million}) > 0.1u(\$2.5\text{ million}) + 0.89u(\$1\text{ million}) + 0.01u(\$0)$. If a utility of 0 is assigned to receiving \$0 and a utility of 1 to receiving \$2.5 million, then $u(\$1\text{ million}) > 1/11$. However, from the preference $A2 > B2$, it follows that $u(\$1\text{ million}) < 1/11$. Obviously, no utility function can satisfy this requirement of assigning a value both greater than and less than 1/11 to \$1 million.

As noted by Savage (1954), the above set of gambles can be reframed in a way that shows that these preferences violate the sure-thing principle. After doing so, Savage found that his initial tendency towards choosing $A1$ over $B1$ and $A2$ over $B2$ disappeared. As noted by Stevenson et al. (1993), this example is one of the first cited cases of a preference reversal caused by reframing a decision, the topic discussed below.

4.2.2. Framing of Decisions and Preference Reversals

A substantial body of research has shown that people's preferences can shift dramatically depending upon the way a decision is represented. The best-known work on this topic was conducted by Tversky and Kahneman (1981), who showed that preferences between medical intervention strategies changed dramatically depending upon whether the outcomes were posed as losses or gains. The following question, worded in terms of benefits, was presented to one set of subjects:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.
>
> Which of the two programs would you favor?

---

* Yates (1992) and Singleton and Hovden (1987) are useful sources for the reader interested in additional details on risk perception, risk acceptability, and risk-taking behavior. Section 5.1.1 is also relevant to this topic.

The results showed that 72% of subjects preferred program A. The second set of subjects was given the same cover story, but worded in terms of costs, as given below:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die.
>
> Which of the two programs would you favor?

The results now showed that 78% of subjects preferred program D. Since program D is equivalent to B and Program A is equivalent to C, the preferences for the two groups of subjects were strongly reversed. Tversky and Kahneman concluded that this reversal illustrated a common pattern in which choices involving gains are risk averse and choices involving losses are risk seeking. The interesting result was that the way the outcomes were worded caused a shift in preference for identical alternatives. Tversky and Kahneman called this tendency the *reflection effect.* A body of literature has since developed showing that the framing of decisions can have practical effects for both individual decision makers (Kahneman et al. 1982; Heath et al. 1994) and group decisions (Paese et al. 1993). On the other hand, recent research shows that the reflection effects can be reversed by certain outcome wordings (Kuhberger 1995); more importantly, Kuhberger provides evidence that the reflection effect observed in the classic experiments can be eliminated by fully describing the outcomes (i.e., referring to the above paragraph, a more complete description would state, "If Program C is adopted, 400 people will die AND 200 WILL LIVE").

Other recent research has explored the theory that perceived risk and perceived attractiveness of risky outcomes are psychologically distinct constructs (Weber et al. 1992). In the latter study, it was concluded that perceived risk and attractiveness are "closely related, but distinct phenomena." Related research has shown weak negative correlations between the perceived risk and value of indulging in alcohol-related behavior for adolescent subjects (Lehto et al. 1994). This latter study also showed that the rated propensity to indulge in alcohol-related behavior was strongly correlated with perceived value ($R = 0.8$), but weakly correlated with perceived risk ($R = -0.15$). Both findings are consistent with the theory that perceived risk and attractiveness are distinct constructs, but the latter finding indicates that perceived attractiveness may be the better predictor of behavior. Lehto et al. conclude that intervention methods attempting to lower preferences for alcohol-related behavior should focus on lowering perceived value rather than on increasing perceived risk.

4.2.3. Prospect Theory

Prospect theory (Kahneman and Tversky 1979) attempts to account for behavior not consistent with the SEU model by including the framing of decisions as a step in the judgment of preference between risky alternatives. Prospect theory assumes that decision makers tend to be risk averse with regard to gains and risk seeking with regard to losses. This leads to a value function that disproportionately weights losses. As such, the model is still equivalent to SEU, assuming a utility function expressing mixed risk aversion and risk seeking. Prospect theory, however, assumes that the decision maker's reference point can change. With shifts in the reference point, the same returns can be viewed as either gains or losses.* This latter feature of prospect theory, of course, is an attempt to account for the framing effect discussed above. Prospect theory also deviates significantly from SEU theory in the way probabilities are addressed. To describe human preferences more closely, perceived values are weighted by a function $\pi(p)$, instead of the true probability, p. Compared to the untransformed form of p, $\pi(p)$ overweights very low probabilities and underweights moderate and high probabilities. The function $\pi(p)$ is also generally assumed to be discontinuous and poorly defined for probability values close to 0 or 1.

Prospect theory assumes that the choice process involves an editing phase and an evaluation phase. The editing phase involves reformulation of the options to simplify subsequent evaluation and choice. Much of this editing process is concerned with determining an appropriate reference point in a step called coding. Other steps that may occur include the segregation of riskless components of the decision, combining probabilities for events with identical outcomes, simplification by rounding off probabilities and outcome measures, and search for dominance. In the evaluation phase, the perceived values are then weighed by the function $\pi(p)$. The alternative with the greatest weighed value is then selected. Several other modeling approaches that differentially weigh utilities in risky decision making have been proposed (Goldstein and Hogarth 1998). As in prospect theory, such models often assume that the subjective probabilities, or decision weights, are a function of outcome sign (i.e. positive, neutral, or negative), rank (i.e., 1st, 2nd, etc), or magnitude. Other models focus on display effects (i.e., single-stage vs. multistage arrangements) and distribution effects (i.e., two outcome lotteries vs multiple-outcome lotteries). Prospect theory and other approaches also address how the value or utility of particular outcomes can change between decision contexts, as discussed below.

---

*The notion of a reference point against which outcomes are compared has similarities to the notion of making decisions on the basis of regret (Bell 1982). Regret, however, assumes comparsion to the best outcome. The notion of different reference points also is related to the well-known trend that buying and selling price of assets often differ for a decision maker (Raiffa 1968).

4.2.4. Labile Preferences

There is no doubt that human preferences often change after receiving some outcome. After losing money, an investor may become risk averse. In other cases, an investor may escalate her commitment to an alternative after an initial loss, even if better alternatives are available. From the most general perspective, any biological organism becomes satiated after satisfying a basic need, such as hunger. Preferences also change over time or between decision contexts. For example, a 30-year-old decision maker considering whether to put money into a retirement fund may currently have a very different utility function than at retirement. The latter case is consistent with SEU theory but obviously complicates analysis.

Economists and behavioral researchers have both focused on mathematically modeling choice processes to explain intransitive or inconsistent preference orderings of alternatives (Goldstein and Hogarth 1997). Game theory provides interesting insight into this issue. From this perspective, preferences of the human decision maker are modeled as the collective decisions obtained by a group of internal agents, or selves, each of which is assumed to have distinct preferences (see Elster 1986). Intransitive preferences and other violations of rationality on the part of the human decision maker then arise from interactions between competing selves.* Along these lines, Ainslie (1975) proposed that impulsive preference switches (often resulting in risky or unhealthy choices) arise as the outcome of a struggle between selves representing conflicting short-term and long-term interests, respectively.

Another area of active research has focused on how experiencing outcomes can cause shifts in preference. One robust finding is that people tend to be more satisfied if an outcome exceeds their expectations and less satisfied if it does not (i.e., Feather 1966; Connolly et al. 1997). Expectations therefore provide a reference point against which obtained outcomes are compared. Numerous studies have also shown that people in a wide variety of settings often consider sunk costs when deciding whether to escalate their commitment to an alternative by investing additional resources (Arkes and Blumer 1985). From the perspective of prospect theory, sunk costs cause people to frame their choice in terms of losses instead of gains, resulting in risk-taking behavior and consequently escalating commitment. Other plausible explanations for escalating commitment include a desire to avoid waste or to avoid blame for an initially bad decision to invest in the first place. Interestingly, some recent evidence suggests that people may deescalate commitment in response to sunk costs (Heath 1995). The latter effect is also contrary to classical economic theory, which holds that decisions should be based solely on marginal costs and benefits. Heath explains such effects in terms of mental accounting. Escalation is held to occur when a mental budget is not set or expenses are difficult to track. Deescalation is held to occur when people exceed their mental budget, even if the marginal benefits exceed the marginal costs.

Other approaches include value or utility as random variables within models of choice to explain intransitive or inconsistent preference orderings of alternatives. Random utility models (Iverson and Luce 1998) describe the probability $P_{a,A}$ of choosing a given alternative a from a set of options A as

$$P_{a,A} = \text{Prob}(U_a \geq U_b, \text{ for all } b \text{ in } A) \tag{27}$$

where U_a is the uncertain utility of alternative a and U_b is the uncertain utility of alternative b. The most basic random utility models assign a utility to each alternative by sampling a single value from some known distribution. The sampled utility of each alternative then remains constant throughout the choice process. Basic random utility models can predict a variety of preference reversals and intransitive preferences for single and multiple attribute comparisons of alternatives (i.e., Tverski 1972).

Sequential sampling models extend this approach by assuming preferences can be based on more than one observation. Preferences for particular alternatives are accumulated over time, by integrating or otherwise summing the sampled utilities. The utility of an alternative, at a particular time, is proportional to the latter sum. A choice is made when the summed preferences for a particular alternative exceed some threshold, which itself may vary over time or depend on situational factors (Wallsten 1995; Busemeyer and Townsend 1993). It is interesting to observe that sequential sampling models can explain speed accuracy trade-offs in signal-detection tasks (Stone 1960), as well as shifts in preferences due to time pressure (Wallsten 1995; Busemeyer and Townsend 1993), if it is assumed that people adjust their threshold downwards under time pressure. That is, under time pressure, people sample less information before making a choice. The following section will further explore how and why decision strategies might change over time and between decision contexts.

---

* As discussed further in Section 6, group decisions, even though they are made by rational members, are subject to numerous violations of rationality. For example, consider the case where the decision maker has three selves that are, respectively, risk averse, risk neutral, and risk seeking. Assume that the decision maker is choosing between alternatives A, B, and C. Suppose the risk-averse self rates the alternatives in the order A, B, C; the risk-neutral self rates them in the order B, C, A; and the risk-seeking self rates them in the order C, A, B. Also, assume the selves are equally powerful. Then two of the three agents always agree that $A > B$, $B > C$, and $C > A$. This ordering is, of course, nontransitive.

5. DYNAMIC AND NATURALISTIC DECISION MAKING

In dynamic decision making, actions taken by a decision maker are made sequentially in time. Taking actions can change the environment, resulting in a new set of decisions. The decisions might be made under time pressure and stress, by groups or by single decision makers. This process might be performed on a routine basis or might involve severe conflict. For example, either a group of soldiers or an individual officer might routinely identify marked vehicles as friends or foes. When a vehicle has unknown or ambiguous marking, the decision changes to a conflict driven process. Naturalistic decision theory has emerged as a new field that focuses on such decisions in real-world environments (Klein et al. 1993; Klein 1998). The notion that most decisions are made in a routine, nonanalytical way is the driving force of this approach.* Areas where such behavior seems prominent include juror decision making, troubleshooting of complex systems, medical diagnosis, management decisions, and numerous other examples.

The following discussion will first address models of dynamic and naturalistic decision making. These models both illustrate naturalistic decision-making strategies and explain their relation to experience and task familiarity. A brief discussion will also be provided on teams and team leadership, in naturalistic settings. Attention will then shift to the issue of time pressure and stress and how this factor influences performance in naturalistic decision making.

5.1. Naturalistic Decision Making

In recent years, it has been recognized that decision making in natural environments often differs greatly between decision contexts (Beach 1993; Hammmond 1993). In addressing this topic, the involved researchers often question the relevance and validity of both classical decision theory and behavioral research not conducted in real-world settings (Cohen 1993). Numerous naturalistic models have been proposed (Klein et al. 1993). These models assume that people rarely weigh alternatives and compare them in terms of expected value or utility. Each model is also descriptive rather than prescriptive. Perhaps the most general conclusion that can be drawn from this work is that people use different decision strategies, depending upon their experience, the task and the decision context. Several of the models also postulate that people choose between decision strategies by trading off effectiveness against the effort required.

The following discussion will briefly review seven modeling perspectives that fit into this framework: (1) levels of task performance (Rasmussen 1983), (2) recognition-primed decisions (Klein 1989), (3) image theory (Beach 1990), (4) contingent decision making (Payne et al. 1993), (5) dominance structuring (Montgomery 1989), (6) explanation-based decision making (Pennington and Hastie 1988), and (7) shared mental models and awareness. Attention will then shift to leadership and its impact on team performance in naturalistic settings.

5.1.1. Levels of Task Performance

There is growing recognition that most decisions are made on a routine basis in which people simply follow past behavior patterns (Rasmussen 1983; Beach 1993; Svenson 1990). Rasmussen (1983) follows this approach to distinguish among skill-based, rule-based, and knowledge-based levels of task performance. Lehto (1991) further considers judgment-based behavior as a fourth level of performance.

Performance is said to be at either a skill-based or a rule-based level when tasks are routine in nature. Skill-based performance involves the smooth, automatic flow of actions without conscious decision points. As such, skill-based performance describes the decisions made by highly trained operators performing familiar tasks. Rule-based performance involves the conscious perception of environmental cues, which trigger the application of rules learned on the basis of experience. As such, rule-based performance corresponds closely to recognition-primed decisions (Klein 1989). The knowledge-based level of performance is said to occur during learning or problem-solving activity during which people cognitively simulate the influence of various actions and develop plans for what to do. The judgment-based level of performance occurs when affective reactions of a decision maker cause a change in goals or priorities between goals (Janis and Mann 1977; Etzioni 1988; Lehto 1991). Distinctive types of errors in decision making occur at each of the four levels (Reason 1989; Lehto 1991).

---

*Drucker (1985), in discussing ways of improving the effectiveness of executive decision makers, emphasizes the importance of establishing a generic principle or policy that can be applied to specific cases in a routine way. This recommendation is interesting because it prescribes a naturalistic form of behavior.

At the skill-based level, errors occur due to perceptual variability and when people fail to shift up to rule-based or higher levels of performance. At the rule-based level, errors occur when people apply faulty rules or fail to shift up to a knowledge-based level in unusual situations where the rules they normally use are no longer appropriate. The use of faulty rules leads to an important distinction between running and taking risks. Along these lines, Wagenaar (1992) discusses several case studies in which people following risky forms of behavior do not seem to be consciously evaluating the risk. Drivers, in particular, seem to habitually take risks. Wagenaar explains such behavior in terms of faulty rules derived on the basis of benign experience. In other words, drivers get away with providing small safety margins most of the time and consequently learn to run risks on a routine basis. Drucker (1985) points out several cases where organizational decision makers have failed to recognize that the generic principles they used to apply were no longer appropriate, resulting in catastrophic consequences.

At the knowledge-based level, errors occur because of cognitive limitations or faulty mental models or when the testing of hypotheses cause unforeseen changes to systems. At the judgment-based levels, errors (or violations) occur because of inappropriate affective reactions, such as anger or fear (Lehto 1991). As noted by Isen (1993), there also is growing recognition that positive affect can influence decision making. For example, positive affect can promote the efficiency and thoroughness of decision making, but may cause people to avoid negative materials. Positive affect also seems to encourage risk-averse preferences. Decision making itself can be anxiety provoking, resulting in violations of rationality (Janis and Mann 1977).

A recent study involving drivers arrested for drinking and driving (McKnight et al. 1995) provides an interesting perspective on how the sequential nature of naturalistic decisions can lead people into traps. The study also shows how errors can occur at multiple levels of performance. In this example, decisions made well in advance of the final decision to drive while impaired played a major role in creating situations where drivers were almost certain to drive impaired. For instance, the driver may have chosen to bring along friends and therefore have felt pressured to drive home because the friends were dependent upon him or her. This initial failure by drivers to predict the future situation could be described as a failure to shift up from a rule-based level to a knowledge-based level of performance. In other words, the driver never stopped to think about what might happen if he or she drank too much. The final decision to drive, however, would correspond to an error (or violation) at the judgment-based level if the driver's choice was influenced by an affective reaction (perceived pressure) to the presence of friends wanting a ride.

5.1.2. Recognition-Primed Decision Making

Klein (1989, 1998) developed the theory of recognition-primed decision making on the basis of observations of firefighters and other professionals in their naturalistic environments. He found that up to 80% of the decisions made by firefighters involved some sort of situation recognition, where the decision makers simply followed a past behavior pattern once they recognized the situation.

The model he developed distinguishes between three basic conditions. In the simplest case, the decision maker recognizes the situation and takes the obvious action. A second case occurs when the decision maker consciously simulates the action to check whether it should work before taking it. In the third and most complex case, the action is found to be deficient during the mental simulation and is consequently rejected. An important point of the model is that decision makers don't begin by comparing all the options. Instead, they begin with options that seem feasible based upon their experience. This tendency, of course, differs from the SEU approach but is comparable to applying the satisficing decision rule (Simon 1955) discussed earlier.

Situation assessment is well recognized as an important element of decision making in naturalistic environments (Klein et al. 1993). Recent research by Klein and his colleagues has examined the possibility of enhancing situation awareness through training (Klein and Wolf 1995). Klein and his colleagues have also applied methods of cognitive task analysis to naturalistic decision-making problems. In these efforts, they have focused on identifying (1) critical decisions, (2) the elements of situation awareness, (3) critical cues indicating changes in situations, and (4) alternative courses of action (Klein 1995). Accordingly, practitioners of naturalistic decision making tend to focus on process-tracing methods and behavioral protocols (Ericsson and Simon 1984) to document the processes people follow when they make decisions.*

---

* Goldstein and Hogarth (1998) describe a similar trend in judgment and decision-making research.

5.1.3. Image Theory

Image theory (Beach 1990) is a descriptive theory of decision making. Beach theorizes that knowledge used to make decisions falls into three categories: value images, trajectory images, and strategic images. The value image describes the decision maker's values, and principles; the trajectory image describes goals; the strategic image describes plans to attain the goals. He also theorizes that there are two types of decisions: adoption decisions and progress decisions. Adoption decisions first involve a screening process where alternatives are eliminated from consideration. The most promising alternative is then selected from the screened set. Progress decisions involve a comparison between goals and the expected result of choosing the alternative.

Two means of evaluating decisions are applied. One test compares the compatibility of the generated alternatives to value images, trajectory images, and strategic images. The profitability test is used to evaluate screened options further in adoption decisions when more than one option survives the initial screening. Beach (1993) argues strongly for the primacy of screening as a characteristic of most real-world decision-making activity.

5.1.4. Contingent Decision Making

The theory of contingent decision making (Beach and Mitchell 1978; Payne et al. 1993) is similar to image theory and cognitive continuum theory (see Section 4.1.4) in that it holds that people use different decision strategies, depending upon the characteristics of the task and the decision context. Payne et al. limit their modeling approach to tasks that require choices to be made (simple memory tasks are excluded from consideration). They also add the assumption that people make choices about how to make choices.*

Choices between decision strategies are assumed to be made rationally by comparing their cost (in terms of cognitive effort) against their benefits (in terms of accuracy). Cognitive effort and accuracy (of a decision strategy) are both assumed to depend upon task characteristics, such as task complexity, response mode, and method of information display. Cognitive effort and accuracy also are assumed to depend upon contextual characteristics, such as the similarity of the compared alternatives, attribute ranges and correlations, the quality of the considered options, reference points, and decision frames. Payne et al. place much emphasis on measuring the cognitive effort of different decision strategies in terms of the number of elemental information elements that must be processed for different tasks and contexts. They relate the accuracy of different decision strategies to task characteristics and contexts and also present research showing that people will shift decision strategies to reduce cognitive effort, increase accuracy, or in response to time pressure.

5.1.5. Dominance Structuring

Dominance structuring (Montgomery 1989) holds that decision making in real contexts involves a sequence of four steps. The process begins with a preediting stage in which alternatives are screened from further analysis. The next step involves selecting a promising alternative from the set of alternatives that survive the initial screening. A test is then made to check whether the promising alternative dominates the other surviving alternations. If dominance is not found, then the information regarding the alternatives is restructured in an attempt to force dominance. This process involves both the bolstering and deemphasizing of information in a way that eliminates disadvantages of the promising alternative.

5.1.6. Explanation-Based Decision Making

Explanation-based decision making (Pennington and Hastie 1986, 1988) assumes that people begin their decision-making process by constructing a mental model that explains the facts they have received. While constructing this explanatory model, people are also assumed to be generating potential alternatives to choose between. The alternatives are then compared to the explanatory model, rather than to the facts from which it was constructed.

Pennington and Hastie have applied this model to juror decision making and obtained experimental evidence that many of its assumptions seem to hold. They note that juror decision making requires consideration of a massive amount of data that is often presented in haphazard order over a long time period. Jurors seem to organize this information in terms of stories describing causation and intent. As part of this process, jurors are assumed to evaluate stories in terms of their uniqueness, plausibility, completeness, or consistency. To determine a verdict, jurors then judge the fit between choices provided by the trial judge and the various stories they use to organize the information.

---

* As such, the theory of contingent decision making directly addresses a potential source of conflict shown in the integrative model of decision making presented earlier (Figure 1). That is, it states that decision makers must choose between decision strategies when they are uncertain how to compare alternatives."

Jurors' certainty about their verdict is assumed to be influenced by both evaluation of stories and the perceived goodness of fit between the stories and the verdict.

5.1.7. Shared Mental Models and Awareness

Orasanu and Salas (1993) discuss two closely related frameworks for describing the knowledge used by teams in naturalistic settings. These are referred to as *shared mental models* and the *team mind*. The common element of these two frameworks is that the members of teams hold knowledge in common and organize it in the same way. Orasanu and Salas claim that this improves and minimizes the need for communication between team members, enables team members to carry out their functions in a coordinated way, and minimizes negotiation over who should do what at what time. Under emergency conditions, Orasanu and Salas claim there is a critical need for members to develop a shared situation model. As evidence for the notion of shared mental models and the team mind, the authors cite research in which firefighting teams and individual firefighters developed the same solution strategies for situations typical of their jobs.

This notion of shared mental models and the team mind can be related to the notion discussed earlier of schemas containing problem-specific rules and facts (Cheng and Holyoak 1985). It also might be reasonable to consider other team members as a form of external memory (Newell and Simon 1972). This approach would have similarities to Wegner's (1987) concept of transactive memory where people in a group know who has specialized information of one kind or another. Klein (1998) provides an interesting discussion of how this metaphor of the team mind corresponds to thinking by individuals. Teams, like people, have a working memory that contains information for a limited time, a long-term or permanent memory, and limited attention. Like people, they also filter out and process information and learn in many ways.

5.1.8. Team Leadership

Torrance (1953) describes retrospective accounts of military survivors lost behind enemy lines indicating that survival depended upon the leader's leadership skills. Important elements of leadership skills included keeping the members of the group focused on a common goal, making sure they knew what needed to be done, and keeping them informed of the current status. Related conclusions concerning the value of keeping people informed have been obtained in retrospective accounts of survivors of mining accidents (Mallet et al. 1993). Orasanu and Salas (1993) cite research in which captains of high-performing air crews explicitly stated more plans, strategies, and intentions to the other members of the crew. They also gave more warnings and predictions to the crew members. Orasanu and Salas cite other work showing that crews performed better with captains who were task oriented and had good personal skills. Performance dropped when captains had negative expressive styles and low task orientation.

A complementary literature has been developed on leadership theory (Chemers and Ayman 1993). Most of this research is based on leaders in organizational contexts. A sampling of factors which have been shown to be related to the effectiveness of leadership include legitimacy, charisma, individualized attention to group members, and clear definitions of goals. These results seem quite compatible with the above findings for leadership in naturalistic, dynamic contexts.

5.2. Time Pressure and Stress

Time pressure and stress are a defining characteristic of naturalistic decision making. Jobs requiring high levels of skill or expertise, such as firefighting, nursing, emergency care, and flying an airplane, are especially likely to involve high stakes, extreme time pressure, uncertainty, or risk to life. The effect of stressors, such as those mentioned above, on performance has traditionally been defined in terms of physiological arousal.* The Yerkes–Dodson law (Yerkes and Dodson 1908) states that the relation between performance and arousal is an inverted U. Either too much or too little arousal causes performance to drop. Too little arousal makes it difficult for people to maintain focused attention. Too much arousal results in errors, more focused attention (and filtering of low-priority information), reduced working memory capacity, and shifts in decision strategies.† One explanation of why performance drops when arousal levels are too high is that arousal consumes cognitive resources that could be allocated to task performance (Mandler 1979).

Time pressure is a commonly studied stressor assumed to impact decision making. Maule and Hockey (1993) note that people tend to filter out low-priority types of information, omit processing information, and accelerate mental activity when they are under time pressure. Variable state activation theory (VSAT) provides a potential explanation of the above effects in terms of a control model of stress regulation (Maule and Hockey 1993). Sequential sampling models provide a compatible perspective on how time pressure can cause changes in performance, such as speed–accuracy trade-offs (see Section 4.2.4). The two approaches are compatible, because VSAT provides a means of modeling how the decision thresholds used within a sequential sampling model might change as a function of time pressure. VSAT also proposes that disequilibriums between control processes and the demands of particular situations can lead to strong affective reactions or feelings of time pressure. Such reactions could, of course, lead to attentional narrowing or reduced working memory capacity and therefore result in poorer task performance. Alternatively, performance might change when decision thresholds are adjusted.

---

*The general adaptation syndrome (Selye 1936, 1979) describes three stages of the human response to stressors. In simplified form, this sequence corresponds to (1) arousal, (2) resistance, and (3) exhaustion.

†The literature on stress and its effects on decision making will not be surveyed here. Books edited by Hamilton and Warburton (1979), Svenson and Maule (1993), Driskell and Salas (1996), and Flin et al. (1997) provide a good introduction to the area.

Time pressure also can cause shifts between the cognitive strategies used in judgment and decision-making situations (Payne et al. 1993; Maule and Hockey 1993; Edland and Svenson 1993). People show a strong tendency to shift to noncompensatory decision rules when they are under time pressure. This finding is consistent with contingency theories of strategy selection (Section 5.1.4). In other words, this shift may be justified when little time is available, because a noncompensatory rule can be applied more quickly. Compensatory decision rules also require more analysis and cognitive effort. Intuitive decision strategies require much less effort because people can rely on their experience or knowledge, and can lead to better decisions in some situations (Hammond et al. 1987). As Klein (1998) points out, stress should impact performance if people use analytical choice procedures.

Novices and experts in novel, unexpected, situations will lack domain experience and knowledge and therefore will have to rely on analytical choice procedures. Consequently, it is not surprising that time pressure and stress have a major negative impact on novice decision makers performing unfamiliar tasks. Interestingly, there is little evidence that stress or time pressure causes experienced personnel to make decision errors in real-world tasks (Klein 1996; Orasanu 1997). The latter finding is consistent with research indicating that experts rely on their experience and intuition when they are under stress and time pressure (Klein 1998). The obvious implication is that training and experience are essential if people are to make good decisions under time pressure and stress.

6. GROUP DECISION MAKING

Much research has been done over the past 25 years or so on decision making by groups and teams. Most of this work has focused on groups, as opposed to teams. In a team, it is assumed that the members are working toward a common goal and have some degree of inderdependence, defined roles and responsibilities, and task-specific knowledge (Orasanu and Salas 1993). Team performance is a major area of interest in the field of naturalistic decision theory (Klein et al. 1993; Klein 1998), as discussed earlier. Group performance has traditionally been an area of study in the fields of organizational behavior and industrial psychology. Traditional decision theory has also devoted some attention to group decision making (Raiffa 1968; Keeney and Raiffa 1976). The following discussion will first briefly discuss some of the ways that group decisions differ from those made by isolated decision makers who need to consider only their own preferences. That is, ethics and social norms play a much more prominent role when decisions are made by or within groups. Attention will then shift to group processes and how they affect group decisions. The last section will address methods of supporting or improving group decision making.

6.1. Ethics and Social Norms

When decisions are made by or within groups, a number of issues arise that have not been touched upon in the earlier portions of this chapter. To start, there is the complication that preferences may vary between members of a group. It often is impossible to maximize the preferences of all members of the group, meaning that trade-offs must be made and issues such as fairness must be addressed to obtain acceptable group decisions. Another complication is that the return to individual decision makers can depend on the actions of others. Game theory* distinguishes two common variations of this situation. In competitive games, individuals are likely to take "self-centered" actions that maximize their own return but reduce returns to other members of the group. Behavior of group members in this situation may be well described by the minimax decision rule discussed in Section 2.1.5. In cooperative games, the members of the group take actions that maximize returns to the group as a whole.

Members of groups may choose cooperative solutions that are better for the group as a whole for many different reasons (Dawes et al. 1988). Groups may apply numerous forms of coercion to punish members who deviate from the cooperative solutions. Group members may apply decision strategies such as reciprocal altruism. They also might conform because of their social conscience, a need for self esteem, or feelings of group identity. Fairness considerations can in some case explain preferences and choices that seem to be in conflict with economic self-interest (Bazerman 1998). Changes in the status quo, such as increasing the price of bottled water immediately after a hurricane, may be viewed as unfair even if they are economically justifiable based on supply and demand. People are often willing to incur substantial costs to punish "unfair" opponents and reward their friends or allies. The notion that costs and benefits should be shared equally is one fairness-related heuristic people use (Messick 1991). Consistent results were found by Guth et al. (1982) in a simple bargaining game where player 1 proposes a split of a fixed amount of cash and player 2 either accepts the offer or rejects it. If player 2 rejects the offer, both players receive nothing. Classical economics predicts that player 2 will accept any positive amount (that is, player 2 should always prefer something to nothing). Consequently, player 1 should offer player 2 a very small amount greater than zero. The results showed that, contrary to predictions of classical economics, subjects tended to offer a substantial proportion of the cash (the average offer was 30%). Some of the subjects rejected positive offers. Others accepted offers of zero. Further research, as summarized by Bolton and Chatterjee (1996), confirms these findings that people seem to care about whether they receive their fair share.

---

* Friedman (1990) provides an excellent introduction to game theory.

Ethics clearly plays an important role in decision making. Some choices are viewed by nearly everyone as being immoral or wrong (i.e., violations of the law, dishonesty, and numerous other behaviors that conflict with basic societal values or behavioral norms). Many corporations and other institutions formally specify codes of ethics prescribing values such as honesty, fairness, compliance with the law, reliability, considerance or sensitivity to cultural differences, courtesy, loyalty, respect for the environment, and avoiding waste. It is easy to visualize scenarios, where it is in the best interest of a decision maker to choose economically undesirable options (at least in the short term) to comply with ethical codes. According to Kidder (1995), the "really tough choices . . . don't center on right versus wrong. They involve right versus right." Kidder refers to four dilemmas of right vs. right he feels qualify as paradigms: (1) truth vs. loyalty (i.e., whether to divulge information provided in confidence), (2) individual vs. community, (3) short term vs. long term, and (4) justice vs. mercy. At least three principles, which in some cases provide conflicting solutions, have been proposed for resolving ethical dilemmas. These include (1) utilitarianism, or selecting the option with the best overall consequences, (2) rule-based, or following a rule regardless of its current consequences (i.e., waiting for a stop light to turn green, even if no cars are coming), and (3) fairness, or doing what you would want others to do for you.

Numerous social dilemmas also occur in which the payoffs to each participant result in individual decision strategies harmful to the group as a whole. The tragedy of the commons (Hardin 1968) is illustrative of social dilemmas in general. For a recent example, discussed in detail by Baron (1998), consider the recent crash of the East Coast commercial fishing industry, brought about by overfishing. Here, the fishing industry as a whole is damaged by overfishing, but individual fishers gain a short-term advantage by catching as many fish as possible. Individual fishers may reason that if they don't catch the fish, someone else will. Each fisher attempts to catch as many fish as possible, even if this will cause the fish stocks to crash. Despite the fact that cooperative solutions, such as regulating the catch, are obviously better than the current situation, individual fishers continue to resist such solutions. Regulations are claimed to infringe on personal autonomy, to be unfair, or to based on inadequate knowledge.

Other similar examples include littering, wasteful use of natural resources, pollution, or social free riding. These behaviors can all be explained, in terms of the choices faced by the offending individual decision maker (Schelling 1978). Simply put, the individual decision maker enjoys the benefits of the offensive behavior, as small as they may be, but the costs are incurred by the entire group.

6.2. Group Processes

A large amount of research has focused on groups and their behavior. Accordingly, many models have been developed that describe how groups make decisions. A common observation is that groups tend to move through several phases as they go through the decision-making process (Ellis and Fisher 1994). One of the more classic models (Tuckman 1965) describes this process with four words: forming, storming, norming, and performing. Forming corresponds to initial orientation, storming to conflict, norming to developing group cohesion and expressing opinions, and performing to obtaining solutions. As implied by Tuckman's choice of terms, there is a continual interplay between socio-emotive factors and rational, task-oriented behavior throughout the group decision-making process. Conflict, despite its negative connotations, is a normal, expected aspect of the group decision process and can in fact serve a positive role (Ellis and Fisher 1994). The following discussion will first address causes and effects of group conflict. Attention will then shift to conflict resolution.

6.2.1. Conflict

Whenever people or groups have different preferences, conflict can occur. As pointed out by Zander (1994), conflict between groups becomes more likely when groups have fuzzy or potentially antagonistic roles, or when one group is disadvantaged (or perceives it is not being treated fairly). A lack of conflict-settling procedures and separation or lack of contact between groups can also contribute to conflict. Conflict becomes especially likely during a crisis and often escalates when the issues are perceived to be important, or after resistance or retaliation occurs. Polarization, loyalty to one's own group, lack of trust, and cultural and socioeconomic factors are often contributing factors to conflict and conflict escalation.

Ellis and Fisher (1994) distinguish between affective and substantive forms of conflict. Affective conflict corresponds to emotional clashes between individuals or groups, while substantive conflict involves opposition at the intellectual level. Substantive conflict is especially likely to have positive effects on group decisions by promoting better understanding of the issues involved. Affective conflict can also improve group decisions by increasing interest, involvement, and motivation among group members and, in some cases, cohesiveness. On the other hand, affective conflict may cause significant ill will, reduced cohesiveness, and withdrawal by some members from the group process. Baron (1998) provides an interesting discussion of violent conflict, and how it is related to polarized beliefs, group loyalty, and other biases.

Defection and the formation of coalitions is a commonly observed effect of conflict, or power struggles, within groups. Coalitions often form when certain members of the group can gain by following a common course of action at the expense of the long-run objectives of the group as a whole. Rapidly changing coalitions between politicians and political parties are obviously a fact of life. Another typical example is when a subgroup of technical employees leave a corporation to form their own small company producing a product similar to one they had been working on. Coalitions, and their formation, have been examined from decision-analytic and game theory perspectives (Bolton and Chatterjee 1996; Raiffa 1982). These approaches make predictions regarding what coalitions will form, depending on whether the parties are cooperating or competing, which have been tested in a variety of experiments (Bolton and Chatterjee 1996). These experiments have revealed that the formation of coalitions is influenced by expected payoffs, equity issues, and the ease of communication. However, Bazerman (1998) notes that the availability heuristic, overconfidence, and sunk cost effects are likely to explain how coalitions actually form in the real world.

6.2.2. Conflict Resolution

Groups resolve conflict in many different ways. Discussion and argument, voting, negotiation, arbitration, and other forms of third-party intervention are all methods of resolving disputes. Discussion and argument is clearly the most common method followed within groups to resolve conflict. Other methods of conflict resolution normally play a complementary, rather than primary, role in the decision process. That is, the latter methods are relied upon when groups fail to reach consensus after discussion and argument, or they simply serve as the final step in the process.

Group discussion and argument is often viewed as being a less than rational process. Along these lines, Brashers et al. (1994) state that the literature suggests ''that argument in groups is a social activity, constructed and maintained in interaction, and guided perhaps by different rules and norms than those that govern the practice of ideal or rational argument. Subgroups speaking with a single voice appear to be a significant force. . . . Displays of support, repetitive agreement, and persistence all appear to function as influence mechanisms in consort with, or perhaps in place of, the quality or rationality of the arguments offered.'' Brashers et al. also suggest that members of groups appear uncritical because their arguments tend to be consistent with social norms rather than the rules of logic: ''[S]ocial rules such as: (a) submission to higher status individuals, (b) experts' opinions are accepted as facts on all matters, (c) the majority should be allowed to rule, (d) conflict and confrontation are to be avoided whenever possible.''

A number of approaches for conflict management have been suggested that attempt to address many of the issues raised by Brashers et al. These approaches include seeking consensus rather than allowing decisions to be posed as win–lose propositions, encouraging and training group members to be supportive listeners, deemphasizing status, depersonalizing decision making, and using facilitators (Likert and Likert 1976). Other approaches that have been proposed include directing discussion toward clarifying the issues, promoting an open and positive climate for discussion, facilitating face-saving communications, and promoting the development of common goals (Ellis and Fisher 1994).

Conflicts can also be resolved through voting and negotiation, as discussed further in Section 6.3. Negotiation becomes especially appropriate when the involved people have competing goals and some form of compromise is required. A typical example would be a dispute over pay between a labor union and management. Strategic concerns play a major role in negotiation and bargaining (Schelling 1960). Self-interest on the part of the involved parties is the driving force throughout a process involving threats and promises, proposals and counterproposals, and attempts to discern how the opposing party will respond. Threats and promises are a means of signaling what the response will be to actions taken by an opponent and consequently become rational elements of a decision strategy

(Raiffa 1982). Establishing the credibility of signals sent to an opponent becomes important because if they are not believed, they will not have any influence.

Methods of attaining credibility include establishing a reputation, the use of contracts, cutting off communication, burning bridges, leaving an outcome beyond control, moving in small steps, and using negotiating agents (Dixit and Nalebuff 1991). Given the fundamentally adversarial nature of negotiation, conflict may move from a substantive basis to an affective, highly emotional state. At this stage, arbitration and other forms of third-party intervention may become appropriate due to a corresponding tendency for the negotiating parties to take extreme, inflexible positions.

6.3. Group Performance and Biases

The quality of the decisions made by groups in a variety of different settings has been seriously questioned. Part of the issue here is the phenomenon of so-called group think, which has been blamed for several disastrous public policy decisions (Hart et al. 1997; Janus 1972). Eight symptoms of groupthink cited by Janis and Mann (1977) are the illusion of invulnerability; rationalization (discounting of warnings and negative feedback); belief in the inherent morality of the group; stereotyping of outsiders; pressure on dissenters within the group; self-censorship; illusion of unanimity; and the presence of mindguards who shield the group from negative information. Janis and Mann proposed that the results of groupthink include failure to consider all the objectives and alternatives, failure to reexamine choices and rejected alternatives, incomplete or poor search for information, failure to adequately consider negative information, and failure to develop contingency plans. Groupthink is one of the most cited characteristics of how group decision processes can go wrong. Given the prominence of groupthink as an explanation of group behavior, it is somewhat surprising that only a few studies have empirically evaluated this theory. Empirical evaluation of the groupthink effect and the development of alternative modeling approaches continue to be an active area of research (Hart et al. 1997).

Other research has attempted to measure the quality of group decisions in the real world against rational, or normative, standards. Viscusi (1991) cites several examples of apparent regulatory complacency and regulatory excess in government safety standards in the United States. He also discusses a variety of inconsistencies in the amounts awarded in product liability cases. Baron (1998) provides a long list of what he views as errors in public decision making and their very serious effects on society. These examples include collective decisions resulting in the destruction of natural resources and overpopulation, strong opposition to useful products such as vaccines, violent conflict between groups, and overzealous regulations, such as the Delaney clause. He attributes these problems to commonly held, and at first glance innocuous, intuitions such as Do no harm, Nature knows best, and Be loyal to your own group, the need for retribution (an eye for an eye), and a desire for fairness.

A significant amount of laboratory research also is available that compares the performance of groups to that of individual decision makers (Davis 1992; Kerr et al. 1996). Much of the early work showed that groups were better than individuals on some tasks. Later research indicated that group performance is less than the sum of its parts. Groups tend to be better than individuals on tasks where the solution is obvious once it is advocated by a single member of the group (Davis 1992; Kerr et al. 1996). Another commonly cited finding is that groups tend to be more willing to select risky alternatives than individuals, but in some cases the opposite is true. One explanation is that group interactions cause people within the group to adopt more polarized opinions (Moscovici 1976). Large groups seem especially likely to reach polarized, or extreme, conclusions (Isenberg 1986). Groups also tend to overemphasize the common knowledge of members, at the expense of underemphasizing the unique knowledge certain members have (Gruenfeld et al. 1996; Stasser and Titus, 1985). A more recent finding indicates that groups were more rational than individuals when playing the ultimatum game (Bornstein and Yaniv 1998).

Duffy (1993) notes that teams can be viewed as information processes and cites team biases and errors that can be related to information-processing limitations and the use of heuristics, such as framing. Topics such as mediation and negotiation, jury decision making, and public policy are now being evaluated from the latter perspective (Heath et al. 1994). Much of this research has focused on whether groups use the same types of heuristics and are subject to the same biases of individuals. This research has shown: (1) framing effects and preference reversals (Paese et al. 1993), (2) overconfidence (Sniezek 1992), (3) use of heuristics in negotiation (Bazerman and Neale 1983), and (4) increased performance with cognitive feedback (Harmon and Rohrbaugh 1990). One study indicated that biasing effects of the representativeness heuristic were greater for groups than for individuals (Argote et al. 1986). The conclusion is that group decisions may be better than those of individuals in some situations but are subject to many of the same problems.

6.4. Prescriptive Approaches

A wide variety of prescriptive approaches have been proposed for improving group decision making. The approaches address some of the above issues, including the use of agendas and rules of order, idea-generating techniques such as brainstorming, nominal group and Delphi techniques, decision structuring, and methods of computer-mediated decision making. As noted by Ellis and Fisher (1994), there is conflicting evidence regarding the effectiveness of such approaches. On the negative side, prescriptive approaches might stifle creativity in some situations and can be sabotaged by dissenting members of groups. On the positive side, prescriptive approaches make the decision process more orderly and efficient, promote rational analysis and participation by all members of the group, and help ensure implementation of group decisions. The following discussion briefly reviews some of these tools for improving group decision making.

6.4.1. Agendas and Rules of Order

Agendas and rules of order are often essential to the orderly functioning of groups. As noted by Welch (1994), an agenda ''conveys information about the structure of a meeting: time, place, persons involved, topics to be addressed, perhaps suggestions about background material or preparatory work.'' Agendas are especially important when the members of a group are loosely coupled or do not have common expectations. Without an agenda, group meetings are likely to dissolve into chaos (Welch 1994). Rules of order, such as *Robert's Rules of Order* (Robert 1990), play a similarly important role, by regulating the conduct of groups to ensure fair participation, by all group members, including absentees. Rules of order also specify voting rules and means of determining consensus. Decision rules may require unanimity, plurality, or majority vote for an alternative.

Attaining consensus poses an advantage over voting, because voting encourages the development of coalitions, by posing the decision as a win–lose proposition (Ellis and Fisher 1994). Members of the group who voted against an alternative are often unlikely to support it. Voting procedures can also play an important role (Davis 1992).

6.4.2. Idea-Generation Techniques

A variety of approaches have been developed for improving the creativity of groups in the early stages of decision making. Brainstorming is a popular technique for quickly generating ideas (Osborn 1937). In this approach, a small group (of no more than 10 individuals) is given a problem to solve. The members are asked to generate as many ideas as possible. Members are told that no idea is too wild and encouraged to build upon the ideas submitted by others. No evaluation or criticism of the ideas is allowed until after the brainstorming session is finished. Buzz group analysis is a similar approach, more appropriate for large groups (Elliss and Fisher 1994). Here, a large group is first divided into small groups of four to six members. Each small group goes through a brainstorming-like process to generate ideas. They then present their best ideas to the entire group for discussion. Other commonly applied idea-generating techniques include focus group analysis and group exercises intended to inspire creative thinking through role playing (Elliss and Fisher 1994; Clemen 1996).

The use of brainstorming and the other idea-generating methods mentioned above will normally provide a substantial amount of, in some cases, creative suggestions, especially when participants build upon each other's ideas. However, personality factors and group dynamics can also lead to undesirable results. Simply put, some people are much more willing than others to participate in such exercises. Group discussions consequently tend to center around the ideas put forth by certain more forceful individuals. Group norms, such as deferring to participants with higher status and power, may also lead to undue emphasis on the opinions of certain members.

6.4.3. Nominal Group and Delphi Technique

Nominal group technique (NGT) and the Delphi technique attempt to alleviate some of the disadvantages of working in groups (Delbecq et al. 1975). The nominal group technique consists of asking each member of a group to write down and think about his or her ideas independently. A group moderator then asks each member to present one or more of his or her ideas. Once all of the ideas have been posted, the moderator allows discussion to begin. After the discussion is finished, each participant rates or ranks the presented ideas. The subject ratings are then used to develop a score for each idea. Nominal group technique is intended to increase participation by group members and is based on the idea that people will be more comfortable presenting their ideas if they have a chance to think about them first (Delbecq et al. 1975).

The Delphi technique allows participants to comment anonymously, at their leisure, on proposals made by other group members. Normally, the participants do not know who proposed the ideas they are commenting on. The first step is to send an open-ended questionnaire to members of the group. The results are then used to generate a series of follow-up questionnaires in which more specific questions are asked. The anonymous nature of the Delphi process theoretically reduces the effect of participant status and power. Separating the participants also increases the chance that members will provide opinions ''uncontaminated'' by the opinions of others.

6.4.4. Structuring Group Decisions

As discussed earlier in this chapter, the field of decision analysis has devised several methods for organizing or structuring the decision-making process. The rational reflection model (Siebold 1992) is a less formal, six-step procedure that serves a similar function. Group members are asked first to define and limit the problem by identifying goals, available resources, and procedural constraints. After defining and limiting the problem, the group is asked to analyze the problem, collect relevant information, and establish the criteria a solution must meet. Potential solutions are then discussed in terms of the agreed-upon decision criteria. After further discussion, the group selects a solution and determines how it should be implemented. The focus of this approach is on forcing the group to confine its discussion to the issues that arise at each step in the decision making process. As such, this method is similar to specifying an agenda.

Raiffa (1982) provides a somewhat more formal decision-analytic approach for structuring negotiations. The approach begins by assessing (1) the alternatives to a negotiated settlement, (2) the interests of the involved parties, and (3) the relative importance of each issue. This assessment allows the negotiators to think analytically about mutually acceptable solutions. In certain cases, a bargaining zone is available. For example, an employer may be willing to pay more than the minimum salary acceptable to a potential employee. In this case, the bargaining zone is the difference between the maximum salary the employer is willing to pay and the minimum salary a potential employee is willing to accept. The negotiator may also think about means of expanding the available resources to be divided, potential trading issues, or new options that satisfy the interests of the concerned parties.

Other methods for structuring group preferences are discussed in Keeny and Raiffa (1976). The development of group utility functions is one such approach. A variety of computer-mediated methods for structuring group decisions are also available, as discussed below.

6.4.5. Computer-Mediated Group Decision Making

Computer tools for helping groups make decisions are now available that implement all of the approaches discussed above to varying degrees. The spectrum of available tools ranges from traditional tools used in decision analysis, such as the analytic hierarchy process (Basak and Saaty 1993; Saaty 1988), to electronic meeting places (Mockler and Dologite 1991; Nunamaker et al. 1991) and group decision support systems (GDSS) (Sage 1997). Some of the many functions provided by GDSS (Johansen 1988; Sage 1997) include computer-supported presentations, project and calendar management, group authoring, electronic meeting places, audio and visual conferencing, screen sharing, group e-mail services, text-filtering services, conversational structuring, group and organizational memory management, and comprehensive work team support.

Computer-mediated group decision making has several potential benefits (Brasher et al. 1994), including (1) enabling all participants to work simultaneously (they don't have to wait their turn to speak), (2) providing a more equal and potentially anonymous opportunity to be heard, (3) providing a more structured environment (that is, a more linear process and control of the agenda). Computer-mediated group decision making also might make it easier to control and manage conflict, through the use of facilitators and convenient voting procedures. As noted by Sage (1997), the purpose of GDSS is to (1) remove communication barriers, (2) provide techniques for the formulation, analysis, and interpretation of decisions, and (3) systematically direct the decision-making process. Successfully attaining these objectives means that a GDSS must provide the right information to decision makers, at an appropriate level of detail, at the time it is needed, in a form that is conveniently applied. Simply put, to be useful, a GDSS must not make group decision making more difficult than it already is.

7. SUMMARY CONCLUSIONS

Beach (1993) discusses four revolutions in behavioral decision theory. The first took place when it was recognized that the evaluation of alternatives is seldom extensive. It is illustrated by use of the satisficing rule (Simon 1955) and heuristics (Tversky and Kahneman 1974) rather than optimizing. The second occurred when it was recognized that people choose between strategies to make decisions. It is marked by the development of contingency theory (Beach 1990) and cognitive continuum theory (Hammond 1980). The third is currently occurring. It involves the realization that people rarely make choices and instead rely on prelearned procedures. This perspective is illustrated by the levels-of-processing approach (Rasmussen 1983) and recognition-primed decisions (Klein 1989). The fourth is just beginning. It involves recognization that decision-making research must abandon a single-minded focus on the economic view of decision making and include approaches drawn from relevant developments and research in cognitive psychology, organizational behavior, and systems theory.

The discussion within this chapter parallels this view of decision making. The integrative model presented at the beginning of the chapter shows how these various approaches fit together as a whole. Each path through the model is distinguished by specific sources of conflict, the methods of conflict resolution followed, and the types of decision rules used to analyze the results of conflict-resolution processes. The different paths through the model correspond to fundamentally different ways of making decisions, ranging from routine situation assessment-driven decisions to satisficing, analysis of single- and multiattribute expected utility, and even obtaining consensus of multiple decision makers in group contexts. Numerous other strategies discussed in this chapter are also described by particular paths through the model.

This chapter goes beyond simply describing methods of decision making by pointing out reasons people and groups may have difficulty making good decisions. These include cognitive limitations, inadequacies of various heuristics used, biases and inadequate knowledge of decision makers, and task-related factors, such as risk, time pressure, and stress. The discussion also provides insight into the effectiveness of approaches for improving human decision making. The models of selective attention point to the value of providing only truly relevant information to decision makers. Irrelevant information might be considered simply because it is there, especially if it is highly salient. Methods of highlighting or emphasizing relevant information, therefore, clearly seem to be warranted. The models of selective information also indicate that methods of helping decision makers cope with working memory limitations will be of value. There also is reason to believe that providing feedback to decision makers in dynamic decision-making situations will be useful. Cognitive, rather than outcome, feedback is indicated as being particularly helpful when decision makers are learning. Training decision makers also seems to offer potentially large benefits. One reason for this conclusion is that the studies of naturalistic decision making revealed that most decisions are made on a routine, nonanalytical basis.

The studies of debiasing also partially support the potential benefits of training and feedback. On the other hand, the many failures to debias expert decision makers imply that decision aids, methods of persuasion, and other approaches intended to improve decision making are no panacea. Part of the problem is that people tend to start with preconceived notions about what they should do and show a tendency to seek out and bolster confirming evidence. Consequently, people show a tendency to develop overconfidence with experience, and strongly held beliefs become difficult to modify, even if they are hard to defend rationally.

REFERENCES

Ainslie, G. (1975), “Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control,” *Psychological Bulletin,* Vol. 82, pp. 463–509.

Allais, M. (1953), “Le comportement de l’homme rationel devant le risque: Critique des postulates et axioms de l’ecole americaine,” *Econometrica,* Vol. 21, pp. 503–546.

Anderson, N. H. (1981), *Foundations of Information Integration Theory,* Academic Press, New York.

Argote, L., Seabright, M. A., and Dyer, L. (1986), “Individual versus Group: Use of Base-Rate and Individuating Information,” *Organizational Behavior and Human Decision Making Processes,* Vol. 38, pp. 65–75.

Arkes, H. R., and Blumer, C. (1985), “The Psychology of Sunk Cost,” *Organizational Behavior and Human Decision Processes,* Vol. 35, pp. 124–140.

Balzer, W. K., Doherty, M. E., and O’Connor, R. O., Jr. (1989), “Effects of Cognitive Feedback on Performance,” *Psychological Bulletin,* Vol. 106, pp. 41–433.

Bar-Hillel, M. (1973), “On the Subjective Probability of Compound Events,” *Organizational Behavior and Human Performance,* Vol. 9, pp. 396–406.

Baron, J. (1985), *Rationality and Intelligence,* Cambridge University Press, Cambridge.

Baron, J. (1998), *Judgment Misguided: Intuition and Error in Public Decision Making,* Oxford University Press, New York.

Basak, I., and Saaty, T. (1993), “Group Decision Making Using the Analytic Hierarchy Process,” *Mathematical and Computer Modeling,* Vol. 17, pp. 101–109.

Bazermen, M. (1998), *Judgment in Managerial Decision Making,* 4th Ed., Wiley & Sons, New York.

Bazerman, M. H., and Neale, M. A. (1983), “Heuristics in Negotiation: Limitations to Effective Dispute Resolution,” in *Negotiating in Organizations,* M. H. Bazerman and R. Lewicki, Eds., Sage, Beverly Hills.

Beach, L. R. (1990), *Image Theory: Decision Making in Personal and Organizational Contexts,* John Wiley & Sons, Chichester.

Beach, L. R. (1993), “Four Revolutions in Behavioral Decision Theory,” in *Leadership Theory and Research,* M. M. Chemers and R. Ayman, Eds., Academic Press, San Diego.

Beach, L. R., and Mitchell, T. R. (1978), “A Contingency Model for the Selection of Decision Strategies,” *Academy of Management Journal,* Vol. 3, pp. 439–449.

Bell, D. (1982), “Regret in Decision Making under Uncertainty,” *Operations Research,* Vol. 30, pp. 961–981.

Bernoulli, D. (1738), *Exposition of a New Theory of the Measurement of Risk,* Imperial Academy of Science, St. Petersburg.

Birnbaum, M. H., and Mellers, B. A. (1983), "Bayesian Inference: Combining Base Rates with Opinions of Sources Who Vary in Credibility," *Journal of Personality and Social Psychology,* Vol. 37, pp. 792–804.

Birnbaum, M. H., Coffey, G., Mellers, B. A., and Weiss, R. (1992), "Utility Measurement: Configural-Weight Theory and the Judge's Point of View," *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 18, pp. 331–346.

Bock, R. D., and Jones, L. V. (1968), *The Measurement and Prediction of Judgment and Choice,* Holden-Day, San Francisco.

Bolton, G. E., and Chatterjee, K. (1996), "Coalition Formation, Communication, and Coordination: An Exploratory Experiment," in *Wise Choices: Decisions, Games, and Negotiations,* R. J. Zeckhauser, R. L. Keeney, and J. K. Sebenius, Eds., Harvard Business School Press, Boston.

Bornstein, G., and Yaniv, I. (1998), "Individual and Group Behavior in the Ultimatum Game: Are Groups More Rational Players?" *Experimental Economics,* Vol. 1, pp. 101–108.

Brashers, D. E., Adkins, M., and Meyers, R. A. (1994), "Argumentation and Computer-Mediated Group Decision Making," 12 in *Group Communication in Context,* L. R. Frey, Ed., Erlbaum, Hillsdale, NJ.

Brehmer, B. (1981), "Models of Diagnostic Judgment," in *Human Detection and Diagnosis of System Failures,* J. Rasmussen and W. Rouse, Eds., Plenum Press, New York.

Brehmer, B., and Joyce, C. R. B. (1988), *Human Judgment: The SJT View,* North-Holland, Amsterdam.

Brookhouse, J. K., Guion, R. M., and Doherty, M. E. (1986), "Social Desirability Response Bias as One Source of the Discrepancy between Subjective Weights and Regression Weights," *Organizational Behavior and Human Decision Processes,* Vol. 37, pp. 316–328.

Brunswick, E. (1952), *The Conceptual Framework of Psychology,* University of Chicago Press, Chicago.

Buck, J. R. (1989), *Economic Risk Decisions in Engineering and Management,* Iowa State University Press, Ames, IA.

Budescu, D., and Weiss, W. (1987), "Reflection of Transitive and Intransitive Preferences: A Test of Prospect Theory," *Organizational Behavior and Human Performance,* Vol. 39, pp. 184–202.

Busemeyer, J. R., and Townsend, J. T. (1993), "Decision Field Theory: A Dynamic–Cognitive Approach to Decision Making in an Uncertain Environment," *Psychological Review,* Vol. 100, pp. 432–459.

Caverni, J. P., Fabre, J. M., and Gonzalez, M. (1990), *Cognitive Biases,* North-Holland, Amsterdam.

Chemers, M. M., and Ayman, R. (eds) (1993), *Leadership Theory and Research,* Academic Press, San Diego.

Cheng, P. E., and Holyoak, K. J. (1985), "Pragmatic Reasoning Schemas," *Cognitive Psychology,* Vol. 17, pp. 391–416.

Christensen-Szalanski, J. J., and Willham, C. F. (1991), "The Hindsight Bias: A Meta-Analysis," *Organizational Behavior and Human Decision Processes,* Vol. 48, pp. 147–168.

Clemen, R. T. (1996), *Making Hard Decisions: An Introduction to Decision Analysis,* 2nd Ed., Duxbury Press, Belmont, CA.

Cohen, M. S. (1993), "The Naturalistic Basis of Decision Biases," in *Decision Making in Action: Models and Methods,* G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ, pp. 51–99.

Connolly, T., Ordonez, L. D., and Coughlan, R. (1997), "Regret and Responsibility in the Evaluation of Decision Outcomes," *Organizational Behavior and Human Decision Processes,* Vol. 70, pp. 73–85.

Davis, J. H. (1992), "Some Compelling Intuitions about Group Consensus Decisions, Theoretical and Empirical Research, and Interperson Aggregation Phenomena: Selected Examples, 1950–1990," *Organizational Behavior and Human Decision Processes,* Vol. 52, pp. 3–38.

Dawes, R. M., and Mulford, M. (1996), "The False Consensus Effect and Overconfidence: Flaws in Judgement or Flaws in How We Study Judgement?" *Organizational Behavior and Human Decision Processes,* Vol. 65, pp. 201–211.

Dawes, R. M., van de Kragt, A. J. C., and Orbell, J. M. (1988), "Not Me or Thee but We: The Importance of Group Identity in Eliciting Cooperation in Dilemma Situations: Experimental Manipulations," *Acta Psychologia,* Vol. 68, pp. 83–97.

Delbecq, A. L., Van de Ven, A. H., and Gustafson, D. H. (1975), *Group Techniques for Program Planning,* Scott, Foresman, Glenview, IL.

Dixit, A., and Nalebuff, B. (1991), "Making Strategies Credible," in *Strategy and Choice,* R. J. Zechhauser, Ed., MIT Press, Cambridge, MA, pp. 161–184.

Dorris, A. L., and Tabrizi, J. L. (1978), "An Empirical Investigation of Consumer Perception of Product Safety," *Journal of Products Liability,* Vol. 2, pp. 155–163.

Driskell, J. E., and Salas, E., Eds. (1996), *Stress and Human Performance,* Erlbaum, Hillsdale, NJ.

Drucker, P. F. (1985), *The Effective Executive,* Harper Row, New York.

Du Charme, W. (1970), "Response Bias Explanation of Conservative Human Inference," *Journal of Experimental Psychology,* Vol. 85, pp. 66–74.

Duda, R. O., Hart, K., Konolige, K., and Reboh, R. (1979), "A Computer-Based Consultant for Mineral Exploration," Technical Report, SRI International, Stanford, CA.

Duffy, L. (1993), "Team Decision Making Biases: An Information Processing Perspective," in *Decision Making in Action: Models and Methods,* G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ.

Edland, E., and Svenson, O. (1993), "Judgment and Decision Making under Time Pressure," in *Time Pressure and Stress in Human Judgment and Decision Making,* O. Svenson, and A. J. Maule, Eds., Plenum, New York.

Edwards, W. (1954), "The Theory of Decision Making," *Psychological Bulletin,* Vol. 41, pp. 380–417.

Edwards, W. (1968), "Conservatism in Human Information Processing," in *Formal Representation of Human Judgment,* B. Kleinmuntz, Ed., John Wiley & Sons, New York, pp. 17–52.

Einhorn, H. J., and Hogarth, R. M. (1978), "Confidence in Judgment: Persistence of the Illusion of Validity," *Psychological Review,* Vol. 70, pp. 193–242.

Einhorn, H. J., and Hogarth, R. M. (1981), "Behavioral Decision Theory: Processes of Judgment and Choice," *Annual Review of Psychology,* Vol. 32, pp. 53–88.

Ellsberg, D. (1961), "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics,* Vol. 75, pp. 643–699.

Elster, J., Ed. (1986), *The Multiple Self,* Cambridge University Press, Cambridge.

Embrey, D. E. (1984), *SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment,* NUREG/CR-3518, Vols. 1 and 2, U.S. Nuclear Regulatory Commission, Washington, DC.

Ericsson, K. A., and Simon, H. A. (1984), *Protocol Analysis: Verbal Reports as Data,* MIT Press, Cambridge, MA.

Etzioni, A. (1988), "Normative-Affective Factors: Toward a New Decision-Making Model," *Journal of Economic Psychology,* Vol. 9, pp. 125–150.

Evans, J. B. T. (1989), *Bias in Human Reasoning: Causes and Consequences,* Erlbaum, London.

Evans, J. B. T., and Pollard, P. (1985), "Intuitive Statistical Inferences about Normally Distributed Data," *Acta Psychologica,* Vol. 60, pp. 57–71.

Feather, N. T. (1966), "Effects of Prior Success and Failure on Expectations of Success and Failure," *Journal of Personality and Social Psychology,* Vol. 3, pp. 287–298.

Fedrizzi, M., Kacprzyk, J., and Yager, R. R., Eds. (1994), *Decision Making under Dempster–Shafer Uncertainties,* John Wiley & Sons, New York.

Fischhoff, B. (1982), "For Those Condemned to Study the Past: Heuristics and Biases in Hindsight," in *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic, and A. Tversky, Eds., Cambridge University Press, Cambridge.

Fischhoff, B., and MacGregor, D. (1982), "Subjective Confidence in Forecasts," *Journal of Forecasting,* Vol. 1, pp. 155–172.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977), "Knowing with Certainty: The Appropriateness of Extreme Confidence," *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 3, pp. 552–564.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978), "Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation," *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 4, pp. 330–344.

Fishburn, P. C. (1974), "Lexicographic Orders, Utilities, and Decision Rules: A Survey," *Management Science,* Vol. 20, pp. 1442–1471.

Flin R., Salas, E., Strub, M., and Martin, L., Eds. (1997), *Decision Making under Stress: Emerging Themes and Applications,* Ashgate, Aldershot, UK.

Fong, G. T., Krantz, D. H., and Nisbett, R. E. (1986), "The Effects of Statistical Training on Thinking about Everyday Problems," *Cognitive Psychology,* Vol. 18, pp. 253–292.

Friedman, J. W. (1990), *Game Theory with Applications to Economics,* Oxford University Press, New York.

Frisch, D., and Clemen, R. T. (1994), "*Beyond Expected Utility: Rethinking Behavioral Decision Research,*" *Psychological Bulletin,* Vol. 116, No. 1, pp. 46–54.

Gertman, D. I., and Blackman, H. S. (1994), *Human Reliability and Safety Analysis Data Handbook,* John Wiley & Sons, New York.

Goldstein, W. M., and Hogarth, R. M. (1997), "Judgment and Decision Research: Some Historical Context," in *Research on Judgment and Decision Making: Currents, Connections, and Controversies,* W. M. Goldstein, and R. M. Hogarth, Eds., Cambridge University Press, Cambridge, pp. 3–65.

Gordon, J., and Shortliffe, E. H. (1984), "The Dempster–Schafer Theory of Evidence," in *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project,* Addison-Wesley, Reading, MA.

Gruenfeld, D. H., Mannix, E. A., Williams, K. Y., and Neale, M. A. (1996), "Group Composition and Decision Making: How Member Familiarity and Information Distribution Affect Process and Performance," *Organizational Behavior and Human Decision Making Processes,* Vol. 67:1, pp. 1–15.

Güth, W., Schmittberger, R., and Schwarze, B. (1982), "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization,* Vol. 3, pp. 367–388.

Hamilton, V., and Warburton, D. M., Eds. (1979), *Human Stress and Cognition,* John Wiley & Sons, New York.

Hammer, W. (1993), *Product Safety Management and Engineering,* 2nd Ed., ASSE, Chicago.

Hammond, K. R. (1980), "Introduction to Brunswikian Theory and Methods," in *Realizations of Brunswick's Experimental Design,* K. R. Hammond, and N. E. Wascoe, Eds., Jossey-Bass, San Francisco.

Hammond, K. R. (1993), "Naturalistic Decision Making From a Brunswikian Viewpoint: Its Past, Present, Future," in *Decision Making in Action: Models and Methods,* G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ, pp. 205–227.

Hammond, K. R, Stewart, T. R., Brehmer, B., and Steinmann, D. O. (1975), "Social Judgment Theory," in *Human Judgment and Decision Processes,* M. F. Kaplan, and S. Schwartz, Eds., Academic Press, New York, pp. 271–312.

Hammond, K. R., Hamm, R. M., Grassia, J., and Pearson, T. (1987), "Direct Comparison of the Efficacy of Intuitive and Analytical Cognition in Expert Judgment," *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-17, pp. 753–770.

Hardin, G. (1968), "The Tragedy of the Commons," *Science,* Vol. 162, pp. 1243–1248.

Harmon, J., and Rohrbaugh, J. (1990), "Social Judgement Analysis and Small Group Decision Making: Cognitive Feedback Effects on Individual and Collective Performance," *Organizational Behavior and Human Decision Processes,* Vol. 46, pp. 34–54.

Hart, P., Stern, E. K., and Sundelius, B. (1997), *Beyond Groupthink: Political Group Dynamics and Foreign Policy-Making,* University of Michigan Press, Ann Arbor.

Heath, C. (1995), "Escalation and De-escalation of Commitment in Response to Sunk Costs: The Role of Budgeting in Mental Accounting," *Organizational Behavior and Human Decision Processes,* Vol. 62, pp. 38–54.

Heath, L., Tindale, R. S., Edwards, J., Posavac, E. J., Bryant, F. B., Henderson-King, E., Suarez-Balcazar, Y., and Myers, J. (1994), *Applications of Heuristics and Biases to Social Issues,* Plenum Press, New York.

Hogarth, R. M., and Einhorn, H. J. (1992), "Order Effects in Belief Updating: The Belief-Adjustment Model," *Cognitive Psychology,* Vol. 24, pp. 1–55.

Holtzman, S. (1989), *Intelligent Decision Systems,* Addison-Wesley, Reading, MA.

Howard, R. A. (1968), "The Foundations of Decision Analysis," *IEEE Transactions on Systems, Science, and Cybernetics,* Vol. SSC-4, pp. 211–219.

Howard, R. A. (1988), "Decision Analysis: Practice and Promise," *Management Science,* Vol. 34, pp. 679–695.

Huber, J., Wittink, D. R., Fiedler, J. A., and Miller, R. (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research,* Vol. 30, pp. 105–114.

Isen, A. M. (1993), "Positive Affect and Decision Making," in *Handbook of Emotions,* M. Lewis, and J. M. Haviland, Eds., Guilford Press, New York, pp. 261–277.

Isenberg, D. J. (1986), "Group Polarization: A Critical Review and Meta Analysis," *Journal of Personality and Social Psychology,* Vol. 50, pp. 1141–1151.

Iverson, G., and Luce, R. D. (1998), "The Representational Measurement Approach to Psychophysical and Judgmental Problems," in *Measurement, Judgement, and Decision Making,* M. H. Birnbaum, Ed., Academic Press, San Diego, pp. 1–79.

Jacoby, J. (1977), "Information Load and Decision Quality: Some Contested Issues," *Journal of Marketing Research,* Vol. 14, pp. 569–573.

Janis, I. L. (1972), *Victims of Groupthink,* Houghton-Mifflin, Boston.

Janis, I. L., and Mann, L. (1977), *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment,* Free Press, New York.

Johansen, R. (1988), *Groupware: Computer Support for Business Teams,* Free Press, New York.

Johnson, E. J., and Payne, J. W. (1985), "Effort and Accuracy in Choice," *Management Science,* Vol. 31, No. 4, pp. 395–414.

Kahneman, D., and Tversky, A. (1973), "On the Psychology of Prediction," *Psychological Review,* Vol. 80, pp. 251–273.

Kahneman, D., and Tversky, A. (1979), "Prospect Theory: An Analysis of Decision under Risk," *Econometrica,* Vol. 47, pp. 263–291.

Kahneman, D., Slovic, P., and Tversky, A. (1982), *Judgment under Uncertainty: Heuristics and Biases,* Cambridge University Press, Cambridge.

Keeney, R. L., and Raiffa, H. (1976), *Decisions with Multiple Objectives: Preferences and Value Tradeoffs,* John Wiley & Sons, New York.

Kerr, L. N., MacCoun, R. J., and Kramer, G. P. (1996), "Bias in Judgment: Comparing Individuals and Groups," *Psychological Review,* Vol. 103, pp. 687–719.

Kieras, D. E. (1985), "The Role of Prior Knowledge in Operating Equipment from Written Instructions," Report No. 19 (FR-85/ONR-19), Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor.

Keren, G. (1990), "Cognitive Aids and Debiasing Methods: Can Cognitive Pills Cure Cognitive Ills?" in *Cognitive Biases,* J. P. Caverni, J. M. Fabre, and M. Gonzalez, Eds., Amsterdam, North Holland.

Klayman, J., and Ha, Y. W. (1987), "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Journal of Experimental Psychology: Human Learning and Memory,* pp. 211–228.

Klein, G. A. (1989), "Recognition-Primed Decisions," in Advances in Man-Machine System Research, W. Rouse, Ed., CT: JAI Press, Greenwich.

Klein, G. A. (1995), "The Value Added by Cognitive Analysis," in *Proceedings of the Human Factors and Ergonomics Society—39th Annual Meeting,* pp. 530–533.

Klein, G. A. (1996), "The Effect of Acute Stressors on Decision Making," in *Stress and Human Performance,* J. E. Driskell, and E. Salas, Eds., Erlbaum, Hillsdale, NJ.

Klein, G. A. (1998), *Sources of Power: How People Make Decisions,* MIT Press, Cambridge, MA.

Klein, G. A., and Wolf, S. (1995), "Decision-Centered Training," *Proceedings of the Human Factors and Ergonomics Society—39th Annual Meeting,* pp. 1249–1252.

Klein, G. A., Orasanu, J., Calderwood, R., and Zsambok, E., Eds. (1993), *Decision Making in Action: Models and Methods,* Ablex, Norwood, NJ.

Kleinmuntz, B. (1984), "The Scientific Study of Clinical Judgment in Psychology and Medicine," *Clinical Psychology Review,* Vol. 4, pp. 111–126.

Koehler, J. J. (1996), "The Base Rate Fallacy Reconsidered: Descriptive, Normative, and Methodological Challenges," *Behavioral and Brain Sciences,* Vol. 19, pp. 1–53.

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980), "Reasons for Confidence," *Journal of Experimental Psychology: Human Learning and Memory,* Vol. 6, pp. 107–118.

Kraus, N. N., and Slovic, P. (1988), "Taxonomic Analysis of Perceived Risk: Modeling Individual and Group Perceptions within Homogenous Hazards Domains," *Risk Analysis,* Vol. 8, pp. 435–455.

Kuhberger, A. (1995), "The Framing of Decisions: A New Look at Old Problems," *Organizational Behavior and Human Decision Processes,* Vol. 62, pp. 230–240.

Lathrop, R. G. (1967), "Perceived Variability," *Journal of Experimental Psychology,* Vol. 23, pp. 498–502.

Lehto, M. R. (1991), "A Proposed Conceptual Model of Human Behavior and Its Implications for Design of Warnings," *Perceptual and Motor Skills,* Vol. 73, pp. 595–611.

Lehto, M. R., and Papastavrou, J. (1991), "A Distributed Signal Detection Theory Model: Implications to the Design of Warnings," in *Proceedings of the 1991 Automatic Control Conference* (Boston), pp. 2586–2590.

Lehto, M. R., Papastavrou, J. P., Ranney, T. A., and Simmons, L. (2000), "An Experimental Comparison of Conservative vs Optimal Collision Avoidance System Thresholds," *Safety Science,* Vol. 36, No. 3, pp. 185–209.

Lehto, M. R., James, D. S., and Foley, J. P. (1994), "Exploratory Factor Analysis of Adolescent Attitudes Toward Alcohol and Risk," *Journal of Safety Research,* Vol. 25, pp. 197–213.

Levin, L. P. (1975), "Information Integration in Numerical Judgements and Decision Processes," *Journal of Experimental Psychology: General,* Vol. 104, pp. 39–53.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982), "Calibration of Probabilities: The State of the Art to 1980," in *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic, and A. Tversky, Eds., pp. 306–334.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., and Coombs, B. (1978), "Judged Frequency of Lethal Events," *Journal of Experimental Psychology: Human Learning & Memory,* Vol. 4, pp. 551–578.

Likert, R., and Likert, J. G. (1976), *New Ways of Managing Conflict,* McGraw-Hill, New York.

Luce, R. D., and Raiffa, H. (1957), *Games and Decisions,* John Wiley & Sons, New York.

Mallet, L., Vaught, C., and Brnich, M. J., Jr. (1993), "Sociotechnical Communication in an Underground Mine Fire: A Study of Warning Messages During an Emergency Evacuation," *Safety Science,* Vol. 16, pp. 709–728.

Martocchio, J. J., Webster, J., and Baker, C. R. (1993), "Decision-Making in Management Information Systems Research: The Utility of Policy Capturing Methodology," *Behaviour and Information Technology,* Vol. 12, pp. 238–248.

Maule, A. J., and Hockey, G. R. J. (1993), "State, Stress, and Time Pressure," in *Time Pressure and Stress in Human Judgment and Decision Making,* O. Svenson, and A. J. Maule, Eds., Plenum, New York, pp. 27–40.

McGuire, W. J. (1966), "Attitudes and Opinions," *Annual Review of Psychology,* Vol. 17, pp. 475–514.

McKnight, A. J., Langston, E. A., McKnight, A. S., and Lange, J. E. (1995), "The Bases of Decisions Leading to Alcohol Impaired Driving," in *Proceedings of the 13th International Conference on Alcohol, Drugs, and Traffic Safety* (Adelaide, August 13–18), C. N. Kloeden and A. J. McLean, Eds., pp. 143–147.

Messick, D. M. (1991), "Equality as a Decision Heuristic," in *Psychological Issues in Distributive Justice,* B. Mellers, Ed., Cambridge University Press, New York.

Mockler, R. L., and Dologite, D. G. (1991), "Using Computer Software to Improve Group Decision-Making," *Long Range Planning,* Vol. 24, pp. 44–57.

Montgomery, H. (1989), "From Cognition to Action: The Search for Dominance in Decision Making," in *Process and Structure in Human Decision Making,* H. Montgomery and O. Svenson, Eds., John Wiley & Sons, Chichester.

Moscovici, S. (1976), *Social Influence and Social Change,* Academic Press, London.

Murphy, A. H., and Winkler, R. L. (1974), "Probability Forecasts: A Survey of National Weather Service Forecasters," *Bulletin of the American Meteorological Society,* Vol. 55, pp. 1449–1453.

Myers, J. L., Suydam, M. M., and Gambino, B. (1965), "Contingent Gains and Losses in a Risky Decision Situation," *Journal of Mathematical Psychology,* Vol. 2, pp. 363–370.

Naruo, N., Lehto, M., and Salvendy, G. (1990), "Development of a Knowledge Based Decision Support System for Diagnosing Malfunctions of Advanced Production Equipment," *International Journal of Production Research,* Vol. 28, pp. 2259–2276.

Navon, D. (1979), "The Importance of Being Conservative," *British Journal of Mathematical and Statistical Psychology,* Vol. 31, pp. 33–48.

Newell, A., and Simon, H. A. (1972), *Human Problem Solving,* Prentice-Hall, Englewood Cliffs, NJ.

Nisbett, R., and Ross, L. (1980), *Human Inference: Strategies and Shortcomings of Social Judgment,* Prentice-Hall, Englewood Cliffs, NJ.

Nunamaker, J. F., Dennis, A. R., Valacich, J. S., Vogel, D. R., and George, J. F. (1991), "Electronic Meeting Systems to Support Group Work: Theory and Practice at Arizona," *Communications of the ACM,* Vol. 34, pp. 40–61.

Orasanu, J. (1997), "Stress and Naturalistic Decision Making: Strengthening the Weak Links," in. *Decision Making under Stress: Emerging Themes and Applications,* R. Flin, E. Salas, M. Strub, and L. Martin, Eds., Ashgate, Aldershot, UK.

Orasanu, J., and Salas, E. (1993), "Team Decision Making in Complex Environments," in *Decision Making in Action: Models and Methods,* G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ.

Osborn, F. (1937), *Applied Imagination,* Charles Scribner & Sons, New York.

Paese, P. W., Bieser, M., and Tubbs, M. E. (1993), "Framing Effects and Choice Shifts in Group Decision Making," *Organizational Behavior and Human Decision Processes,* Vol. 56, pp. 149–165.

Payne, J. W. (1980), "Information Processing Theory: Some Concepts and Methods Applied to Decision Research," in *Cognitive Processes in Choice and Decision Research,* T. S. Wallsten, Ed., Erlbaum, Hillsdale, NJ.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993), *The Adaptive Decision Maker,* Cambridge University Press, Cambridge.

Pennington, N., and Hastie, R. (1986), "Evidence Evaluation in Complex Decision Making," *Journal of Personality and Social Psychology,* Vol. 51, pp. 242–258.

Pennington, N., and Hastie, R. (1988), "Explanation-Based Decision Making: Effects of Memory Structure on Judgment," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* Vol. 14, pp. 521–533.

Pitz, G. F. (1980), "The Very Guide of Life: The Use of Probabilistic Information for Making Decisions," in *Cognitive Processes in Choice and Decision Behavior,* T. S. Wallsten, Ed., Erlbaum, Hillsdale, NJ.

Raiffa, H. (1968), *Decision Analysis,* Addison-Wesley, Reading, MA.

Raiffa, H. (1982), *The Art and Science of Negotiation,* Harvard University Press, Cambridge, MA.

Rasmussen, J. (1983), "Skills, Rules, Knowledge: Signals, Signs, and Symbols and Other Distinctions in Human Performance Models," *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-13, No. 3, pp. 257–267.

Reason, J. (1990), *Human Error,* Cambridge University Press, Cambridge.

Rethans, A. J. (1980), "Consumer Perceptions of Hazards," in *PLP-80 Proceedings,* pp. 25–29.

Robert, H. M. (1990), *Robert's Rules of Order Newly Revised,* 9th Ed., Scott, Foresman, Glenview, IL.

Saaty, T. L. (1988), *Multicriteria Decision Making: The Analytic Hierarchy Process,* T. Saaty, Pittsburgh.

Sage, A. (1981), "Behavioral and Organizational Considerations in the Design of Information Systems and Processes for Planning and Decision Support," *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-11, pp. 61–70.

Savage, L. J. (1954), *The Foundations of Statistics,* John Wiley & Sons, New York.

Schelling, T. (1960), *The Strategy of Conflict,* Harvard University Press, Cambridge, MA.

Schelling, T. (1978), *Micromotives and Macrobehavior,* W. W. Norton, New York.

Selye, H. (1936), "A Syndrome Produced by Diverse Noxious Agents," *Nature,* Vol. 138, p. 32.

Selye, H. (1979), "The Stress Concept and Some of Its Implications," in *Human Stress and Cognition,* V. Hamilton and D. M. Warburton, Eds., John Wiley & Sons, New York.

Shafer, G. (1976), *A Mathematical Theory of Evidence,* Princeton University Press, Princeton, NJ.

Siebold, D. R. (1992), "Making Meetings More Successful: Plans, Formats, and Procedures for Group Problem-Solving," in *Small Group Communication,* 6th Ed., R. Cathcart and L. Samovar, Eds., Brown, Dubuque, IA, pp. 178–191.

Simon, H. A. (1955), "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics,* Vol. 69, pp. 99–118.

Simon, H. A. (1983), "Alternative Visions of Rationality," in *Reason in Human Affairs,* Stanford University Press, Stanford, CA.

Singleton, W. T., and Hovden, J. (1987), *Risk and Decisions,* John Wiley & Sons, New York.

Slovic, P. (1978), "The Psychology of Protective Behavior," *Journal of Safety Research,* Vol. 10, pp. 58–68.

Slovic, P. (1987), "Perception of Risk," *Science,* Vol. 236, pp. 280–285.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1977), "Behavioral Decision Theory," *Annual Review of Psychology,* Vol. 28, pp. 1–39.

Sniezek, J. A. (1992), "Groups under Uncertainty: An Examination of Confidence in Group Decision Making," *Organizational Behavior and Human Decision Processes,* Vol. 52, pp. 124–155.

Sniezek, J. A., and Henry, R. A. (1989), "Accuracy and Confidence in Group Judgment," *Organizational Behavior and Human Decision Processes,* Vol. 43, pp. 1–28.

Stanoulov, N. (1994), "Expert Knowledge and Computer-Aided Group Decision Making: Some Pragmatic Reflections," *Annals of Operations Research,* Vol. 51, pp. 141–162.

Stasser, G., and Titus, W. (1985), "Pooling of Unshared Information in Group Decision Making: Biased Information Sampling during Discussion," *Journal of Personality and Social Psychology,* Vol. 48, pp. 1467–1478.

Stevenson, M. K., Busemeyer, J. R., and Naylor, J. C. (1993), "Judgment and Decision-Making Theory," in *Handbook of Industrial and Organizational Psychology,* 2nd Ed., Vol. 1, M. D. Dunnette and L. M. Hough, Eds., Consulting Psychologists Press, Palo Alto, CA.

Stone, M. (1960), "Models for Reaction Time," *Psychometrika,* Vol. 25, pp. 251–260.

Strat, T. M. (1994), "Decision Analysis Using Belief Functions," in *Decision Making under Dempster–Shafer Uncertainties,* M. Fedrizzi, J. Kacprzyk, R. R. Yager, Eds., John Wiley & Sons, New York.

Stukey, E., and Zeckhauser, R. (1978), "Decision Analysis," in *A Primer for Policy Analysis,* W. W. Norton, New York, pp. 201–254.

Svenson, O. (1990), "Some Propositions for the Classification of Decision Situations," in *Contemporary Issues in Decision Making,* K. Borcherding, O. Larichev, and D. Messick, Eds., pp. 17–31, North Holland, Amsterdam.

Svenson, O., and Maule, A. J. (1993), *Time Pressure and Stress in Human Judgment and Decision Making,* Plenum, New York.

Swain, A. D., and Guttman, H. (1983), *Handbook for Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications,* NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington, DC.

Tanner, W. P., and Swets, J. A. (1954), "A Decision Making Theory of Visual Detection," *Psychological Review,* Vol. 61, pp. 401–409.

Torrance, E. P. (1953), "The Behavior of Small Groups under the Stress Conditions of 'Survival,'" *American Sociological Review,* Vol. 19, pp. 751–755.

Tuckman, B. W. (1965), "Development Sequence in Small Groups," *Psychological Bulletin,* Vol. 63, pp. 289–399.

Tversky, A. (1969), "Intransitivity of Preferences," *Psychological Review,* Vol. 76, pp. 31–48.

Tversky, A. (1972), "Elimination by Aspects: A Theory of Choice," *Psychological Review,* Vol. 79, pp. 281–289.

Tversky, A., and Kahneman, D. (1973), "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology,* Vol. 5, pp. 207–232.

Tversky, A., and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science,* Vol. 185, pp. 1124–1131.

Tversky, A., and Kahneman, D. (1981), "The Framing of Decisions and the Psychology of Choice," *Science,* Vol. 211, pp. 453–458.

Valenzi, E., and Andrews, I. R. (1973), "Individual Differences in the Decision Processes of Employment Interviews," *Journal of Applied Psychology,* Vol. 58, pp. 49–53.

Viscusi, W. K. (1991), *Reforming Products Liability,* Harvard University Press, Cambridge, MA.

von Neumann, J., and Morgenstern, O. (1947), *Theory of Games and Economic Behavior,* Princeton University Press, Princeton, NJ.

Wagenaar, W. A. (1992), "Risk Taking and Accident Causation," in *Risk-Taking Behavior,* J. F. Yates, Ed., John Wiley & Sons, New York, pp. 257–281.

Wallsten, T. S. (1972), "Conjoint-Measurement Framework for the Study of Probabilistic Information Processing," *Psychological Review,* Vol. 79, pp. 245–260.

Wallsten, T. S. (1976), "Using Conjoint-Measurement Models to Investigate a Theory about Probabilistic Information Processing," *Journal of Mathematical Psychology,* Vol. 14, pp. 144–185.

Wallsten, T. S. (1995), "Time Pressure and Payoff Effects on Multidimensional Probabilistic Inference," in *Time Pressure and Stress in Human Judgment,* O. Svenson and J. Maule, Eds., Plenum Press, New York.

Wallsten, T. S., Zwick, R., Kemp, S., and Budescu, D. V. (1993), "Preferences and Reasons for Communicating Probabilistic Information in Verbal and Numerical Terms," *Bulletin of the Psychonomic Society,* Vol. 31, pp. 135–138.

Weber, E. (1994), "From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events," *Psychological Bulletin,* Vol. 115, pp. 228–242.

Weber, E., Anderson, C. J., and Birnbaum, M. H. (1992), "A Theory of Perceived Risk and Attractiveness," *Organizational Behavior and Human Decision Processes,* Vol. 52, pp. 492–523.

Wegner, D. (1987), "Transactive Memory: A Contemporary Analysis of Group Mind," in *Theories of Group Behavior,* B. Mullen and G. R. Goethals, Eds., Springer-Verlag, New York, pp. 185–208.

Weinstein, N. D. (1979), "Seeking Reassuring or Threatening Information about Environmental Cancer," *Journal of Behavioral Medicine,* Vol. 2, pp. 125–139.

Weinstein, N. D. (1980), "Unrealistic Optimism about Future Life Events," *Journal of Personality and Social Psychology,* Vol. 39, pp. 806–820.

Weinstein, N. D. (1987), "Unrealistic Optimism about Illness Susceptibility: Conclusions from a Community-Wide Sample," *Journal of Behavioral Medicine,* Vol. 10, pp. 481–500.

Weinstein, N. D., and Klein, W. M. (1995), "Resistance of Personal Risk Perceptions to Debiasing Interventions," *Health Psychology,* Vol. 14, pp. 132–140.

Welch, D. D. (1994), *Conflicting Agendas: Personal Morality in Institutional Settings,* Pilgrim Press, Cleveland.

Welford, A. T. (1976), *Skilled Performance,* Scott, Foresman, Glenview, IL.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance,* HarperCollins, New York.

Winkler, R. L., and Murphy, A. H. (1973), "Experiments in the Laboratory and the Real World," *Organizational Behavior and Human Performance,* Vol. 10, pp. 252–270.

Winterfeldt, D. V., and Edwards, W. (1986), *Decision Analysis and Behavioral Research,* Cambridge University Press, Cambridge.

Yates, J. F., Ed. (1992), *Risk-Taking Behavior,* John Wiley & Sons, New York.

Zander, A. (1994), *Making Groups Effective,* 2nd Ed., Jossey-Bass, San Francisco.

Zimmer, A. (1983), "Verbal versus Numerical Processing of Subjective Probabilities," in *Decision Making under Uncertainty,* R. W. Scholtz, Ed., North Holland, Amsterdam.

Zey, M., Ed. (1992), *Decision Making: Alternatives to Rational Choice Models,* Sage, London.

CHAPTER 85

Design of Experiments

H. SAMUEL WANG
Chung Yuan Christian University

CHUNG-PU CHANG
Eureka Consulting Co.

1. INTRODUCTION 2225
- 1.1. Perspective 2225
- 1.2. Statistical Experiments 2225
- 1.3. Basic Definitions 2225

2. PLANNING FOR EXPERIMENTS 2226
- 2.1. Program and Activities: Steps and Checkpoints 2226
  - 2.1.1. Stage 1: PLAN 2226
  - 2.1.2. Stage 2: DO 2226
  - 2.1.3. Stage 3: STUDY 2227
  - 2.1.4. Stage 4: ACT 2227
- 2.2 Size of Experiments 2227

3. GOOD EXPERIMENTAL PRACTICES 2228
- 3.1. Randomization 2228
- 3.2. Blocking 2228
- 3.3. Replication 2228

4. PRECAUTIONS FOR EXPERIMENTAL DESIGNS 2228
- 4.1. Ten Commandments for Experimental Designs 2228

5. FUNDAMENTAL DESIGNS AND CONCEPTS 2229
- 5.1. A Case: Weight Watch Program 2229
- 5.2. Fixed-Effect and Random-Effect Models 2229
- 5.3. Completely Randomized Design (CRD) 2230
- 5.4. Randomized Complete Block Design (RCBD) 2230
- 5.5. Latin Square Designs and Interactions 2230
- 5.6. Factorial Design 2230
- 5.7. 2^k and 3^k Factorial Designs 2231
- 5.8. Fractional Factorial Designs 2231
- 5.9. Orthogonal Arrays 2232

6. ANALYSIS OF A BASIC DESIGN 2232
- 6.1. Hypotheses and Models 2232
- 6.2. ANOVA: Analysis of Variance 2233
- 6.3. Marginal Averages 2234
- 6.4. Rationale of ANOVA Analysis 2234

7. SCREENING DESIGNS 2235
- 7.1. Strategy of Screening Design 2235
- 7.2. Weight Watch Experiment Using $L_8(7)$ 2235
- 7.3. ANOVA 2235
- 7.4. Recommendations 2236

8. PARAMETER DESIGNS 2237
- 8.1. Strategy of Parameter Design 2237
- 8.2. Concepts of Parameter Design 2237
- 8.3. Weight Watch Experiments 2238

9. THE STRATEGIES OF EXPERIMENTS 2238

10. CONCLUSION 2239

REFERENCES 2239

ADDITIONAL READING 2240

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

1.1. Perspective

Experimentation is common in every aspect of life. As part of a problem-solving program, experiments are carried out in order to observe the effects of changes under a controlled framework. Through one or more iterations of experiments, adequate knowledge is acquired or confirmed. Know-how, sometimes coupled with know-why, is gathered and used for decision making. Experiments are indispensable to the learning process.

The need for learning through experiments is particularly obvious in industry, whether in manufacturing or the service sector. Consider the development and marketing of a new drug. After a new drug is found to be effective for treating a certain kind of cancer, a series of experiments is usually conducted before the drug is formally marketed. For instance, a laboratory scientist performs experiments to identify other supportive constituents. With the aid of these experimental results, he or she picks the most effective composition. A manufacturing engineer uses experiments to determine process conditions such as pressure, temperature, flow rate, the catalyst quantity, and so forth. Thus the goal of fabricating quality medicine at the lowest possible cost is achieved. A marketing staff relies on computer simulation, which in fact is a form of numerical experimentation, to pick the most profitable marketing strategy. An FDA officer carries out experiments to detect potential adverse effects on different consumers determined by age, sex, and ethnicity. Experimentation is also important to the consumer organization. It is relied upon to compare the effectiveness of this new drug against others existing in the market.

Other examples illustrating when and where experiments are performed are numerous (Diamond 1989, 1997; John 1998).

1.2. Statistical Experiments (Box et al. 1978; Du Pont Co. 1988)

An experiment is often confused with a trial or a test, which in practice takes no account of experimental errors due to inherent variations. In fact, variations occur in every component and stage of experimentation, including variation in experiment parameters, due to inaccurate setting of machines and instruments, in methods and handling, in measurements, and due to analysis.

Moreover, experiments are often run by intuition with factors varied one at a time. This is not only ineffective costwise, it also causes the risk of reaching incorrect conclusions due to negligence of the potential interactions among factors

In the following, we are concerned with the design and analysis of experiments based on statistical considerations. These are often referred to as statistical experiments.

A statistical experiment serves as a means to compare and choose the most effective treatment, identify significant factors, reveal cause-and-effect mechanisms, and determine optimal process conditions. It therefore plays an important role in quality improvement, productivity increases, cost saving, and management decisions.

In essence, a statistical experiment implies a systematic varying of process, observation of change in response, collection and analysis of data, and extraction of information to arrive at a conclusion. Experiments are designed so that the appropriate decision can be arrived at in the shortest time and within cost constraints.

1.3. Basic Definitions (Anderson and McLean 1974; Du Pont Co. 1988; Hunter 1998; Montgomery 1996)

An experimental design is a formal plan for execution of the experiment. It includes the choice of response, factors, designation of levels, and assignment of blocks as well as application of treatment on experimental units.

These commonly used terms are defined and explained below:

- *Response:* A response is the dependent variable that corresponds to the outcome or resulting effects of interest in the experiment. One or more response variables may be studied simultaneously.
- *Factors:* A factor is a variable contribute to the response. A factor can be controllable or uncontrollable. It may be quantitative, such as pressure in psi or duration time in minutes. It may be qualitative, such as different methods, different operators, or different suppliers.
- *Levels:* The levels are the chosen conditions of the factor under study. They may be quantitative values such as 5%, 8%, and 10% alcohol concentration. They also take categorical forms such as supplier A, B, C, and D.
- *Blocks:* A block is a homogenous portion of the experimental environment or materials that bears certain variation effects on the response(s). A block may be a batch of material supplied

by a vendor or products manufactured in a shift on a production floor. The term *block* is sometimes associated with *factor* and called *block variable*.

- *Treatments:* A treatment is the condition or a factor associated with a specific level in a specific experiment.
- *Experimental units:* Experimental units are the objects or entities that are used for application of treatments and measurements of resulting effects.

2. PLANNING FOR EXPERIMENTS (Du Pont Co. 1988; Hunter 1998)

A complete experimental process involves five stages: including (1) design, (2) data collection, (3) data analysis, (4) interpretation of results, and (5) communication of results. In order to achieve an effective and efficient experimentation, it is of utmost importance to take effort to work out a comprehensive experimental plan.

In the planning stage of the experiment, the design of the experiment implies the careful and thorough consideration of the following issues:

- Global environment of the problem
- Objectives of the study
- Properties to be studied
- Variables to be controlled
- The environment of concern
- Size of experimental units
- Number of experimental runs
- Conduct of the experiments
- Approach for data analysis

Each of the technical, statistical, and administrative aspects of experiments are to be taken into consideration.

2.1. Program and Activities: Steps and Checkpoints (Du Pont Co. 1988; Hunter 1998)

2.1.1. Stage 1: PLAN

1. *Problem Recognition*
    - **(a)** Formation of task force
    - **(b)** Evaluation of strengths, weaknesses, opportunities, and threats
    - **(c)** Identification of problem area(s)
2. *Statement of problem and objective:*
    - **(a)** Determination of ultimate goal(s)
    - **(b)** Determination of en route objectives
    - **(c)** Determination of immediate objectives
    - **(d)** Identification of the cost and time constraints
3. *Design of experiment:*
    - **(a)** Definition of experimental units
    - **(b)** Determination of response variable(s)
    - **(c)** Selection of factors
    - **(d)** Determination of factor levels
    - **(e)** Appraisal of possible interaction
    - **(f)** Choice of design
    - **(g)** Definitions of data and effect models
    - **(h)** Decision of number of replicates
    - **(i)** Setup of execution plan (timetable, sequence schedule, facilities allocation)
    - **(j)** Setup of data-collection plan

2.1.2. Stage 2: DO

1. *Mission orientation:*
    - **(a)** Explanation of objectives and tasks to be achieved
    - **(b)** Attention to precautions in execution and data recording

2. *Physical preparations:*
    - **(a)** Preparation of experimental units
    - **(b)** Development of methods and facilities needed
3. *Execution of experiments:*
    - **(a)** Execution in accordance with prescribed conditions
    - **(b)** Control of schedules
    - **(c)** Varying and control of treatments in terms of randomization
4. *Observation and measurement:*
    - **(a)** Surveillance of condition of facilities
    - **(b)** Control of measurement procedures
5. *Recording:*
    - **(a)** Recording of program by date, run numbers, etc.
    - **(b)** Recording of abnormal situations
    - **(c)** Recording of change of designs
    - **(d)** Recording of measured data
    - **(e)** Recording of data from extra or missing experiments

2.1.3 Stage 3: STUDY

1. *Analysis of data:*
    - **(a)** Diagnosis of data
    - **(b)** Application of appropriate statistical methods
    - **(c)** Graphical analysis
    - **(d)** Checking of model adequacy
2. *Interpretation of results:*
    - **(a)** Identification of substantial factors
    - **(b)** Selection of desired levels
    - **(c)** Estimation of factor effects
    - **(d)** Account for limitations in data acquisition or analysis
    - **(e)** Interpretation in terms of statistical, technical, and economical significance

2.1.4. Stage 4: ACT

1. *Confirmation of conclusion*
2. *Presentation of results:*
    - **(a)** Preparation of report
    - **(b)** Use of graphical and tabular forms
    - **(c)** Indication of implications for potential applications
3. *Recommendations:*
    - **(a)** For process change
    - **(b)** For further experiments
    - **(c)** For modification in strategies of experiment
4. *Process change and standardization*
5. *New situation appraisal*
6. *Preparation for further experiments*

2.2. Size of Experiments

Each experimenter should be concerned with the size of the experiment. A large enough experiment enables the detection of the significant effects of the factors in the response and thus ensures obtaining some know-how from the study. Yet it must be small enough to ensure that the cost of the experimentation is within the allocated budget and can be completed within the assigned time frame.

The number of experimental runs or experimental units is to be determined beforehand. Its precise determination involves statistical computation that requires prestated probability of committing type I and II errors, the desired accuracy in detecting the difference between the means of the responses resulting from different treatments. Besides, other statistical parameters are also required. As a rule of thumb, the appropriate size for an experiment is between 8 and 60. For more elaborate evaluation, consult statistical handbooks (Box et al. 1978; Daniel 1976; Wadsworth 1990; Winer et al. 1991).

3. GOOD EXPERIMENTAL PRACTICES (Hicks 1982; Montgomery 1996)

In the process of experimentation, there exist two types of errors: random errors and bias errors. Random error is experimental error for which the numerical values change from one run to another without a consistent pattern. It can be thought of as inherent noise in measured responses. Bias error is experimental error for which the numerical values tend to follow a consistent pattern over a number of experimental runs. It is attributed to an assignable cause. To reduce the effects of both types of errors, it is strongly adviced that the following good experimental practices be taken into consideration.

Replication, randomization, and blocking, the three measures for ensuring a successful experiment as presented below, are called the three Fisherean principles of experimental designs. They are attributed to R. A. Fisher, the forerunner of the modern design of experiments.

3.1. Randomization

Randomization is the procedure of assigning the experimental units to various treatments in a purely chance manner. It is also used for arrangement of the experiments in random order. It is intended to balance out the effect of uncontrollable variables. Because bias errors are not confused with the effect of the factors, the quality of data is improved and statistical inferences are made possible.

To achieve randomization, the order of experiment is scrambled so that any bias present will be mixed up and become a part of the random variation. One of the following options can be taken

The trial numbers can be written on small slips of paper and selected at random.

A table of random numbers can be employed to assign a run order and the trials.

3.2. Blocking

The source of bias error in an experiment may accompany differences among blocks, namely batches of raw materials, production machine, hours within a day, or seasons of the year. It is necessary to reduce their influence by proper design of the experiment. Blocking means running the experiment in a specially chosen subgroup that allows removal of the effect of bias errors that are confounded with the main factors.

To achieve blocking, the run order is broken up into smaller units so that the bias is negligible within the block. Note that under these circumstances a separate randomization is needed in each block.

3.3. Replication

Replication involves the repetition of experimental runs so that more than one observation for each treatment combination is available for statistical analysis.

The benefit of replication is that the average of several observations comes closer to the true value than a single observation. Replication helps balance out the bias due to the effect of nuisance factors. It also helps to detect gross errors in the measurements. It therefore improves the precision of the statistical inferences.

Different randomization applies to different replications of the experiment.

4. PRECAUTIONS FOR EXPERIMENTAL DESIGNS

4.1. Ten Commandments for Experimental Designs (Gryna and Juran 1993; Montgomery 1996)

1. *Don't set out without a clear problem definition and objective statement:* An unplanned experiment often ends up in total loss of time and money.
2. *Do keep the design and analysis of experiment as simple as possible:* A comprehensive but complicated design of experiment may cost much more and end nowhere.
3. *Don't rely solely on statistical experts:* Interaction with subject matter specialists makes professional insight invaluable.
4. *Don't underestimate the importance of randomization:* The adverse effect of systematic errors can be of vital importance.
5. *Don't start statistical analysis without first challenging the validity of the data:* What can you expect out of garbage input?
6. *Don't throw away outliers without solid reasoning:* Every piece of data stores a hidden story waiting to be opened.
7. *Do make full use of graphical presentations:* A picture can be worth more than hundreds of words.

TABLE 1 A Balanced Design

DIET	D_1	D_2	D_3
	y_{11}	y_{21}	y_{31}
	y_{12}	y_{22}	y_{32}
	y_{13}	y_{23}	y_{33}

8. *Do avoid statistical jargon in conclusion and report writing:* Problem language is the only thing that is universal in a corporation.
9. *Don't blindly follow statistical conclusions without taking into account their practical significance and economic considerations:* Negligence of the nonstatistical aspects of the experimental design can prove to be vital.
10. *Don't think that one iteration of a time experiment can solve the problem once and for all:* The outcome of one experiment often provides a direction for further iterations of exploration.

5. FUNDAMENTAL DESIGNS AND CONCEPTS (Anderson and McLean 1974; Box and Draper 1987; Gryna and Juran 1993; Hunter 1998; Montgomery 1996)

Statistical experiments can have various objectives and constraints. Identifying the most influential factor(s) or independent variables and their respective effect on the response or dependent variable(s) is one of the most common objectives. The nature and number of factors of interest, the number of levels a factor can vary, the limitation in time span and budget, and so forth are some of the common constraints. Depending on its specific objective and constraints, an experiment can have various designs. In the following we take a weight-watch program as an example to illustrate various alternative designs and their underlying principles.

5.1. A Case: Weight-Watch Program

One response variable of concern is the weight loss (WTLOSS) in kilograms measured in the first month. The main factor DIET has three levels, that is, options, D_1, D_2, and D_3. Suppose nine persons volunteer to participate in this experiment program. One natural arrangement is to assign three persons to each of the three DIETs, as shown in Table 1. Note that this is a balanced design in the sense that the same number of persons received the same treatment. For a simple design like this, balancing is not a must but is desired. However, in other cases it is almost a requirement for easy analysis and guarantees the same degree of precision in estimation of different treatments.

Under this arrangement, one diet is assigned to three volunteers, that is, each has three replications. Replication is used to estimate the experimental error. It also helps to increase the precision.

5.2. Fixed-Effect and Random-Effect Models

In this experiment, the three DIETs may be the only ones of concern to the experimenter. In this case it is called a fixed-effect model, and the conclusion drawn is applicable only to these specific three options. However, situations may arise in which the experimenter is seeking for a conclusion applicable to all possible DIET options and yet he or she can only handle three options. Then, three of the available options should be drawn from all possible at random. This is called a random-effect model. In addition to the difference in the scope of conclusion, different types of effect models also imply a somewhat different approach of data analysis. In what is to follow, only the fixed-effect model is considered.

An advertent error can easily occur at this point. Very often, experimental units are assigned to treatments by convenience. For example, the WTLOSS may vary from one ethnic group to another. Yet persons of the same ethnic group (ETHNIC) are assigned the same diet. One extreme case could appear like the one shown in Table 2, where all Caucasians (C) are assigned to D_1, all Africans (A)

TABLE 2 Confounding of Factors

DIET	D_1	D_2	D_3
	C	A	O
	C	A	O
	C	A	O

to D_2, and all Orientals (O) to D_3. One problem arises: Whenever one of the DIETs is found to be more effective than the others, we cannot separate the confounding of the two factors and thus can hardly claim which is the truly dominant one.

5.3. Completely Randomized Designs (CRD)

To overcome this problem, the experimental units are usually assigned to treatments by randomization, that is, in a purely chance manner. This way of completely randomized allocation of DIETs to volunteers, as shown in Table 3, is called completely randomized design (CRD).

5.4. Randomized Complete Block Design (RCBD)

Another way to guard against any possible bias due to the effect of ETHNIC is to carry out the experiment by the randomized complete block design (RCBD), as shown in Table 4. Here, each of the three ethnic groups constitutes a block and receives all of the three treatments in random order.

A RCBD has the advantage of eliminating the contamination of the block factor on the main factor. It permits the removal of the block effect from the experimental error and thus provides a more decisive conclusion. Moreover, the effect of the block factor can usually be tested and evaluated.

5.5. Latin Square Designs and Interactions

At this point one may suspect that different amount of EXERCISE, say light (L), medium (M), and heavy (H), may have different degrees of effect on WTLOSS and should be treated as another block factor. Under these circumstances, the experiment is often carried out according to the Latin square design, as shown in Table 5. Note that each DIET is assigned only once to each ETHNIC and only once to each EXERCISE. It enables the evaluation of three factors with only nine observations. However, it requires that no interaction exist between the factors.

Two factors are said to have interaction when the effect of one factor varies under the different levels of another factor. The concept of interaction is illustrated in Figure 1.

5.6. Factorial Designs

Whenever interaction exists between factors, the experiment must be run according to the factorial design (FD) shown in Table 6 to ensure accurate and precise conclusion. Note that this experiment covers each of the 27 combinations of the levels of the three factors.

One may note that the factorial design differs from the conventional one-factor-at-a-time approach. In the latter the experiment is run in several iterations, and in each iteration only one factor is varied while the others are held constant. As a result, the factorial design allows the test and evaluation of

TABLE 3 Completely Randomized Design (CRD)

DIET	D_1	D_2	D_3
	W	A	A
	B	B	A
	W	B	W

TABLE 4 Randomized Complete Block Design (RCBD)

DIET	D_1	D_2	D_3
	C	A	A
	A	O	C
	O	C	O

TABLE 5 Latin Square Design

DIET	B	W	A
	D_1	D_2	D_3
	D_2	D_3	D_1
	D_3	D_1	D_2

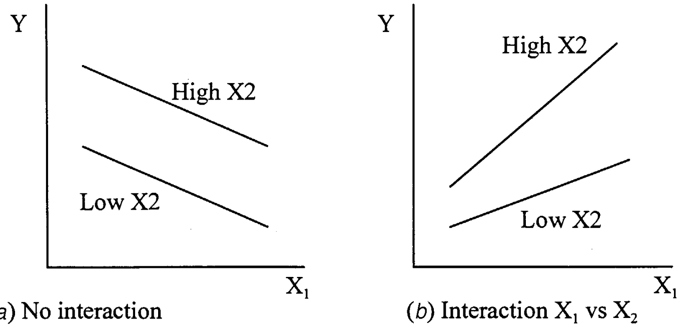


Figure 1 Concept of Interaction.

the effects of all three factors at one iteration. It is therefore more effective in the sense that it requires a shorter time to reach a complete conclusion. It also guards against the risk of missing the optimum of a surface, as is often seen in other approaches.

The FD also has disadvantages. The number of observations increases exponentially with the number of factors and also with their number of levels.

5.7. 2^k and 3^k Factorial Designs

As long as the experiment is still in its exploratory stage and one is mainly interested in screening out the less effective factors, the preceding problem can be partially solved by reducing the number of levels of each factor to two or three. Under such arrangement, the k-factor FD is then denoted as 2^k or 3^k FD, respectively. Although the two-level FD (see Table 7) requires a lower number of observations than its three-level counterpart, it is applicable only when it is believed that none of the factors has a nonlinear effect on the response.

5.8. Fractional Factorial Designs (FFD)

Further reduction of the number of observations can be achieved by employing the fractional factorial design (FFD). The notation 2^{k-p} is used to denote a 2^p fraction of a 2^k fractional design. See Table 8 for the one-half fraction of the 2^3 FD of the weight-watch experiment.

TABLE 6 Factorial Design

DIET (1)		D_1			D_2			D_3		
EXERCISE (2)		L	M	H	L	M	H	L	M	H
ETHNIC (3)	C	y_{111}	y_{121}	y_{131}	y_{211}	y_{221}	y_{231}	y_{311}	y_{321}	y_{331}
	O	y_{112}	y_{122}	y_{132}	y_{212}	y_{222}	y_{232}	y_{312}	y_{322}	y_{332}
	A	y_{113}	y_{123}	y_{133}	y_{213}	y_{223}	y_{233}	y_{313}	y_{323}	y_{333}

TABLE 7 Factorial Design (FD)

Run	DIET	ETHNIC	EXERCISE
1	1	1	1
2	2	1	1
3	1	2	1
4	2	2	1
5	1	1	2
6	2	1	2
7	1	2	2
8	2	2	2

TABLE 8 Fractional Factorial Design (FFD)

Run	DIET	ETHNIC	EXERCISE
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

TABLE 9 The Orthogonal Array $L_8(2^7)$

Factor Interaction	a	b	$a \times b$	c	$a \times b$	$b \times c$	$a \times b \times c$
Column	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

5.9. Orthogonal Arrays (OA) (Taguchi 1986)

For each FD there exist several alternative FFDs. One specific class of FFDs is called an orthogonal array (OA), often referred to as the Taguchi method. Reduced to its simplest level in allocating factor-level combinations, it has become increasingly popular.

Depending on the number of levels and the number of rows, a cluster of different OAs is available from which to choose.

Shown in Table 9 is one typical two-level OA with eight rows and seven columns, denoted as $L_8(2^7)$. The subscript denotes the number of rows or the number of factor-level combinations this OA provides. The lower number between the parentheses stands for the number of factors it should have, while its superscript stands for the number of columns this OA has or the maximum number of factors the experimenter is allowed to allocate.

Inside the array, numbers 1 and 2 label the low and high levels of the factors. However, each of the rows indicates the factor-level combination to be applied to the specific experimental unit. Note that in each column, both levels 1 and 2 appear an equal number of times. It is also true for all combinations (1,1), (1,2), (2,1) (2,2) appearing in any two columns. Hence the name ‘‘orthogonal array.’’

The OA is a balanced design in nature. As a form of fractional factorial design, it has an economic advantage, and it allows evaluation of main effects and the interactions between factors as well.

Attached to the bottom of the table is the component row, which tells how the main effects and interaction effects are associated with columns. In other words, if factors A, B, and C are assigned to columns, 1, 2, and 4, the component row points out that the interactions A X B, A X C, B X C, and A X B X C, are associated with columns 3, 5, 6, and 7. It can therefore be used as an aid to allocate factors to the array.

Besides $L_8(2^7)$, other OAs of the two-level family include $L_4(2^3)$, $L_{16}(2^{15})$, and $L_{32}(2^{31})$. Their counterparts in the three-level family include $L_9(3^4)$ and $L_{27}(1^3)$.

A set of linear graphs is available for each of the OAs to facilitate allocation of factors and interactions to columns of the arrays.

6. ANALYSIS OF A BASIC DESIGN (Box et al. 1978; Hicks 1982; Montgomery 1996)

6.1. Hypotheses and Models

In the weight-watch experiment we are concerned with the problem of comparing the effects of three diets. The hypotheses under test are therefore

H_0: $\mu_1 = \mu_2 = \mu_3$.
H_I: At least one of the three is different from the others.

The true mean of any one treatment can, in fact, be looked upon as the sum of a grand mean μ and the specific effect of the ith treatment α_i, that is, $\mu_i = \mu + \alpha_i$.
The hypotheses under test hence become (Box et al. 1978):

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$.
H_1: At least one of α_i's is unequal to zero.

Realizing the existence of variation due to environment, experimental units, execution process, and also measurement errors, any measured WTLOSS taken from the ith DIET group is unlikely equal to its treatment mean$_1$ and is usually decomposed as (Bhote 1991):

$$\begin{aligned} Y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij} \end{aligned}$$

This equation is usually referred to as the data model. The data model of the measured WTLOSS data acquired under the CRD setup shown in Table 3 can be represented as

$$\begin{array}{ccccccc} \text{Observation} & & \text{Mean} & & \text{Effect} & & \text{Error} \\ \begin{bmatrix} 6.53 & 3.23 & -0.11 \\ 6.72 & 2.19 & 0.35 \\ 3.91 & 4.72 & 2.61 \end{bmatrix} & = & \begin{bmatrix} 4 & 4 & 4 \\ 4 & 4 & 4 \\ 4 & 4 & 4 \end{bmatrix} & + & \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} & + & \begin{bmatrix} 1.53 & -0.77 & -3.11 \\ 1.72 & -1.81 & -2.65 \\ -1.09 & 0.72 & -0.039 \end{bmatrix} \end{array}$$

To enable valid statistical analysis, the error component ε_{ij} is required to follow independent identical normal distribution of the mean zero and common variance, or iid $N(0, \sigma^2)$.

6.2. ANOVA: Analysis of Variance

The test of the hypotheses uses the analysis of variance (ANOVA) approach, which is based on the decomposition principle of sum of squares. In other words, the variation of the observations from the grand average can be decomposed into two components: the variation around their group average, or the within-group variation, and the variation between the group average, or the between-group variation.

If the between-group variation is larger than what is expected from the variation that occurs within the groups, we would suspect that group means μ_1, μ_2, μ_3 are not the same. The F distribution is used for checking this point.

The F statistic is computed from

$$F = \frac{\mathrm{SSB}/(k-1)}{\mathrm{SSW}/(N-k)}$$

where SSB = sum of squares due to between-group variation.

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (\overline{Y}_{ij} - \overline{Y})^2$$

where SSW = sum of squares due to within-group variation.
= SSTO − SSB

$$\mathrm{SSTO} = \sum_{i=1}^{k} \sum_{j=1}^{n} (\overline{Y}_{ij} - \overline{Y})^2$$

where k = number of groups
n = number of observations in each group
$N = kn$

The whole computation procedure is usually summarized in an ANOVA table, as shown in Table 10.

The ANOVA table is provided by almost every statistical software.

TABLE 10 ANOVA Table for WTLOSS Data in CRD

Source	SS	*df*	MS	F_s
Between DIETs	34.13	2	17.07	8.25
Within DIETs	12.41	6	2.07	
Total	46.54	8		

The computed F statistic F_s is now compared to its corresponding critical value at the specified significance level $\alpha = 0.05$, that is, $F_{0.05}$ (2,6) = 5.14. Since $F_s > 5.14$, we accept H_1 and conclude that at least one of the DIETs is unequal to the others.

6.3. Marginal Averages

By comparing the marginal averages (see Figure 2) of the three DIETs, those of DIET D_1 have an average 5.72, which is significantly larger than DIETs D_2 and D_3. We are therefore assured that D_1 is the most effective program for a keen weight watcher.

6.4. Rationale of ANOVA Analysis

According to the statistical theory, the mean squares SSB/(k − 1) and SSB/(N − k) have their respective expected mean squares (EMS) as (Box et al. 1978):

$$\mathrm{E(MSB)} = \sigma^2 + n\sigma_A^2$$

$$\mathrm{E(MSE)} = \sigma^2$$

where

$$\sigma_A^2 = \sum_{i=1}^{k} \alpha_i^2/(\mathrm{k} - 1) \text{ for the fixed-effect model}$$

The test of the hypotheses can also be understood by viewing F approximately, as the ratio of $\sigma^2 + \sigma_A^2$ to σ^2. If all α_i's are the same and equal to zero, the ratio E(MSB)/E(MSW) is likely to be close to 1 and therefore leads us to accept H_0. On the other hand, if one of the α_i's is not equal to the others, the same ratio is likely to take a quantity much greater than 1 and therefore leads us to accept H_1.

For the applicable occasions and the analysis of some basic designs, please refer to the summary that follows.

Whenever there are more than one source or variation to be compared to, the EMS can be used as a quick decision aid. As a principle, a designate is always compared with a shorter row having common elements but one.

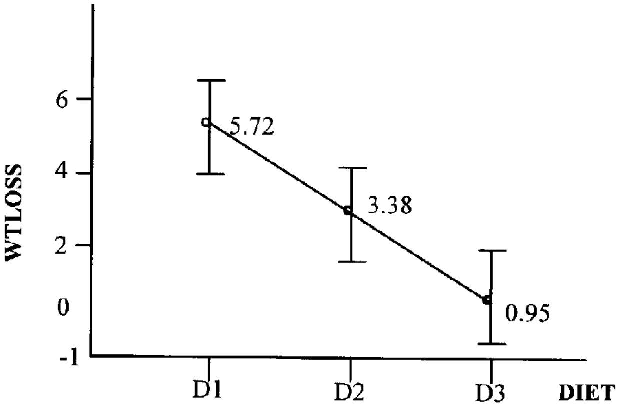


Figure 2 Marginal Average Chart with 95% Confidence Interval.

7. SCREENING DESIGNS (Barker 1990; Box et al. 1978; Dean and Voss 1999)

7.1. Strategy of Screening Design (Du Pont Co. 1988)

Many of the factors initially considered in the early stage of an experimental project may have little or no effect on response. The purpose of a screening experiment is to reduce experimental time and cost by identifying the factors that deserve thorough investigation in the subsequent stages. Therefore, in designing an experiment with many factors, it is useful to start with a screening experiment before going on to the more in-depth studies described in the preceding section.

7.2. Weight-Watch Experiment Using $L_8(2^7)$

Let us return to the weight-watch experiment to illustrate how an OA such as $L_8(2^7)$ is used to assign the factor levels. Suppose the factors and their levels of concern at this stage are as listed here.

Factor	Level 1	Level 2
DIET	DIET 1(D_1)	DIET 2(D_2)
EXERCISE	Medium (M)	High (H)
SEX	Female (F)	Male (M)
ETHNIC	White (W)	Black (B)

Suppose it is anticipated that interaction between DIET and EXERCISE and between DIET and SEX may exist. With the aid of the component row of the $L_8(2^7)$, we choose to assign the factors DIET, EXERCISE, SEX, and ETHNIC to columns 1, 2, 4, and 6. The treatment conditions can then be readily read from the array and are shown in Table 11.

Taking the randomization principle into account, the treatment condition D_1LFW, which is decoded as DIET 1, low exercise on a white female, and so on should be assigned in a random manner.

7.3. ANOVA

Table 11 also serves as a worksheet for data analysis. Recorded in the far-right column of Table 11 are the WTLOSS data resulting from the assigned treatment combinations. The T_1 of any column stands for the total of all observed data that are associated with level 1. Likewise, T_2 stands for the same types of data for level 2. For example, the T_1 for the DIET column is obtained by adding the four observations corresponding to the l's in column 1. Hence,

$$T_1\ (\text{DIET}) = 6.21 + 8.82 + 2.23 + 4.85 = 22.11$$

The task of subsequent analysis is to test the null hypotheses that different levels of the factors DIET, EXERCISE, SEX, and ETHNIC do not have any effect on WTLOSS. It is the same as in analyzing the CRD data; the ANOVA is used to test the previously mentioned hypotheses. The

TABLE 11 Orthogonal Array for Weight-Watch Experiment

Factor Interaction	DIET	EXR	$D \times E$	SEX	$D \times S$	ETH			
Column	1	2	3	4	5	6	7	Treatment Combination	Observed WTLOSS
1	1	1	1	1	1	1	1	D_1MFW	6.21
2	1	1	1	2	2	2	2	D_1MMB	8.82
3	1	2	2	1	1	2	2	D_1FMB	2.23
4	1	2	2	2	2	1	1	D_1HMW	4.85
5	2	1	2	1	2	1	2	D_2MFW	1.06
6	2	1	2	2	1	2	1	D_2MMB	2.75
7	2	2	1	1	2	2	1	D_2HFB	2.75
8	2	2	1	2	1	1	2	D_2HMW	9.26
T_1	22.11	18.84	27.04	12.25	20.45	21.38	16.56		
T_2	15.82	19.09	10.89	25.68	17.48	16.55	21.37		
SS	4.946	0.008	32.602	22.545	1.103	2.916	2.892		

TABLE 12A ANOVA for Weight-Watch Experiment

Source	SS	DOF	MS	F_c
DIET	4.95	1	4.95	1.71
EXERCISE	0.01	1	0.01	0.00
SEX	22.54	1	22.54	7.80
ETHNIC	2.92	1	2.92	1.01
Interaction				
DIET × EXR	32.60	1	32.60	11.27
DIET × SEX	1.1	1	1.10	0.38
ERROR	2.87	1	2.87	
TOTAL	67.01	7		

$F_{0.05}(1,1) = 161$

variation due to the effect of a specific factor is compared to the variation due to random error. The sum-of-squares for the *i*th column and the corresponding F statistic are computed by

$$SS_i = \frac{(T_1 - T_2)^2}{8}$$

$$F = \frac{SS_i/df_i}{SSE/df_E}$$

where SSE, the sum-of-squares due to errors, is found by summing up the SS's of the undesignated columns. For the weight-watch experiment, SSE $=$ SS$_7$. The ANOVA table obtained through this method is shown in Table 12A.

The computed F statistics are now compared to their common critical value at the specified significance level, say, $\alpha = 0.05$, that is, $F_{0.05}(1,1) = 161$. Since none of these is greater than 161,we do not have sufficient evidence to conclude that any of the main effects and interactions are significant at $\alpha = 0.05$. However, the computed F statistics do reveal that while the interaction DIET × EXR is relatively larger than the others, the factor SEX and another interaction DIET × SEX are negligible and thus can be merged with the error term. The resulting ANOVA table is listed in Table 12B.

As the computed F statistics of the interaction DIET × EXR, that is $F_c = 14.15 > 10.128 = F_{0.05}(1,3)$, we do have sufficient evidence to claim that two DIETs do affect WTLOSS differently at the two different levels of EXERCISE amount. It also suggests that SEX may have different WTLOSS. Further examination of the marginal average chart (see Figure 3) reveals that the male under this experimental setup attains greater WTLOSS than the female. It is noted that DIET 1 has effect on WTLOSS when it is used with a medium amount of EXERCISE. However, DIET 2 also produces a sizable WTLOSS when it is used with a higher amount of EXERCISE.

7.4. Recommendations

As mentioned earlier, the OA mainly serves the purpose of screening the vital few from the trivial many. In order to reach a solid and meaningful conclusion, more thorough confirmatory experiments are required. Take the weight-watch experiment as an example. It is recommended that the focus be

TABLE 12B ANOVA for Weight Watch Experiment

Source	SS	DOF	MS	F_c
DIET	4.95	1	4.95	2.15
EXERCISE	0.01	1	0.01	0.00
SEX	22.54	1	22.54	9.79
Interaction				
DIET × EXR	32.60	1	32.60	14.15
ERROR	6.91	3	2.30	
TOTAL	67.01	7		

$F_{0.05}(1,3) = 10.128$

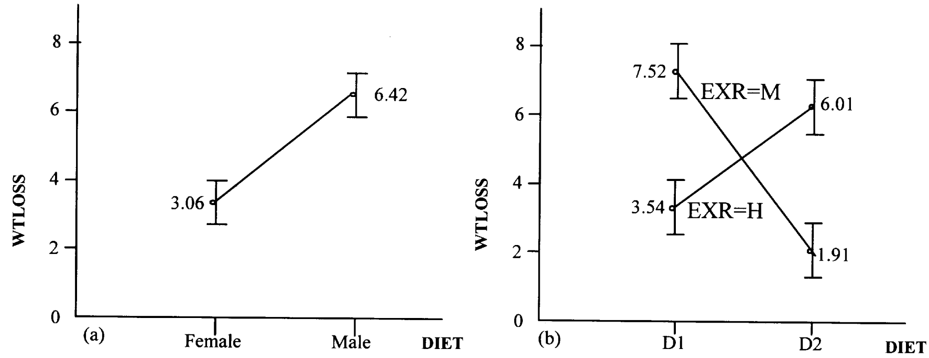


Figure 3 Marginal Average Chart with 95% Confidence Level.

placed at this point on only DIET and EXERCISE as the two main factors. Adding a few more levels may be worthwhile. A replicated two-way factorial design run on both sexes may be the experimenter's best choice. Some statistical tools, such as multiple comparison, confidence interval estimation, residual analysis, normality check, and response surface methodologies, are commonly used.

8. PARAMETER DESIGN (Peace 1989; Phadke 1989; Taguchi and Wu 1980; Taguchi 1986, 1987)

8.1. Strategy of Parameter Design

Traditionally, when performing an experimental design, the experimenter places his or her focus on finding the product or process condition that yields the best mean performance. However, this approach may not satisfy the demands of modern marketing strategies. This strategy requires a robust product or process that satisfies a wide range of customer interests.

For example, the combination of DIET I and the medium exercise amount in the preceding weight-watch experiment may, in fact, yield the greatest amount of WTLOSS, though this conclusion is valid for only some combinations of SEX and ETHNIC. However, a weight-watch service salesperson is more interested in a program that is robust or resistant to noise factors such as SEX and ETHNIC, that is, a program that is suitable for all possible combinations of SEX and ETHNIC.

The same is true for a production engineer. In designing a product or process, in addition to the major parameter settings, noise factors such as manufacturing variation, component tolerance, customer use conditions, and product deterioration need to be taken care of. A good product or process is one that is robust to variations due to these noise factors.

It is this requirement for robustness that prompted Taguchi to develop the concept of parameter design and produced a great impact on the world of experimental design.

8.2. Concepts of Parameter Designs (Barker 1990; Taguchi and Wu 1980; Taguchi 1986)

In the context of parameter design, the simple response is no longer of major interest. Rather, a composite performance measure that integrates both the mean and variance of the response plays the role of a dependent variable. Depending on the nature of the problem under study, various performance measures are developed for situations including "the larger the better," "the smaller the better," and "the specified target value is the best."

The objective of the parameter design is a matter of choosing a product or process condition that yields the best performance measure. In other words reducing the variation of response from the target while controlling the mean response toward the target is the ultimate goal of the parameter design.

These settings are determined by (1) systematically varying the settings of design parameters in the experiment and (2) comparing the effect of noise factors for each test run.

The parameter design achieves this goal by setting up an inner array and an outer array that constitute an orthogonal array such as $L_8(2^7)$ or $L_9(3^4)$. This array is assigned with control factor parameters, while the outer array is also an orthogonal array. This latter array is assigned with noise factor parameters.

TABLE 13 Parameter Design for Weight Watch Experiments

Outer Array			Row		1	2	3	4	
			Column						
			1	SEX	1	1	2	2	
			2	ETH	1	2	1	2	
			3		1	2	2	1	
Column Row	Inner 1 DIET	Array 2 EXR	3	4	Observed WTLOSS				Performance Statistics Z_i
1	1	1	1	1	y_{11}	y_{12}	y_{13}	y_{14}	Z_1
2	1	2	2	2	y_{21}	y_{22}	y_{23}	y_{24}	Z_2
3	1	3	3	3	y_{31}	y_{32}	y_{33}	y_{34}	Z_3
4	2	1	2	3	y_{41}	y_{42}	y_{43}	y_{44}	Z_4
5	2	2	3	1	y_{51}	y_{52}	y_{53}	y_{54}	Z_5
6	2	3	1	2	y_{61}	y_{62}	y_{63}	y_{64}	Z_6
7	3	1	3	2	y_{71}	y_{72}	y_{73}	y_{74}	Z_7
8	3	2	1	1	y_{81}	y_{82}	y_{83}	y_{84}	Z_8
9	3	3	2	1	y_{91}	y_{92}	y_{93}	y_{94}	Z_9

8.3. Weight-Watch Experiments

Table 13 illustrates how the parameter design is set up for the weight-watch experiment.

Placed in the lower left corner is a $L_9(3^4)$. Columns 1 and 2 are assigned as DIET and EXERCISE, respectively, both having three levels. Placed in the upper right corner is a transposed outer array $L_4(2^3)$. In this outer array the transposed column 1 is assigned the noise factor SEX while column 2 is assigned another noise factor ETHNIC. The two arrays are arranged in such a manner that each of the nine control factor parameter combinations cross all of the four noise factor parameter combinations. Thus, a total of 36 observed WTLOSS pieces of date data are obtained and displayed as shown.

The subsequent analysis of data consists of computing performance statistics based on a formula such as

$$Z_i = (-10)\log 1/n \sum_{i=1}^{n} (1/y_i^2)$$

for a ‘‘greater-the-better’’ case like this.

The next steps follow the same flow of analysis using an ANOVA table as seen in the preceding sections.

9. THE STRATEGIES OF EXPERIMENTS (Du Pont Co. 1988)

The practice of experimentation is a matter of problem solving. It is also a learning process rendered through step-by-step development of know-how as well as know-why. However, it takes sound and smart strategies to reach this goal effectively and efficiently.

A complete set of strategies consists of three major components: strategy for screening designs, strategy for parameter designs, and strategy of response surface designs.

The main purpose of the strategy of screening designs is to screen out the significant factors from various possible factors selected for experiment.

However, in order to keep the cost of experimentation from being too high, two levels for each factor are used in general. Consequently, 2^k orthogonal arrays are most oftenly used in the screening designs.

The strategy of parameter designs is practiced after the strategy of screening designs. At this stage, only the factors that are found to be significant are included in the experiments. As the number of factors is reduced, the experiment is now focused at the finding and determination of the optimal condition. The 3^k orthogonal arrays are commonly used in this stage.

The strategy of response surface designs takes place after the parameter design strategy. All of the data collected in the repeatability experiments are then utilized in the database used in conjunction with the response surface designs.

The elaboration of response surface designs is beyond the scope of this article. For more treatment of the subject, see Cornell (1990), Khuri and Cornell (1987), and Montgomery (1991).

10. CONCLUSION

The success of experimentation relies on the following issues:

1. Thorough consideration of all technical, statistical, and administrative aspects
2. Sound planning of the experimentation
3. Effective implementation
4. Proper observation of good experimental practices
5. Smart adoption of strategies of experimentation

REFERENCES

Anderson, V. L., and McLean, R. A. (1974), *Design of Experiments,* Marcel Dekker, New York.

Barker, T. B. (1990), *Engineering Quality by Design: Interpreting the Taguchi Approach,* Marcel Dekker, New York.

Bechhofer R. E., Santner, T. J., and Goldsman, D. M., *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons,* Wiley Series in Probability and Statistics, John Wiley & Sons. New York, 1995.

Bhote, K. R. (1991), *World Class Quality: Using Design of Experiments to Make It Happen,* AMACOM, New York.

Box, G. E. P., and Draper, N. R. (1987), *Empirical Model Building and Response Surface,* John Wiley & Sons, New York.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters,* John Wiley & Sons, New York.

Cornell. J. A. (1990c), *How to Apply Response Surface Methodology (rev. ed.).* American Society for Quality. Milwaukee.

Daniel, C. (1976), *Application of Statistics to Industrial Experiments,* John Wiley & Sons, New York.

Dean, A., and Voss, D. T. (1999), *Design and Analysis of Experiments,* Springer, New York.

Diamond, W. (1989), *Practical Experimental Designs.* Van Nostrand Reinhold, New York.

Diamond, W. J. (1997), *Practical Experiment Designs for Engineers and Scientists,* John Wiley & Sons, New York.

Du Pont Co. (1988), *Strategy of Experimentation,* Du Pont Quality and Technology Center, Newark, NJ.

Gryna, F. M., and Juran, J. M. (1993), *Planning and Analysis of Quality: From Product Development Through Use,* 3rd Ed., McGraw-Hill, New York.

Hicks, C. R. (1982), *Fundamental Concepts in the Design of Experiments,* 3rd Ed., Holt, Rinehart & Winston, New York.

Hunter, J. S. (1998), "Design and Analysis of Experiments," in *Juran's Quality Handbook,* 5th Ed. J. M. Juran and A. B. Godfrey, Eds., McGraw-Hill, New York.

John, W. M. (1998), *Statistical Design and Analysis of Experiments,* Society for Industrial and Applied Mathematics, Philadelphia.

Khuri, A. I., and Cornell, J. A. (1987), *Response Surfaces: Designs and Analyses,* Marcel Dekker, New York.

Montgomery, D. C. (1991), *Introduction to Statistical Quality Control,* 2nd Ed., John Wiley & Sons, New York.

Montgomery, D. C. (1996), *Design and Analysis of Experiments,* 4th Ed., John Wiley & Sons, New York.

Peace, G. S. (1989), *Taguchi Methods: A Hands-on Approach,* Addison-Wesley, Reading, MA.

Phadke, M. S. (1989), *Quality Engineering Using Robust Design,* Prentice Hall, Englewood Cliffs, NJ.

Taguchi, G. (1986), *Introduction to Quality Engineering,* Asian Productivity Association, Tokyo.

Taguchi, G. (1987), *Systems of Experimental Design,* Vol. 1, UNIPUB, Kraus International, New York.

Taguchi, G., and Wu, Y. I. (1980), *Introduction to Off-line Quality Control,* Central Japan Quality Control Association, Nagoya, Japan.

Wadsworth, H., Ed. (1990), *Handbook of Statistical Methods for Engineers and Scientists,* McGraw-Hill, New York.

Winer, B. J., Brown, D. R., and Michels, K. M. (1991), *Statistical Principles in Experimental Design,* McGraw-Hill, New York.

ADDITIONAL READINGS

Barker, T. B., *Quality by Experimental Design,* 2nd Ed., Marcel Dekker, New York, 1994.

Boniface, D. R., *Experiment Design and Statistical Methods: For Behavioral and Social Research,* CRC Press, Boca Raton, FL, 1995.

Cleveland, W. S., *Visualizing Data,* Hobart Press, Summit, NJ, 1993.

Cobb, G. W., *Introduction to Design and Analysis of Experiments,* Springer, New York.

Cornell, J. A., *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, 2nd Ed., John Wiley & Sons, New York, 1990.

Cornell, J. A., *How to Run Mixture Experiments for Product Quality,* Rev. Ed., American Society for Quality Press, Milwaukee, 1990.

Fleiss, J. L., *The Design and Analysis of Clinical Experiments,* John Wiley & Sons, New York, 1986.

Gad, S. C., *Statistics and Experimental Design for Toxicologists,* CRC Press, Boca Raton, FL, 1998.

Grove, D. M., and Davis, T. P., *Engineering Quality and Experimental Design,* John Wiley & Sons, New York, 1991.

Gunst, R. F., and Mason, R. L., *How to Construct Fractional Factorial Experiments,* American Society for Quality Press, Milwaukee, 1991.

Haaland, P. D., *Experimental Design in Biotechnology,* Marcel Dekker, New York, 1989.

Luftig, P. D., and Jeffrey, T., *Design of Experiments in Quality Engineering,* McGraw-Hill, New York, 1998.

Mason, R. L., Gunst, R. F., and Hess, J. L., *Statistical Design and Analysis of Experiments with Applications to Engineering and Science,* John Wiley & Sons, New York, 1989.

Moen, R. D., Nolan, T. W., and Provost, L. P., *Improving Quality through Planned Experimentation,* McGraw-Hill, New York, 1991.

Myers, R. H., and Montgomery, D. C., *Response Surface Methodology,* John Wiley & Sons, New York, 1995.

Schmidt, S. R., and Launsby, R. G., *Understanding Industrial Designed Experiments,* Air Academy Press, Colorado Springs, CO, 1991.

CHAPTER 86

Statistical Inference and Hypothesis Testing

DON T. PHILLIPS
ALBERTO GARCIA-DIAZ
Texas A&M University

1. INTRODUCTION 2241
2. STATISTICAL INFERENCE 2242
3. STATISTICAL HYPOTHESIS TESTING 2243
4. TESTING A MEAN VALUE (μ) WITH σ^2 KNOWN 2244
5. TESTING A MEAN VALUE (μ) WITH σ^2 UNKNOWN 2248
6. HYPOTHESIS TESTING: SINGLE VARIANCE 2249
7. HYPOTHESIS TESTING: TWO POPULATION MEANS WITH VARIANCES KNOWN 2249
8. HYPOTHESIS TESTING WITH TWO MEANS: POPULATION VARIANCES UNKNOWN BUT ASSUMED EQUAL 2250
9. HYPOTHESIS TESTING FOR EQUALITY OF TWO POPULATION VARIANCES 2251
10. EQUALITY OF TWO MEANS WITH VARIANCES UNKNOWN AND NOT EQUAL 2252
11. CONFIDENCE INTERVAL ESTIMATION 2253
12. MAXIMUM-LIKELIHOOD ESTIMATORS 2254
13. TESTING FOR EQUALITY OF MEANS AND VARIANCES FOR *K* POPULATIONS 2255
    - 13.1. Testing Means 2255
    - 13.2. Testing the Homogeneity of Variances 2255
    - 13.3. Cochran's Test 2255
    - 13.4. Bartlett's Test 2255
    - 13.5. Levene's Test 2256
14. OTHER USES OF HYPOTHESIS TESTING 2256
    - 14.1. Nonparametric Tests 2256
    - 14.2. Goodness of Fit Test 2256
15. HYPOTHESIS TESTING IN THE ANALYSIS OF DESIGNED EXPERIMENTS 2260
    - 15.1. One-Factor Experiments 2260
    - 15.2. After-ANOVA Range Tests 2261
    - 15.3. Factorial Experiments 2262
    - 15.4. Hypothesis Testing in Regression Analysis 2262

REFERENCES 2263

1. INTRODUCTION

One might define the function of applied statistics as the art and science of collecting and processing data in order to make inferences about the parameters of one or more populations associated with random phenomena. These inferences are made in such a way that the conclusions reached are consistent and unbiased. When properly applied and executed, statistical procedures depend entirely on specific methodologies, definitions, and parameters required by the statistical test chosen.

The science of statistics is purely mathematical with probability theory as the cornerstone. Because all statistical methods are based on probability concepts, it is necessary for one to understand the

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

basic concept of probability measure before undertaking statistical analysis. In order to proceed with the development of procedures for hypothesis testing and estimation, a fundamental knowledge of statistical measures, random variables, probability density functions, and statistical sampling procedures will be assumed.

2. STATISTICAL INFERENCE

Each (and every) random variable has a unique probability distribution. For the most part statisticians deal with the theory of these distributions. Engineers, on the other hand, are mostly interested in finding factual knowledge about certain random phenomena, by way of probability distributions of the variables directly involved, or other related variables.

In basic statistics we learn that probability density functions can be defined by certain constants called *distribution parameters.* These parameters in turn can be used to characterize random variables through measures of location, shape, and variability of random phenomena. The most important parameters are the mean μ and the variance σ^2. The parameter μ is a measure of the center of the distribution (an analogy is the center of gravity of a mass) while σ^2 is a measure of its spread or range (an analogy being the moment of inertia of a mass). Hence, when we speak of the mean and the variance of a random variable, we refer to two statistical parameters (constants) that greatly characterize or influence the probabilistic behavior of the random variable. The mean or expected value of a random variable x is defined as

$$\mu = \sum_x xp(x) \quad \text{or} \quad \mu = \int_x xf(x)dx$$

where $p(x)$ represents probabilities of a *discrete* random variable and $f(x)$ represents the probability density function of a *continuous* random variable. The parameters of interest are embedded in the form of the probability density functions. As an illustration, the variance of a random variable is defined as

$$\sigma^2 = \sum_x (x - \mu)^2 p(x) \quad \text{or} \quad \sigma^2 = \int_x (x - \mu)^2 f(x)dx$$

In mathematical statistics we can show that many random variables that occur in nature follow the same general form of distribution with differences only in the parameters and the statistical quantities μ and σ^2 Some of these recurring distributions have been given special names, such as:

Binomial
Hypergeometric
Poisson
Geometric
Negative binomial
Gamma
Beta
Normal
Chi-square
Student's-t
F distribution
Exponential
Uniform
Cauchy
Rayleigh
Maxwell
Weibull
Erlang

Thus, it is not difficult to see why the field of ''probability and statistics'' is a discipline within itself, nor is it difficult to see why almost every discipline in existence needs a working knowledge of statistics. Random phenomena (variables) exist in all phases of activity.

When one is interested in certain random phenomena, the first requirement seems to be that one develop some means of measuring it. Upon doing so, one often collects a number of observations of the random phenomena. Statistics deals with developing tools and techniques for choosing those observations (a sample) and manipulating them in such a way that useful information is gained about the underlying random variable(s). This information is generally derived from studying probability distributions of the random variables or functions of the random variables. The average (or mean) and/or the variance (or spread) of the probability distribution of the random variable obviously yield useful information.

While the parameters of a statistical distribution are constant, any computation based on the numerical values of the random observations in a sample may yield different quantities from sample to sample. These quantities are known as ''statistics.'' The two most widely used statistics are the mean of a sample of n observations

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

and the variance of the sample

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n - 1}$$

Note that these descriptors are not theoretical in nature but are calculated from a set of n data points.

Mathematical developments have proven that $\bar{x}$ is usually the best single (point) estimate of μ, and S^2 is usually the best single (point) estimate of σ^2. Normally, the nature of these parameters is totally unknown, and statistics is used to draw inferences about their true values.

Since knowledge of the mean and variance is of utmost importance, statistics deals extensively with developing tools and techniques for studying their behavior. Two basic objectives will be dealt with in the remainder of this chapter:

1. Ways of testing to see whether or not some assumed value of μ or σ^2 is "reasonable" under normal operating or assumed conditions.
2. Ways of using observed values of $\bar{x}$ and S^2 so that one may state with a given measure of confidence that the population parameters of these estimates fall within a given interval. For example, we can be 95% confident that the interval from I_1 to I_2 includes the true value of μ.

   The first objective is dealt with using statistical *hypothesis testing,* while the second one gives rise to *confidence interval estimation.*

Statistical tools also exist for estimating the difference between, or testing an assumption about, the means or variances of two or more probability distributions. These tools are natural extensions of the tools developed for estimating and testing hypotheses about single populations.

All statistical methods have, as a basis, a sample of n observations on the random phenomena of interest. Such methods require that the random sample (of size n) be "representative" of the outcomes that could occur. Because of this, much is said about making sure that the sample is a random sample. Many statistical calculations become invalid when the data used are not representative.

3. STATISTICAL HYPOTHESIS TESTING

Working with a representative (random) sample of n observations, statisticians have shown that the function $Z = \sqrt{n}(\bar{x} - \mu)/\sigma$ for sufficiently large values of n is a random variable that follows (or has) a normal probability distribution with $\mu = 0$ and $\sigma^2 = 1$. This fact is the result of the well-known central limit theorem, which can be stated as follows (Bowker and Lieberman 1972; Devore 1987; Dougherty 1990; Hogg and Craig 1978):

If $\bar{x}$ is the mean of a random sample of size n taken from a population having a mean μ and a finite variance σ^2, then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

is a random variable whose probability distribution approaches that of the standard normal distribution ($\mu = 0$, $\sigma^2 = 1$) as n approaches infinity.

This is undoubtedly the most amazing theorem in statistics for it does not require that one know anything about the shape of the probability distribution of the individual observations. It only requires that the distribution of those random observations have a finite mean, μ, and variance, σ^2. The standard normal density function is defined as

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-1/2z^2} \quad \text{for} \quad -\infty < z < \infty$$

This distribution is graphically represented in Figure 1.

If we use the notation $Z_{\alpha/2}(-Z_{\alpha/2})$ to indicate the value of Z corresponding to an area $\alpha/2$ under the distribution falling to the right (left) of $Z_{\alpha/2}$ $(-Z_{\alpha/2})$, then we can associate the cross-hatched area shown in Figure 2 with the region in which $100(1 - \alpha)$% of all random variables, characterized by the standard normal density function $f(Z)$ with mean $\mu_z = 0$ and variance $\sigma_z^2 = 1$, are expected to lie. Within the context of a hypothesis test, this area will be called the *acceptance region* and the

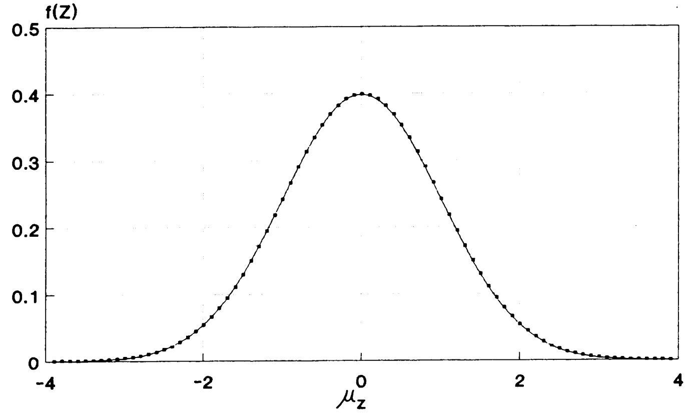


Figure 1 Standardized Normal Distribution.

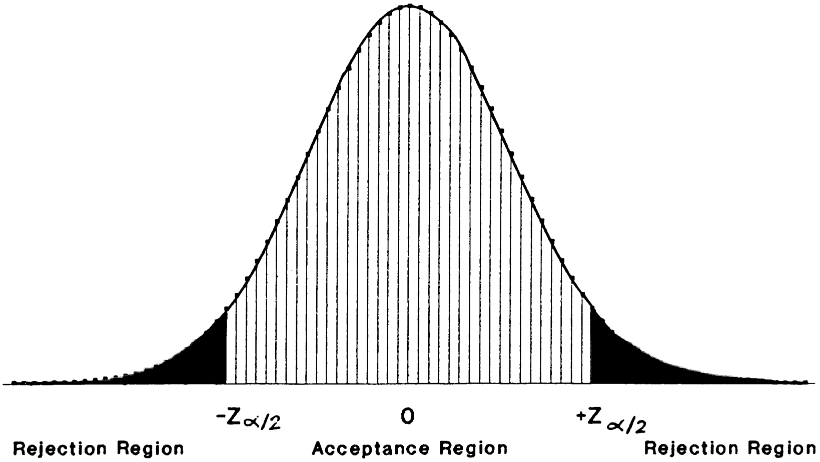


Figure 2 Acceptance and Rejection Regions.

tail areas called the *rejection region.* The points $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ will be called the *rejection points* for reasons which will shortly become evident. Although the underlying probability density function $f(Z)$ might change, these concepts will remain the same. The procedure of statistical hypothesis testing will now be described.

4. TESTING A MEAN VALUE (μ) WITH σ^2 KNOWN

Given a value of $\bar{x}$ computed using a sample of size n from an infinite population with known mean (μ) and variance (σ^2), the probability that the random variable

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

falls between the points $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ is $1 - \alpha$. Note that α is a value between zero and one and represents the probability that a random variable $\bar{x}$, which approximates the mean μ, will naturally fall outside the points $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. This interpretation of the natural behavior of the random variable $\bar{x}$, along with the distribution of the (transformed) variable Z, allows one to structure a hypothesis test concerning the true mean, μ. Assume that the value of $\alpha = 0.05$, and that the following statement is to be tested:

$$H_0: \quad \mu = \mu_0$$
$$H_1: \quad \mu \neq \mu_0$$

The value μ_0 is the numerical value of μ which is assumed known or is hypothesized. H_0 is called the *null* or *primary* hypothesis and H_1 is called the *alternative* or *secondary* hypothesis. If a random sample of size n is extracted from the population under study, then

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

can be calculated. Since $\bar{x}$ is the best point estimator of μ, and μ is assumed to be equal to μ_0 one would expect the random variable

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

to fall between the points $-Z_{\alpha/2} = -Z_{0.025}$ and $Z_{\alpha/2} = Z_{0.025}$ 95% of the time. The values $-Z_{0.025}$ and $Z_{0.025}$ can be determined by using a standard normal table. We can see that they are equal to ± 1.96. These values are called *critical values* and obviously depend upon α. Hence, calculation of Z yields a statistic that will cause H_0 to be believed 95% of the time and H_1 to be believed only 5% of the time, when H_0 is actually *true.* Therefore, α can be interpreted as the magnitude of the error of *rejecting* the *null hypothesis* when in fact it is true. This error is often referred to as an error of type I. Additionally, if the null hypothesis is *false,* there is still a chance that the calculated value of Z will lie between ± 1.96 (when $\alpha = 0.05$). This result will cause the decision analyst to *accept* the null hypothesis when in fact it is *false.* The magnitude (likelihood) of this error is commonly denoted by β, and this error is called an error of type II. Table 1 characterizes the decision process.

In order to illustrate the basic procedures of statistical hypothesis testing, consider the following example:

An oil investment cartel is considering the purchase of an oil well from Blow Hard, Inc. in Texas. Current owners claim that the well produces on the average 100 barrels of oil per day, with a standard deviation of 10 barrels per day. In order to test this claim, the cartel chooses $\alpha = 0.05$ and observes daily production for 16 days. Total production over this period of time is 1690 barrels of oil. Can the owner's claim be disputed?

Assumptions: $\mu = 100$
$\sigma^2 = 100$
Infinite population
$\alpha = 0.05$

Hypothesis test: H_0: $\mu = 100$
H_1: $\mu \neq 100$

Test statistic: $Z - \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Critical values: $\pm Z_{0.025} = \pm 1.96$

Calculated Z statistic: $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{105.63 - 100}{10/4} = 2.252$

Since the value of $Z = 2.252$ is greater than $Z_{\alpha/2} = 1.96$, one would choose to reject H_0 in favor of

TABLE 1 Decision Based upon Sampling Evidence

True State or Nature	H_0 True	H_0 False
Hypothesis H_0 True	No Error	Type I Error
Hypothesis H_0 False	Type II Error	No Error

H_1. In other words, if H_0 is true, it is highly unlikely that a sample of size $n = 16$ would yield a value of $\bar{x}$ equal to 105.63, resulting in a value of Z as large as $Z = 2.252$. What value of $\bar{x}$ would cause the decision maker to accept H_0?

Since $\pm Z_{\alpha/2} = = \pm 1.96$ one can define the following relationship to find the limits of the acceptance region:

$$\pm 1.96 = \frac{\bar{x}_c - 100}{2.5}$$

It can be verified that the solution for $\bar{x}_c$ is given by $\bar{x}_c = 95.1$ and $\bar{x}_c = 104.9$. Therefore, any sample average between these two values will result in the acceptance of H_0.

Next, let us assume that the true population mean (daily production rate) is equal to $\mu_1 = 110$ and not 100. What is the probability that the null hypothesis will be erroneously accepted? Assuming that the variance remains constant, consider the two distributions shown in Figure 3.

Note that the probability that the null hypothesis will be accepted (erroneously) is given by the area marked β in Figure 3. This area can be calculated as follows:

Test statistic: $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Critical value: $Z_1 = \dfrac{104.9 - 110}{2.5} = -2.04$

Type II error: $\beta = P(Z \leq -2.04)$

Using a standard normal table, the probability corresponding to the β area is given by $\beta = 0.0217$. Note, however, that the β error should not include that area to the *left* of $\bar{x}_c = 95.1$. The probability that $\bar{x}$ will be less than this value given that $\mu = 110$ is determined by $P(Z \leq Z_2)$, where

$$Z_2 = \frac{95.1 - 110}{2.5} = -5.96$$

This probability is almost 0. Therefore, $\beta = 0.0217$ is the correct value to four significant digits. In other words the probability of accepting the null hypothesis when μ is actually $\mu_1 = 110$ is only 0.0217. Note that this probability is actually based on the values $\bar{x}_c = 95.1$ and $\bar{x} = 104.9$, which were uniquely determined by the chosen value of α. Clearly, one would never know what the true population mean (μ) actually is in any meaningful application. Hence, the value of β cannot be calculated except in reference to values of μ different from that specified in the null hypothesis. For this example several values of β were calculated for the alternative values of μ_1 indicated in Table 2.

For clarity, one should note that *in the limit* as μ approaches 100, the probability of a type II error approaches $\beta = 1 - \alpha$. At the exact point $\mu_0 = 100$, the null hypothesis is true and a type II error does not exist; hence, $\beta = 0$ at that single point. Figure 4 graphically depicts the behavior of the β error as a function of the true (unknown) population mean, under the original rejection criterion specified by the null hypothesis and the chosen α error. This curve is called an *operating characteristic curve,* or simply an *OC curve.*[1]

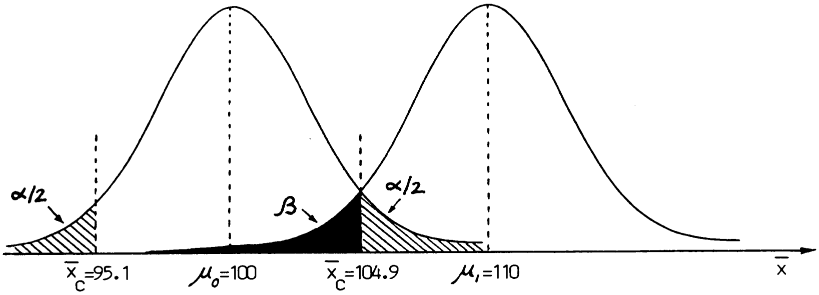


Figure 3 Probabilities of Type I and Type II Errors.

TABLE 2 Probability of Type II Error

					Population Mean Values								
91	93	95	96	97	98	99	101	102	103	104	105	107	109
0.025	0.200	0.484	0.641	0.776	0.874	0.932	0.932	0.874	0.776	0.641	0.484	0.200	0.025

The calculations shown in Table 2 were performed relative to an alternative hypothesis that included two rejection points $(-Z_{\alpha/2}, Z_{\alpha/2})$. Such a hypothesis test is called a *two-tailed hypothesis test.*

Consider once again our numerical example. Note that the null hypothesis states that well production is *exactly* 100 barrels per day. In reality, the purchase of the well would be desirable if the daily production met or exceeded 100 barrels per day. In that case the null and alternative hypotheses would be

Case I: H_0: $\mu \geq 100$
H_1: $\mu < 100$

or

Case II: H_0: $\mu \leq 100$
H_1: $\mu > 100$

Both hypothesis statements reflect the same objective, but there are significant differences in the decision criterion utilized in each hypothesis. The null hypothesis of case I assumes that the well is producing 100 or more barrels per day unless statistical evidence proves otherwise, resulting in rejection of H_0. The null hypothesis of case II assumes that the well production is inferior unless production records indicate that daily output is more than 100 barrels per day, which will result in rejection of H_0. Both tests are valid and are called *one-tailed* hypothesis tests under a given type I error. Consider again the original data with $\alpha = 0.05$. Table 3 summarizes the calculations for the significance tests associated with cases I and II.

Both test statistics in Table 3 use the value of 100 in calculating a Z value, for it is at this single point that the type I error and the type II errors are the greatest. It should also be noted that the type II error now exists in only one direction, and hence the operating characteristic curve will be one sided. For completeness, one should note that the null hypothesis is an a priori state of nature that one chooses to believe, unless statistical evidence indicates otherwise. In statistics, one does not "prove" the null hypothesis but rather "fail to reject" the hypothesis. These concepts are consistent

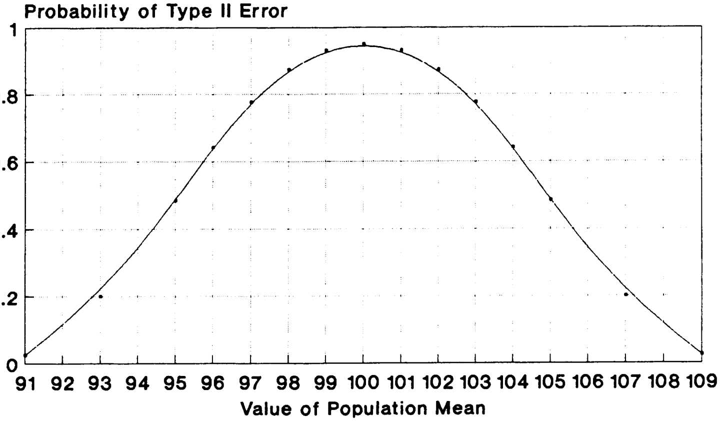


Figure 4 Operating Characteristic Curve. Level of significance = 0.05.

TABLE 3 One-Tail Hypothesis Tests for Means

Procedure	Case I	Case II
Assumptions	$\mu \geq 100$ $\sigma^2 = 100$ Normal population $\alpha = .05$	$\mu \leq 100$ $\sigma^2 = 100$ Normal population $\alpha = .05$
Hypothesis test	H_0: $\mu \geq 100$ H_1: $\mu < 100$	H_0: $\mu \leq 100$ H_1: $\mu > 100$
Test statistic	$Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$	$Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$
Critical value	$-Z_\alpha = -1.645$	$Z_\alpha = 1.645$
Calculated value	$Z = 2.252$	$Z = 2.252$
Conclusion	Accept H_0 Buy the well	Reject H_0 Buy the well

with the underlying uncertain (stochastic) nature under which hypothesis testing is conducted. One can never be absolutely certain of statistical inference. Along these same lines of thought, it is critical that the hypothesis test be chosen *before* statistical sampling and not after. Selection of the test (one or two tailed) and the associated a error should never be chosen a posteriori to "statistically confirm" any belief. Such statistical inference is obviously improper.

In summary, a decision maker can apply either a one-tailed or two-tailed hypothesis test with a chosen type I error probability equal to α. The true probability of type II error (β) is always unknown since the actual population mean is unknown. However, the β risk can be characterized by the construction of an OC curve.

5. TESTING A MEAN VALUE (μ) WITH σ^2 UNKNOWN

Consider again the oil well example. After detailed examination, it was discovered that the theoretical variance (σ^2) was really not known but had been "guessed at" by the seller. In order to estimate σ^2, the sample of 16 days was used to calculate S^2 in the following manner:

$$S^2 = \frac{\sum_{i-1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Using $n = 16$, a value of $S^2 = 225$ was calculated. Under the assumption that σ^2 is unknown, the Z statistic is no longer a valid test statistic. It can be shown that the following test statistic should be used in this case (Bowker and Lieberman 1972; Devore 1987; Hogg and Craig 1978).

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

This is called simply the student t test statistic. The rejection points $t_{\alpha/2}$ and $-t_{\alpha/2}$ for a two-tailed test can be determined from an appropriate t table, but they now depend on a parameter called the "degrees of freedom," which is defined by $df = n - 1$. Consider the original two-tailed hypothesis test under the new assumption (σ^2 unknown):

Assumptions: σ^2 unknown
$\mu = 100$
Infinite population
Population is normal
$\alpha = 0.05$

Hypothesis test: H_0: $\mu = 100$
H_1: $\mu \neq 100$

Test statistic: $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$

Critical values: $t_{\alpha/2,df} = t_{0.025,15} = 2.131$
$-t_{\alpha/2,df} = -t_{0.025,15} = -2.131$

Calculated t statistic: $t = \dfrac{105.63 - 100}{15/4} = 1.501$

Therefore, the null hypothesis cannot be rejected and one is led to believe that the true well production is actually 100 barrels per day.

This example illustrates the use of the t statistic for testing a hypothesis related to a population mean value (μ) when the variance (σ^2) is *unknown.* As before, both two-tailed and one-tailed tests are possible, whichever the problem situation demands. Rejection limits are set based on a chosen type I (α) error, and operating characteristic curves (OC curves) can be constructed to reflect the associated type II (β) error risk. Finally, one should note that the t test was necessitated due to the fact that σ^2 was being estimated by S^2 from a sample of size n. If n is large enough, then one would expect S^2 to closely approximate σ^2 and the Z test can be used anyway. For $n \geq 30$ this is generally an acceptable procedure.

6. HYPOTHESIS TESTING: SINGLE VARIANCE

Continuing with the example about oil well production, the potential buyer was quite perplexed at the result obtained from the two-tailed t test, since it differed from that previously obtained via the original two-tailed Z test. An engineer explained that the difference was probably a result of lack of knowledge concerning the population variance, σ^2. Since σ^2 was unknown, the induced uncertainty caused failure to reject the null hypothesis (note that the rejection points were wider).

Reflecting upon this logic, management requested a statistical examination of the sample variance to determine if the original (specified) σ^2 had indeed changed, since the estimated value of σ^2 (S^2) was greater than the original specified value. The proper statistical test is a chi-square test. The chi-square test is described as follows:

Assumptions: $\sigma^2 = 100$
Infinite population
Population is normal
$\alpha = .05$

Hypothesis test: H_0: $\sigma^2 = 100$
H_1: $\sigma^2 \neq 100$

Test statistic: $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2}$

Critical values: $\chi^2_{df,1-\alpha/2}$; $\chi^2_{df,\alpha/2}$

Degrees of freedom: $df = n - 1$

As in previous tests, the χ^2 rejection values are both functions of the chosen type 1 error (α) and the sample size ($df = n - 1$). Critical values are easily obtained from χ^2 tables. For the numerical example the procedure is as follows:

Hypothesis test: H_0 $\sigma^2 = 100$
H_1 $\sigma^2 \neq 100$

Test statistic: $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{(15)(225)}{100} = 33.75$

Critical values: $\chi^2_{15,0.975} = 6.262$, $\chi^2_{15,0.025} = 27.488$

Since $\chi^2 = 33.75$ is greater than the critical value $\chi^2_{15,0.025} = 27.488$, one is led to believe that the underlying statistical variance of well production has changed from 100 to something else. Of course, based on the value of S^2, one might conclude that it has increased to somewhere around 225. Although further investigation would be up to decision maker (buyer), it appears that a good course of action could be to obtain more sample data and reexamine the entire situation.

7. HYPOTHESIS TESTING: TWO POPULATION MEANS WITH VARIANCES KNOWN

After a rather lengthy discussion, company management reached the decision that the well should be purchased. Due to optimistic market projections, it was decided that a second well should also be purchased if one could be found that produced 100 more barrels per day on the average than the single established well. Further testing led management to believe that, for comparison purposes, the first well did indeed produce at a rate of 100 barrels per day. At this point Blow Hard, Inc. presented a new well that it claimed produced at a rate of 100 barrels per day more than the first well. Management again insisted on statistical investigation and in order to provide new, current data, a separate independent evaluation was undertaken. Samples from both wells were obtained to compare daily production rates. Blow Hard, Inc. assured the purchasing cartel that the variance in daily production for the first well was actually 240 and for the second well 276.

Additionally, a sample of production records for $n_1 = 12$ days from well 1 and $n_2 = 18$ days from well 2 provided average daily production of $\bar{x}_1 = 102$ and $\bar{x}_2 = 212$ barrels per day, respectively.

Since the variances of production, $\sigma_1^2 = 240$ and $\sigma_2^2 = 276$, are both assumed known, the proper test statistic is a Z test for the difference in the means of populations. The procedure is as follows:

Assumptions: σ_1^2 and σ_2^2 known
Infinite populations
Normal populations
$\alpha = 0.05$

Hypothesis test: H_0: $\mu_2 - \mu_1 = \delta$
H_1: $\mu_2 - \mu_1 \neq \delta$
$\delta = 100$

Test statistic: $Z = \dfrac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Critical values: $-Z_{\alpha/2}, Z_{\alpha/2}$

The numerical calculations are as follows:

$$H_0: \quad \mu_2 - \mu_1 = 100$$

$$H_1: \quad \mu_2 - \mu_1 \neq 100$$

$$Z = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{(212 - 102) - 100}{\sqrt{\dfrac{240}{12} + \dfrac{276}{18}}}$$

$$Z = \frac{10}{5.94} = 1.684$$

The critical value for this test is given by $\pm Z_{\alpha/2} = \pm Z_{0.025} = \pm 1.96$. Hence, there is no statistical evidence to support the rejection of the null hypothesis, and the proper decision would be to purchase both wells under the management interpretation of these results. However, the same clever engineer who questioned the first test results again questioned the validity of using σ_1^2 and σ_2^2 in the calculations. The question was then posed, "Can we perform a similar test without knowing the population variances?" The statistician responded yes as the same example used to calculate $\bar{x}_1$ and $\bar{x}_2$ could be used to estimate σ_1^2 and σ_2^2 with S_1^2 and S_2^2, respectively.

8. HYPOTHESIS TESTING WITH TWO MEANS: POPULATION VARIANCES UNKNOWN BUT ASSUMED EQUAL

The test statistic, assumptions, and hypothesis test are as follows:

Assumptions: σ_1^2 and σ_2^2 (both are unknown)
Infinite populations
Normal populations
$\alpha = .05$

Hypothesis test: H_0: $\mu_2 - \mu_1 = \delta$
H_1: $\mu_2 - \mu_1 \neq \delta$

Test statistic: $t = \dfrac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{\sqrt{(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2}} \sqrt{\dfrac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$

Critical values: $-t_{\alpha/2,df}, t_{\alpha/2,df}$

Degrees of freedom: $df = n_1 + n_2 - 2$

Using the same set of data, it was found that $S_1^2 = 165$ and $S_2^2 = 453$. The calculations related to this example are as follows:

Hypothesis test: H_0: $\mu_2 - \mu_1 = 100$
H_1: $\mu_2 - \mu_1 \neq 100$

The calculated value of the test statistic is given by

$$t = \frac{(212 - 102) - (100)}{\sqrt{11(165) + 17(453)}} \sqrt{\frac{(12)(18)(28)}{30}} = 1.454$$

The critical values (rejection points) for this test are $t_{0.025,28} = 2.048$ and $-t_{0.025,28} = -2.048$. Since the calculated value of $t = 1.454$ is less than the upper rejection value, there is insufficient evidence to reject the null hypothesis. Finally, one should note that under the assumption that $\sigma_1^2 = \sigma_2^2$, a pooled estimate of the variance is used from a total sample of $N = n_1 + n_2$. As in the single parameter t test, if n_1 and n_2 are both greater than 30, one can simply use the two parameter Z test directly.

Our same clever engineer now observes that these results can be reached only if it can be assumed that the two population variances are equal. At this point, the manager asks our statistician, "Can we test this assumption?" The answer is yes, using an F test for equality of variances.

9. HYPOTHESIS TESTING FOR EQUALITY OF TWO POPULATION VARIANCES

The following assumptions, hypothesis test, and test statistic should be observed when conducting an F test:

Assumptions: Normal populations
Infinite populations
Hypothesis test: H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_2^2 \neq \sigma_2^2$
Test statistic: $F = S_1^2/S_2^2$
Critical values: $F_{1-\alpha/2,v_1,v_2}$, $F_{\alpha/2,v_1,v_2}$
v_1 = degrees of freedom in numerator
$= n_1 - 1$
v_2 = degrees of freedom in denominator
$= n_2 - 1$

The critical values $F_{\alpha/2,v_1,v_2}$ for commonly used values of α are easily found in statistical tables. Once these values are obtained, the values of $F_{1-\alpha/2,v_1,v_2}$ are easily calculated for the left-hand rejection point according to the following formula

$$F_{1-\alpha/2,v_1,v_2} = (F_{\alpha/2,v_2,v_1})^{-1}$$

For example, if S^2 was calculated using a sample of size $n_1 = 13$ and S_2^2 was calculated using a sample of size $n_2 = 21$; the rejection points for the quantity $F = S_1^2/S^2$ for $\alpha = .10$ would be given by

$$F_{\alpha/2,v_1,v_2} = F_{.05,12,20} = 2.28$$

and

$$F_{1-\alpha/2,v_1,v_2} = (F_{0.05,20,12})^{-1} = (2.54)^{-1} = 0.394$$

In order to illustrate the F test, consider the previous example. Recall that $S_1^2 = 165$, $n_1 = 12$, $S_2^2 = 453$, and $n_2 = 18$.

Assumptions: $\alpha = 0.10$
Normal populations
Hypothesis test: H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 \neq \sigma_2^2$
Test statistic: $F = S_1^2/S_2^2$

Critical values: Using $\alpha = 0.10$, $n_1 = 12$, $n_2 = 18$, we obtain the critical values shown:

$$F_{0.05,11,17} = 2.41$$

$$F_{0.95,11,17} = (F_{0.05,17,11})^{-1} = 0.372$$

Calculated F statistic: $F = S_1^2/S_2^2 = 165/453 = 0.364$

Hence, the null hypothesis would be rejected and one is led to assume that $\sigma_1^2 \neq \sigma_2^2$. Our astute manager, still seeking statistical evidence, now inquires as to the availability of a statistical test when $\sigma_1^2 \neq \sigma_2^2$. Fortunately, such a test is available, and is called the t' test.

10. EQUALITY OF TWO MEANS WITH VARIANCES UNKNOWN AND NOT EQUAL

The proper statistical test for this procedure is the t' test and is based on the following:

Assumptions: $\sigma_1^2 \neq \sigma_2^2$
Normal populations
Infinite populations

Hypothesis test: H_0: $\mu_1 - \mu_2 = \delta$
H_1: $\mu_1 - \mu_2 \neq \delta$

Test statistic: $$t' = \frac{(\bar{x}_2 - \bar{x}_1) - \delta}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

Critical values: $t_{\alpha/2,df}$, $t_{1-\alpha/2,df}$

Degrees of freedom: $$df = \left[\frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}\right]$$

For the oil well example, the following calculations illustrate the procedure:

Assumptions: $\sigma_1^2 \neq \sigma_2^2$
Normal populations
Infinite populations
$\alpha = 0.05$

Hypothesis test: H_0: $\mu_2 - \mu_1 = 100$
H_1: $\mu_2 - \mu_1 \neq 100$

Critical values: $t_{\alpha/2,df}$, $-t_{\alpha/2,df}$

Degrees of freedom: $$df = \left[\frac{\left(\dfrac{165}{12} + \dfrac{453}{18}\right)^2}{\dfrac{\left(\dfrac{165}{12}\right)^2}{11} + \dfrac{\left(\dfrac{453}{18}\right)^2}{17}}\right] = [27.82] = 27$$

Critical values: $t_{0.025,27} = 2.052$
$-t_{0.025,27} = -2.052$

Calculated t statistic: $$t' = \frac{(212 - 102) - 100}{\sqrt{\dfrac{165}{12} + \dfrac{453}{18}}} = \frac{10}{\sqrt{38.92}}$$

$$t' = 1.603$$

Since $t' = 1.603$ is less than the critical value of $t = 2.052$, statistical evidence does not support rejection of the null hypothesis.

11. CONFIDENCE INTERVAL ESTIMATION

A subject closely related to hypothesis testing is that of confidence intervals. The basic concepts are best understood by considering once again the test statistic:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

If, as before, we use the notation $Z_{\alpha/2}$ and $-Z_{\alpha/2}$ to indicate the values of the random variable Z where $\alpha/2$ of the area lies to the right of $Z_{\alpha/2}$ and $\alpha/2$ to the left of $-Z_{\alpha/2}$, then the following probability statement must be true:

$$P\left\{-Z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right\} = 1 - \alpha$$

If we rearrange the inequality in brackets and solve for the true population parameter μ, then we obtain

$$\bar{x} - Z_{\alpha/2}\, \sigma/\sqrt{n} \leq \mu \leq \bar{x} + Z_{\alpha/2}\, \sigma/\sqrt{n}$$

This last expression says that if we take a random sample of n observation on the random phenomenon of interest and calculate the interval,

$$\bar{x} - Z_{\alpha/2}\, \sigma/\sqrt{n} \quad \text{to} \quad \bar{x} + Z_{\alpha/2}\, \sigma/\sqrt{n}$$

we can be $100(1 - \alpha)\%$ confident that the interval will include the true mean, μ. In other words, if we repeatedly took samples of size n and calculated the interval from each sample, in the long run $100(1 - \alpha)\%$ of the intervals will include μ. The interval is obviously calculated only once, and the resulting values of the endpoints constitute a $100(1 - \alpha)\%$ confidence interval estimate of μ. For a numerical illustration consider once again the first example given in this chapter, that is, the single oil well production problem:

Assumptions: $\alpha = 0.05$
$\sigma^2 = 100$
Critical values: $\pm Z_{\alpha/2} = \pm Z_{0.025} = \pm 1.96$
Input data: $\bar{x} = 105.63$, $n = 16$

The 95% confidence interval for the true (unknown) population mean (μ) is given by

$$105.63 - (1.96)(5/2) \leq \mu \leq 105.63 + (1.96)(5/2)$$

or

$$100.73 \leq \mu \leq 110.53$$

Note that this interval assumes that one knows σ^2, the variance of the random variable. Usually this is not the case, and we have to estimate σ^2 with S^2. If we change the expression for Z accordingly, we get a different distribution. As already mentioned in this chapter,

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

is a random variable that follows a t distribution.

Now, utilizing the knowledge that we have provided about the t distribution, we write a probability statement similar to the one for the previous case:

$$P\left\{-t_{\alpha/2} \leq \frac{\bar{x} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}\right\} = 1 - \alpha$$

and upon rearranging the inequality in the brackets, we obtain

$$\bar{x} - t_{\alpha/2}\, S/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2}\, S/\sqrt{n}$$

which is a 100(1 − α)% confidence interval estimate of μ when σ is not known.

It should be pointed out that when the sample size from which S is calculated exceeds 30, it makes little difference whether one uses the normal distribution with $\sigma = S$ or the more precise distribution. The probability distributions of t and Z become the same at $n = \infty$.

If one is interested in calculating a 100(1 − α)% confidence interval for the quantity $\mu_1 - \mu_2$, the difference between the means of two different distributions, one would take a random sample from each distribution, n_1 from the first and n_2 from the second. From these samples one would calculate $\bar{x}_1$ and S_1^2 as well as $\bar{x}_2$ and S_2^2. In a manner identical to that used earlier, the following 100(1 − α)% confidence interval could be constructed:

$$P\left\{(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} = 1 - \alpha$$

Clearly, this procedure could be repeated for *any* test statistic previously discussed in this section. The reader is referred to any of a number of engineering statistics texts for developments of χ^2, F, and t' confidence intervals.

12. MAXIMUM-LIKELIHOOD ESTIMATORS

A well-known procedure for finding estimators of unknown parameters is the method of maximum likelihood (Devore 1987; Dougherty 1990; Hogg and Craig 1978). Maximum-likelihood estimators are consistent and have minimum variance but are not always unbiased. A summary of the procedure follows.

Let X be a random variable with density function $f(x,\theta)$, where the parameter θ is unknown. Given a random sample of independent observations $x_1, x_2, \ldots, x_n$, the likelihood function is defined as

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1, \theta) f(x_2, \theta), \ldots, f(x_n, \theta)$$

The likelihood function is actually the joint probability density function (for continuous variables) or the joint mass probability function (for discrete variables) of the n random variables. Therefore, the value of θ for which the observed sample would have the highest probability of being extracted, can be found by maximizing the likelihood function over all possible values of the parameter θ. As shown in elementary calculus, this can be achieved by setting the first derivative of the likelihood function with respect to the parameter equal to zero, and then solving for θ:

$$dL/d\theta = 0$$

The solution to this equation can be more efficiently found by considering the logarithm (base e) of the likelihood function, instead of the function itself:

$$d \ln L/d\theta = 0$$

To illustrate the fundamental steps of the procedure followed to obtain a maximum-likelihood estimator, we will consider a random variable X having the exponential density function:

$$f(x; \theta) = \theta e^{-\theta x}$$

where x is nonnegative. For a sample size n the likelihood function for this example is given by

$$L(x_1, x_2, \ldots, x_n) = \Pi_i \theta e^{-\theta x_i}$$

where the index i takes on the values $i = 1, 2, \ldots, n$. Taking the natural logarithm of the preceding likelihood function, we obtain

$$\ln L(x_1, x_2, \ldots, x_n) = n \ln \theta - \theta \sum_i x_i$$

After differentiating ln L with respect to θ, setting the derivative equal to zero, and solving for θ we obtain the final result shown:

$$\hat{\theta} = n / \sum_i x_i$$

which is equal to the inverse of the sample mean. Now let us assume that the random sample consists of the following values for $n = 5$: 0.9, 1.7, 0.4, 0.3, 2.4. In this case the maximum-likelihood estimator would be equal to $\hat{\theta} = 5/5.7 = 0.88$. It can be verified that $\overline{x}$ and $S^2 = \Sigma_i(x_i - \overline{x})^2/n$ are maximum-likelihood estimators for the mean and variance, respectively, of a normal distribution (Bowker and Lieberman 1972).

13. TESTING FOR EQUALITY OF MEANS AND VARIANCES FOR *K* POPULATIONS

13.1. Testing Means

In the case where there are k populations under consideration, and the test of hypothesis is equality of population means, a different type of procedure is necessitated. This area of statistics is called *experimental design* or *analysis of variance.* This topic is covered in Chapter 85 of this Handbook.

13.2. Testing the Homogeneity of Variances

A problem frequently encountered in applied statistics is that of testing the equality of variances of several normal populations. Let us assume that we have collected a random sample of size n_i from a normal population with mean μ_i and variance σ_i^2, repeating this basic procedure for $i = 1, 2, \ldots, k$. The hypothesis to be tested in this case can be formulated as

$$H_0: \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

As already explained earlier in this chapter, an F test can be used for $k = 2$. However, a different procedure is required for larger values of k.

Among several methodologies for testing this hypothesis, the following three are perhaps the best known (Snedecor and Cochran 1980): (a) Cochran's test, (b) Barlett's test, and (c) Levene's test. Each of these methods is described below.

13.3. Cochran's Test

In the case where there are k populations and the test of equality of population variances is required, the most commonly applied test is the Cochran's test for homogeneity of variances.

Assumptions: Samples are independent
Populations are normal

Test statistic: $R = \dfrac{\max\{S_1^2, S_2^2, \ldots, S_k^2\}}{\sum_{i=1}^{k} S_i^2}$

In the relationship defined for the test statistic R, S_i^2 is the unbiased point estimator of σ_i^2 for $i = 1, 2, \ldots, k$. Each S_i^2 is calculated from a sample of size n. The corresponding test of significance is conducted according to the following rule:

Test rule: Accept the null hypothesis that $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ if $R \leq RC_{\alpha,n,k}$, where $RC_{\alpha,n,k}$ is a critical value for chosen values of the type I error (α), the sample size (n), and the number of populations (k). Tables of critical values for $\alpha = 0.05$ and 0.01, for values of k up to 120 and n up to 145, are given in Bowker and Lieberman (1972).

13.4. Barlett's Test

Barlett developed in 1937 a testing procedure that can be used when all sample sizes n_i for $i = 1, 2, \ldots, k$, are not equal. The procedure is described as follows, where N indicates the total number of observations collected from the k populations:

Assumptions: Samples are independent
Populations are normal

Test statistic: $M = -\dfrac{1}{c}\,\Sigma_i(n_i - 1)\ln\dfrac{S_i^2}{S^2}$

$S^2 = \Sigma_i(n_i - 1)S_i^2/(N - k)$

$c = (\Sigma_i\, 1/f_i - 1/f)/3(k - 1) + 1$

TABLE 4 Sample Calculations for Goodness of Fit Testing

Sample	O_i	e_i	$(O_i - e_i)^2/e_i$
1	18	19.4	0.101
2	17	19.4	0.297
3	19	19.4	0.008
4	17	19.4	0.297
5	18	19.4	0.101
6	20	19.4	0.019
Total	—	—	0.823

Test rule: The statistic M has approximately a χ^2 distribution with $k - 1$ degrees of freedom. This approximation is more appropriate for sample sizes larger than 3. If the observed value of M exceeds the critical value of the chi-square statistic for a level of significance α and $k - 1$ degrees of freedom, the hypothesis is rejected.

13.5. Levene's Test

An approximate test, which is less sensitive to the lack of normality in the data than Barlett's test, was developed by H. Levene in 1960. The procedure assumes that all sample sizes are equal to *n*. This testing method is described as follows:

Assumptions: Samples are independent
Populations are normal (or approximately normal)
Test statistic: F

Test rule: For Levene's test we conduct an analysis of variance (ANOVA) of the absolute deviations from each sample average. Details on the ANOVA procedure are given in another chapter of this handbook. If the observed mean square ratio exceeds the appropriate critical value of the F statistic, we reject the hypothesis that all variances are equal.

14. OTHER USES OF HYPOTHESIS TESTING

Finally, it should be noted that the concepts of type 1(α) error, type II (β) error, critical values, OC curves, and one/two-tailed hypothesis tests are common to all hypothesis testing. This chapter has illustrated only tests concerning means and variances since they are most common to industrial engineering. One might also find occasions to test hypotheses in quality control applications, proportions, percentages, and in goodness-of-fit testing. These applications, and others, are well documented in a host of applied engineering textbooks.

14.1. Nonparametric Tests

One should note that for most of the hypothesis tests we have discussed, the assumption of normally distributed random variables is required. In actual practice this may not be justified. One has two choices by which this assumption can be ignored. First, one can obtain enough samples to use a Z test (normal tables) rather than a t test (t tables). If the sample is large enough, then the assumption of normality is not required. However, in the case where larger samples cannot be obtained or cost prohibits larger samples, the use of *nonparametric statistical tests* is necessitated (Hollander and Wolfe 1973; Lehmann 1975; Marascuilo and McSweeney 1977).

A nonparametric test is one that requires no assumptions regarding the form or shape of the underlying random variables. Usually, all that is required is knowledge of the scale of measurement used in the experiment and whether the random variable is discrete or continuous. The treatment of nonparametric statistics is well beyond the scope of this introductory section. However, the reader should be aware of its role in hypothesis testing. A test frequently used in industrial organizations is the goodness of fit test, which is described in the next section.

14.2. Goodness of Fit Test

The assumption of normality is not an unusual one in problems dealing with hypothesis testing about means. However, it becomes an important one when testing hypotheses about variances. For this reason, before testing the hypotheses, one should verify if there is sufficient statistical evidence indicating that the population is normal. A statistical method used for this purpose is known as ''test

TABLE A **Cumulative Normal Distribution** $\theta(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$

t	0.00	0.01	0.02	0.03	0.04
0.0	0.500 00	0.503 99	0.507 98	0.511 97	0.515 95
0.1	0.539 83	0.543 79	0.547 76	0.551 72	0.555 67
0.2	0.579 26	0.583 17	0.587 06	0.590 95	0.594 83
0.3	0.617 91	0.621 72	0.625 51	0.629 30	0.633 07
0.4	0.655 42	0.659 10	0.662 76	0.666 40	0.670 03
0.5	0.691 46	0.694 97	0.698 47	0.701 94	0.705 40
0.6	0.725 75	0.729 07	0.732 37	0.735 65	0.738 91
0.7	0.758 03	0.761 15	0.764 24	0.767 30	0.770 35
0.8	0.788 14	0.791 03	0.793 89	0.796 73	0.799 54
0.9	0.815 94	0.818 59	0.821 21	0.823 81	0.826 39
1.0	0.841 34	0.843 75	0.846 13	0.848 49	0.850 83
1.1	0.864 33	0.866 50	0.868 64	0.870 76	0.872 85
1.2	0.884 93	0.886 86	0.888 77	0.990 65	0.892 51
1.3	0.903 20	0.904 90	0.906 58	0.908 24	0.909 88
1.4	0.919 24	0.920 73	0.922 19	0.923 64	0.925 06
1.5	0.933 19	0.934 48	0.935 74	0.936 99	0.938 22
1.6	0.945 20	0.946 30	0.947 38	0.948 45	0.949 50
1.7	0.955 43	0.956 37	0.957 28	0.958 18	0.959 07
1.8	0.965 07	0.964 85	0.965 62	0.966 37	0.967 11
1.9	0.971 28	0.971 93	0.972 57	0.973 20	0.973 81
2.0	0.977 25	0.977 78	0.978 31	0.978 82	0.979 32
2.1	0.982 14	0.982 57	0.983 00	0.983 41	0.983 82
2.2	0.986 10	0.986 45	0.986 79	0.987 13	0.987 45
2.3	0.989 28	0.989 56	0.989 83	0.990 10	0.990 36
2.4	0.991 80	0.992 02	0.992 24	0.992 45	0.992 66
2.5	0.993 79	0.993 96	0.994 13	0.994 30	0.994 46
2.6	0.995 34	0.995 47	0.995 60	0.995 73	0.995 85
2.7	0.996 53	0.996 64	0.996 74	0.996 83	0.996 93
2.8	0.997 44	0.997 52	0.997 60	0.997 67	0.997 74
2.9	0.998 13	0.998 19	0.998 25	0.998 31	0.998 36
3.0	0.998 65	0.998 69	0.998 74	0.998 78	0.998 82
3.1	0.999 03	0.999 06	0.999 10	0.999 13	0.999 16
3.2	0.999 31	0.999 34	0.999 36	0.999 38	0.999 40
3.3	0.999 52	0.999 53	0.999 55	0.999 57	0.999 58
3.4	0.999 66	0.999 68	0.999 69	0.999 70	0.999 71
3.5	0.999 77	0.999 78	0.999 78	0.999 79	0.999 80
3.6	0.999 84	0.999 85	0.999 85	0.999 86	0.999 86
3.7	0.999 89	0.999 90	0.999 90	0.999 90	0.999 91
3.8	0.999 93	0.999 93	0.999 93	0.999 94	0.999 94
3.9	0.999 95	0.999 95	0.999 96	0.999 96	0.999 96
0.0	0.519 94	0.523 92	0.527 90	0.531 88	0.535 86
0.1	0.559 62	0.563 56	0.567 49	0.571 42	0.575 34
0.2	0.598 71	0.602 57	0.606 42	0.610 26	0.614 09
0.3	0.636 83	0.640 58	0.644 31	0.648 03	0.651 73
0.4	0.673 64	0.677 24	0.680 82	0.684 38	0.687 93
0.5	0.708 84	0.712 26	0.715 66	0.719 04	0.722 40
0.6	0.742 15	0.745 37	0.748 57	0.751 75	0.754 90
0.7	0.773 37	0.776 37	0.779 35	0.782 30	0.785 23
0.8	0.802 34	0.805 10	0.807 85	0.810 57	0.813 27
0.9	0.828 94	0.831 47	0.833 97	0.836 46	0.838 91
1.0	0.853 14	0.855 43	0.857 69	0.859 93	0.862 14
1.1	0.874 93	0.876 97	0.879 00	0.881 00	0.882 97
1.2	0.894 35	0.896 16	0.897 96	0.899 73	0.901 47
1.3	0.911 49	0.913 08	0.914 65	0.916 21	0.917 73
1.4	0.926 47	0.927 85	0.929 22	0.930 56	0.931 89

TABLE A (*Continued*)

t	0.00	0.01	0.02	0.03	0.04
1.5	0.939 43	0.940 62	0.941 79	0.942 95	0.944 08
1.6	0.950 53	0.951 54	0.952 54	0.953 52	0.954 48
1.7	0.959 94	0.960 80	0.961 64	0.962 46	0.963 27
1.8	0.967 84	0.968 56	0.969 26	0.969 95	0.970 62
1.9	0.974 41	0.975 00	0.975 58	0.976 15	0.976 70
2.0	0.979 82	0.980 30	0.980 77	0.981 24	0.981 69
2.1	0.984 22	0.984 61	0.985 00	0.985 37	0.985 74
2.2	0.987 78	0.988 09	0.988 40	0.988 70	0.988 99
2.3	0.990 61	0.990 86	0.991 11	0.991 34	0.991 58
2.4	0.992 86	0.993 05	0.993 24	0.993 43	0.993 61
2.5	0.994 61	0.994 77	0.994 92	0.995 06	0.995 20
2.6	0.995 98	0.996 09	0.996 21	0.996 32	0.996 43
2.7	0.997 02	0.997 11	0.997 20	0.997 28	0.997 36
2.8	0.997 81	0.997 88	0.997 95	0.998 01	0.998 07
2.9	0.998 41	0.998 46	0.998 51	0.998 56	0.998 61
3.0	0.998 86	0.998 89	0.998 93	0.998 97	0.999 00
3.1	0.999 18	0.999 21	0.999 24	0.999 26	0.999 29
3.2	0.999 42	0.999 44	0.999 46	0.999 48	0.999 50
3.3	0.999 60	0.999 61	0.999 62	0.999 64	0.999 65
3.4	0.999 72	0.999 73	0.999 74	0.999 75	0.999 76
3.5	0.999 81	0.999 81	0.999 82	0.999 83	0.999 83
3.6	0.999 87	0.999 87	0.999 88	0.999 83	0.999 89
3.7	0.999 91	0.999 92	0.999 92	0.999 92	0.999 92
3.8	0.999 94	0.999 94	0.999 95	0.999 95	0.999 95
3.9	0.999 96	0.999 96	0.999 96	0.999 97	0.999 97

Source: From W. M. Hines and D. C. Montgomery, *Probability and Statistics in Engineering Science,* 2nd Ed., John Wiley & Sons, New York, 1980, pp. 474–475. Reprinted by permission.

of goodness of fit.'' Actually, this test can be used to verify if a random sample comes from a specified theoretical distribution (such as the binomical, Poisson, uniform, and normal distributions). The procedure can be described as follows:

1. A theoretical distribution is specified in the null hypothesis H_0.
2. A level of significance α is assumed.
3. An empirical distribution with m frequency classes is formed with the observations in a random sample of size n. In this distribution 0_i is the observed frequency (number of observations) in class i, for $i = 1, 2, \ldots, m$.
4. The test statistic used is defined as

$$G = \sum_{i=1}^{m} (O_i - e_i)^2 / e_i$$

where e_i is the expected frequency for class i according to the theoretical distribution specified in the null hypothesis H_0.

It is shown that this test statistic approximately follows a chi-square distribution, with $m - 1 - r$ degrees of freedom, where r is the number of population parameters estimated by sample statistics.

5. The decision rule is simply stated as follows:

H_0 is rejected if the calculated value of G exceeds the critical value $\chi^2_{m-1-r,\alpha}$

The following numerical example illustrates these steps of the goodness of fit procedure. In a test of industrial welding robots 6 samples of 20 robots each were taken at random. Each robot in the sample operated continuously until it either failed or reached 1000 hr of operation. It is desired to verify if the true proportion of the entire population of robots that can operate over 1000 hr is 0.97.

TABLE B Percentage Points of the *t* Distribution

	α										
v^a	0.45	0.40	0.35	0.30	0.25	0.125	0.05	0.025	0.0125	0.005	0.0025
1	0.158	0.325	0.510	0.727	1.000	2.414	6.314	12.71	25.45	63.66	127.3
2	0.142	0.289	0.445	0.617	0.817	1.604	2.920	4.303	6.205	9.925	14.09
3	0.137	0.277	0.424	0.584	0.765	1.423	2.353	3.183	4.177	5.841	7.453
4	0.134	0.271	0.414	0.569	0.741	1.344	2.132	2.776	3.495	4.604	5.598
5	0.132	0.267	0.408	0.559	0.727	1.301	2.015	2.571	3.163	4.032	4.773
6	0.131	0.265	0.404	0.553	0.718	1.273	1.943	2.447	2.969	3.707	4.317
7	0.130	0.263	0.402	0.549	0.711	1.254	1.895	2.365	2.841	3.500	4.029
8	0.130	0.262	0.399	0.546	0.706	1.240	1.860	2.306	2.752	3.355	3.833
9	0.129	0.261	0.398	0.543	0.703	1.230	1.833	2.262	2.685	3.250	3.690
10	0.129	0.260	0.397	0.542	0.700	1.221	1.813	2.228	2.634	3.169	3.581
11	0.129	0.260	0.396	0.540	0.697	1.215	1.796	2.201	2.593	3.106	3.500
12	0.128	0.259	0.395	0.539	0.695	1.209	1.782	2.179	2.560	3.055	3.428
13	0.128	0.259	0.394	0.538	0.694	1.204	1.771	2.160	2.533	3.012	3.373
14	0.128	0.258	0.393	0.537	0.692	1.200	1.761	2.145	2.510	2.977	3.326
15	0.128	0.258	0.393	0.536	0.691	1.197	1.753	2.132	2.490	2.947	3.286
20	0.127	0.257	0.391	0.533	0.687	1.185	1.725	2.086	2.423	2.845	3.153
25	0.127	0.256	0.390	0.531	0.684	1.178	1.708	2.060	2.385	2.787	3.078
30	0.127	0.256	0.389	0.530	0.683	1.173	1.697	2.042	2.360	2.750	3.030
40	0.126	0.255	0.388	0.529	0.681	1.167	1.684	2.021	2.329	2.705	2.971
60	0.126	0.254	0.387	0.527	0.679	1.162	1.671	2.000	2.299	2.660	2.915
120	0.126	0.254	0.386	0.526	0.677	1.156	1.658	1.980	2.270	2.617	2.860
∞	0.126	0.253	0.385	0.524	0.674	1.150	1.645	1.960	2.241	2.576	2.807

[a] v = degrees of freedom.

Source: From W. M. Hines and D. C. Montgomery, *Probability and Statistics in Engineering Science,* 2nd Ed., John Wiley & Sons, New York, 1980, p. 477. Reprinted by permission.

In other words, we want to test the hypothesis that the number of robots satisfying the specified requirements follows a binomial distribution with parameter equal to 0.97.

The computations for the appropriate chi-square test with $\alpha = 0.05$, $m = 6$, and $r = 0$ are summarized in Table 4, where O_i is the observed number of robots in each sample of 20 operating over 1000 hr, and $e_i = (20)\ (0.97) = 19.4$. Since $G = 0.823$ is significantly smaller than $\chi^2_{5,.05} = 11.1$, we accept the null hypothesis that the true proportion of robots operating continuously over 1000 hr is indeed 0.97.

Other non-parametric techniques frequently used are:

1. The sign test to compare two treatments. We assume that there are several independent pairs of observations on the two treatments. The hypothesis to be tested states that each difference has a probability distribution having mean equal to zero. For each difference the algebraic sign is noted and then the number of times the less frequent sign is considered as the test statistic. There are specialized tables for the critical value of this quantity once a level of significance is chosen.
2. The Wilcox signed-rank test is used to test the hypothesis that observations come from symmetrical populations having a specified common median. For each observation the hypothesized median is substracted and then all differences are ranked from lowest to highest in order of magnitude, omitting the algebraic sign. The test statistic is the sum of ranks for all differences originally having positive signs. The significance test is performed by means of a statistic known as the signed-rank statistic.
3. Run tests to test the hypothesis that observations have been randomly collected from a single population. In this procedure, positive signs are assigned to observations above the median, and negative signs to those below. If the number of runs associated with plus and minus signs is larger or smaller than expected by chance, the hypothesis is rejected. The critical values for the number of runs comes from specialized tables.

There are entire books devoted to nonparametric statistical testing. We have only tried to alert the reader to several common examples.

TABLE C Percentage Points of the χ^2 Distribution

v^a	α	0.995	0.990	0.975	0.950	0.500	0.050	0.025	0.010	0.005
1		0.00+	0.00+	0.00+	0.00+	0.45	3.84	5.02	6.63	7.88
2		0.01	0.02	0.05	0.10	1.39	5.99	7.38	9.21	10.60
3		0.07	0.11	0.22	0.35	2.37	7.81	9.35	11.34	12.84
4		0.21	0.30	0.48	0.71	3.36	9.49	11.14	13.28	14.86
5		0.41	0.55	0.83	1.15	4.35	11.07	12.83	15.09	16.75
6		0.68	0.87	1.24	1.64	5.35	12.59	14.45	16.81	18.55
7		0.99	1.24	1.69	2.17	6.35	14.07	16.01	18.48	20.28
8		1.34	1.65	2.18	2.73	7.34	15.51	17.53	20.09	21.96
9		1.73	2.09	2.70	3.33	8.34	16.92	19.02	21.67	23.59
10		2.16	2.56	3.25	3.94	9.34	18.31	20.48	23.21	25.19
11		2.60	3.05	3.82	4.57	10.34	19.68	21.92	24.72	26.76
12		3.07	3.57	4.40	5.23	11.34	21.03	23.34	26.22	28.30
13		3.57	4.11	5.01	5.89	12.34	22.36	24.74	27.69	29.82
14		4.07	4.66	5.63	6.57	13.34	23.68	26.12	29.14	31.32
15		4.60	5.23	6.27	7.26	14.34	25.00	27.49	30.58	32.80
16		5.14	5.81	6.91	7.96	15.34	26.30	28.85	32.00	34.27
17		5.70	6.41	7.56	8.67	16.34	27.59	30.19	33.41	35.72
18		6.26	7.01	8.23	9.39	17.34	28.87	31.53	34.81	37.16
19		6.84	7.63	8.91	10.12	18.34	30.14	32.85	36.19	38.58
20		7.43	8.26	9.59	10.85	19.34	31.41	34.17	37.57	40.00
25		10.52	11.52	13.12	14.61	24.23	37.65	40.65	44.31	46.93
30		13.79	14.95	16.79	18.49	29.34	43.77	46.98	50.89	53.67
40		20.71	22.16	24.43	26.51	39.34	55.76	59.34	63.69	66.77
50		27.99	29.71	32.36	34.76	49.33	67.50	71.42	76.15	79.49
60		35.53	37.48	40.48	43.19	59.33	79.08	83.30	88.38	91.95
70		43.28	45.44	48.76	51.74	69.33	90.53	95.07	100.42	104.22
80		51.17	53.54	57.15	60.39	79.33	101.88	106.63	112.33	116.32
90		59.20	61.75	65.65	69.13	89.33	113.14	118.14	124.12	128.30
100		67.33	70.06	74.22	77.93	99.33	124.34	129.56	135.81	140.17

[a] v = degrees of freedom.

Source: From W. M. Hines and D. C. Montgomery, *Probability and Statistics in Engineering Science,* 2nd Ed., John Wiley & Sons, New York, 1980, p. 476. Reprinted by permission.

15. HYPOTHESIS TESTING IN THE ANALYSIS OF DESIGNED EXPERIMENTS

15.1. One-Factor Experiments

The simplest experiment corresponds to one factor and no restrictions on randomization. The statistical model of the experiment response is formulated as

$$Y_{if} = \mu + \tau_j + \varepsilon_{ij}, \qquad j = 1, 2, \ldots, k; \qquad i = 1, 2, \ldots, n_j$$

where μ = common effect, μ_j = population mean for jth treatment, and τ_j = effect of jth treatment, defined as $\mu_j - \mu$. There are two types of completely randomized one-factor models:

Fixed-effect model, H_0: $\tau_j = 0$, for all j
Random-effect model, H_0: $\sigma_\tau^2 = 0$, τ_j distributed as $N(0, \sigma_\tau^2)$

The hypothesis testing is carried out in each case by means of an F-test.

Other one-factor experiments can include one or more restrictions on randomization. Examples of these designs are the randomized block design, the Latin square design, the Graeco-Latin square design, and the Youden square design. In all these design the effect of the restrictions on randomization can be tested using F-tests.

TABLE D Percentage Points of the *F* Distribution ($\alpha = 0.10$)

v_2	1	2	3	4	5	6	7	8	9
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
10	3.28	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
12	3.13	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63

Source: From W. M. Hines and D. C. Montgomery, *Probability and Statistics in Engineering Science,* 2nd Ed., John Wiley & Sons, New York, 1980, pp. 482–483. Reprinted by permission.

15.2. After-ANOVA Range Tests

The purpose of the range test is to investigate which *pairs* of treatments are significantly different to cause H_0 to be rejected. Consider the sample averages $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$, computed from k random samples drawn from the k populations corresponding to the k levels of the factor. Furthermore, let us assume that $n_j = n$ for all j. These sample averages can be rearranged in increasing order of magnitude, from smallest to largest, as $\bar{Y}_{(1)}, \bar{Y}_{(2)}, \ldots, \bar{Y}_{(k)}$. The hypothesis to be tested is formulated as H_0: $\mu_{(j)} = \mu_{(i)}$, for any two values of i and j such that $j > i$.

The statistic to be used is the *studentized range statistic* as $Q_{rf} = \dfrac{\bar{Y}_{(j)} - \bar{Y}_{(i)}}{(MS_{\text{error}}/n)^{1/2}}$ where $r = j - i + 1$ is number of steps (on an ordered scale) associated with the range defined by the ith and the jth treatments, or more specifically, treatments (i) and (j). There are several versions of the range test, the following ones being the most popular. These tests use specialized tables of critical values of the studentized range statistic:

1. Newman-Keusl range test, $r = 2, \ldots, k$
2. Tukey's Range Procedure, $r = k$
3. Duncan's Multiple Range Test, $r = 2, \ldots, k$

TABLE D (*Continued*)

10	12	15	20	24	30	40	60	120	∞	v_2
60.20	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.83	1
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	2
5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13	3
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76	4
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10	5
2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72	6
2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47	7
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29	8
2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16	9
2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06	10
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97	11
2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90	12
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85	13
2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80	14
2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76	15
2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72	16
2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69	17
1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66	18
1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63	19
1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61	20
1.92	1.88	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59	21
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57	22
1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55	23
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53	24
1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52	25
1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50	26
1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49	27
1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48	28
1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47	29
1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46	30
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38	40
1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29	60
1.65	1.60	1.54	1.48	1.45	1.41	1.37	1.32	1.26	1.19	120
1.60	1.55	1.49	1.42	1.33	1.34	1.30	1.24	1.17	1.00	∞

15.3. Factorial Experiments

Instead of assuming that the experimental response can be affected by one factor, a factorial experiment assumes that it is affected by two or more factors, each having an arbitrary number of levels. In a factorial experiment a level combination (one level per factor) is known as an experimental condition. All experimental conditions are sampled, and all observations are collected at random. It is possible to test the significance of each main effect (due to a factor) or and interaction effect (due to a group of factors) by means of an F-test.

15.4. Hypothesis Testing in Regression Analysis

A regression model is a *fitting* relationship that allows the estimation of a dependent variable or experimental response for given settings of a specified group of independent variables or factors. The parameters of the model are known as regression coefficients. Typical tests include the following:

1. Testing the hypothesis that a regression coefficient is equal to zero. The test statistic is the t-statistic.
2. Testing for linearity of regression using an analysis-of-variance technique. The test statistic is the F-statistic.

REFERENCES

Bowker, A. H., and Lieberman, G. J. (1972), *Engineering Statistics,* 2nd Ed., Prentice-Hall, Englewood Cliffs, NJ.

Devore, J. L. (1987), *Probability and Statistics for Engineering and the Sciences,* 2nd Ed., Brooks/Cole, Monterey, CA.

Dougherty, E. R. (1990), *Probability and Statistics for the Engineering, Computing, and Physical Sciences,* Prentice-Hall, Englewood Cliffs, NJ.

Hogg, R. V., and Craig, A. T. (1978), *Introduction to Mathematical Statistics,* 4th Ed., Macmillan, New York.

Snedecor, G. W., and Cochran, W. G. (1980), *Statistical Methods,* 7th Ed., Iowa State University Press, Ames, IA.

Hollander, M., and Wolfe, D. (1973), *Nonparametric Statistical Methods,* John Wiley & Sons, New York.

Lehmann, E. (1975), *Nonparametrics: Statistical Methods Based on Ranks,* Holden-Day, San Francisco.

Marascuilo, L., and McSweeney, M. (1977), *Nonparametric and Distribution-Free Methods for the Social Sciences,* Brooks/Cole, Monterey, CA.

CHAPTER 87

Regression and Correlation*

Raja M. Parvez
Lucent Technologies

Donald Fusaro
Lucent Technologies

1. **INTRODUCTION TO REGRESSION ANALYSIS** **2265**
   - 1.1. General Linear Model 2265
   - 1.2. Utility and Dangers 2266
   - 1.3. Importance of Goals 2266
   - 1.4. Kinds of Models 2267
   - 1.5. Appropriate Use of Statistics 2267
   - 1.6. Role of Assumptions 2267
2. **RELATING TWO VARIABLES** **2268**
   - 2.1. Least-Squares Method 2268
   - 2.2. Residual Variance 2270
   - 2.3. Correlation 2271
   - 2.4. Model Specification 2272
   - 2.5. Model Validation 2272
   - 2.6. Coefficient Estimation 2273
   - 2.7. Interval Estimation for a Point on the Line 2274
   - 2.8. Predicting a Future Value 2274
3. **MULTIPLE LINEAR REGRESSION** **2275**
   - 3.1. Intercorrelation Effects 2275
     - 3.1.1. Potentially Enlarged Variances 2275
     - 3.1.2. Intercorrelated Estimates 2276
     - 3.1.3. Ambiguity in Assessing Contributions 2276
   - 3.2. Detection of Intercorrelation 2277
   - 3.3. Meaning of Partial Correlation 2277
   - 3.4. Multiple Correlation 2278
   - 3.5. Relating *t* and *F* in Modeling 2278
   - 3.6. Dealing with Interactions 2279
   - 3.7. Basics for Attribute Modeling 2279
   - 3.8. Dealing with Covariates 2280
   - 3.9. Application to the Example 2280
4. **RELATING DIAGNOSTIC QUESTIONS TO GOALS** **2282**
   - 4.1. Motivation 2282
   - 4.2. Summary of Interrelated Diagnostic Questions 2282
5. **AN INTRODUCTION TO MODERN DIAGNOSTICS** **2282**
   - 5.1. Notation 2283
   - 5.2. Getting the Catcher and the Hat 2283
   - 5.3. Row Deletion 2284
   - 5.4. Internal Validation 2284
   - 5.5. Examining Residual Errors and Influence 2284
   - 5.6. Partial Regression Leverage Plots 2286
6. **DIAGNOSTICS FOR THE EXAMPLE** **2286**
   - 6.1. Leverage and Influence 2286
   - 6.2. Final Results 2287

---

*Parts of this chapter were originally presented by the late Douglas C. Crocker in his chapter in the second edition of this Handbook.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

7. OTHER REGRESSION TOPICS 2289
- 7.1. Variable Selection 2289
  - 7.1.1. All Possible Regressions 2289
  - 7.1.2. Forward Selection 2289
  - 7.1.3. Backward Elimination 2290
  - 7.1.4. Stepwise 2290
- 7.2. Ridge Regression 2290

8. SOME PRACTICAL CONCERNS 2291
- 8.1. Model Use and Maintenance 2292
- 8.2. Helpful Hints in Practice 2292

REFERENCES 2292

ADDITIONAL READING 2293

1. INTRODUCTION TO REGRESSION ANALYSIS

Regression analysis is:

A technique for measuring and explaining (reducing unexplained) variability in a system
An aid to understanding interrelationships in complex systems
A process for building a useful model of a system
A method for improving forecasting or prediction
A mechanism for focusing on important phenomena
A system for evaluating theories or beliefs
An aid in formulating new theory
A method for obtaining better control of variation
A technique for estimating equation parameters

Regression modeling involves practical problems, problems of judgment, and a good deal of art. This chapter is not intended to be a recipe book or a catalog of rules of thumb. It is intended to introduce the reader to some basic principles involved in statistical modeling while at once exposing the dangers. In this spirit, this chapter discusses many of the difficulties that may be encountered in attempting to model systems displaying statistical variation. It is intended to serve as a good blend of theoretical structure, philosophical outlook, and practical guidance.

1.1. General Linear Model

An equation of the form

$$Y_i = \sum_{j=0}^{P} b_j X_{ij} \tag{1}$$

is sometimes referred to as the general linear model. In this equation, Y is a variable whose behavior is of interest. It was once common to refer to Y as the dependent variable, taken from the mathematical concept of a function. In statistical modeling, most authors have come to call Y the response variable. This is the convention adopted here.

In Eq. (1), Y is a linear additive function of the X variables, which are P in number, $P \geq 1$. These X's were formerly often referred to as independent variables, again using the mathematical sense. They are now sometimes called regressors or explanatory variables but are more commonly called predictors (although prediction may not be the goal). The subscript j denotes which predictor. In this general form of Eq. (1), there is a dummy variable, $X_0 = 1$ ("dummy" because it does not vary), which is not counted as a predictor but is included in the summation. Its coefficient, b_0 is the constant term or intercept. It is in units of Y. The other regression coefficients, b_j ($j = 1$ to P), are the slopes (multipliers of their respective predictors) and are expressed in units of Y/X_j. These b_j's are unknowns that are to be determined from the analysis. The values obtained are estimates of the "true" unknown coefficients, β_j. Geometrically, Eq. (1) represents a line, a plane, or a hyperplane in $P + 1$ dimensional space. This process is known as multiple linear regression (MLR) analysis. The subscript, i, denotes the series of observations in the sample going from 1 to n. Each observation provides a value for each of the variables—the predictors and the response—for each of the units or individuals in the sample. The unit of observation may be, for example, a day, a person, an automobile, a task, an event, or a batch.

The *X*'s can, of course, represent quite complicated transformations of originally observed bits of information about each unit. Reciprocals, powers, and logs are examples, and so are ratios or products of two (or more) predictors. It is astounding to witness how often this linear additive equation form gives a very good representation of the underlying physics that relate the response to the predictors.* That the variability, that is, the behavior, of so many things in nature can be so well described (predicted) by this simple summation process is truly profound.

1.2. Utility and Dangers

Variation is the essence of statistical modeling. Variation is the problem. Information is contained in variation. In fact, without variation, there is no information. The activities of industrial engineers virtually always involve dealing with variation in multiple-variable systems. The goal may be to evaluate or explain previous events or to predict or control future events. Modeling a response variable in such systems is usually complex and difficult. Part of the difficulty arises in most cases because the data come from the existing system as it normally operates rather than being generated during a designed experiment. Such data might be called nonexperimental or clinical.

Some major difficulties found in dealing with nonexperimental data result from the interrelationships naturally present among the predictors. The unwanted intercorrelations are avoided in controlled experiments by keeping the predictors uncorrelated with (orthogonal to) each other. This difficulty in dealing with clinical data is shared with many other disciplines. In fact, the exception is the analyst who is able to operate with "scientific" laboratory technique.

The great power of MLR lies in its ability to relate simultaneously the many intercorrelated predictors to the response—to deal with nonexperimental data. Herein also lies the main source of danger. Successful modeling of nonexperimental data is a tricky business. But not all the dangers are associated with the natural intercorrelations of nonexperiments. The variety of ways in which the analyst can encounter trouble is nearly as great as the variety of problem situations. Perhaps no other technique suffers more misuse and abuse than regression analysis. Because of this, much criticism of the general technique is offered by those who apparently do not understand its power or proper use and who misrepresent it. The dangers can be avoided or treated if they are recognized and understood. Much of the balance of this chapter deals directly or indirectly with establishing appropriate safeguards.

1.3. Importance of Goals

Multiple linear regression should not be a process that follows a fixed, predetermined path or employs an established ritual for achieving a goal. That is because different goals require different analytic behavior. As illustrated by this chapter's opening list, regression goals are various. Before attempting to model a system, it is important to know what the model is supposed to do. What is the question the analysis is supposed to answer?

Because we are dealing with practice in industrial engineering, it is important first to make the distinction between science and decision making (see Healy 1978). The statistical requirements for establishing scientific truth are much more stringent than for decision making. The manager cannot wait for the discovery of ultimate truth but must decide today. Ordinarily, the industrial engineer operates in support of that process and will serve the manager best if the decision process is supported in a timely manner. This is not to suggest carelessness or disregard for theory. It is to suggest recognition of the basic fact that the manager will make the decision with or without the potential help. A responsibly derived, yet imperfect, model can be very much better than no model at all.

Very broadly, the various goals can be put into five categories. These represent a natural evolutionary sequence of four steps, any one of which may be the intended end use.

1. *Exploration:* fishing, hypothesis finding (see Finch 1979)
2. *Specification:* hypothesis testing, confirmation of the model form (rarely an end use)
3. *Estimation:* estimating model parameters with sufficient precision (estimating future events is referred to in this chapter as "prediction")
4. *Prediction:* use of the model for anticipation or for "inverse estimation" (calibration)
5. *Control:* use of the model to prescribe change or to direct or guide policy or the behavior of a system

---

* A known underlying causal relationship is not a requirement for useful statistical modeling.

TABLE 1 Simple Classification of Regression Models

Class	Kind	Basis
Associative	Concomitant, precursory	Observation
Physical	Empirical	
	Causal, mechanistic	Theory

1.4. Kinds of Models

Kinds of models seem to lie more along a continuum and are therefore less easily classified. The main continuum is closeness to causality. The scale slides from loose empiricism to exact causal representation (mechanism). How far along the scale the analyst moves may depend on either the maturity of the corresponding physical discipline or the needs imposed by the goals.

There is another subset where causality is not an issue. These models might be called "associative." Here the response and the predictors may both be "caused" by some outside force. They behave concomitantly. An example is the use of leading indicators in economic models. Another example is the precursory use of animal characteristics or behavior to predict the severity of the winter. Presumably no one would claim that the extra hair on the woolly bear caterpillar causes snow to fall (causation might be suspected in the other direction in such cases if it were not for our belief that cause must precede in time its effect).

This simple classification scheme thus takes the form shown in Table 1.

1.5. Appropriate Use of Statistics

Statistical measures and diagnostics can and do serve an essential role in regression modeling, but they must be used appropriately. Their use must be related to goals. In general, any adequate MLR computer program system will list many statistics that may not be relevant in any given situation. For example, the multiple correlation coefficient, R, is universally printed. It may be of no interest. Further, even if it is of interest, its value must be judged in the context of the problem. It depends on the question. The analyst must know what questions need to be answered and must use relevant statistical measures accordingly.

1.6. Role of Assumptions

Assumptions are, in most regression articles and texts, listed as a sort of litany to precede the analysis as if they universally apply. Moreover, they are treated as if they describe the problem setting. They are really descriptions of the mathematical model whose behavioral properties are known and that is to be used as an analog to the system under study. Assumptions (model characteristics) relate to goals just as statistics do. Those that are relevant are rarely ever exactly met by the problem system. The severity of trouble the analyst may expect because of the remaining differences ("violations of assumptions") is a matter for judgment and experience and cannot be removed from the problem context.

Table 2 offers a skeleton relationship of assumptions to goals in a hierarchical order (for a more complete discussion, see Eisenhart 1947). Random residual variation in Y is associated with a host of small, unimportant (in context) contributions. Notice that the usual assumptions of homoscedasticity and normality are not imposed for specification and estimation. The least-squares estimates of the regression coefficients provided by MLR are the most efficient, unbiased linear estimates among all linear estimates for uniform error variance and are still unbiased for nonuniform error variance. The central limit theorem will give very good protection—just as with ordinary averaging—allowing

TABLE 2 Cumulative Relationships of Assumptions to Goals

Goals	Desirable Data Characteristics	Model and Process Characteristics
Exploration	Random Y for given X	Least-squares fitting
Specification	"Complete" X set	"Correct" model form
Estimation	Spread, balanced X's	b's normal by central limit theorem
Prediction and Control	Typical X space	Specified error distribution

the normal model to be used with nonnormal data for establishing confidence intervals. Stated characteristics are cumulative descending the table.

2. RELATING TWO VARIABLES

The actual use of the simple (single-predictor) model is rare (real systems are rarely that simple.) However, for examining the principles involved in regression modeling, the simple model serves well.

For illustration, hypothetical data representing steam consumption (Y) for a particular building are modeled here. This response variable was chosen because (1) energy use has universal relevance and global importance, (2) such a wide variety of goals can be authentically represented in an energy system, and (3) this same problem setting can be expanded in following sections to represent more complex modeling ventures. The structure under study might be an office complex, a factory, a warehouse, a hospital, a hotel, or even a home. For demonstration, only 20 observations are contained in the sample. Each observation represents a four-week period. Weekly data would be preferred in most cases, but to cover extremes of weather in only 20 data points, four-week periods were chosen. The goal is to establish control of steam consumption for this building. "Excessive" use is now dismissed as being weather related.

At first it is assumed that comfort heating in this building is the major use of steam. Its use (measured in giga-British thermal units, gBtu), should be reasonably well related to degree-days (X). This is measured relative to 65°F (degree-days F/1.8 = degree-days C) and is also reported here on a per-period basis. The 20 observations are shown in Table 3, with the periods numbered within year from 1 to 13 and years numbered 1, 2, and 3. The method of least squares will be employed to relate steam use to degree-days.

Table 3 also shows the fitted values, called *YHAT,* and the residual fitting errors, called *Y-YHAT*. The names and their symbolic forms are discussed following Eq. (3). Diagnostic attention will be given later to the values in these two columns.

2.1. Least-Squares Method

In MLR, understanding variation is the basis for problem solving. Variation in the response is made up (theoretically) of two parts:

TABLE 3 Hypothetical Data for Modeling Steam Consumption

i	Period Number	X (k degree-days)[a]	Y (gBtu)[b]	Fitted *YHAT* (gBtu)	Residual *Y-YHAT* (gBtu)
1	10-1	0.156	7.991	6.990	1.001
2	11-1	0.419	8.589	8.095	0.494
3	12-1	0.658	9.145	9.100	0.045
4	13-1	1.009	11.212	10.575	0.637
5	1-2	1.380	11.754	12.134	−0.380
6	2-2	1.103	11.469	10.970	0.499
7	3-2	1.000	10.584	10.537	0.047
8	4-2	0.703	9.509	9.289	0.220
9	5-2	0.207	7.457	7.204	0.253
10	6-2	0.086	6.989	6.696	0.293
11	7-2	0.024	6.537	6.435	0.102
12	8-2	0.005	4.938	6.355	−1.417
13	9-2	0.026	5.275	6.444	−1.169
14	10-2	0.161	7.452	7.011	0.441
15	11-2	0.307	7.962	7.625	0.337
16	12-2	0.664	8.915	9.125	−0.210
17	13-2	1.039	9.758	10.701	−0.943
18	1-3	1.275	11.183	11.693	−0.510
19	2-3	1.193	11.523	11.348	0.175
20	3-3	0.953	10.426	10.340	0.086

$n = 20$; $\Sigma X_i = 12.369$; $\Sigma Y_i = 178.67$

$\Sigma X_i^2 = 11.915$; $\Sigma Y_i^2 = 1678.7$; $\Sigma X_i Y_i = 128.42$

[a] k degree-days = 10^3 degree-days.

[b] gBtu = giga-British thermal units = 10^9 Btu.

1. The systematic variation (signal), which is associated with or is in response to changes in the predictors
2. Leftover variation (noise), which is called "residual error" or "experimental error."

The distinction is really not so sharp. The leftover error is actually associated with a great many things that, in practice, might be measured (and included in the model) if analysts had sufficient time, wisdom, patience, and money. They simply choose not to try to identify all sources of variation. They will discontinue the search when there seems to be no regular pattern of errors left over and when either all the reasonable predictors have been adequately tested or the residual error variance is small enough—again depending on goals. In terms of the true coefficients and residual error of the theoretical model, the observed response variable may be expressed as

$$Y_i = \sum_{j=0}^{P} \beta_j X_{ij} + \varepsilon_i \tag{2}$$

where ε_i is the "residual error" associated with Y and (theoretically) has variance σ_ε^2. The fitted model containing the estimates of the β_j's, then, is

$$\hat{Y}_i = \sum_{j} b_j X_{ij} \tag{3}$$

where the circumflex or "hat" on Y denotes the predicted or estimated value of the response. It is like an average (where a "bar" is used). In fact, it is the conditional average, given the location in the space defined by the X_{ij}'s. It is an estimate of the expected or true value of the response for that location or set of conditions.*

The differences between the observed and fitted values of Y are the residual errors or, simply, "residuals,"

$$e_i = Y_i - \hat{Y}_i = \hat{\varepsilon}_i \tag{4}$$

where e_i is an estimate of the "true error" ε_i. In practice, e_i may contain anything the analyst chooses to omit from the model. It has sample variance

$$s_{Y \cdot X}^2 = s_e^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n - P - 1} = \frac{\sum_{i=1}^{n} e_i^2}{n - P - 1} = \hat{\sigma}_\varepsilon^2 \tag{5}$$

which, for the theoretical case, is an estimate of the "experimental" error variance. The subscript $Y \cdot X$ ("Y dot X") means "for Y, given the model containing a particular set of X's." Thus $s_{Y \cdot X}^2$ is the sample estimate of the residual variance in Y, given the model.

The least-squares method chooses values for the b_j's of Eq. (3), which are unbiased estimates of the β_j's of Eq. (2). The least-squares estimates are universally minimum variance unbiased estimates for normally distributed residual errors and are minimum variance among all linear estimates (linear combinations of the observed Y's), regardless of the residual error distribution shape (see Eisenhart 1964). The b_j's (as well as the $\hat{Y}_i$'s) are linear combinations of the observed Y_i's. The least-squares method determines the weight given to each Y value. The derivations of the least-squares solution and/or associated equations used later in this chapter are shown in other sources (see Additional Reading). In essence, the b_j's are chosen to minimize the numerator of Eq. (5)—the sum of squares of e_i's of Eq. (4)—hence "least squares."

In returning to the example problem, a geometric interpretation is presented first. Figure 1 is a plot of steam consumption vs. degree-days from Table 3. The regression coefficient, b_1 is represented by the slope of the least-squares line. It is the tangent of the angle θ. The e_i's whose squares are to be summed to a minimum are distances measured in the Y direction from the points to the line. They are illustrated by typical distances, e_4 and e_{17}.

The least-squares solutions for the simple model are

---

* It is important to realize that, although the model may be used for predicting future $\hat{Y}$ values, Y does not predict their individual behavior but estimates the conditional average about which those individuals are expected to vary.

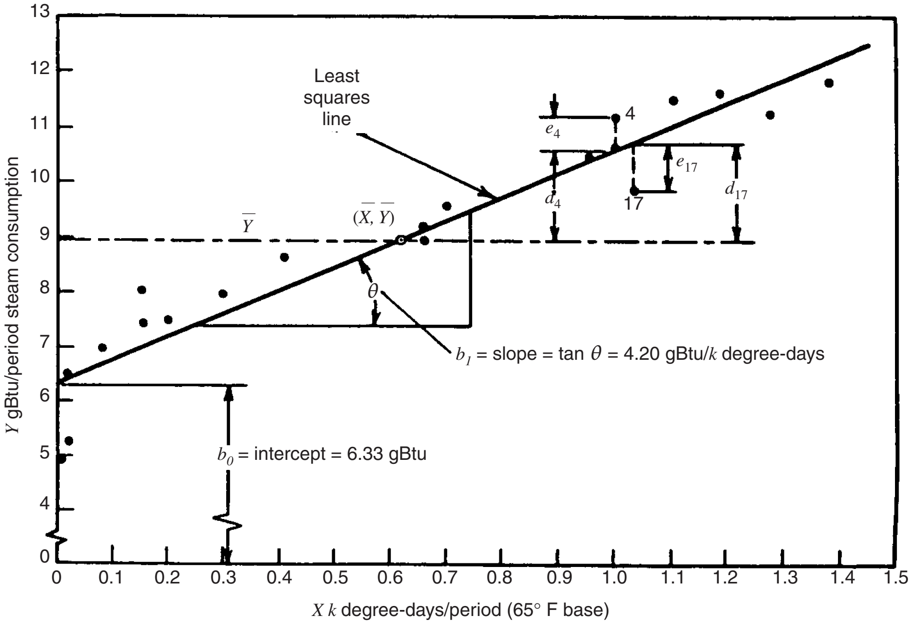


Figure 1 Relationship of Steam Use to Degree Days.

$$b_1 = \frac{\text{SPXY}}{\text{SSX}} \quad b_0 = \bar{Y} - b_1 \bar{X} \tag{6}$$

where SPXY = the (corrected) sum of products of the XY pairs
SSX = the (corrected) sum of squares of X's
$\bar{Y}$ and $\bar{X}$ = the arithmetic averages (which are also least-squares estimators) of the two variables

These averages and sums (with all sums taken for $i = 1$ to n) are

$$\begin{aligned} \bar{X} &= \sum X_i/n & \bar{Y} &= \sum Y_i/n \\ \text{SPXY}^* &= \sum (X_i - \bar{X})(Y_i - \bar{Y}) & &= \sum X_iY_i - \text{n}\overline{XY} \\ \text{SSX} &= \sum (X_i - \bar{X})^2 & &= \sum X_i^2 - \text{n}\bar{X}^2 \end{aligned} \tag{7}$$

Equations (6) and (7) yield the following values for the example:

$$\begin{aligned} \bar{X} &= 12.369/20 = 0.618 \qquad \bar{Y} = 178.67/20 &&= 8.93 \\ \text{SPXY} &= 128.42 - 20(0.618)(8.93) &&= 17.93 \\ \text{SSX} &= 11.915 - 20(0.618)^2 &&= 4.23 \\ b_1 &= 17.93/4.27 = 4.20 \qquad b_0 = 8.93 - 4.20(0.618) &&= 6.33 \end{aligned}$$

Equation 3 then takes the form $\hat{Y}_i = 6.33 + 4.20X_i$.

2.2. Residual Variance

For $P = 1$, Eq. (5) reduces to $s^2_{Y \cdot X} = \text{SSRes}/(n - 1)$, where SSRes is the residual sum of squares given by

---

*These are not shown in the computationally easiest form. This form demonstrates meaning.

$$\text{SSRes} = \text{SSY} - \text{SSReg} \tag{8}$$

and where SSY is the (corrected) sum of squares of Y's and SSReg is the regression sum of squares. In Eq. 8, SSY is exactly parallel in form to SSX in Eq. (7), and when divided by its $n - 1$ DOF, it yields the Y mean square, which might be used to estimate the variance of Y. Regardless of the appropriateness of such an interpretation, the expression $\text{SSY}/(n - 1)$ is a measure of the raw variability in the response whose explanation is the goal. The contribution to SSY that is associated with X is SSReg. This is given by

$$\text{SSReg} = b_1\text{SPXY} \tag{9}$$

and is the sum of squared distances from $\overline{Y}$ to the regression line as shown by typical distances d_4 and d_{17} in Figure 1. For this example,

$$\text{SSY} = 82.55 \qquad s_Y = \left(\frac{\text{SSY}}{19}\right)^{1/2} = 2.08$$

$$\text{SSReg} = (4.20)(17.93) = 75.38$$

$$\text{SSRes} = 82.55 - 75.38 = 7.17$$

$$s_{Y\cdot X} = \left(\frac{\text{SSRes}}{18}\right)^{1/2} = 0.631$$

It can be seen that $s_{Y\cdot X}$ is only 30% of s_Y. That is, the regression equation provided a 70% reduction in variation in Y. Another way to evaluate this residual standard deviation or residual standard error is to compare it to the mean of Y. In this case it is 100 (0.631/8.93) = 7.0% of the mean and, depending on the goal, might represent a satisfactory reduction in variability. Still another way to measure the association between Y and X, and hence the residual lack of association, is to use the correlation coefficient that is developed next.

2.3. Correlation

The theoretical concept of correlation arises in conjunction with the bivariate normal distribution function. That function has five parameters. If the two variables are X and Y, the parameters are the means (μ_x, μ_y) and the variances (σ_x^2, σ_y^2) of each variate and a measure of covariation, the correlation coefficient, ρ (rho). This chapter does not deal with the theoretical bivariate (or multivariate) normal distribution. However, in practice, the sample correlation coefficient, r, is a useful measure of linear association. It is a dimensionless ratio ranging from -1.0 (perfect inverse linear agreement) through zero (orthogonal or linearly unrelated) to $+1.0$ (perfect direct linear agreement). The value can be obtained from Eq. (10) and used as an index without any assertion whatever being made about distribution form.

$$r_{XY} = \frac{\text{SPXY}}{[(\text{SSX})(\text{SSY})]^{1/2}} = \frac{s_{XY}}{s_X s_Y} \tag{10}$$

The first form of the expression for r_{XY} has the same numerator as b_1 in Eq. (6), which shows that it is just a rescaling of the same basic information. It is easily shown that $r_{XY} = b_1\, s_X/s_Y$.* In the second form in Eq. 10, s_{XY} is the sample covariance (not standard deviation). It has the same sign as SPXY and r and is $\text{SPXY}/(n - 1)$.

The square of r is called the coefficient of determination. It ranges from 0 to 1 and can be interpreted as the fraction of the variation in Y (with variation represented by SSY) that is accounted for or "explained" by variation in X. Thus and from Eq. (8):

$$r_{XY}^2 = \frac{\text{SSReg}}{\text{SSY}} = 1 - \frac{\text{SSRes}}{\text{SSY}} \tag{11}$$

Using Eq. (10) for the example data, $r_{XY} = 17.93/[(4.27)(82.55)]^{1/2} = 0.955$; $r_{XY}^2 = 0.912$. This is seen to be equal to the result of Eq. 11, where $r_{XY}^2 = 75.38/82.55 = 0.913$ (slight rounding error).

---

*The subscript order "XY" on r is arbitrary; $r_{XY} = r_{YX}$. But the ratio s_X/s_Y implies that b_1 is for "Y on X." With $s_Y/s_{XY}\, b_1$ would be "X on Y," with minimization of squared errors taken in the X direction, a different line except where $r = 1.0$.

So variation in X has accounted for 91% of SSY. This is approximately the same as claiming 91% reduction in the variance of Y (from s_Y^2 to $s_{Y \cdot X}^2$).

2.4. Model Specification

In other circumstances, where the physics and chemistry are not so well understood (e.g., in studying the "cause" of a disease), the question may focus on the statistical significance of the relationship. The analyst is attempting to decide whether the relationship seen in the sample is something real or just the result of chance association. This decision is appropriate along all goal sequences except where existing theory permits prior specification* of the model.

Model specification is the process of choosing an adequate representation of reality. To decide this question of reality, the analyst would want a test model for the behavior of the estimator, b_j, when the association is just chance. One way would be to use the t test model with the null hypothesis that $\beta_j = 0$ (or some other appropriate value). The alternative hypothesis might be $\beta_j > 0$. The t distribution is appropriate by the central limit theorem. Then

$$t_j = \frac{b_j}{s_{bj}} \tag{12}$$

with the critical value for t of $t_{n-2,\alpha}$, where α represents the specified degree of risk of rejecting a true null hypothesis (claiming a nonexistent association). The standard errors for b_0 and b_1 are given by

$$s_{b_0} = s_{Y \cdot X} \left(\frac{\Sigma X_i^2}{n\mathrm{SSX}} \right)^{1/2} \tag{13}$$

$$s_{b_1} = \frac{s_{Y \cdot X}}{(\mathrm{SSX})^{1/2}} \tag{14}$$

For the example data, $s_{b_0} = 0.236$ and s_{b_1}, $= 0.306$. The corresponding t ratios are $t_0 = 26.9$ and $t_1 = 13.8$, indicating, as was "known" in advance, that both constants are statistically well removed from zero (highly "significant" compared to a critical value of $t_{18,0.05} = 2.10$). This information is put to more appropriate use later in this section. Statgraphics (a statistical software package) output for this analysis is shown in Figure 2. An intermediate precaution should concern the analyst, that of model validation.

2.5. Model Validation

The least-squares method has permitted each of the data points to play a role in determining the constants b_0 and b_1. It is entirely possible (and nearly always true) that some observations in the data set contain errors (mistakes) in one or more of the variables or arise from unusual conditions that the model is not intended to represent. The "back substitution" (obtaining $Y_i - \hat{Y}_i$, values for the development data set, as shown in Table 3) may reveal suspicious points. Generally, residual errors in excess of $\pm 2s_{Y \cdot X}$ should be viewed with mild suspicion, although about 1 in 20 is expected to be in these regions. More sensitive measures will be developed in Section 5. There is an extensive literature associated with this problem of dealing with "outliers". One discussion that might serve as a starting point is Barnett (1978).

Perhaps the most useful graphic for examining residuals is a plot of *Y-YHAT* vs. *YHAT*. That will be illustrated later for the multiple regression model. In the simple case, one need not bother. Such a plot is equivalent to tipping Figure 1 so that the regression line is horizontal. No sophisticated techniques are required to see that the fit is poor. Ten of the first 11 points lie above the line. It may be tempting to suggest, say, a quadratic in X to achieve a better fit. However, there is no theoretical reason to expect curvature in the relationship. More properly, additional predictors might be sought, as will be shown presently. The development for the simple case is continued later, ignoring the lack of fit.

One possibility for testing the model's stability is to validate it on fresh data, that is, by back-substituting data points that were not used in developing the model.

This method of validation is called external validation. A method known as internal validation (which is arguably far superior) will be presented in Section 5.

---

*The distinction between specification and estimation is rarely made. See Hunter and Box (1965) for further discussion. Also see Healy (1978) regarding significance testing.

```
Simple Regression - Steam vs. Heat

Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Steam
Independent variable: Heat
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        6.33429       0.235869        26.8551         0.0000
Slope            4.20295       0.305585        13.7538         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     75.3754      1      75.3754     189.17       0.0000
Residual                   7.1723     18     0.398461
-----------------------------------------------------------------------------
Total (Corr.)             82.5477     19

Correlation Coefficient = 0.95557
R-squared = 91.3113 percent
Standard Error of Est. = 0.631238
```

Figure 2 Statgraphics Regression Analysis for Steam vs. Heat.

2.6. Coefficient Estimation

Suppose the model has been specified from existing theory, or by exploration, testing, and validation, and is judged adequate. Back-substitution residual errors are well behaved. Now, whether the values of the b_j's themselves are of interest or whether they are simply to be used in the equation for predicting future values of Y, the precision with which they estimate the β_j's is of concern.* Point estimates were obtained from Eq. (6), but coefficient estimation is not complete without obtaining interval estimates. It is not sufficient, where estimation is the goal (or a step on the path to the goal), just to have "significant" t ratios.

Use of the confidence interval (CI) concept helps to contrast these two steps, specification and estimation. With $\alpha = 0.05$ risk that the CI will not contain β as is claimed, the interval is

$$100(1 - \alpha)\ \text{CI} = 95\%\ \text{CI} = b_j \pm t_{n-2,0.05}\, s_{bj} \tag{15}$$

Notice that (with prescribed t) ts_b represents the maximum probable error associated with the estimate of β. This can be expressed as a percentage error (where engineers very often seek estimates that are within 5 or 10%). Using b as the base (in the absence of knowing β), let E represent the potential percentage error associated with a t value of 2, an approximate value that is never more than 2% in error for $df \geq 30$. (For $df < 30$, substitution of the correct value is advised.) Then

$$E = \frac{100ts_b}{b} = \frac{200s_b}{b} \tag{16}$$

But notice that s_b/b is just the inverse of the t ratio calculated from the sample using Eq. (12).

$$E = \frac{200}{t} \tag{17}$$

This implies the need for calculated t values of 20 or even 40 to meet our common expectation of a 10 or 5%, respectively, error of estimation!

---

*In general, the requirements for precision will be greatest for control, least for prediction (see Section 3.1.2).

From Eqs. (14) and (16) it can be seen that the error in estimating the slope is directly proportional to $s_{Y \cdot X}$ and inversely proportional to $(\mathrm{SSX})^{1/2}$. Thus, to achieve a prescribed value of E, either of two things must be done: (1) an improved (less noisy) model must be found to reduce residual error, or (2) a larger sample must be obtained to increase SSX (see Crocker 1985 for a more complete discussion of this issue). In general, precision will improve approximately as the square root of n.

2.7. Interval Estimation for a Point on the Line

The regression equation can be used to estimate the "true" value of the response for some specified value of the predictor. This is estimating a conditional population mean of Y and is analogous to estimating (unconditionally) the population mean in a univariate setting. The CI for this case is

$$\begin{aligned} 100(1 - \alpha)\ \mathrm{CI} &= \hat{Y}_c \pm t_{n-2,\alpha}\, s_{\hat{Y}c} \\ s_{\hat{Y}c} &= s_{Y \cdot X} \left[\frac{1}{n} + \frac{(X_c - \overline{X})^2}{\mathrm{SSX}} \right]^{1/2} \end{aligned} \tag{18}$$

where the subscript c denotes the condition, the location in X, at which the estimate is to be made. Notice that the square root of n (again) determines the interval width at the mean of X and that the interval grows wider the greater distance X_c is from $\overline{X}$, the sample mean.

In most texts this CI is presented as a pair of curved lines, implying a confidence band for the entire line. Equation (18) is meant to be used for one specified location. To sustain α as the risk of not containing the true value, the entire procedure of selecting n observations, computing the coefficients, and so on would need to be followed for each X_c. Wider limits would be needed if the analyst desired limits for the entire true line. Acton (1959) gives a good discussion of this and many related concepts.

2.8. Predicting a Future Value

For predicting a future value at X_c, $\hat{Y}_c$ is obtained from the regression equation just as in the CI. Here it is the estimate of the mean about which individual values are expected to vary. The expression for prediction limits for a single future value of Y must recognize this extra source of variation associated with individuals. The interval for prediction is here abbreviated PI and called, for example, a "95% PI" for $\alpha = 0.05$.

$$100(1 - \alpha)\mathrm{PI} = \hat{Y}_c \pm t_{n-2,\alpha}\, s_{Y \cdot X} \left[1 + \frac{1}{n} + \frac{(X_c - \overline{X})^2}{\mathrm{SSX}} \right]^{1/2} \tag{19}$$

The use of t in this expression implies the additional requirement that the individuals be normally distributed around the line. If this is not the case, some other constant (possibly with asymmetry) representing the actual distribution will be substituted for t. Statgraphics output for confidence intervals and predictions limits is shown in Table 4.

Again, this process applies for a single prediction. If some fraction of all future values is to be included within the limits, the limits would be called tolerance limits. (The reader is referred again to Crocker 1985 for more detailed discussion.) Table 5 offers selected values of K_c in Eq. (20) for obtaining tolerance intervals (TI) around the line at $\overline{X}(K_1)$ and at $\overline{X} \pm 2s_X(K_2)$. Linear interpolation may be employed to obtain straight-line approximations of the curved tolerance limits. The values of K were obtained by inverse interpolation of the normal distribution for 0.95 confidence of including at least 95% of all future values.

$$0.95/95\%\ \mathrm{TI} = \hat{Y}_c \pm K_c\, s_{Y \cdot X} \tag{20}$$

TABLE 4 95% Prediction and Confidence Limits

		95.00% Prediction Limits		95.00% Confidence Limits	
X	Predicted *Y*	Lower	Upper	Lower	Upper
0.005	6.35531	4.94046	7.77015	5.86233	6.84828
0.250	7.38503	6.00567	8.76440	7.00572	7.76435
0.500	8.43577	7.07471	9.79683	8.12964	8.74190
0.750	9.48651	8.12495	10.84810	9.17816	9.79485
1.000	10.53720	9.15641	11.91810	10.15260	10.92190
1.250	11.58800	10.16980	13.00610	11.08560	12.09030
1.380	12.13440	10.69010	13.57860	11.56250	12.70620

TABLE 5 Coefficients for 0.95/95% Tolerance Limits[a] for $P = 1$

n	K_1 (at $\bar{X}$)	K_2 (at $\bar{X} \pm 2s_x$)	n	K_1 (at $\bar{X}$)	K_2 (at $\bar{X} \pm 2s_x$)
5	6.25	8.00	18	2.85	3.14
6	5.01	6.24	20	2.78	3.03
7	4.37	5.32	25	2.65	2.85
8	3.98	4.76	30	2.56	2.72
9	3.71	4.37	40	2.45	2.57
10	3.52	4.09	50	2.38	2.48
12	3.25	3.71	100	2.23	2.28
14	3.07	3.45	200	2.14	2.16
16	2.95	3.27	500	2.07	2.08

[a] Safe as approximate "simultaneous" limits within the $4s_x$ range.

3. MULTIPLE LINEAR REGRESSION

The ability of today's computers and software to handle large data sets has provided the analysts with many opportunities and dangers. Many of these dangers are associated with the intercorrelations found among the predictor variables in nonexperimental data sets. The predictor matrix is said to be "ill conditioned" or is carelessly referred to as "multicollinear." ("Multicollinear" really means the polar condition where some of the X's enter into linear combinations, resulting in an indeterminant system. "Intercorrelation" is used here to describe the general case of nonorthogonality among the predictors.)

The basic relationships and computational forms, represented in matrix notation, are shown here paralleling the equations of the simple case given earlier ($v = n - P - 1$).

("true" Y)
$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \tag{21}$$

(observed Y)
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{22}$$

(b_j's)
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{23}$$

(predicted or fitted Y)
$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} \tag{24}$$

$$\mathbf{SSRes} = \mathbf{Y}'\mathrm{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} \tag{25}$$

$$\sum_{j=1}^{p} \mathrm{SSReg}_j = \mathbf{b}'\mathbf{X}'\mathbf{Y} = \sum_{j=1}^{P} b_j\mathbf{SPX}_j\mathbf{Y} \tag{26}$$

(variance–covariance)
$$\hat{V}(\mathbf{b}) = [\mathbf{X}'\mathbf{X}]^{-1}\hat{\sigma}^2_\varepsilon \tag{27}$$

$$s^2_{Y\cdot X} = \hat{\sigma}^2_\varepsilon = \frac{\mathrm{SSRes}}{n - P - 1} \tag{28}$$

(joint CI for b's)
$$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \le (P + 1)s^2_{Y\cdot X}F_{(p+1),v,\alpha} \tag{29}$$

(CI for $\hat{Y}$)
$$100(1 - \alpha)\ \mathrm{CI} = \mathbf{X}_c'\mathbf{b} \pm t_{v,\alpha}s_{Y\cdot X}[\mathbf{X}_c'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c]^{1/2} \tag{30}$$

(PI)
$$100(1 - \alpha)\mathrm{PI} = \mathbf{X}_c'\mathbf{b} \pm t_{v,\alpha}s_{Y\cdot X}[\mathbf{X}_c'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c + 1]^{1/2} \tag{31}$$

3.1. Intercorrelation Effects

In regression modeling, intercorrelation affects the process in three basic ways. These three have many secondary and corollary consequences, which will be easily perceived if the basic three are understood. They are:

1. Potentially enlarged variances of the b_j's
2. Intercorrelated estimates of the b_j's
3. Ambiguity in assessing the individual contributions to the regression sums of squares

3.1.1. Potentially Enlarged Variances

In the theoretical case—with "correct" model and fixed residual variance—the variances of the b_j's will grow larger as intercorrelated predictors are added to the model (see Snee 1973) as a consequence of the inverse matrix in Eq. (27). (Notice that estimates of the variances of the b_j's are obtained because an estimate of residual error variance is used. The true theoretical variances result from using σ_ε^2). In many cases in practice also, the s_b^2's will grow larger with the addition of intercorrelated predictors. This is because the increase due to the inverse matrix will more than offset the decrease due to a smaller $s_{Y \cdot X}^2$, which results from additional regression sums of squares. However, in practice, $s_{Y \cdot X}^2$ is often reduced enough by the extra predictor(s) to offset the intercorrelation effect in the inverse matrix. These considerations are at the heart of the burgeoning variety of (predictor) variables selection schemes currently appearing in the literature. A brief discussion of this topic will provided later (see Hocking 1976).

3.1.2. Intercorrelated Estimates

In the left side of Eq. (27), in addition to the diagonal variances, there are off-diagonal covariances of pairs of b_j's. Just as with correlation between variables in Eq. (10), covariance of b_j's implies correlation of b_j's. Figure 3 depicts the joint sampling distribution for a pair of positively correlated b_j's. The distribution results from repeated samplings of n values of Y for a given X matrix. For the case of $P = 2$, the correlation of the b_j's is equal in magnitude but opposite in sign to the intercorrelation of the X_j's (they tend to be equal and opposite also for $P > 2$). Notice that the unconditional sampling range of, for example, b_2 (shown by distance A) is very large compared to the conditional range of b_2 (shown by distance B), given the particular estimate of β_1. The important consequence of these two considerations is that errors in estimating the β_j's tend to be compensating among intercorrelated predictors. So intercorrelations may adversely affect the precision of estimate of the β_j's but may have little adverse effect on the use of the model for prediction. This last conclusion depends, of course, on the intercorrelations among the predictors staying about the same in prediction as they were in the sample.

3.1.3. Ambiguity in Assessing Contributions

The underlying nature of the problem is easy to comprehend (for an introductory geometric interpretation of these phenomena, see Crocker 1967, 1969). Interpreting the specific consequences in a particular problem can be extremely complicated. This is true because the ambiguity can be of up to Pth order. The problem is further complicated by the existence of two basic classes of intercorrelated ambiguity, which, for $P \geq 3$, can simultaneously be present in all sorts of hierarchical combinations. Here, the surface will only be scratched with an illustration contrasting the two classes for $P = 2$, the least complex intercorrelation situation. (See also Sections 3.9 and 6.)

In most references, "intercorrelated" and "confounded" are regarded as synonymous. Actually, confounding is only one of the two classes just mentioned. The other has not been given a name by others but is here titled "resolving." This name was chosen because the separate effects of the two or more (resolving) predictors are not "resolved" (clearly seen) until they appear in the model together. The contrast between confounding and resolving is shown in Table 6. The circles at the bottom left represent the two predictors. Area is proportional to regression sums of squares with values as shown. The shaded area of overlap represents intercorrelation. The table shows the allocation

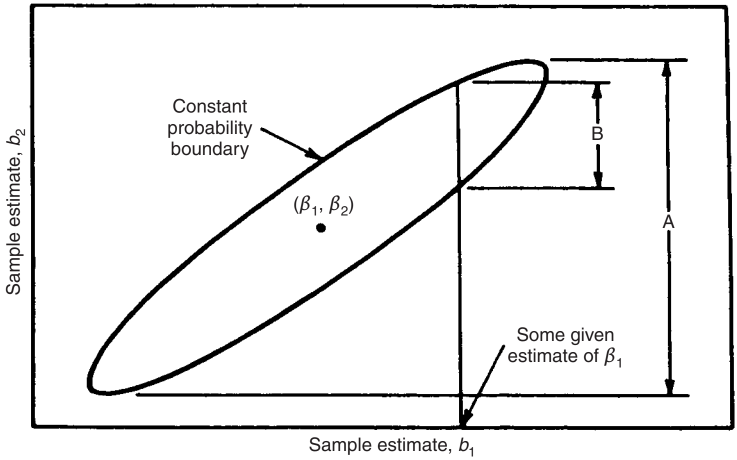
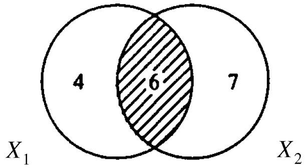


Figure 3 Correlation of Estimates of Slopes.

TABLE 6 Demonstration of Intercorrelation Effects[a]

	Class			
	Confounding		Resolving	
Definition	$R^2 < r_{1Y}^2 + r_{2Y}^2$		$R^2 > r_{1Y}^2 + r_{2Y}^2$	
Model order	X_1 first	X_2 first	X_1 first	X_2 first
SSReg_1	10	④	4	⑩
SSReg_2	⑦	13	⑬	7



[a]For assumed $n = 103$, $\text{SSY} = 27$, $F_j = 10\ \text{SSReg}_j$.
For last predictor (circled entries), $F_j = t_j^2$.

of the 17 SSReg units to the two predictors for the two classes and for the two possible orderings of predictors in the model.

In the confounding case the ambiguous six units are allocated to the first predictor; the second predictor accounts for the balance. In resolving, the six units are available to the second predictor only after the first has clarified the picture. Notice that the total information, 17 units, is always the same. (other features of this table are discussed in subsequent sections as other diagnostic measures are presented). Relevant to this allocation process, care must be taken in interpreting Eq. (26). This equation says that the total regression sum of squares can be obtained from the sums of products of the b_j's with their corresponding SPX_jY's. It does not assert that the individual SSReg_j's can be found this way. As can be seen from the foregoing discussion, the individual SSReg_j's will depend on the order of appearance in the model. The total is order independent.

Extreme confounding is frequently encountered in nonexperimental data sets. It is important to recognize two quite different situations that may arise. Essentially, there is a duplication of information (i.e., a redundancy in the system). In one situation it may be that the same information is presented twice in slightly different forms (such as two different price indexes). This represents model redundancy and is dealt with by removing the redundant predictor. By contrast, it might be that two really different effects are present, but because in nature they are highly intercorrelated, their separate contributions cannot be discriminated statistically. This represents data redundancy and can clearly present a danger if one or the other predictor is arbitrarily excluded from the model, if estimation is the goal. An example is the use of R&D and capital expenditures to assess the number of technical staff needed in a business. Both effects are real, yet it would not be surprising to see them highly intercorrelated and thus inseparably confounded. This true dilemma motivates the current development of biased estimation techniques such as ridge regression (e.g., see Wichern and Churchill 1978) and will be discussed briefly later.

3.2. Detection of Intercorrelation

Several techniques have been proposed for detecting intercorrelation. These include examination of the off-diagonal elements of $\mathbf{X}'\mathbf{X}$, the examination of eigenvalues of $\mathbf{X}'\mathbf{X}$, the use of principal components, and the use of variance inflation factors (VIFs). Where VIFs are the diagonal elements of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix. Most current regression software will display these VIFs.

The VIF for each estimated coefficient b_j can be computed as $\text{VIF}_j = 1/(1 - R_j^2)$, where R_j^2 is the coefficient of determination obtained from regressing X_j on the other predictor variables. As R_j^2 approaches 1 (i.e., nearly linear dependent) the VIF for the estimated coefficient will tend to infinity. VIFs larger than 10 suggest problems with intercorrelation.

3.3. Meaning of Partial Correlation

For the two-predictor case, the (first-order) partial correlation is given by

$$r_{2Y\cdot 1} = \frac{r_{2Y} - r_{12}\, r_{1Y}}{[(1 - r_{12}^2)(1 - r_{1Y}^2)]^{1/2}} \tag{32}$$

This gives the correlation of X_2 with Y, given X_1 (or "while holding X_1 constant " or "while first removing the effect of X_1''). For a given pair of correlations of X_1 and X_2 with Y, r_{12} can influence this expression to be larger or smaller in absolute value than it would be in the orthogonal case ($r_{12} = 0$). When the partial is diminished compared to the orthogonal case, confounding exists. When the partial is increased, it is resolving. Partial correlations relating to the example in Table 6 would, in each case, be based on the circled (last-position) values. Ordinary correlations would be based on the uncircled (first-position) values. The coefficient of determination [Eq. (11)] can be used to represent these two views. The ordinary r^2 would use Eq. (11) as is. The partial coefficient of determination would place the circled value in the numerator and the net amount of SSY remaining, after removing the effect of the first predictor, in the denominator. Table 7 shows these ratios based on SSY = 27.

3.4. Multiple Correlation

The multiple correlation is represented by R. It is in fact the correlation of $\hat{Y}$ with Y, where $\hat{Y}$ a linear combination of the X's. Of course, the X's may individually have correlations with Y of either sign. Hence R is arbitrarily defined as being positive. Direct practical interpretation of R is difficult. Two transformations help to improve interpretation. One is R^2. As with the simple model, R^2 is the "coefficient of determination" and represents the fraction of SSY accounted for by the model (R^2 = SSReg/SSY). For orthogonal predictors, $R^2 = \sum_{j=1}^{P} r_{jY}^2$. For $P = 2$, if $R^2 > r_{1Y}^2, + r_{2Y}^2 X_1$ and X_2 represent a resolving pair, where $R^2 < r_{1Y}^2 + r_{2Y}^2$, X_1 and X_2 are confounded. These relationships were shown in Table 6 and evaluated in Table 7. A second transformation is "$\%s_Y$ removed." The percentage reduction in s_y is related to R as follows:

$$\%\ s_y \text{ removed} = 100\left\{1 - \left[\frac{(1 - R^2)(n - 1)}{n - P - 1}\right]^{1/2}\right\} \tag{33}$$

For a more extensive discussion of R and a graph of Eq. 33, see Crocker 1972. Another related statistic is the "adjusted" R^2 and is used because the ordinary R^2 will never decrease when a new predictor is added to the model. The adjusted R^2, $\overline{R}^2$, is estimated by replacing SSRes and SSY with their mean squares (MS). The resulting equation is

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - P - 1} = \frac{s_Y^2 - s_{Y \cdot X}^2}{s_Y^2} \tag{34}$$

3.5. Relating *t* and *F* in Modeling

As shown in Section 2, the t ratio [Eq. (12)] gives the number of standard errors the estimated value of the coefficient is away from zero. That is still a correct interpretation in the multiple case. It is still useful in assessing the precision of the estimate as per Eq. (17). The t ratio does not, however, measure the contribution, the importance, the practical significance, or even the statistical significance of the associated term in the model! To use this statistic for assessing the contribution of a predictor, it must be carefully qualified. It answers the question "What is the impact of the unique contribution of this predictor?" "Unique" is taken here to mean "impact after resolving." Hence, it is the same as asking what the impact is for this predictor put last in the model.

For answering scientific questions about truth, this gives the t ratio a conservative interpretation. In terms of its influence in reducing $s_{Y \cdot X}^2$, $|t| = 1.0$ is the break-even value for any one predictor. With $|t| > 1.0$, $s_{Y \cdot X}^2$ is reduced by including this predictor. To have (unique) statistical significance, $|t|$ should exceed some appropriate critical value. For excellent precision in estimating β,$|t|$ should be near, say, 20 or 40 (see Section 2.6). The analyst must be wary not to exclude an important term with small t resulting from confounding. What action is appropriate depends heavily on goals (see Section 1) and upon intimate system knowledge.

The ordered F ratio for a single predictor is defined by the ratio of mean squares, regression/residual:

TABLE 7 Coefficients of Determination for Table 6 Values

	r_{1Y}^2	$r_{1Y \cdot 2}^2$	r_{2Y}^2	$r_{2Y \cdot 2}^2$	$R_{Y \cdot 12}^2$	$r_{1Y}^2 + r_{2Y}^2$
Confounding	0.370	0.286	0.481	0.412	0.630 <	0.851
Resolving	0.148	0.500	0.259	0.565	0.630 >	0.407

$$F_j = \frac{\text{MSReg}_j}{\text{MSRes}} = \frac{\text{SSReg}_j/1}{\text{SSRes}/(n - P - 1)} \tag{35}$$

It is called "ordered" because it contains the SSReg of the associated predictor, and this quantity is order dependent, as illustrated in Table 6. When $j = P$, $F_j = t_j^2$. Thus, the t ratios are all proportional to the square roots of the respective SSReg obtained for each predictor as if it were in last position.

For the example of Table 6, the denominator of Eq. (35) is $(27 - 17)/100 = 0.1$. Hence the ordered F values are the Table 6 entries multiplied by 10, and for the circled values these are t^2. So it is seen that t ratios are really "partial" t ratios and are best interpreted in terms of their relationship to last-position SSReg contributions.

3.6. Dealing with Interactions

Sometimes intercorrelation is carelessly referred to as "interaction." Care should be taken to distinguish these two very different concepts. Intercorrelation is a data phenomenon and is not determined by the form of the regression equation, but rather by the particular set of observed values of the predictor variables. Interaction is a model characteristic. It is represented in the model by the product of two or more predictors. It is put there in an attempt to measure interactive behavior in the system represented by the model. Equation (36) shows an interactive model where $X_3 = X_1X_2$ represents a third predictor created from the first two (subscript i is omitted for simplicity).

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \tag{36}$$

The meaning of "interaction" is this: The effect of one predictor depends on the value of another predictor. This is easily seen to be the case for Eq. (36) by factoring either X_1 or X_2. For illustration, X_1 is used.

$$\hat{Y} = b_0 + (b_1 + b_3X_2)\,X_1 + b_2X_2 \tag{37}$$

Here the coefficient of X_1 is $(b_1 + b_3X_2)$. Therefore, the effect of X_1 (its coefficient, $b_1 + b_3X_2$) depends on the value of X_2. By symmetry, the reverse is also true.

No special steps need to be taken to evaluate an interaction. Its t ratio will assess its additional contribution to SSReg, as was previously discussed. However, care is needed in interpreting the associated main effects. In general, where the X's are in their raw original forms, the interaction term will be highly confounded with the associated main effects—the predictors from which it is formed. This will tend to depress the t ratios of these main effects even where the interaction contributes a sizable SSReg (thereby reducing $s_{Y\cdot X}^2$). This should be of no concern. It is purely an arbitrary scaling problem. If desired, the interaction can be made approximately orthogonal to the main effects by subtracting their respective means before forming the product. This has no effect on the statistical assessment of the interaction.

3.7. Basics for Attribute Modeling

Regression modeling is not limited to using quantitative predictors. Any categorization, classification, or logical distinction can be represented. If there is a single class, no distinction is needed. If there are two classes, (e.g., male, female), an additional X is provided to give an attribute code to distinguish the two classes: $X = 0$ if the individual is in the first class (male), $X = 1$ for the second (female). The value chosen is arbitrary, but 0, 1 coding is easiest to interpret. This is "differential coding," which means that the intercept, b_0, will represent the level in Y of the $X = 0$ group, and the coefficient for this code will estimate the difference in Y between the two classes.

To measure, for example, differences of each working day compared to Monday (arbitrarily chosen as the base of comparison), four extra predictors will be needed. Each will be given the value 1 only if the observation represents the associated day; otherwise it will be given the value 0. In general, the number of predictors added will be one less than the number of classes $(c - 1)$. In statistically evaluating the contribution of such a categorical coding scheme, a single test statistic should be used for the $c - 1$ DOF. This is because individual (single-DOF) SSReg contributions depend on the arbitrary choice of the base of comparison and the order. The total, however, is independent of the choice of base and order. The total can be evaluated using the F ratio as shown in Eq. (38), assuming that these terms are last in the model.

$$F_{c-1,n-P-1,\alpha} = \frac{\sum_{j=p-c+2}^{P} \text{SSReg}_j/(c-1)}{\text{MSRes}} \tag{38}$$

Variables selection programs that operate on individual DOF effects are clearly inappropriate for dealing with categorical structures.

Figure 4 illustrates a model with a single quantitative predictor, a two-class attribute shift, and an interaction of these two. Equation (36) applies here and implies that the slope in X_1 is different for the two classes.

3.8. Dealing with Covariates

A covariate is a source of variation contributing to SSY that may not be of particular interest but whose effect must be removed (1) in order to get unbiased estimates of other predictors of interest and (2) in order to reduce the noise level of the system so that predictors of interest can be more clearly seen. It may be that a covariate is confounded with a predictor of interest. The use of the t ratio in evaluating the reality of that predictor's contribution will then quite properly be conservative—discounting the information held in common with the covariate.

Where a categorical structure of three or more classes is involved in a covariate situation, special care must be taken. If the categorical group is the focus of interest, then it must be placed at the end of the model so that its apparent contribution, evaluated by Eq. (38), will have been reduced according to confounding with a covariate. If the categorical group is the covariate, then the term with which it is confounded will have a properly deflated t ratio independent of model position. Hence attribute code groups can always be safely placed at the end of the model.

3.9. Application to the Example

In the application of the simple model to the data of Table 3, a poor fit was obtained (see Section 2). This motivated an inquiry to find additional predictor variables. The plant engineer suggested testing steam used in production processes. He also later recalled a change in policy regarding comfort heating—a change that was coincident with the data set in hand. Table 8 extends the original data set of Table 3 by adding two more predictors. Production (X_2) is in units per period associated with a process which uses steam for heat in manufacturing. The change in policy is represented by the attribute code (variable 3), which starts at 0 when the heating level was 72°F (22°C) and goes through an adjustment (estimated by the engineer) over several periods to the new level of 65°F (18°C). The policy change should (only) affect the heating coefficient, β_1, and so is introduced as an interaction,

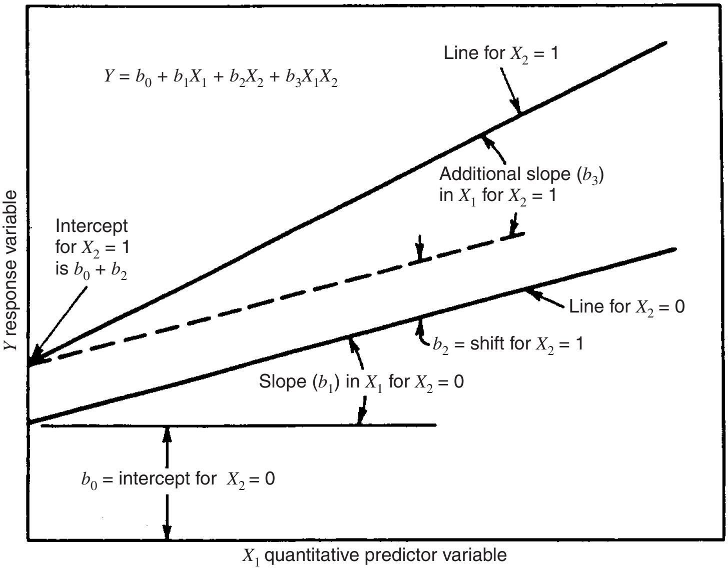

$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$
Line for $X_2 = 1$
Additional slope (b_3) in X_1 for $X_2 = 1$
Intercept for $X_2 = 1$ is $b_0 + b_2$
Line for $X_2 = 0$
b_2 = shift for $X_2 = 1$
Slope (b_1) in X_1 for $X_2 = 0$
b_0 = intercept for $X_2 = 0$
Y response variable
X_1 quantitative predictor variable

Figure 4 Illustration of an Attribute Code Shift and Slope Change (Interaction).

TABLE 8 Example Steam Consumption—Extended Data Set

i	Period Number	Heat (X_1) k degree-days	Production (X_2) Units	Policy (V_3) Attributes	$X_3 = X_1^* V_3$ k-degree-days	Steam (Y) (gBtu)
1	1-10	0.156	413	0.00	0.000	7.991
2	1-11	0.419	396	0.00	0.000	8.589
3	1-12	0.658	385	0.00	0.000	9.145
4	1-13	1.009	243	0.00	0.000	11.212
5	2-01	1.380	391	0.00	0.000	11.754
6	2-02	1.103	407	0.00	0.000	11.469
7	2-03	1.000	411	0.00	0.000	10.584
8	2-04	0.703	379	0.00	0.000	9.509
9	2-05	0.207	402	0.00	0.000	7.457
10	2-06	0.086	406	0.00	0.000	6.989
11	2-07	0.024	383	0.00	0.000	6.537
12	2-08	0.005	227	0.10	0.001	4.938
13	2-09	0.026	265	0.25	0.007	5.275
14	2-10	0.161	384	0.40	0.064	7.452
15	2-11	0.307	400	0.55	0.169	7.962
16	2-12	0.664	379	0.70	0.465	8.915
17	2-13	1.039	354	0.85	0.883	9.758
18	3-01	1.275	392	1.00	1.275	11.183
19	3-02	1.193	412	1.00	1.193	11.523
20	3-03	0.953	408	1.00	0.953	10.426

X_3 = degree-days × policy. This was suggested when the residuals from the two-predictor model displayed a slight downward trend over time. Essentially, this corrective third predictor is a covariate.

Analysis using Eq. (36) provided residual errors that were examined for pattern and excessive deviance. Figure 5 shows the residual plot (*Y-YHAT* vs. *YHAT*) produced by Statgraphics for the data of Table 8 using the $P = 3$ model and the full data set ($n = 20$). It may be noted that the residuals (*Y-YHAT*) form an arc over the range of *YHAT* and that one point (observation 4) plotted just below $P = 3$ in the title is distinctly away from the group. This fourth point was found to be at $+2.6s_{Y \cdot X}$ (Additional diagnostics relating to this are discussed in Section 6.) Further investigation revealed the malfunctioning of a steam trap in the production system. This would account for an indeterminate excess consumption of steam during the fourth period. Therefore the point was excluded.

Using the remaining 19 observations, the regression analysis was repeated for $P = 3$ [see Eq. (36)]. Statgraphics residual plot for this case is shown in Figure 6. The graph no longer displays any obvious lack of fit, outliers, or nonuniformity of variance. Additional diagnostics for this example will be displayed in Section 6 after some additional concepts have been introduced.

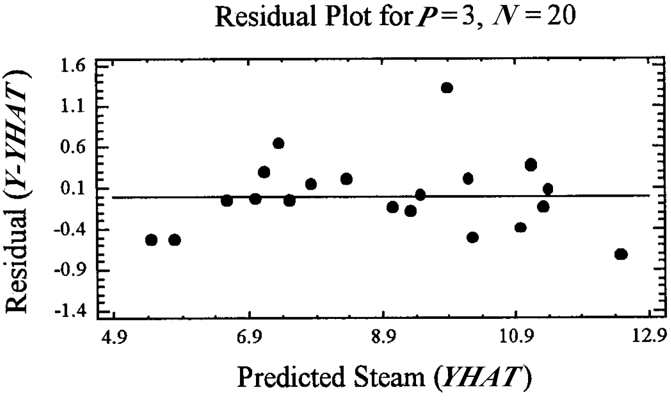


Figure 5 Representation of the Statgraphics Residual Plot.

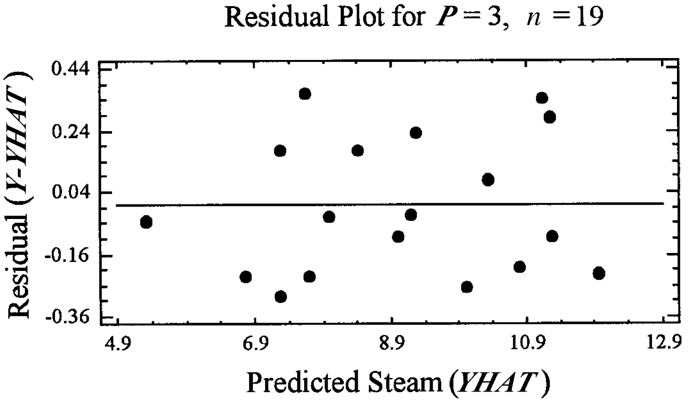


Figure 6 Representation of the Statgraphics Residual Plot.

4. RELATING DIAGNOSTIC QUESTIONS TO GOALS

4.1. Motivation

Regression analysis programs offer a great variety of analytic and diagnostic devices to assist the user. Some statistics that are automatically provided are not always relevant. Sometimes inexperienced users ask what a particular statistic is supposed to "do" or complain that it "isn't any good." Others try to discover some ritualized procedure that can be "followed" and wonder which are the key indicators belonging to such a procedure. Still others, seeking simplistic answers to complex questions, will attempt to impose rules of thumb or employ model selection algorithms. The analyst must learn to understand the complex relationships between diagnostic devices and analytic goals in order to make intelligent use of a program. Not all diagnostics need be examined just because they are there. There is no single, fixed path to follow in doing regression modeling, but rather a complicated cyclical, evolutionary behavioral process that requires simultaneous examination of many relevant diagnostics whose interpretation will frequently generate additional questions.

Space does not permit a complete display relating diagnostics and goals. It is hoped that the following summary of questions diagnostically related to goals will be helpful to the analyst in choosing (or choosing to examine) appropriate diagnostics. In addition to the five project goal-categories presented in Section 1, the analyst has two additional goals while performing regression analysis: data evaluation and model validation. They represent interconnecting analytic steps that motivate the generation of additional diagnostics.

4.2. Summary of Interrelated Diagnostic Questions

1. Are data typical, of adequate range, properly transformed, and error free (especially those of high influence)? Is the data sample large enough? Can a more balanced (less intercorrelated) sample be obtained?
2. Is there evidence of missing predictors or other lack of fit?
3. Is the residual standard deviation small enough to indicate probable model utility?
4. Are intercorrelations substantial? If so, are they reasonable and understood? What effects may they have on interpretation and decisions with respect to specific goals?
5. Do model coefficients have expected signs and reasonable magnitudes? Are they estimated with adequate precision?
6. Is there evidence of overfitting or instability?
7. Is the model to be used in regions of predictor space not seen in the development data set?

5. AN INTRODUCTION TO MODERN DIAGNOSTICS

Many modern regression diagnostics are magically intertwined—mathematically and computationally—and derive from a single germ. The germ idea can best be phrased as a question: What changes occur if a particular observation (one row of the data matrix) is deleted from the data set? Four aspects of change are of particular interest: change in the residual error of the deleted point and changes in the regression surface, regression coefficients, and residual variance. Another closely related powerful diagnostic is the partial plot (also known as the partial-regression leverage plot). This involves deleting an X column from the data matrix. These row and column deletion constructs will be developed next after proposing some special notation.

5.1. Notation

In addition to the notation shown in Sections 2 and 3, the following conventions will be used in this development:

$\hat{Y}_i(i)$	The ith fitted response value using the regression equation derived with the ith observation row deleted (“$Y - YHAT$ sub i not i”)
$e_i(i)$	The deleted residual, $Y_i - \hat{Y}_i(i)$ (“e sub i not i”)
$\mathbf{b}(i)$	$(P + 1) \times 1$ column vector of estimated regression coefficients derived with the ith row deleted (“**b** not i”)
$\mathbf{X}_j$	$n \times 1$ column vector, jth column of $\mathbf{X}$, one predictor
$\mathbf{X}_i$	$1 \times (P + 1)$ row vector, ith row of $\mathbf{X}$, one observation’s X’s
$\mathbf{X}(i)$	$(n - 1) \times (P + 1)$ matrix of predictor variables excluding the ith observation row
$\mathbf{Y} \cdot (j)$	$n \times 1$ column vector of residuals in $\mathbf{Y}$ found when $\mathbf{Y}$ is fitted using all but the jth predictor
$\mathbf{X}_j \cdot (j)$	$n \times 1$ column vector of residuals in $\mathbf{X}_j$ found when $\mathbf{X}_j$ is fitted using all but the jth predictor
$s_{Y \cdot X}(i)$	Residual standard deviation for a model fitted to a data set excluding the ith observation row

5.2. Getting the Catcher and the Hat

Equation (23) can be rewritten in condensed notation as

$$\mathbf{b} = \mathbf{C}'\mathbf{Y} \tag{39}$$

where $\mathbf{C}'$ is called the matrix of catchers by Mosteller and Tukey (1977) and is defined as

$$\mathbf{C}' = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}' \tag{40}$$

So an individual regression coefficient can be found from

$$b_j = \sum_j c_{ij}\, Y_i = \mathbf{C}_j'\mathbf{Y} \tag{41}$$

which shows that the estimated coefficients are just linear combinations of the observed responses. Because each observation potentially contributes differently to this estimation process, each one also has a potentially different expectation for how well it will be fitted, depending on this leverage. This in turn gives an expected error variance for *Y-HAT* for each location in the sample:

$$s_{e_i}^2 = (1 - H_i)\, s_{Y \cdot X}^2 \tag{42}$$

where H_i, the leverage, is the ith diagonal of

$$\mathbf{H} = \mathbf{X}\mathbf{C}' \tag{43}$$

Substituting Eqs. (39) and (43) into Eq. (24) gives

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{C}'\mathbf{Y} = \mathbf{H}\mathbf{Y} \tag{44}$$

which gives $\mathbf{H}$ its name: the “hat” matrix (it puts the hat on Y). H_i sums to $P + 1$ and so represents the “consumption” of fractional DOF associated with individual observations. H_i ranges from $1/n$ at the centroid to a maximum of 1.0 for a shift parameter representing a single observation.

Equation (42) expresses the fraction of $s_{Y \cdot X}^2$ associated with the residual error around the surface. The error variance for the surface at the ith location is the complement,

$$s_{\hat{Y}_i}^2 = H_i s_{Y \cdot X}^2 \tag{45}$$

[Notice that for the simple case, H_i is just the familiar $1/n + (X_i - \overline{X}^2)/\mathrm{SSX}$ of Eq. (18)]. The estimated prediction-error variance for future observations recognizes the potential for influence on the model measured by the leverage. The consequence is the augmentation of the unit variance by H; just the reverse of Eq. (42).

$$s^2_{\text{PRED}i} = s^2_{Y\cdot X}\,(1 + H_i) \tag{46}$$

5.3. Row Deletion

One way to examine the effect of deleting the ith observation is to relate the deleted residual, $e_i(i)$, to the residual, e_i. It seems counterintuitive that this should be simply

$$e_i(i) = \frac{e_i}{1 - H_i} \tag{47}$$

because the leverage is dependent only on the X's: This relationship is unaffected by the observed value of Y. It is important to notice, however, that while the ratio of residuals, $e_i(\mathrm{i})/e_i$, grows larger as H_i, the leverage, grows larger, both residuals may be very small. Leverage is a measure of potential influence; actual influence does indeed depend on observed Y. Unfortunately, some authors treat leverage and influence as being synonymous. By some, high-influence points are regarded as a great danger whereas correct high-influence points are to be cherished as the most informative points. (From very large data sets, leverage can be used for selection of a working subset—analogous to a designed experiment with extremes and center points.)

A proof of Eq. (47) was given by Allen (1971). Another (perhaps more accessible) path to Eq. (47) starts with the basic deletion formulas known as the Sherman–Morrison–Woodbury theorem, illustrated by Rao (1973) in an exercise.

5.4. Internal Validation

Conceptually, n regressions can be performed, each with one of the n observations deleted. From each regression the residual error for the deleted point can be calculated. This permits the validation process to be performed on all n observations while using the full available data set to estimate the regression function. How much more sensible this is than holding out some valuable data (information) for external validation! Fortunately, as seen from Eq. (47), only one regression computation need be performed to obtain all the desired information.

One useful statistic that can be derived from the $e_i(i)$ is called PRESS, an acronym for prediction sum of squares.

$$\text{PRESS} = \sum_i [Y_i - \hat{Y}_i(i)]^2 = \sum_i e_i^2(i) \tag{48}$$

It is axiomatic that the model that fits the development data set the best (minimum $s_{Y\cdot X}$) will not be the best prediction model. This is the consequence of the tendency to “overfit” to fit noise. The minimization of PRESS as a criterion for the choice of a predictive model, as suggested by Allen (1971), tends to counter this overfitting and is regarded by many as being superior to the C_P statistic (see, e.g., Mallow 1973), whose use is similarly motivated.

5.5. Examining Residual Errors and Influence

A variety of diagnostics for examining individual observations for extreme deviance and extreme influence can be derived using H_i. Four are chosen here. The following developments [except Eq. (57)] are given by or derived from the work of Belsley et al. (1980). The first diagnostic, the standardized residual for the ith observation, is

$$t_i = \frac{e_i}{s_{Y\cdot X}(1 - H_i)^{1/2}} \tag{49}$$

This statistic (sometimes called the “internalized t ratio”) is often examined in screening for outliers (observations that may contain mistakes or may represent unusual conditions). Observations that generate values exceeding 2.0 in absolute value might be routinely examined.

Other analysts prefer the second of the four, the studentized residual (the “externalized t ratio”),

$$t_i(i) = \frac{e_i}{s_{Y\cdot X}(i)(1 - H_i)^{1/2}} \tag{50}$$

where

$$s_{Y\cdot X}(i) = \frac{[(n - P - 1)\,s^2_{Y\cdot X} - e_i^2/(1 - H_i)]^{1/2}}{(n - P - 2)^{1/2}} \tag{51}$$

[The original SSRes is reduced by the product of e_i and $e_i(i)$.] Here, one might usefully examine a listing of the $s_{Y \cdot X}(i)$ as well. Here, one might also argue for the use of $e_i(i)$ in Eq. (50) with the prediction form (from Eq. (46) in the denominator. It turns out to be exactly the same thing! For the predicted point, the predictive location has $H_i(i)$ found by

$$H_i(i) = \mathbf{X}_i \, [\mathbf{X}'\mathbf{X}(i)]^{-1} \, \mathbf{X}_i' = \frac{H_i}{1 - H_i} \tag{52}$$

which provides the equivalence for the studentized residual. For Gaussian errors the distribution of $t_i(i)$ would closely follow the t distribution with $n - P - 2$ DOF.

One intuitively appealing measure of influence would be the deleted regression surface shift, $\hat{Y}_i - \hat{Y}_i(i)$. In developing a standardized form of this shift for assessment, it is discovered that the result is identical to t_i, the standardized residual of Eq. (49)! Hence, that statistic may be retained to provide this additional meaning. And in passing it is noted that the surface shift is

$$\hat{Y}_i - \hat{Y}_i(i) = \frac{e_i H_i}{1 - H_i} \tag{53}$$

An alternative scaling of this difference may be obtained by answering the question "Compared to the uncertainty with which the position of the surface has been established at this location, how big is the shift created by this point's inclusion?" This would use the standard error of the surface instead of the standard error of the shift in the surface. This scaling, the third measure of influence /deviance, has been labeled DFITS_i, defined as

$$\text{DFITS}_i = \frac{\hat{Y}_i - \hat{Y}_i(i)}{s_{\hat{Y}_i}(i)} \tag{54}$$

where

$$s_{\hat{Y}_i}(i) = s_{Y \cdot X}(i) \, H_i^{1/2} \tag{55}$$

Then substituting Eqs. (53) and (55) into (54), DFITS_i becomes

$$\text{DFITS}_i = \frac{H_i^{1/2}}{(1 - H_i)} \, \frac{e_i}{S_{Y \cdot X}(i)} \tag{56}$$

Cook (1977) developed his index of influence (the fourth examined here) in terms of the shift in the vector **b** associated with the deletion of the ith observation [see Eq. (59)]. It is structured so that the shift can be evaluated using $F(P + 1, n - P - 1, 1 - \alpha)$ with $1 - \alpha = 0.10$ giving an upper bound for "an uncomplicated analysis," according to Cook. The index can be reduced to the following form:

$$\text{COOKD}_i = (t_i^2) \, \frac{1}{P + 1} \, \frac{H_i}{1 - H_i} \tag{57}$$

The third factor in this expression is the surface shift factor of Eq. (53) and is the "deleted leverage" of Eq. (52). It is also the ratio of the two partitioned parts of the residual variance [Eqs. (45) and (42)]:

$$\frac{s_{Y_i}^2}{s_{e_i}^2} = \frac{H_i}{1 - H_i} \tag{58}$$

This says that the larger the leverage, the larger the uncertainty of the location of the surface, the more the shrinkage of the residual. In Eq. (57), this ratio amplifies the measure of deviance given by t_i^2. Squaring DFITS and dividing by $P + 1$ gives the same measure as Eq. except that $t_i^2(i)$ is used instead of t_i^2.

It is noted in passing—for examining the deletion impact on the regression coefficients—that

$$\mathbf{b} - \mathbf{b}(i) = \frac{\mathbf{C}_i' e_i}{1 - H_i} \tag{59}$$

This result is also derived from the work of Beckman and Trussell (1974) and the basic deletion formulas discussed after Eq. (47).

5.6. Partial Regression Leverage Plots

A partial plot reveals the underlying relationship between the response and the *j*th predictor with the influence of all other predictors removed. That is, it plots $\mathbf{Y} \cdot (j)$ vs. $\mathbf{X} \cdot (j)$. Such a plot may reveal curvature, discontinuities, extreme influence, or other aberrations often not readily detected by plotting *Y-YHAT* or $\mathbf{Y} \cdot (j)$ vs. $\mathbf{X}_j$. This is especially true for resolving predictors. The simple regression of these residual variables gives the partial regression slope for the *j*th predictor in the full model—the usual multiple regression slope not always qualified as being partial. Deviations around this regression line are the full-model residuals. The correlation of these two residual variables is the $(P - 1)$st-order partial correlation of Y with X_j. One need not perform $2P$ multiple regressions to obtain the required vectors. Mosteller and Tukey (1977) and Velleman and Welsch (1981) discuss details leading to the following results. The starting point is the identity

$$\begin{aligned}\mathbf{Y} &= \sum_{(j)} b_k \mathbf{X}_k + b_j \mathbf{x}_j \cdot (j) + \mathbf{e} \\ &= \hat{\mathbf{Y}}(j) + b_j \mathbf{X}_j \cdot (j) + \mathbf{e} \qquad (60)\end{aligned}$$

Then

$$\mathbf{Y} - \hat{\mathbf{Y}}(j) = \mathbf{Y} \cdot (j) = b_j \mathbf{X}_j \cdot (j) + \mathbf{e} \qquad (61)$$

From Eq. (61) it is evident that the ordinate of the partial plot is simply $b_j \mathbf{X}_j \cdot (j) + \mathbf{e}$. The abscissa is $\mathbf{X}_j \cdot (j)$. Since the b_j and $\mathbf{e}$ are available from the complete multiple regression, all that is needed is $\mathbf{X}_j \cdot (j)$. This is obtained from

$$[X_j \cdot (j)]_i = \frac{c_{ij}}{\sum_k c_{kj}^2} \qquad (62)$$

where the denominator sums of squares of c_{kj} are just the diagonal elements of $[\mathbf{X}'\mathbf{X}]^{-1}$.

6. DIAGNOSTICS FOR THE EXAMPLE

6.1. Leverage and Influence

Available space is insufficient to permit demonstrating the partial plots for the example. However, Table 9 lists the leverages and the four influence diagnostics associated with individual observations

TABLE 9 Printing of Diagnostics for $P = 3$ $n = 20$

Row	Steam	Predicted	Residual	Studentized Residual	DFITS	COOKD	Leverage
1	7.991	7.357	0.634	1.406	0.592	0.083	0.151
2	8.589	8.376	0.213	0.430	0.131	0.005	0.086
3	9.145	9.333	−0.188	−0.378	−0.114	0.003	0.083
4	11.212	9.900	1.312	7.530	7.013	2.744	0.465
5	11.754	12.482	−0.728	−1.921	−1.375	0.405	0.339
6	11.469	11.396	0.073	0.159	0.083	0.002	0.213
7	10.584	10.979	−0.395	−0.858	−0.402	0.041	0.180
8	9.509	9.487	0.022	0.045	0.014	0.000	0.086
9	7.457	7.503	−0.046	−0.095	−0.035	0.000	0.121
10	6.989	7.009	−0.020	−0.041	−0.018	0.000	0.157
11	6.537	6.589	−0.052	−0.107	−0.044	0.001	0.146
12	4.938	5.469	−0.531	−1.440	−1.248	0.365	0.429
13	5.275	5.809	−0.534	−1.271	−0.791	0.151	0.279
14	7.452	7.149	0.303	0.625	0.219	0.123	0.110
15	7.962	7.825	0.137	0.277	0.090	0.002	0.097
16	8.915	9.056	−0.141	−0.282	−0.074	0.001	0.065
17	9.758	10.269	−0.511	−1.131	−0.531	0.069	0.181
18	11.183	11.318	−0.135	−0.318	−0.225	0.013	0.334
19	11.523	11.144	0.379	0.888	0.570	0.082	0.292
20	10.426	10.219	0.207	0.445	0.216	0.122	0.190

for the example data of Table 8. Observation 4 has the highest leverage (at nearly a half a *df* "consumed" by this one point) and is also shown by all four diagnostics to have high influence. As noted earlier, the steam consumption for this observation was found to be anomalous and so the point was removed from the data set. For contrast, it is noted that observation 12 has nearly as high a leverage (at the low end of the ranges of both X_1 and X_2) but fits rather well and so has much smaller measures of influence (although COOKD_{12} at 0.365 is above the 0.10 value of F, which is 0.259). The regression analysis from which these diagnostics were obtained ($n = 20$, $P = 3$) is shown in Table 10 as produced by Statgraphics.

6.2. Final Results

With the fourth observation removed, the $P = 3$ model was repeated for the $n = 19$ remaining periods. The final results are summarized in Tables 11 and 12. The following comments deal with the main conclusions demonstrated in the table and with some diagnostics not mentioned earlier.

1. The coefficients were judged reasonable in sign and size, and the two main predictors are estimated with reasonable precision [see Eq. (17)].
2. The covariate third predictor does not show strong influence but was retained in the model to avoid biasing the estimates of the other coefficients. (The intercept represents the average consumption of all other steam uses uncorrelated with the model predictors.)
3. The final value for $s_{Y \cdot X}$ of 0.238 is the residual standard deviation for this model and data set.

TABLE 10 Statgraphics Regression Analysis Results for $P = 3$ $n = 20$

```
Multiple Regression Analysis
Dependent variable: Steam

                                          Standard           t
Parameter             Estimate            Error          Statistic       P-Value
CONSTANT              3.93783000        0.77098700         5.10752        0.0001
Heat                  4.30712000        0.29227900        14.73630        0.0000
Production            0.00665141        0.00209879         3.16917        0.0060
Heat × Policy        -0.56407600        0.31144900        -1.81114        0.0889

                         Analysis of Variance

                  Sum of
Source            Squares       Df      Mean Square       F-Ratio        P-Value
Model             78.48270       3        26.160900        102.97         0.0000
Residual           4.06502      16         0.254064
Total (Corr.)     82.5477       19
R-squared = 95.0756 percent
R-squared (adjusted for d.f.) = 94.1522 percent
Standard Error of Est. = 0.504047
Mean absolute error = 0.328052
Durbin-Watson statistic = 2.31226

Unusual Residuals

                            Predicted                             Studentized
Row            Y                Y              Residual             Residual
 4          11.212             9.9              1.312                  7.53

Influential Points

                                       Mahalanobis
Row                Leverage              Distance                       DFITS
  4                0.464520              14.66730                     7.01308
  5                0.338859               8.27828                    -1.37505
 12                0.428938              12.57280                    -1.24796
Average leverage of single data point = 0.2
```

TABLE 11 Statgraphics Regression Analysis Results for $P = 3$ $n = 19$

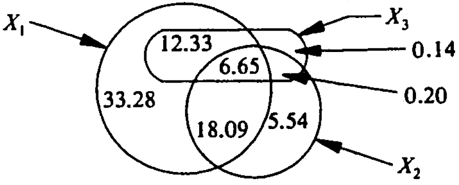
```
Multiple Regression Analysis
Dependent variable: Steam
                                         Standard            t               P-Value
Parameter              Estimate           Error          Statistics

CONSTANT                1.9782800       0.44763600          4.41940          0.0005
Heat                    3.7680100       0.15553200         24.22660          0.0000
Production              0.0122702       0.00124091          9.88810          0.0000
Heat × Policy          −0.2403990       0.15328000         −1.56837          0.1376

                            Analysis of Variance

                     Sum of
Source               Squares          Df     Mean Square       F-Ratio      P-Value

Model              76.232000           3      25.4107000        448.18       0.0000
Residual            0.850463          15       0.0566975

Total (Corr.)           77.0824       18
R-square = 98.8967 percent
R-squared (adjusted for d.f.) = 98.676 percent
Standard Error of Est. = 0.238112
Mean absolute error = 0.188343
Durbin-Watson statistic = 1.96748

95.0% confidence intervals for coefficient estimates (Steam)

                                    Standard          Lower           Upper
Parameter             Estimate        Error           Limits          Limit       V.I.F.

CONSTANT             3.93783000   0.77098700     2.30341000     5.5722500
Heat                 4.30712000   0.29227900     3.68751000     4.9267200   1.43474
Production           0.00665141   0.00209879     0.00220216     0.0111006   1.06577
Heat × Policy       −0.56407600   0.31144900    −1.22432000     0.0961673   1.42454
```

4. The first predictor accounts for about 91% of the variability in Y (as measured by ordered SSReg/SSY). Therefore, it might be tempting to dismiss the use of steam for production (X_2) as unimportant. But X_2 accounts for 53% of the average steam use. That use simply does not vary as severely as weather.
5. The drop from $s_Y = 2.069$ to $s_{Y \cdot X} = 0.238$ represents an 88.5% reduction, and with $s_{Y \cdot X}$ at less than 3% of the mean, this model promises to be effective in monitoring consumption.
6. The VIF for each X is shown to help assess the intercorrelation effects.
7. The SSReg total of 76.23 is allocated among the three predictors according to model position. These SSReg_j can also be subdivided according to the effects of intercorrelations of the predictors. In this uncomplicated system the confounded portions are easily determined to be as shown in Table 12 by both the diagram and the listing to its left.

TABLE 12 SSReg Allocation and Diagram for $P = 3$, $n = 19$

SSReg Class	X_1	X_2	X_3
Unique	33.28	5.54	0.14
Two-Way confounding	18.09(2)	0.20(3)	—
Two-Way confounding	12.33(3)	—	—
Three-Way confounding	6.65	—	—
SSReg_j Totals	70.35	5.74	0.14

All overlap is *confounding*.



7. OTHER REGRESSION TOPICS

7.1. Variable Selection

Variable selection in regression arises when the set of variables to include into the model is not predetermined. The problem to be addressed is from the list of potential candidates to include in the model which ones should be included and in what form. The objectives here are to include as many predictors that can influence the prediction while in addition including as few as possible because the variance of the prediction increases as the number of predictors increase. Hence the goal of variable selection is to find an ‘‘appropriate subset’’ regression model.

Several criteria have been proposed to compare and evaluate the adequacy of the subset regression models. These include using R^2, adjusted R^2, MSE, Mallows’ C_p, and PRESS. A brief description of the adjusted R^2 method will be provided here. Adjusted R^2 was previously defined as

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - P - 1} \tag{34}$$

The adjusted R^2 is used because the ordinary R^2 defined earlier will always increase when new terms are added to the regression model. The adjusted R^2 will not necessarily increase as new terms are added. This helps prevent over fitting the model and determining the ‘‘appropriate subset’’ regression model. Therefore one criterion associated with determining the appropriate subset model is to maximize the adjusted R^2. Note that this is equivalent to selecting the model with a minimum MSE.

7.1.1. All Possible Regressions

This procedure involves fitting all possible subset of regression models and choosing the ‘‘best’’ model based on suitable criteria. If we include the intercept in each model and there are q predictors, then there will be 2^q total equations to be fitted. Thus, if the number of predictors is 5, the total number of equations to be fitted is 32. As can be seen, the number of equations to fit grows rapidly as the number of predictors increase. In the steam example there are 3 predictor variables and thus 8 possible equations to fit. The results of fitting these 8 equations are shown in Table 13 and Figure 7. The equations have been listed by order of maximum adjusted R^2.

For cases where the number of possible equations is large, there are several procedures developed to evaluate only a small subset of these equations by adding or deleting predictors one at a time. These procedures can be classified into three groups (1) forward selection, (2) backward elimination, and (3) stepwise and are briefly described below.

7.1.2. Forward Selection

The procedure begins with the assumption that there are no predictors in the model other than the intercept. An optimal subset is determined by adding predictors into the model one at time with the first one to enter being the predictor with the largest simple correlation with the response variable. This predictor will be entered if it exceeds a predetermined F value (*Fin*). The second predictor to enter the model is then determined by the one that has the largest correlation with the response after adjusting for the effect of the first predictor (i.e., largest partial correlation). This predictor will enter the model if it exceeds *Fin*. This process continues until a predictor does not exceed the *Fin* or when all predictors have been added to the model.

TABLE 13 Statgraphics Output for Regression Model Selection for Steam

Models with Largest Adjusted R-Squared Results

MSE	R-Squared	Adjusted R-Squared	Cp	Included Variables
0.254064	95.0756	94.15220	4.00000	X_1 X_2 X_3
0.288141	94.0660	93.36790	5.28021	X_1 X_2
0.389220	91.9843	91.04130	12.04360	X_1 X_3
0.398461	91.3113	90.82860	12.23030	X_1
3.484540	28.2389	19.79640	219.15900	X_2 X_3
3.676710	19.8271	15.37310	244.48900	X_3
3.934500	14.2060	9.43971	262.75300	X_2
4.344620	0.0000	0.00000	306.90900	

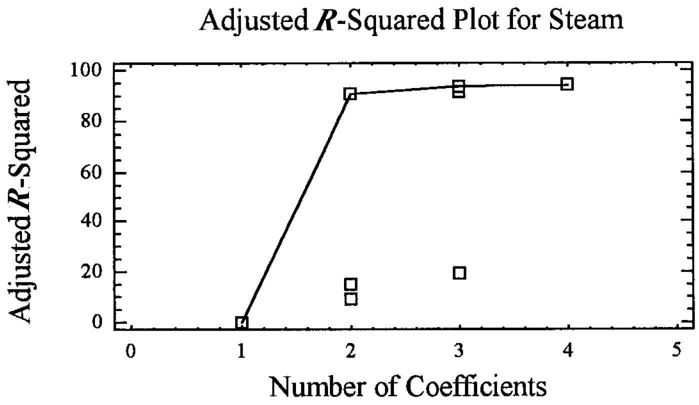


Figure 7 Statgraphics Adjusted R^2 Plot for Steam.

7.1.3. Backward Elimination

This procedure begins with the full equation fitted and successively drops one predictor out of the equation at a time. The predictor that is the first candidate to be eliminated is the one with the smallest contribution to the reduction of the error sum of squares. Based on the full model, a partial F statistic is computed for each predictor as if it were the last variable to enter the model. The predictor with the smallest partial F is eliminated if it is less then a predetermined critical F value (*Fout*). This process continues until no predictor has a partial F less than *Fout.*

7.1.4. Stepwise

The stepwise method is basically a combination of the forward selection and backward elimination methods. Thus, at each stage of the forward selection the possibility of deleting a predictor is also considered. Therefore, a variable that enters at an earlier stage may later be removed.

For all three methods the final model selected like any regression model should be evaluated to the regression diagnostics described earlier.

As an illustrative example of the method, the data for the steam example was analyzed using the forward selection method. The results are shown in Table 14. In this example the first predictor to enter was heat. Next the predictor production was added to the model. The last predictor, heat × policy, did not exceed the *Fin* value and was not included in the final model. However, as discussed earlier, the covariate third predictor was retained in the model to avoid biasing the estimates of the other coefficients.

7.2. Ridge Regression

Ridge regression is employed to combat intercorrelation between the regressors. A set of variables is exactly collinear if one of them is a linear combination of the others. The presence of intercorrelation is given by the variance inflection factors (VIF).

As discussed, least squares provides unbiased estimates with minimum variance of all linear unbiased estimators without upper limit on the variance of the estimators, and if intercorrelation exists, this may produce large variance. Therefore, in the presence of intercorrelation, a penalty is paid for the unbiasness property that is usually attained via least squares. Biased estimation procedures attempt to find a biased estimator of a regression coefficient that has smaller variance than the unbiased coefficient. Ridge regression is a biased estimation procedure to address this. In ridge regression, the analyst would like to select a bias, k, such that the reduction in variance is greater than the increase in the squared bias introduced. The ridge regression estimator, b_r is given by

$$b_r = (\mathbf{X}'\mathbf{X} + kI)^{-1}\,\mathbf{X}'y \tag{63}$$

The choice of k belongs to the analyst and should be chosen where strong evidence shows more stable estimates or improved prediction. One method suggested by Hoerl and Kennard (1970) is the use of a ridge trace. The ridge trace is a plot of the b_r vs. k, usually in the interval (0, 1). For values close to $k = 0$, intercorrelation will cause rapid changes in b_r. The objective is to select a small value of k where the b_r's stabilize.

TABLE 14 Statgraphics Output of Forward Selection Method for Steam

```
Multiple Regression Analysis
Dependent variable: Steam
                                       Standard           t
Parameter          Estimate             Error          Statistic        P-Value
CONSTANT         4.11739000        0.81425000          5.05667          0.0001
Heat             4.03422000        0.26671300         15.12570          0.0000
Production       0.00624244        0.00222214          2.80920          0.0121
Analysis of Variance
                Sum  of                 Mean
Source          Squares       Df        Square             F-Ratio      P-Value
Model           77.6493        2        38.824700           134.74       0.0000
Residual         4.8984       17         0.288141
Total           82.5477       19
(Corr.)

R-squared = 94.066 percent
R-squared (adjusted for d.f.) = 93.3679 percent
Standard Error of Est. = 0.536788
Mean absolute error = 0.34515
Durbin-Watson statistic = 1.82542

Stepwise Regression
Method: forward selection
F-to-enter: 4.0
F-to-remove: 4.0
Step 0:
0 variables in the model.  19 d.f. for error.
R-squared = 0.00%   Adjusted R-squared = 0.00%   MSE = 4.34462

Step 1:
Adding variable Heat with F-to-enter = 189.166
1 variables in the model. 18 d.f. for error.
R-squared = 91.31%   Adjusted R-squared = 90.83%   MSE = 0.398461

Step 2:
Adding variable Production with F-to-enter = 7.89159
2 variables in the model. 17 d.f. for error.
R-squared = 94.07%   Adjusted R-squared = 93.37%   MSE = 0.288141
Final model selected.
```

8. SOME PRACTICAL CONCERNS

The analyst faces a variety of dangers in practice in addition to those discussed earlier. For example, there is often pressure to "keep it simple." The danger is that, by avoiding complexity, the analyst may be seriously misled or fail to develop an adequate model. Simplicity is not necessarily a virtue.

It is also easy to acquire more faith in a complex regression model than it deserves. Even a good model is at best a crude approximation of reality. Yet, by being computer born, it takes on a special aura that may encourage undeserved faith in its utility.

Another danger is the predictive use of a regression model in regions of the joint predictor space not observed in the development sample, even though within all the observed predictor ranges. Equation (52) provides one index of the presence of this condition. "Interpolation" seems to fit, as the result is compared to the distribution of sample leverages.

From initial questions of which variables to collect to the final checking of a suspicious residual, data management is a major constituent of any modeling venture. The collection and "scrubbing" process very often consumes 70% or more of project funds and elapsed time. It bears repeating that a data set will rarely ever be free of mistakes. Some process for auditing the data must be devised at the outset or much effort will be wasted on false starts.

8.1. Model Use and Maintenance

The model is a waste if it is not used. The person who will have responsibility for its use must be involved early enough to gain understanding and develop faith. The model's use must serve an ongoing function that is desired and expected by the user's superiors or it will not survive.

Provision must be made for the timely reporting of the predictors. It is of no use to develop a prediction or control model if the necessary data cannot be obtained in a timely fashion. Results must then be reported to those who can take action. Good predictions kept in a desk serve no one.

Invariably, in practice, the β's estimated are not in fact constant but are creeping and shifting overtime. Additionally, there will inevitably be other systems changes, which, for example, may require the inclusion of additional predictors. So, if the model is to continue in use, provision must be made for updating it. Failing this, the model will begin to miss until it loses credibility and its use is discontinued.

8.2. Helpful Hints in Practice

The following is a summary list of prerequisites for successful use of regression modeling techniques that an analyst should have and/or use:

1. Reasonably specific goals
2. An understanding of statistical procedures
3. Reasonable familiarity with the system modeled
4. Restraint in transforming variables
5. Facility for adequate diagnostic analysis and data scrubbing
6. A cyclical approach with documentation of decisions and choices made
7. Good judgment instead of model selection algorithms
8. Great care when excluding important predictors that were not permitted to vary
9. Great care when including "discovered" relationships
10. A willingness to validate the model and/or anticipate model instability
11. Recognition of the need for maintenance of the model

Computer Software

Statgraphics, Manugistics, Rockville, MD.

REFERENCES

Acton, F. S. (1959), *Analysis of Straight Line Data,* John Wiley & Sons, New York.

Allen, D. M. (1971), "Mean-Square Error of Prediction as a Criterion for Selecting Variables," *Technometrics,* Vol. 13, pp. 469–475.

Barnett, V. (1978), "The Study of Outliers: Purpose and Model," *Applied Statistics,* Vol. 27, No. 3, pp. 242–250.

Beckman, R. J., and Trussell, H. J. (1974), "The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple-Regressions," *Journal of the American Statistical Association,* Vol. 66, pp. 199–201.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* John Wiley & Sons, New York.

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics,* Vol. 19, pp. 15–18.

Crocker, D. C. (1967), "Intercorrelation and the Utility of Multiple Regression in Industrial Engineering," *Journal of Industrial Engineering,* Vol. 18, No. 1, pp. 79–85.

Crocker, D. C. (1969), "Linear Programming Techniques in Regression Analysis: The Hidden Danger," *AIIE Transactions,* Vol. 1, No. 2, pp. 112–126.

Crocker, D. C. (1972), "Some Interpretations of the Multiple Correlation Coefficient," The American Statistician, Vol. 26, No. 2, pp. 31–33.

Crocker, D. C. (1985), *Volume 9: How to Use Regression Analysis in Quality Control,* American Society for Quality Control, Milwaukee.

Eisenhart, C. (1947), "The Assumptions Underlying the Analysis of Variance," *Biometrics,* Vol. 3, No. 1, pp. 1–21.

Eisenhart, C., "The Meaning of 'Least' in Least Squares," *Journal of the Washington Academy of Sciences,* Vol. 54, February, pp. 24–32.

Finch, P. D. (1979), "Description and Analogy in the Practice of Statistics," *Biometrika,* Vol. 66, No. 2, pp. 195–208.

Healy, M. J. R., "Is Statistics a Science?" *Journal of the Royal Statistical Society,* Vol. 141, A1978, Part 3, pp. 385–393.

Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics,* Vol. 32, No. 1, pp. 1–50.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics,* Vol. 12, pp. 69–82.

Hunter, W. G., and Box, G. E. P. (1965), "Experimental Studies of Physical Systems," *Technometrics,* Vol. 7, No. 1, pp. 23–42.

Mallows, C. L., "Some Comments on C_p," Technometrics, Vol. 15, pp. 661–675.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications,* 2nd Ed., John Wiley & Sons, New York, p. 33.

Snee, R. R. (1973), "Some Aspects of Nonorthogonal Data Analysis," *Journal of Quality Technology,* Vol. 5, No. 2, pp. 67–79.

Velleman, P. F., and Welsch, R. E. (1981), "Efficient Computing of Regression Diagnostics," *American Statistician,* Vol. 35, No. 4, November, pp. 234–242.

Wichern, D. W., and Churchill, G. A. (1978), "A Comparison of Ridge Estimators," *Technometrics,* Vol. 20, No. 3, 1978, pp. 301–311.

ADDITIONAL READING

Allen, D. M., and Cady, F. B., *Analyzing Experimental Data by Regression,* Lifetime Learning, Belmont, CA, 1982.

Chatterjee, S., and Price, B., *Regression Analysis by Example,* John Wiley & Sons, New York, 1977.

Dobson, A. J., *An Introduction to Statistical Modelling,* Chapman & Hall, New York, 1986.

Draper, N. R., and Smith, H., *Applied Regression Analysis,* 2nd Ed., John Wiley & Sons, New York, 1981.

Farebrother, R. W., *Linear Least Squares Computations,* Marcel Dekker, New York, 1988.

Freund, R. J., and Minton, P. D., *Regression Methods,* Marcel Dekker, New York, 1979.

Guttman, I., *Linear Models: An Introduction,* John Wiley & Sons, New York, 1982.

Horton, R. L., *The General Linear Model: Data Analysis in the Social and Behavioral Sciences,* McGraw-Hill, New York, 1978.

Kleinbaum, D. G., and Kupper, L. L., *Applied Regression Analysis and Other Multivariate Methods,* Duxbury Press, North Scituate, MA, 1978.

Montgomery, D. C., and Peck, E. A. (1992), *Introduction to Linear Regression Analysis,* John Wiley & Sons, New York.

Myers, R. H., *Classical and Modern Regression with Applications,* PWS-Kent, Boston, 1990.

Neter, J., and Wasserman, W., *Applied Linear Statistical Models,* Richard D. Irwin, Homewood, IL, 1974.

Rice, J. R., *Matrix Computations and Mathematical Software, McGraw-Hill, Tokyo, 1983.*

Wesolowsky, G. O., *Multiple Regression and Analysis of Variance,* John Wiley & Sons, New York, 1976.

Younger, M. S., *A Handbook for Linear Regression,* Duxbury Press, North Scituate, MA, 1979.

V.B

Economic Evaluation

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 88

Product Cost Analysis and Estimating

Phillip F. Ostwald
University of Colorado at Boulder

1. ESTIMATING: IT HAPPENS ALL THE TIME 2298

2. WHY ESTIMATES OF COST ARE MADE 2298
- 2.1. New Product Cost 2298
- 2.2. Make or Buy 2298
- 2.3. Selling Price Determination 2298
- 2.4. Equipment and Technology Acquisition 2298
- 2.5. Cost Control 2299
- 2.6. Temporary Work Standards 2299
- 2.7. Vendor Quote Checks 2299

3. MEASURES OF ECONOMIC WANT 2299

4. REQUEST FOR ESTIMATE 2299
- 4.1. Direct Labor 2299
- 4.2. Indirect Labor 2300
- 4.3. Direct Materials 2300
- 4.4. Indirect Materials 2300
- 4.5. Overhead 2300
- 4.6. General and Administrative 2300
- 4.7. Profit 2300

5. PRELIMINARY AND DETAILED METHODS 2300
- 5.1. Judgment and Conference Method 2300
- 5.2. Unit Method 2301
- 5.3. Comparison 2301
- 5.4. Factor Method 2302
- 5.5. Cost- and Time-Estimating Relationships 2302
  - 5.5.1. Learning Curve 2302
  - 5.5.2. Power Law and Sizing Model 2303
- 5.6. Probability and Statistics Techniques 2304
  - 5.6.1. Expected Value 2304
  - 5.6.2. Percentile Method 2305
  - 5.6.3. PERT-Based Beta Distribution 2305
  - 5.6.4. Computer Simulation 2306
- 5.7. Standard Data 2306
  - 5.7.1. Time Study 2307
  - 5.7.2. Predetermined Time Standards 2307
  - 5.7.3. Work Sampling 2307
- 5.8. Historical Data 2307

6. LABOR ANALYSIS 2307

7. MATERIAL ANALYSIS 2308

8. NEED FOR ACCOUNTING DATA 2309

9. FORECASTING 2310

10. INDEXES 2310

11. OPERATIONS ESTIMATING FOR MANUFACTURING 2311
- 11.1. Preparing the Operations Sheet 2312
- 11.2. Setup Hours Column 2312
- 11.3. Cycle Minutes Column 2314

12. PRODUCT ESTIMATING 2314

13. BILL OF MATERIAL EXPLOSION FOR PRODUCT ESTIMATES 2314

14. COMPUTERS AND ESTIMATING 2316

15. CONCLUSION 2316

REFERENCE 2316

ADDITIONAL READING 2316

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. ESTIMATING: IT HAPPENS ALL THE TIME

Cost estimating is a popular activity within engineering. Even though the professional person may be titled a cost estimator, cost engineer, cost analyst, labor estimator, or material planner, the emphasis remains the same. He or she is required to answer a familiar question: "How much will it cost?" Although the purposes that underlie this question are varied, we find that businesses, government, and not-for-profit organizations desire timely and reliable measures of economic want. For it is an engineer who does the appraisal, analysis, forecasting, and compiling of a pro forma document that extends from the basic cost ingredients to the bottom line of an estimate. Using this evaluation, other management people may determine a price, make vs. buy decision, return on investment (ROI), or public fiscal-year budget. Thus the engineer finds a future value of a product that responds to a specified need.

This historical trail of the development of formal cost estimating is intimately tied to industrial engineering. The original concept of a labor "standard" was seminal in the development of standard cost plans found widely throughout business. The genesis of formal cost estimating is circa 1900 since it was connected closely with manufacturing and construction. Cost estimating is a long-established job and an everyday occurrence for many engineers.

2. WHY ESTIMATES OF COST ARE MADE

Every size and type of organization needs cost estimates to make intelligent decisions. Some organizations employ persons skilled in the area of cost estimating whose primary function is developing estimates. But employees in most functional areas should understand good cost-estimating techniques. With concurrent engineering practices, teamwork philosophies, and total employee involvement, more people need cost-estimating knowledge and skills.

Cost-estimating procedures must be performed quickly and accurately because of tough customer demands and global competition. Listed below are several types of cost estimates that organizations routinely make.

2.1. New Product Cost

When new product concepts or product changes are being considered, detailed estimates of cost aid management in making proper decisions. Detailed estimates include costs of material, processing of material, fabrication, assembly, labor, and purchased components. The processing, fabrication, and assembly costs include estimates for tooling, dies, fixtures, inspection instruments, and so on. Costs for capital equipment investments, space, and facilities are also major estimate areas. If a management decision is to proceed with the new product, the detailed estimate may likely become the budget for the project. This type of estimate should be detailed and cover needs and costs from cradle to grave. Today the end of the product life has been extended to include recycling and disposal of the product and components of the product. It is not uncommon for companies first to determine the market selling price and then work backwards to determine how much cost can be absorbed by different areas of the company. Within each organizational area, costs must be constrained to limits allowed.

2.2. Make or Buy

Companies should consider whether to make components or the final product "in-house" or purchase them from "outside" vendors. Price is usually the deciding factor, but other factors can affect the final decision. Some of these factors are: can production demand requirements be fulfilled, can quality expectations be met, and can delivery schedules be met? Likewise, it might be better to use a supplier because they may have been producing similar parts for years and have the expertise to produce better parts than the company making the estimate. It is always wise to develop these estimates for comparison.

2.3. Selling Price Determination

These estimates can work two ways. First, estimates are used to determine selling price. The estimate establishes the cost to produce, market, deliver, and so on. Then a profit margin can be attached to establish a selling price. If entering an existing market, the competitive selling price can be used to work backwards to determine if producing the product is appropriate.

2.4. Equipment and Technology Acquisition

Companies frequently make decisions about purchasing new equipment, software, or complete systems to replace or add to the present resources. Often this involves comparing different alternatives that comprise new technology and/or changing from manual to automated procedures. Developing accurate cost estimates for new and unfamiliar areas is not easy.

2.5. Cost Control

Some companies, especially job shop-type organizations, use cost estimates as a form of cost control. Lot sizes vary and are usually small and almost every job is different. For these and other reasons, job shops seldom develop work standards to help determine costs. If a management decision is made to proceed with the new product, the detailed estimate may likely become the budget for the project. This type of estimate is not to be considered temporary work standards, because the objective is to determine whether the job can be done profitably and less expensively than by the competition.

2.6. Temporary Work Standards

Flow shop companies producing products in high volume use estimates as temporary work standards. It is to be hoped that these temporary standards will be replaced as soon as possible with accurate time studies, work sampling, or predetermined time standards.

2.7. Vendor Quote Checks

Cost estimates are sometimes used to check vendor bid quotations on outsourced work. This estimate can be used to not only verify appropriate costs for outsourced work but also as a part of the total product cost estimate.

3. MEASURES OF ECONOMIC WANT

The task facing the engineer is to provide a fact or number that represents the economic want of the design. A ‘‘want’’ is a value exchanged between competing and selfish interests. The price a consumer is willing to pay for an item stocked on the grocery shelf, a contractor–owner agreement on the bid value of a building project, and the fiscal-year budget value for a weapons system that the U.S. Department of Defense proposes and Congress accepts are typical examples of wants exchange.

The measure of want for a product estimate is called *cost*. It usually means full cost, as all items contributing to the manufacture and purchase of material, subcontract materials, and services must be included.

4. REQUEST FOR ESTIMATE

It is not common practice for cost engineers to initiate a request for an estimate. The request is typically generated from sales and marketing sources. Another source is engineering design from a potential customer. A request for quotation (RFQ) or request for proposal (RFP) is received by engineering design or generated in sales or marketing. A customer usually does not communicate with cost engineers; usually external communication goes through another function before coming to the cost engineer. Therefore, a request for estimate (RFE) is generated internally after an RFQ, RFP, or production inquiry is received. The customary form is not intended here, but the image is a generated signal on the computer screen.

Information needed varies for each RFE, but there are general areas of information that every engineer seeks. Some of these are status of the design, quantity and production rate expectations, quality specifications, legal requirements including environmental impact, delivery requirements and location. Information necessary to the nature of the design and needed to make a complete and accurate estimate should be provided to the engineer. But it is the engineer’s responsibility to request proper information to develop the estimate. As in all decision making, the cost estimate can be no better than the quality and completeness of data used to create the estimate.

Sources of estimating information are both internal and external to the organization. If the product is going to be produced within the organization, the product estimation is probably internal. Project data, which usually involve capital types of designs, are typically external sources of information. Commercial data and published and private indexes are sources of external data.

Before starting an estimate, it is essential that analysis of elements of cost be understood. Analysis of labor, material, and overhead costs must be undertaken. Once again, the estimate will not be better than the quality and thoroughness of the analysis that precedes the estimating calculations. It is also vital that timely, up-to-date information be used.

The internal elements of cost details making up the estimate are primarily obtained from the accounting department. Cost accounting is the function that collects actual cost data on the various internal elements needed to develop the estimate. Listed below are the primary elements of a cost estimate and a brief description of each.

4.1. Direct Labor

Direct labor is the labor expended to add value to the product, sometimes described as the cost related to individuals who ‘‘touch’’ the product. Process operators, assemblers, and inspectors are included in this area.

4.2. Indirect Labor

Indirect labor supports direct labor. These people are essential to the operation of an organization, but they add no value to the product being produced. Material handlers, tool room employees, shipping and receiving employees, and maintenance people are some in this category.

4.3. Direct Materials

Direct materials consist of both manufactured and purchased components that are part of the product being produced.

4.4. Indirect Materials

Indirect materials are necessary to manufacture, test, and ship the product. Indirect materials are not part of the finished product. Sand used to build a sand-cast mold is an example of indirect material. There is a cost associated with indirect material, and in some situations the indirect material can be used over.

4.5. Overhead

This is an accounting term. Included in this category are salary and management costs. Overhead also includes all costs not covered in categories above. Elements such as machinery costs, shop and office supplies, and insurance are included in this area. Often in developing estimates, overhead is expressed as a percent of direct labor cost. For information on allocating overhead cost refer to Chapter 89.

4.6. General and Administrative

Many companies list general and administrative costs as part of overhead. Other companies list these elements separately. Usually G&A are added to the estimate in the form of a percentage factor developed in the organization. As part of this category, such items as sales commissions and top executive salaries might be included. These costs are provided by the accounting department and not by the cost engineer.

4.7. Profit

Profit that must be obtained from the product must be included in the cost estimate. This margin above production cost is provided by the marketing and accounting department and by top management.

5. PRELIMINARY AND DETAILED METHODS

There are many methods used to make estimates. They range from techniques that are quick and crude, or preliminary estimates, to the comprehensive and more accurate methods, detailed estimates. Regardless of the type of design, the methods used in estimating are similar.

Preliminary methods are used in the formative stages of design, are meant to be fast, and are not expected to be as accurate as those used to prepare detailed estimates. Detailed methods, at the other extreme, are used to set prices, make competitive bids, or allow organizational decisions to be made on what type of economic action to take. As might be expected, detailed methods are much more quantitative. Arbitrary and judgmental factors are suppressed though not fully eliminated.

Quantitative estimating is desirable because it tends to provide more accurate estimates than do nonquantitative methods. The quantitative area of estimating with the use of mathematical formulas is called parametric estimating, or sometimes statistical modeling. Although parametric estimating methods have been used for many years, they are becoming more favorable for estimating because many of the calculation techniques and estimating procedures have been developed into computer software. Several methods are discussed in some detail below. They are presented in order from preliminary to detailed methods. This order also goes from nonquantitative to quantitative, or parametric. When broadly defined, these methods can be used for a variety of designs.

5.1. Judgment and Conference Method

Judgment is an important part of any estimating process. In the absence of data, and when time is of the essence, guesstimates may be the only way to find some cost components of an estimate. The best-suited engineer for the task should be the person developing the cost estimate. This means that the engineer has qualified experience, common sense, and knowledge of the design. Time, cost, and/or quantities, with regard to minor or major line elements, are chosen using the engineer's experience. The engineer must remain objective in properly measuring all the present and future factors that could affect costs. When possible, judgmental estimating should be done collectively.

If time and resources allow, the nonquantitative consensus method of estimating, called conference estimating, can be used. The more pertinent knowledge that can be obtained from various sources about a particular detail of the estimate, the better the chance of the decision being correct. In addition to cost information, other information such as savings, potential, and marginal revenue, can be included. The conference method relies on the collective judgment of the differences between previously determined estimates and their associated relationships with the new designs being considered.

Conference estimating usually involves bringing together representatives from various departments conferring with the engineers in round table discussions. These groups of people determine costs for the features of the design for which they have been given responsibility. These conference estimates might be limited to specific areas such as direct labor, materials, and processing equipment. Overhead, distribution, selling price, and profit are later added, using the organization's various values and formulas. The engineers can add these indirect type costs to the estimate later if the various functional representatives helping develop estimate costs do not allow access to specific organization costing data.

The conference method is not typically analytical, and verifiable facts are usually lacking. When using the conference method, proper group managing techniques should be applied to ensure that the decisions are group decisions and are developed properly in the group setting.

5.2. Unit Method

The unit method, or some variation of it, is the most widely used preliminary estimating tool. This method may also be known as the order of magnitude method, lump sum method, module estimating, or flat rate method. Individuals often use the unit estimating method to estimate costs for their private needs. For example, for estimating what a new home may cost to build, the cost per square foot estimate can provide a good ballpark figure. If construction cost in a geographical area is valued at \$105/ft^2, then a family could calculate the rough cost of having a 2075 ft^2 house built = \$217,850 for the estimate of the cost of the house. Some other examples of unit estimates are:

- Cost of components per kilogram of casting
- Manufacturing cost per machine shop man-hour
- Chemical plant cost per barrel of oil capacity

All of these examples for estimating are per something. The information for these types of estimates can be obtained from the Internet, technical literature, government, banks, data files of cost engineering or accounting, and the service providers.

Contributing to the popularity of unit estimating techniques is their ease of use. Consider the manufacturing machining operation of turning. Using similar parts routings, the total time for several jobs and many part types for a lathe can be compiled. Taking averages of length of cut and time to cut, and knowing the direct labor charge, a cost per unit of length of cut can be determined.

5.3. Comparison

The comparison method is similar to the previously discussed unit method, the difference being that formal logic is applied. If an extremely difficult design is being estimated or part of the design has an unsolvable section, it is given an identifying name such as design *A*. A simpler design problem is then constructed so an estimate can be made. The simpler problem is given a title such as design *B*. The simpler design might be developed from creative and clever manipulations of the original, more difficult design. The simpler estimate may also be made up of relaxed technical constraints from the original problem. If known facts already exist about design *B*, the engineer can gain information useful in developing an estimate for *A*. The alternative design problem *B* must be selected to relate to the original design by the following inequality:

$$C_A(D_A) \leq C_B(D_B) \tag{1}$$

where C_A and C_B are the cost values of the estimate for designs *A* and *B*, respectively. Likewise, D_A and D_B are the designs for *A* and *B*. Obviously, estimates are better when *B* approximates *A* as closely as possible. The cost value C_A of the estimate should be something less than C_B. *A* conservative position may be taken initially, as can be construed from Eq. (1). It may be management's policy to estimate the cost a little high at the beginning. Once the detailed estimate of design *A* is thoroughly explored, it may be found that $C_A(D_A)$ is less than the original comparison estimate.

A comparison estimate can be developed where high and low bounds are placed on either side of the estimate for design *A*. If a similar design is known for, or approximately known for, a design *A*, the logic from above can be used to expand the comparison inequality to the following:

$$C_C(D_C) \leq C_A(D_A) \leq C_B(D_B) \tag{2}$$

The assumption is made that designs *B* and *C* satisfy the technical requirements and bond the economic estimate for *A*. In practice, many engineers use comparison logic to develop estimates. Standard cost plans can provide "similar-to" approaches, and analogy plans and computer retrieval schemes use this technique.

5.4. Factor Method

The factor method is an important method used for project estimating. Methods such as ratio, percentage, and parameter are approximately the same. The factor method is an extension of the unit method discussed previously. The unit cost estimating method was limited to a single factor for calculating overall costs. A natural extension of the unit method achieves improved accuracy by using separate factors for different cost items. For example, the estimate for the house construction from the unit method could be enhanced by added factors for certain types of heating and cooling units, tiled or wood floors, landscaping costs, etc. All the various unit costs can be summed and a more accurate estimate than using the unit method can be obtained. The equation takes the form

$$C = C_e + \sum f_i C_e (I + f_I) \tag{3}$$

where C = cost of design being evaluated
C_e = cost driver or subdesign used as base
f_i = factor for estimating instruments, structures, site clearing, etc.
f_I = factor for estimating indirect expense such as engineering, contractor's profit, and contingencies
$i = 1, 2, \ldots, n$ factor index

The general idea is that C_e is chosen as the cost driver. In the example, the house would be the cost driver. Where in a community the house is desirable to be built would be a contributing factor as would the specific design chosen and the amount of land the house would sit on. These factors can all be correlated and historical data, design parameters, and indexes can be referred to for factor estimate.

5.5. Cost- and Time-Estimating Relationships

Cost estimating relationships (CERs) and time-estimating relationships (TERs) are mathematical or graphical models that estimate cost or time. CERs and TERs are formulated to give estimates in either final or line item form for a cost estimate. Rule-of-thumb estimates are not to be confused with CERs and TERs, which are analytical.

5.5.1. Learning Curve

An excellent example of a CER and TER is the learning curve. There are two types of learning curves; unit cost and average cumulative cost, shown as

$$T_U \quad \text{or} \quad T_{AC} = KN^S \tag{4}$$

where T_U = cost or time value per unit of production, such as dollars, or man-hours required to produce the Nth unit
T_{AC} = average cumulative cost, time, or value of N units.
N = unit number, 1, 2, 3, . . . N
K = constant or estimate for $N = 1$, dimensions compatible with T
S = slope parameter of the improvement rate, equal to $\log L/\log 2$, where L = learning as percent of time (S is negative)

For example, if learning improvement requires only 85% of the previous time, then $S = \log 0.85/\log 2 = -0.2345$. The learning theory is based on the percentage of time or cost to build a quantity when doubled from the known time or cost. For example, assuming an 85% learning curve and assuming it takes 10 hours to build the first unit, then doubling that quantity, which is 2, it would require only 8.5 hours to build the second unit. Doubling the quantity again, it could be estimated that it would require only 7.225 hours to build the fourth unit, which is 85% of 8.5 hours.

Learning curves have either the unit or the cumulative average line as the linear line when drawn on a log–log graph format. In one presentation, on log–log paper, the cumulative average line is straight and the unit line curves under from unit 1 until 10 or 20 units. From then on, the unit line parallels the cumulative line. The other presentation form allows the unit line to be straight or linear when plotted on log–log paper, and the cumulative average line, though starting together with the unit line at unit 1, curves above the unit line, and at about the 10th to 20th unit the two curves will run parallel. Either way is acceptable, but it is important that the engineer understand and clarifys for the other readers of the estimate which presentation is being used. For estimating to build N number of units, the cumulative average time may be more meaningful. A company is more likely to want to know how much time it will take to build N units (N times the cumulative average time) rather than know how long it will take to build the *N*th unit. Table 1 shows the factors for an 85% learning curve. The two approaches are shown.

Different types of manufacturing have general learning curve slopes that are peculiar to them. Electronic manufacturers, ship builders, and so on have learning curve rates that generally apply to their area of production. Each company should have historical data for various types of products to obtain data for developing cost and time estimates for any size production demand. With knowledge of these slopes, or other learning experiences, the engineer determines the appropriate factor for the job being estimated.

Ostwald (1992) is one source of information on the development and application of learning curves. Many cost-estimating and work-measurement books give information on learning curve techniques.

5.5.2. Power Law and Sizing Model

Another application of the CER is the power law technique. The power law and sizing model is frequently used when estimates for equipment or components are given as a lump sum. This concept is concerned with designs that vary in size but are similar in type. An example might be estimating the design of a new and larger size electric motor. The cost to produce a 50 hp motor can be estimated from data for manufacturing a 25 hp motor, provided both are similar in design. Anyone familiar with manufacturing cost or the law of economy of scale would not necessarily expect the larger 50 hp motor to be twice the cost of the smaller 25 hp motor. The power law and sizing model can be expressed as

$$C = C_r \left(\frac{Q_C}{Q_r} \right)^m \tag{5}$$

TABLE 1 Sample Learning Theory table for 85% for Two Methods of Learning: Unit and Average

	Learning Table $\phi = 85\%$				
N	T_U or T'_a	T_c	T_a	T'_c	T'_u
1	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.8500	1.8500	0.9250	1.7000	0.7000
3	0.7729	2.6229	0.8743	2.3187	0.6187
4	0.7225	3.3454	0.8364	2.8900	0.5713
5	0.6857	4.0311	0.8062	3.4284	0.5384
6	0.6570	4.6881	0.7813	3.9419	0.5135
7	0.6337	5.3217	0.7602	4.4356	0.4937
8	0.6141	5.9358	0.7420	4.9130	0.4774
9	0.5974	6.5332	0.7259	5.3766	0.4636
10	0.5828	7.1161	0.7116	5.8282	0.4516
11	0.5699	7.6860	0.6987	6.2693	0.4411
12	0.5584	8.2444	0.6870	6.7012	0.4318
13	0.5480	8.7925	0.6763	7.1246	0.4235
14	0.5386	9.3311	0.6665	7.5405	0.4159
15	0.5300	9.8611	0.6574	7.9495	0.4090
16	0.5220	10.3831	0.6489	8.3521	0.4026
17	0.5146	10.8977	0.6410	8.7489	0.3968
18	0.5078	11.4055	0.6336	9.1402	0.3913
19	0.5014	11.9069	0.6267	9.5264	0.3863
20	0.4954	12.4023	0.6201	9.9079	0.3815
21	0.4898	12.8920	0.6139	10.2850	0.3771
22	0.4844	13.3765	0.6080	10.6579	0.3729
23	0.4794	13.8559	0.6024	11.0268	0.3689
24	0.4747	14.3306	0.5920	11.7536	0.3616
30	0.4505	17.0907	0.5697	13.5141	0.3462
40	0.4211	21.4252	0.5356	16.8435	0.3233
50	0.3996	25.5131	0.5103	19.9811	0.3066
100	0.3397	43.7539	0.4375	33.9680	0.2603
500	0.2329	151.4504	0.3029	116.4542	0.1783

where C = total cost sought for design size Q_c
C_r = known cost for a reference size Q_r
Q_C = design size expressed in engineering units
Q_r = reference design size expressed in engineering units
m = correlating exponent, $0 < m \leq 1$

An equation expressing unit cost C/Q_C can be used as

$$\frac{C}{Q_C} = \left(\frac{C_r}{Q_r}\right)\left(\frac{Q_C}{Q_r}\right)^{m-1} \tag{6}$$

As total cost varies as the mth power of capacity, C/Q_C will vary as the $(m - 1)$th power of the capacity ratio. When $m = 1$, a linear relationship exists and the law of economy of scale is ignored. For chemical processing equipment, for example, m is frequently approximately 0.6 and is sometimes called the "sixth-tenth model." The units of Q are required to be consistent since it enters only as a ratio. For situations such as inflation and deflation, the model can be altered to consider price change. A change factor C_I is placed in the equation along with index factors I_C and I_r, as follows:

$$C = C_r\left(\frac{Q_C}{Q_r}\right)^m\left(\frac{I_C}{I_r}\right) + C_1 \tag{7}$$

where C_I is the constant unassociated cost.

For estimating projects, a CER that can be used is $C = KQ^m$. K is a constant for a project that might be a processing plant, new computer system, or a highway bridge. The concept of economy of scale is derived from this CER, where capital cost per unit produced reduces as the plant size increases. The scale factor m is not constant for all project designs. General scale-up or scale-down by more than a factor of 10 should be avoided.

Multivariable CERs are also possible. For instance, where symbols have been previously supplied, an equation such as the one immediately following could be used:

$$C = KQ^mN^S \tag{8}$$

5.6. Probability and Statistical Techniques

There are a range of estimating methods that are based on applying probability and statistics. Cost is usually treated as a single-point value under conditions of uncertainty. Engineers, knowing the weakness of information and techniques applied, recognize that there are probable errors in the developed estimates. Knowing the cost determined while developing the estimate is a random variable; using probability to estimate is appropriate. In the realm of statistics, a random variable is a numerically valued function of the outcomes of a sample of data. Four probabilistic techniques are discussed below.

5.6.1. Expected Value

When an engineer can assign a probability estimate to elements of uncertainty, as represented by the economics of the design, the method of expected value can be applied. Nonnegative numerical weights associated with design elements are assigned in accordance with the likelihood of the event occurring. The probability of the occurrence must equal 1. The probabilities describe the likelihood of that the predicted event occurring. The method incorporates the effect of risk on potential outcomes by means of a weighted average. Each outcome of an alternative is multiplied by the probability that the outcome will occur. The sum of the products for each alternative becomes the expected value. It is mathematically stated as

$$C(i) = \sum_{i} P_i x_{ij} \tag{9}$$

where $C(i)$ = expected cost of the estimate for alternative I
p_j = probability that x takes on value x_j
x_{ij} = design event

The p_j represents the independent probabilities that their associative x_{ij} will occur with $\Sigma p_j = 1$. For example, it may be predicted that the cost of fuel for use in the design might be charged at the following discrete cost pattern: 20% probability that fuel will cost \$3.00/gallon, 30% probability that fuel will cost \$3.50/gallon, and 50% probability that fuel will cost \$3.75/gallon. Multiplying the discrete probability rates times their related fuel costs and summing gives the expected cost of \$3.525/gallon.

5.6.2. Percentile Method

Estimates reflecting uncertainty may be specified by three values representing the 10th, 50th, and 90th percentiles of an unstated probability distribution. The best value for an engineer to use is the 50th percentile. The 10^{th} percentile cost is the best-chance scenario and represents a 1-in-10 chance that the actual cost will be lower. The 90^{th} percentile cost is the worst-case scenario and represents a 1 in 10 chance the cost will be greater. An example is shown below.

	Percentile			Difference	
Item	10th	50th	90th	(50 − 10)	(90 − 50)
1	\$25	\$33	\$44	\$8	\$11
2	9	13	15	4	2
3	3	4	7	1	3

These costs can be assumed to combine independently, that is, a low cost with a mid-cost with another low cost. After estimating, the 10th and 90th percentiles are expressed as differences from the 50th (or mid-value). The next steps are to square the differences and sum.

	(50 − 10)	Mid-value	(90 − 50)
	\$64	\$33	\$121
	16	13	4
	1	4	9
Total	81	50	134
Square root	9		11.58

Total estimate at 10th percentile = \$50 − 9 = \$41
Total estimate at 50th percentile = \$50
Total estimate at 90th percentile = \$50 + 11.58 = \$61.58

Sensitivity analysis can be applied to the percentile method in a simple way, as shown next.

Item	Contribution to Low Uncertainty	Contribution to Total Cost	Contribution to High Uncertainty
1	79% (64/81 × 100)	66% (33/50 × 100)	90.3% (121/134 x 100)
2	19.8% (16/81 × 100)	26% (13/50 × 100)	3% (4/134 × 100)
3	1.2% (1/81 × 100)	8% (4/50 × 100)	6.7% (9/134 × 100)

This simple sensitivity analysis will identify items to be monitored for possible cost reduction.

5.6.3. PERT-Based Beta Distribution

Project evaluation review technique (PERT), was developed for use in predicting the expected duration of projects and monitoring the progress of the project's activities. It is based on using the most likely cost estimate, optimistic estimate (lowest cost) and pessimistic estimate (highest cost). These estimates are assumed to correspond to the beta distribution, which can be symmetrical or skewed left or right. Using the three estimates, a mean and a variance for the cost element can be calculated as

$$E(C_i) = \frac{L + 4M + H}{6} \tag{10}$$

$$\text{var}(C_i) = \left(\frac{(H - L)}{6}\right)^2 \tag{11}$$

where $E(C_i)$ = expected cost for element I
L = lowest cost, dollars (optimistic)
M = modal value of cost distribution, dollars (most likely cost)
H = highest cost, dollars (pessimistic)

If several elements are estimated using this method, and if their costs are assumed to be independent of each other and are summed together, the distribution of the total cost is approximately normal. This follows from the central limit theorem. Figure 1 illustrates the use of the PERT method. The example shows how to find the contingency effects for a project design. Several elements must be combined when making the estimate to satisfy the conditions of the central limit theorem.

$$E(C_T) = E(C_1) + E(C_2) + \dots + E(C_n) \tag{12}$$

$$\mathrm{var}(C_T) = \mathrm{var}(C_1) + \mathrm{var}(C_2) + \dots + \mathrm{var}(C_n) \tag{13}$$

$E(C_T)$ represents the expected total cost in dollars, and $\mathrm{var}(C_T)$ is the variance of total cost in dollars.

5.6.4. Computer Simulation

Simulation techniques are more acceptable as tools for engineers developing costs of projects and systems. As computer simulation packages become more user friendly and computers' memory grow larger and computation speeds become faster, simulation as a tool for estimating is becoming more popular.

Simulation is defined as the manipulation and observation of a synthetic (logical and mathematical) model representative of a real design, that, for technical or economic reasons, is not susceptible to direct experimentation. The simulation model is developed to represent the essential characteristics of the real system, with many minor details omitted. Product estimates are not suited generally for simulation techniques, although simulation could be applied to determine costs and times for manufacturing systems being estimated to produce the product.

5.7. Standard Data

Standard data are defined as standard time values for all manual work in an estimate. Standard data provide the opportunity to be consistent when developing an estimate.

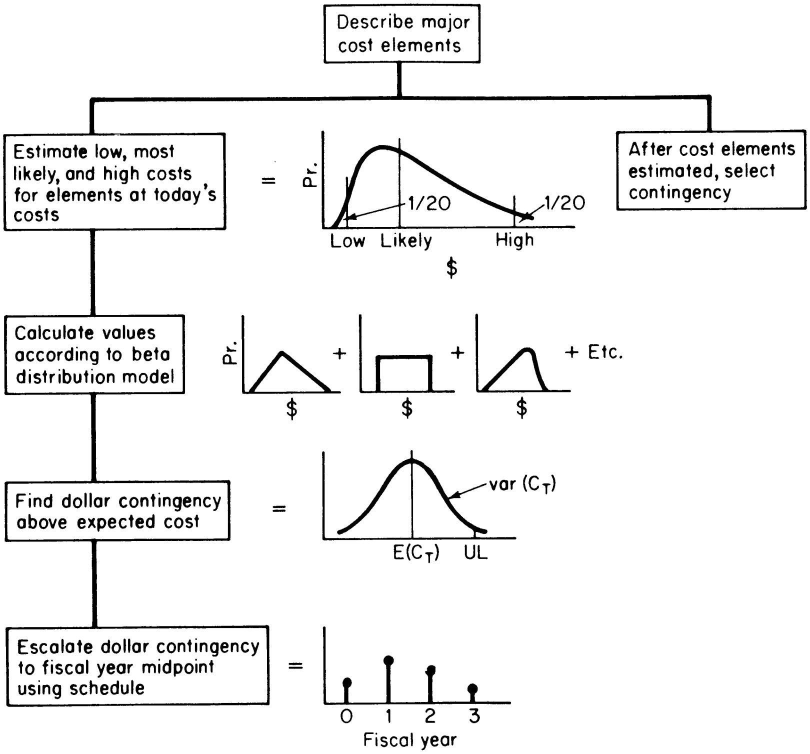

Describe major cost elements
Estimate low, most likely, and high costs for elements at today's costs
=
Pr.
1/20
1/20
Low
Likely
High
$
After cost elements estimated, select contingency
Calculate values according to beta distribution model
Pr.
+
$
+
$
+ Etc.
$
Find dollar contingency above expected cost
=
var (C_T)
E(C_T)
UL
Escalate dollar contingency to fiscal year midpoint using schedule
=
0
1
2
3
Fiscal year

Figure 1 Flowchart of PERT-Based Estimating.

The most accurate way to estimate direct labor cost is with standard time data developed from one of the formal time-measurement techniques. (see Chapters 53 and 54). It is not the original time measurements that the engineer desires, but a set of engineering performance data or standard time data that are needed to make the estimate. Frequently, raw data or times for specific methods are incorrect because methods are altered, equipment is replaced, environmental conditions change, and so on. Industrial engineers may use regression analysis or other techniques to extend these raw data into a more usable form, such as standard data. It is easier to calculate standard time data for processes such as machining operations than fabrication processes. Most of the work content in metal removing on a machine tool is fixed machine time. In fabrication, much of the time may be manual and subject to variation depending on individuals performing work activities.

Standard time data may be divided into preliminary or detailed data. As with preliminary and detailed estimates, the engineer is more likely to be interested in preliminary standard data early in estimating, and later detailed data will become more important. Standard time data are ordinarily determined from any of the various methods of observing work.

5.7.1. Time Study

Some companies develop standard data from stopwatch time studies. Time studies are used to establish rates of production. When time studies are used to establish standard data, care must be taken in defining element content so that work content can be isolated.

5.7.2. Predetermined Time Standards

Usually, one of the commercial systems, such as MTM or MOST, will be used, but sometimes companies will develop their own system. The main advantage of predetermined time standards is the consistency of the data. The major disadvantage is the amount of time necessary to develop the data. The major commercial systems are computerized, and this allows for much faster development time.

5.7.3. Work Sampling

This technique of work measurement used the fundamentals of probability and statistics to develop work standards by making random observations on jobs over a specific period of time. This method is widely used in white-collar environments. It is a desirable technique for studying team activities and long-duration activities.

5.8. Historical Data

Past history or actual performance on jobs produced can be used to develop standard data. A disadvantage of this technique is that it rarely considers the best method of organizing work. This method is popular in smaller companies that do not have the resources to use the other work-measurement methods to develop standard time data.

In manufacturing, time study and predetermined motion time data are the major sources for obtaining standard data. In construction and white-collar environments, work sampling and man-hour reports are the principal means of information. Likewise, in certain government agencies, such as the post office and the military depots, work sampling and man-hour methods are used.

Computer databases allow for the easy and readily accessible time standard data. Often charts and tables are used in hard copy, especially when the engineer is familiar with such estimating tools. Charts and production information can be found in Ostwald (1992). When developed properly, standard time data are considered to be accurate and relatively inexpensive for labor estimating.

6. LABOR ANALYSIS

Labor constitutes one of the most important items of operation designs. Labor has received intensive study, and many recording, measuring, and controlling schemes have been developed in an effort to manage it. Labor can be classified in a number of ways, including direct–indirect, recurring–nonrecurring, designated–nondesignated, exempt–nonexempt, wage–salary, blue collar–management, and union–nonunion. Other ways in which to classify labor are according to social, political, and educational divisions and type of work. Payment of wages may be based upon attendance or performance. For cost–estimating operation designs, the direct–indirect classification is the most appropriate.

For operation designs, there is an unquestioned dependence upon the simple qualitative formula

$$\text{Labor cost} = \text{time} \times \text{wage} \tag{14}$$

The selection of time matches the requirements of the operation design. Time is expressed relative to a unit of measure, which is denoted by terms such as *piece, bag, bundle, container, 100 units,* or *1000 board feet*. The usual ways to measure labor are by time study, predetermined motion–time systems work sampling, or man-hour reports.

Job tickets, especially for smaller organizations, are analyzed and allocated to units of work. For instance, a job ticket may state, "136 units turned of part number 8641" and list "6 person-hour." Simple analysis would show 0.044 hr/unit. The engineer would use 0.044 hr the next time this part was run. Although hardly accurate because of the nature of historical work reports, man-hour reports are used because of their simplicity. Man-hour estimating data are especially popular in construction work.

Direct observation and measurement of labor are of little use to the engineer except for guesstimates of similar work or reruns of the same work. Although the cost engineer may not be directly involved with the measurement of labor, he or she does depend on work measurement. The engineer is satisfied if such labor measurements are objective, as far as that is possible, and is willing to use the information, provided that engineering techniques were used in the determination of time. Although the time measurements are of value, it is immensely more important that work-measurement data be transformed into information that can be applied prior to the time of the operation design.

The time measurements are more valuable when expressed as standard time data (see Chapter 54) and presented in a table or computer format. The estimating data may be described in terms of elements, which are the subwork descriptors of operations, or be expressed as time-estimating relationships (TERs) for operations.

Standard data expressed at the predetermined motion–time level are too detailed for much cost-estimating work. But a typical TER is satisfactory for much cost-estimating work. A typical TER for a drill press operation of sheet metal parts is for setup 0.2 + 0.05/tool hr and for run time 0.015 + 0.003/tool + 0.001/hole hour per unit. Thus, if a sheet metal part requires two different countersinks for 22 holes, setup would be 0.3 hr and run time 0.043 hr/unit.

In some situations the estimate of time may be done from a guesstimate and be unrelated to measured, referenced, and analyzed data. A guesstimate is based on the engineer's observational experience. There are circumstances where these judgmental numbers are unavoidable.

The second part of Eq. (14), wage, is defined in the context of the operation design that is being estimated. The operation design may be for one worker and one machine, for a crew with one machine, or for a crew with several machines or processes. In the simplest case, one on one, the job description and job design are specifications available to the engineer. The number used for the wage corresponds to the time period of work and is money out of pocket. Regression methods, labor contract, and personnel planning are sources for wage trend information.

The practice of what is included in the wage amount is coordinated with the finding of the overhead. Fringe additions could include effects of paid holidays and vacations, health insurance and retirement benefits, Federal Insurance Contributions Act (FICA) benefits, workers' compensation, bonuses, gifts, uniforms, special benefits, profit sharing costs, education, and so on.

7. MATERIAL ANALYSIS

The term *direct materials* includes raw materials, purchased parts, standard commercial items and interdivisional transfers, and subcontracted items required for the design. Direct material cost is the cost of material used in the design. The cost should be significant enough to warrant the cost of estimating it as a direct cost. Some material, by virtue of the difficulty of computation and estimating, may be classified as either indirect or direct costs. The latter estimates are preferred. Paint material of irregularly shaped objects is an example of material that can be classified either way.

The engineer begins by calculating the final exact quantity or shape required for a design. To this quantity, losses for scrap, waste, and shrinkage are added. The general model for cost of direct material is

$$S_a = S_t(I + L_1 + L_2 + L_3) \tag{15}$$

where: S_a = actual shape in units of area, length, mass, volume, count, etc.
L_1 = loss due to scrap, decimal
L_2 = loss due to waste, decimal
L_3 = loss due to shrinkage, decimal

Scrap is material that is lost because of human mistakes, whereas waste is necessary because of the design. Shrinkage losses are due to theft or physical law deterioration. In estimating of foodstuffs, if direct material is not processed at the appropriate time or if it is mishandled, shrinkage of the quantity will result. It is required that these three losses be estimated and that their percentages be added to the theoretical finished requirement.

An example of material estimating is given by the 12 oz. (355 ml) beverage can, which is composed of the body, top, and pull ring. The container body is blanked from 3004 H19 aluminum coils, with the layout given in the Figure 2. An intermediate cup is formed, without any significant change in thickness. The cup is drawn in a horizontal drawing machine, and metal is squeezed to side-wall thickness of 0.0055 in. (0.140 mm), while bottom thickness remains unchanged. The can is trimmed to final height to give an even edge for later rolling to the lid. Various mensuration formulas are used to find first the volume and weight of the object. For a popular soda drink can, there is about 25% waste. These calculations eventually relate to the amount of coil aluminum stock.

8. NEED FOR ACCOUNTING DATA

Cost accounting has always been important to the performance of diverse estimating functions. As colleagues in the gathering, analysis, and reporting of business data, accountants provide overhead rates, standard costs, and budgeting data. The engineer reciprocates with labor and material estimates for the several designs. In many situations the estimate can serve as a mini-profit-and-loss statement for special products. Thus, there is interdependence between these two professions. The engineer is less interested in balance sheets, profit-and-loss statements, and the intimate details of the structure of accounts. Overhead rates are vital for the estimating functions, however, since the engineer may apply these rates in the estimate.

By definition, overhead methods would include the following:

- Indicating whether the rate includes fixed costs, as in absorption costing, or not, as in direct costing.

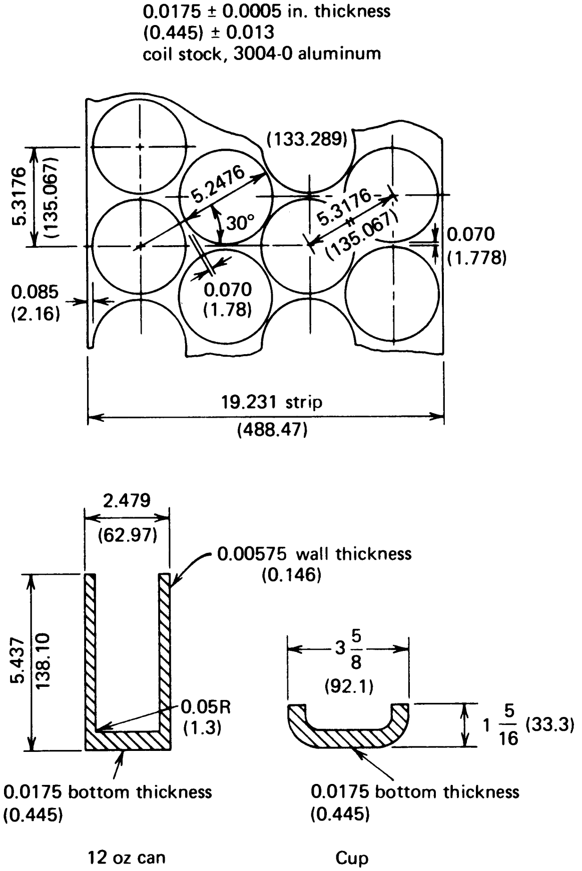

0.0175 ± 0.0005 in. thickness
(0.445) ± 0.013
coil stock, 3004-0 aluminum
(133.289)
5.3176
(135.067)
5.2476
30°
5.3176
(135.067)
0.070
(1.778)
0.085
(2.16)
0.070
(1.78)
19.231 strip
(488.47)
2.479
(62.97)
0.00575 wall thickness
(0.146)
5.437
138.10
0.05R
(1.3)
0.0175 bottom thickness
(0.445)
$3\frac{5}{8}$
(92.1)
$1\frac{5}{16}$ (33.3)
0.0175 bottom thickness
(0.445)
12 oz can
Cup

Figure 2 Simple Design for a Beverage Container, the Common 12 oz Can.

- The base used to distribute overhead, such as direct labor dollars, direct labor hours, or machine hours.
- The scope of the application of the rate, whether it is for the plant, cost center, machine, or design.
- Whether the rate applies to all designs (such as product lines) or to one line of the design.

9. FORECASTING

Many forecasting techniques have been developed to handle a variety of problems. Each has its special advantage, and care is necessary in choosing techniques for cost estimating. Selection of a method depends on the context of the forecast, availability of historical data, accuracy desired, time period to be forecast, and value to the company. The engineer should adopt a technique that makes the best use of the data. He or she should initially use the simplest technique and not expect more from the advanced technique than is justified.

For estimating requirements, we are concerned with data about labor and material overhead and their quantities and cost. The forecast should reflect those values under the proposed actions of the company and environment. It is necessary to recall that forecasting is a future prediction about line elements of the estimate. Forecasts should not deal in overall or grand average cost, time, and quantities, but should be matched to line items required by the pro forma estimate. Forecasting is not estimating, as the term is used here, since forecasting takes data and frames it in a new picture, and judgment is suppressed as much as possible.

10. INDEXES

Cost-estimating indexes are useful for a variety of purposes. Principally, they are multipliers to update an old cost to a new cost. Some examples of indexes are material, labor, material and labor, regional effects, and design type. C is the reference cost associated with a reference index I_r. The cost C to be determined is linked in terms of time to the index I.

$$C = C_r \left(\frac{I}{I_r} \right) \tag{16}$$

Indexes are prepared and published by the government, private industry, banks, consultants, associations, and trade magazines. It is important to determine one's own indexes, especially for materials or labor not charted by other groups.

A cost index is meaningful only in that it expresses a change in price level between two specific times. A cost index for steel in the year 2008 alone is meaningless. An index for material A has no relationship to the index for material B. Similarly, the cost indexes for material A in two geographical areas may not be directly comparable.

To compute a price index for a single material, a series of prices must be gathered covering a period for a specific quantity and quality of the material. Index numbers are usually computed on a periodic basis. The federal government gathers data and calculates and divulges index numbers for periods as short as a month. The prices gathered for the material may be average for the period (month, quarter, half year, or year) or may be a single observed value as found on invoice records for one purchase.

Assume that the following prices have been collected for a standardized unit of silicon laser glass material:

Period	0	1	2	3	4	5
Price	\$43.75	\$44.25	\$45.00	\$46.10	\$47.15	\$49.25
Index, %	94.9	96.0	97.6	100.0	102.3	106.8

Index numbers are computed by relating each period price to one of the prices that has been selected as the base. If period 3 is the benchmark period, because the index is 100 or 1, period 2 price divided by period 3 price = \$45.00/\$46.10 = \$0.976. When period 3 price is expressed as 100.0, period 2 price can be expressed as 97.6.

Movements of indexes from one period to another are expressed as percent changes rather than as changes in index points. Period 6, 7, and so on can be projected, and if a reference price is known, a future price can be calculated. For instance, if C_2 = \$3700, and if we want to know C_7, then projecting C_7 = \$3700 × (110.0/97.6) = \$4170.

Assume that a product called “10-cm disk aperture laser amplifier” is selected for a composite index. Although the 10-cm disk amplifier was produced only during period 0, tracing of selected cost items has continued. To worry about all amplifier components is too involved, so major items were picked for individual tracking, and spot prices have been gathered for 4 years. The quantity of each of the five materials is in proportion to the initial one-time cost of the material to the total cost. Some materials have declined in price, whereas others have increased. Prices for each material have been gathered (or imputed for periods where no information was available) and are shown in Table 2.

The prices conform to quantity and quality specifications. With the index at 100 for benchmark period 0, the following indexes are calculated as 94.1, 89.6, and 93.3. If the unit cost is $43,650 during period 0, the estimated cost is equal to $37,953 at period 5.

One may argue that cost facts, materials, quantities, and qualities are not consistent as given in Table 2. Indeed, if technology is active, a decline in the cost and index is possible. Indexes should reflect basic price movements alone. Index creep results from changes in quality, quantity, and the mix of materials or labor. Table 2 is an example of a product index. The components in this case are selected on the basis of their contribution to the product value. Selection of components could be 100%, random, or stratified in accordance with the needs of cost estimating. Quantity is determined proportional to the design requirements. Specifications provided by engineering are used to fix quality characteristics. Product indexes can be maintained by noting the changes when they occur, inputting all previous data, and recalculating the previous year’s indexes. Every so often it may be necessary to reset the benchmark year whenever delicate effects are influencing the index and are not being removed.

11. OPERATIONS ESTIMATING FOR MANUFACTURING

The operations sheet is fundamental to manufacturing estimating. It is also called a route sheet, traveler, or planner. There are many styles, and each plant has its own form. The purpose of the operations sheet, however, is the same:

- To select the machine, process, or bench that is necessary for converting the material into other forms
- To provide a description of the operations and tools
- To indicate the time for the operation

The order of the operations is special too, as this sequence indicates the various steps in the manufacturing conversion.

Each operations sheet has a title block indicating the material part number, date, quantity, engineer, and other information that may be essential to the company. Following the writing of this information on the form or its entry into the computer, the instructions to the plant are provided.

TABLE 2 Simple Calculation of Index for Composite of Several Materials

			Period				
Material	Quantity	Quality	0	1	2	3	4
Laser glass	3- to 10-cm disk	Silicate	$26,117	$24,027	$22,345	$21,228	$22,713
Stainless steel tunings	18 kg	AISI 304	1,913	2,008	2,129	2,278	2,460
Aluminum extrusion	4 kg	3004	418	426	439	456	479
Fittings	3 kg	Mil Std 713	637	643	656	657	689
Harness cable	4 braid	Mil Std 503	2,103	2,124	2,134	2,305	2,466
Annular glass tubing	4 m	Tempered	$ 4,317	4,187	4,103	4,185	4,311
Total sample cost, $	12 m	3/16 in. wall PPG # 27	$35,505	$33,416	$31,808	$31,112	$33,122
Index, %			100.0	94.1	89.6	87.6	93.3

Suppose that we want to machine an aluminum casting that is on material consignment (meaning that the material is being supplied at no transfer cost). The casting is called SOHO and the part number is unknown. This casting is a consignment material and is part of a larger product. This casting will have a bored hole enlarged and deburred and it will be packaged in a carton. Your parts and assemblies are more complicated than this, but we only want to identify the process of estimating. A typical and simple operations sheet would appear as shown in Figure 3.

11.1. Preparing the Operations Sheet

The operations sheet (1) begins in the upper left-hand corner of Figure 3. The final product name (2) is often given along with the assembled product (3). The operations sheet is for a specific part name (4), which in this case is called SOHO. A part number (5) can be identified and listed if available. The part number and name are removed from the design and repeated on the operations sheet title block. The engineer will enter lot quantity (7) and material specification (8). Knowing the final amount of material required by the design, the engineer will add material to cover losses for scrap, waste, and shrinkage and multiply by the cost per pound rate of the material. Material cost, using the formula given earlier, is used to enter the value. A unit material cost is required for entry (9).

The sequence of the operation number and selection of the machine, process, or bench are made to manufacture the part. These are required at circle 10 and are shown specifically at (18) and (19). A complete operations sheet will show this column. Even though they are vital in operations planning, their importance is less in detailed estimating once that operation has been selected.

The column titled ‘‘Table’’ (12) corresponds to the equipment number for the plant. For example, ‘‘Table 7-4’’ refers to the ram milling machine class. The operations sheet column titled ‘‘Process Description’’ (13) are instructions to the shop that they will follow in making the part or subassembly. The instructions for operation number 10 to the shop are listed in Figure 3. The process description gives a listing of the elements that are pertinent to estimating the operation.

The ‘‘Process Description’’ lists the elements of the operation. These correspond to the element description listed in the estimating tables. It is an elaboration of the description given in the operations planning sheet described earlier; there is no basic difference, except that the number of lines or elements are greater for estimating than for planning. The ‘‘Process Description’’ may also indicate additional information such as length of cut, tooling used, type of NC manuscript order, and so forth.

The ‘‘Table Time’’ column (14) is a listing of the values removed from the estimating tables. These time values are posted in this column.

The ‘‘Table Time’’ column identifies the estimating table and element number. For example, if a sheet metal operation of braking were necessary, the number 3.6 would be first posted. Similarly, for a drill press operation, 9.3 would be written on the row corresponding to the machine selection. Notice, for any estimating table, that clusters of elements have a number too, starting with 1, 2, and so on. These clusters are generally related. The element ‘‘handle’’ may have many possibilities and be listed as element 1. Following the machine number, we list the element number, and it is preceded by a dash. For example, 3.6-1 is a power press brake element called ‘‘brake.’’ Also, 9.3-2 is a cluster of elements for ‘‘clamp and unclamp’’ for the upright drilling machine.

The adjustment factor (15) column operates upon the time column. Once adjusted, the time is either entered into the cycle minutes (16) or the setup hours (17) column. There is more discussion on the adjustment factor column later.

The columns titled ‘‘Cycle Minutes’’ and ‘‘Setup Hours’’ are very important, and the instructions that follow describe the methods and selection of the elements and time that are necessary to manufacture the part for that operation.

The sequence number (18) of the operation is given in the left-hand column along with the equipment (19) necessary for the operation.

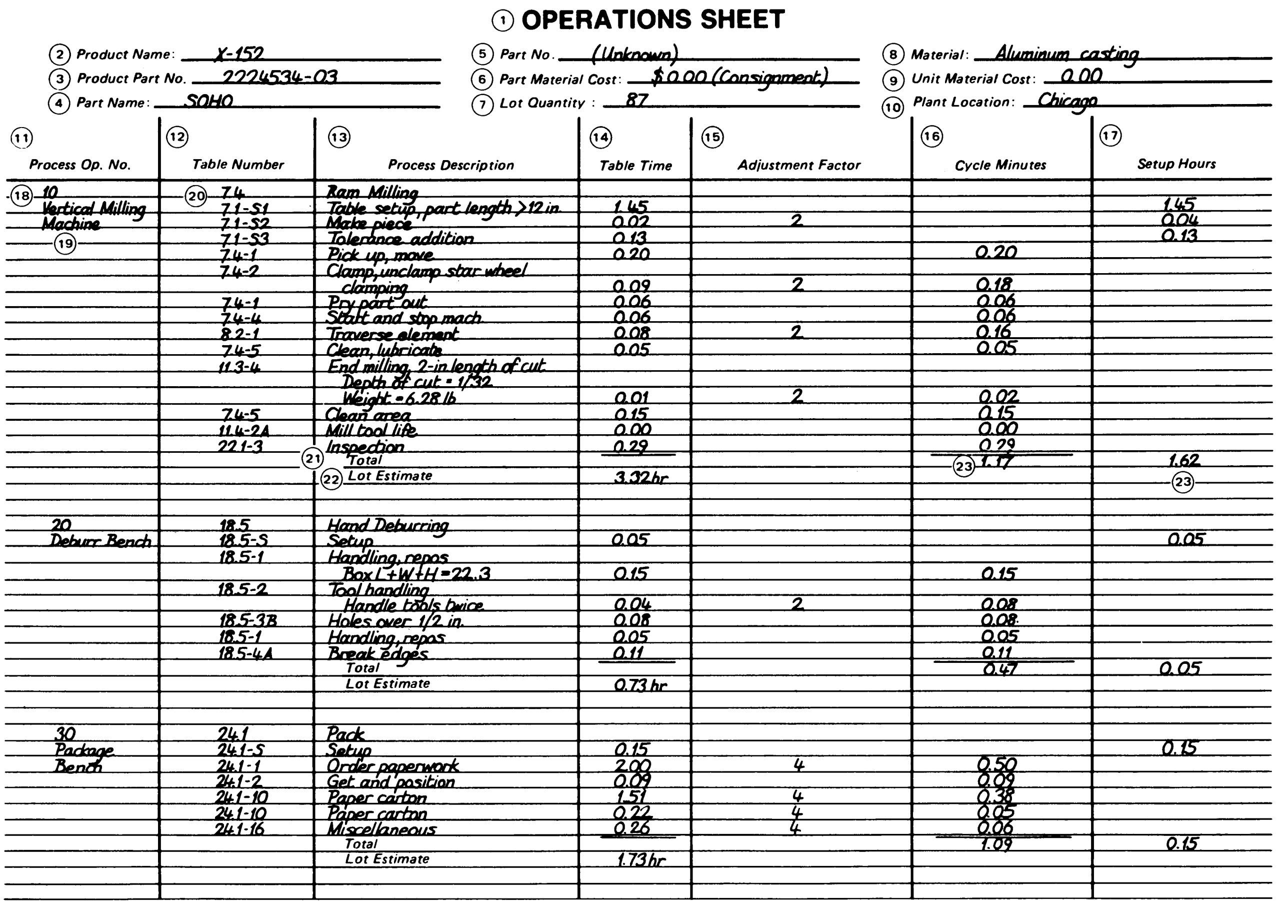
The total (21) of the cycle minutes column and the total (22) of the setup hours column are summed. The lot estimate is calculated and presented (23) with the dimensions in hours. Lot time is calculated as follows:

The total of ‘‘Lot Estimate’’ is a computation that is shown on the operations sheet. The calculation is made using the setup, unit estimate, and lot quantity.

This operations sheet can be altered to consider simple assemblies or complicated products, but the approach remains the same. The purpose of estimating is to provide time or cost for the direct labor or material component of the product. The preparation of the operations sheet is important for the finding of part operational costs. Notice that the part cost is the sum of the operational costs, and this fact allows us to concentrate on the important steps that are necessary for estimating operations. Once the operational sequence, the selection of the machine, process, or bench, and a basic description of the work have been roughed out, cost estimating begins.

11.2. Setup Hours Column

Setup includes work to prepare the machine, process, or bench for product parts or the cycle. Starting with the machine, process, or bench in a neutral condition, setup includes punch in/out, paperwork,



① OPERATIONS SHEET

② Product Name: X-152
③ Product Part No. 2224534-03
④ Part Name: SOHO
⑤ Part No. (Unknown)
⑥ Part Material Cost: $0.00 (Consignment)
⑦ Lot Quantity: 87
⑧ Material: Aluminum casting
⑨ Unit Material Cost: 0.00
⑩ Plant Location: Chicago

⑪ Process Op. No.	⑫ Table Number	⑬ Process Description	⑭ Table Time	⑮ Adjustment Factor	⑯ Cycle Minutes	⑰ Setup Hours
⑱ 10	⑳ 74	Ram Milling				
Vertical Milling	71-S1	Table setup, part length >12 in	1.45			1.45
Machine	71-S2	Make piece	0.02	2		0.04
⑲	71-S3	Tolerance addition	0.13			0.13
	74-1	Pick up, move	0.20		0.20	
	74-2	Clamp, unclamp star wheel				
		clamping	0.09	2	0.18	
	74-1	Pry part out	0.06		0.06	
	74-4	Start and stop mach.	0.06		0.06	
	82-1	Traverse element	0.08	2	0.16	
	74-5	Clean, lubricate	0.05		0.05	
	113-4	End milling, 2-in length of cut				
		Depth of cut = 1/32				
		Weight = 6.28 lb	0.01	2	0.02	
	74-5	Clean area	0.15		0.15	
	114-2A	Mill tool life	0.00		0.00	
	221-3	Inspection	0.29		0.29	
		㉑ Total			㉓ 1.17	1.62
		㉒ Lot Estimate	3.32 hr			㉓
20	18.5	Hand Deburring				
Deburr Bench	18.5-S	Setup	0.05			0.05
	18.5-1	Handling, repos				
		Box L+W+H = 22.3	0.15		0.15	
	18.5-2	Tool handling				
		Handle tools twice	0.04	2	0.08	
	18.5-3B	Holes over 1/2 in.	0.08		0.08	
	18.5-1	Handling, repos	0.05		0.05	
	18.5-4A	Break edges	0.11		0.11	
		Total			0.47	0.05
		Lot Estimate	0.73 hr			
30	241	Pack				
Package	241-S	Setup	0.15			0.15
Bench	241-1	Order paperwork	2.00	4	0.50	
	241-2	Get and position	0.09		0.09	
	241-10	Paper carton	1.51	4	0.38	
	241-10	Paper carton	0.22	4	0.05	
	241-16	Miscellaneous	0.26	4	0.06	
		Total			1.09	0.15
		Lot Estimate	1.73 hr			

Figure 3 Example of Process Sheet with Balloons for Instruction Sequence.

obtaining tools, positioning unprocessed materials nearby, adjusting, and inspecting. It also includes return tooling, cleanup, and tear-down of the machine, process, or bench to a neutral condition ready for the next job. Unless otherwise specified, the setup does not include the time to make parts or perform the repetitive cycle. If scrap is anticipated as a consequence of setup, the engineer may optionally increase the time allotment for unproductive material.

Setup estimating is necessary for job shops and companies whose parts or products have small to moderate quantity production. As production quantity increases, the effect of the setup value lessens its prorated unit importance, although its absolute value remains unchanged. Setup values may not be estimated for some very large quantity estimating. In these instances, setup is handled through overhead practices. Our recommendation is to estimate setup and to allocate it to the operation because it is a more accurate practice than costing by overhead methods. This recommendation applies equally to companies manufacturing their own parts or products and vendors bidding for contract work.

Some operations may not require setup. Flexible manufacturing systems, continuous production, or combined operations may not require setup time. Nonetheless, even a modest quantity may be appropriate in these circumstances. Discussion of the details regarding setup is given for each machine, process, or bench.

11.3. Cycle Minutes Column

Cycle time or run time is the work needed to complete one unit after the setup work is concluded. It does not include any element involved in setup. Besides finding a value for the operational setup, the engineer finds a unit estimate for the work from the listed elements, which is called estimating minutes.

The term *estimating minutes* implies a national norm for trained workers. These times include allowances, in addition to the work time, that take into account personal requirements, fatigue where work effort may be excessive due to job conditions and environment, and legitimate delays for operation-related interruptions. Since the allowances are included in the time for the described elements, and therefore part of the allowed time for several elements and hence several or many operations, then the allowed time is fair. The concept of fairness implies that a worker can generally perform the work throughout the day.

12. PRODUCT ESTIMATING

The cost summary (24) is for the part SOHO and the header information is repeated (see Figure 4). It is an important principle that estimating for manufacturing requires that each operation be estimated. Each operation (18) is identified and these correspond to the basic operations sheet. The table number (12) identifies the basic data set for that operation. Balloon (25) specifies the description of the machine, process, or bench necessary to perform the operation. Lot hours (28) are transferred from the operations sheet. These lot hours differentiate between quantity. For low quantity the setup becomes more important, while the cycle minute influences the lot hours if the quantity is large. Whether the part is for small or large quantity, the method is acceptable. The system is acceptable even with very large quantities that mechanization would require.

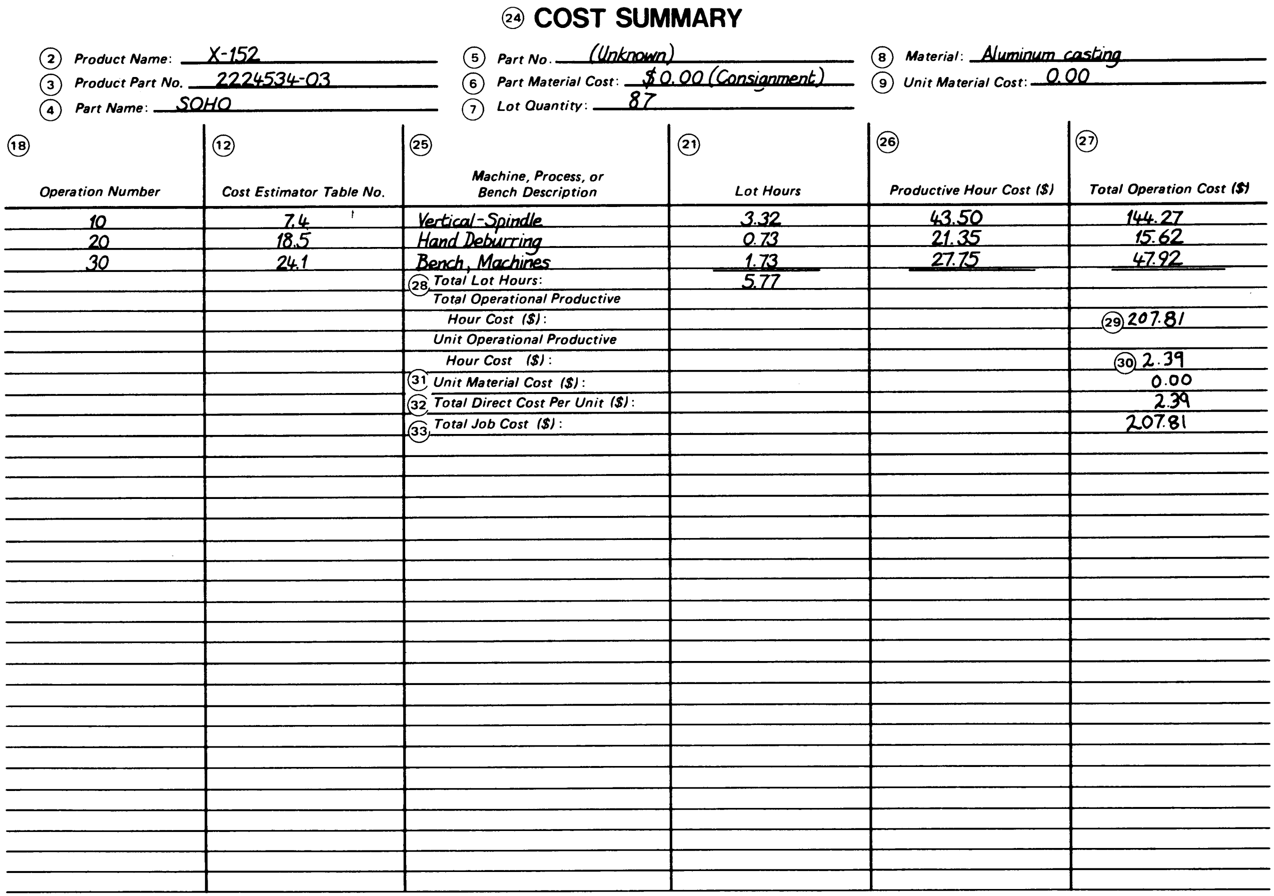
Productive hour cost (PHC) is entered as balloon (26). These company values are the costs for the labor and the machine. Overhead is included for this case. These company values are calculated by accounting.

The lot hours are multiplied by the PHC and given the total operational cost shown as column 27. For example, $3.32 \times 43.50 = \$144.27$. The PHC includes the cost of overhead, and the method can include absorption- or activity-based methods. The sum of the operation cost is given by the total operational productive hour cost, identified as column 29. The value of $207.81, when divided by the lot quantity of 87 (7), gives the unit operational productive hour cost of $2.39. In this case, this value is for labor and the machine process cost. Because the material is a consignment between the buyer and the manufacturer, no cost is assessed for unit material cost identified by column 31. If a unit cost exists, it is entered here. The sum of the material and the unit operational productive hour cost gives the total direct cost per unit (32). This value is multiplied by the lot quantity and the total job cost is entered as column 33.

This cost summary is used to provide information to the bill of material cost summary, which is the means of collecting all costs to obtain the full cost. Except in the case of a single part, the bill of materials is a vital and important document. For those manufacturers who only produce a single part, the cost summary is adequate since it provides the total job cost.

13. BILL OF MATERIAL EXPLOSION FOR PRODUCT ESTIMATES

The estimating of labor and material cost and its extension by overhead calculations will lead to the quantity known as full cost. This in turn will be increased for profit to give price. Before that routine



㉔ COST SUMMARY

② Product Name: X-152
③ Product Part No. 2224534-03
④ Part Name: SOHO
⑤ Part No. (Unknown)
⑥ Part Material Cost: $0.00 (Consignment)
⑦ Lot Quantity: 87
⑧ Material: Aluminum casting
⑨ Unit Material Cost: 0.00

⑱ Operation Number	⑫ Cost Estimator Table No.	㉕ Machine, Process, or Bench Description	㉑ Lot Hours	㉖ Productive Hour Cost ($)	㉗ Total Operation Cost ($)
10	7.4	Vertical-Spindle	3.32	43.50	144.27
20	18.5	Hand Deburring	0.73	21.35	15.62
30	24.1	Bench, Machines	1.73	27.75	47.92
		㉘ Total Lot Hours:	5.77		
		Total Operational Productive Hour Cost ($):			㉙ 207.81
		Unit Operational Productive Hour Cost ($):			㉚ 2.39
		㉛ Unit Material Cost ($):			0.00
		㉜ Total Direct Cost Per Unit ($):			2.39
		㉝ Total Job Cost ($):			207.81

Figure 4 Example of Product Cost Summary.

is executed, it is necessary to find the total bill of material cost for several or many parts, subassemblies, and major assemblies. The bill of material explosion is unnecessary if the manufacturer only sells single-item parts, as the cost estimate serves as the principal summary document for price setting. But in the case of several or many parts and assemblies, it is necessary to organize the cost estimates effectively. A bill of material handles a scheme of this organization.

14. COMPUTERS AND ESTIMATING

Very few cost estimates are done without the aid of a computer. At least a microprocessor is used for word processing, spreadsheet calculations, database queries, and small, engineer-developed programs. At the other end of the spectrum, companies have developed their own in-house estimating software systems. Some companies with cost-estimating expertise have developed commercial cost-estimating packages for organizations wishing to use turnkey-type estimating packages.

Computer estimates are very consistent. Because of this consistency, they have an advantage of being able to be made with more accuracy. Estimates can be adjusted higher or lower as needed, or observed from previous cost estimates.

More detail concerning an estimate can be done because of computers. Details that might be tedious and time consuming if done long-hand can be done quickly and accurately on a computer. Work standard data and machining data can be accessed and inserted into an estimate easily. Also, level of detail relating to risk can easily be determined with the aid of a computer.

Cost-estimating software can provide refinements that would not be possible for an engineer to handle. For example, tool types, tool materials, material conditions, and so on. can easily and quickly be factored into cost, making the estimate more accurate and reliable.

15. CONCLUSION

Of the many paper or paperless documents a manufacturer will prepare, few are as important as the product estimate. This is the principal method for pricing that the firm will use. If the cost estimate is such that a profit will ensue, the enterprise continues the development of bringing the product to market. If the estimate gives an indication that a profit is unfeasible in the competitive market, the firm will cancel the product development or return the design to engineering for reconsideration, redesign, value engineering, or outsourcing. This chapter considers the techniques for bringing small pieces of information from data warehouses that the manufacturer will cultivate. This preparation of the cost estimates answers the question, “What will this product cost?”

REFERENCE

Ostwald, P. F. (1992), *Engineering Cost Estimating,* 3rd Ed., Prentice Hall, Englewood Cliffs, NJ.

ADDITIONAL READING

Ostwald, P. F., *Construction Cost Analysis and Estimating,* Prentice Hall, Upper Saddle River, NJ, 2001.

Ostwald, P. F., and Muñoz, J., *Manufacturing Processes and Systems,* 9th Ed., John Wiley & Sons, 1997.

Stewart, R. D., Wyskida, R. M., and Johannes, J. D., *Cost Estimator’s Manual,* 2nd Ed., Wiley Interscience, New York, 1995.

CHAPTER 89

Activity-Based Management in Manufacturing

Keith V. Smith
Purdue University

1. INTRODUCTION 2317
2. CONVENTIONAL COSTING SYSTEMS 2317
3. THE EVOLVING MANUFACTURING ENVIRONMENT 2318
4. ACTIVITY-BASED COSTING SYSTEMS 2319
5. CASE STUDY OF AN AUTOMOTIVE MANUFACTURER 2319
6. IMPROVED PRODUCT PRICING AND PROFITABILITY 2322
7. IMPROVED CAPITAL INVESTMENT JUSTIFICATION 2324
8. IMPROVED PERFORMANCE MEASUREMENT 2327
9. SUMMARY 2329
REFERENCES 2329

1. INTRODUCTION

During the last 15 years of the 20th century, the industrial sector of the United States made remarkable progress in restoring its competitiveness. The need to do so resulted from the trend toward a global economy, the increasing pace of technological innovation and development, and rapid advances in informational technology that made it possible. Progress occurred when industrial engineers and corporate managers began to *understand* the reasons why the U.S. industrial sector had slipped competitively. They began to *focus* on specific areas within the organization where things could be done more efficiently and effectively. And they began to develop a series of technological and management *improvements* in how work is planned, organized, and controlled in the modern-day organization. Improvements included material requirements planning (MRP), total quality management (TQM), computer-integrated manufacturing (CIM), just-in-time inventory (JIT), and activity-based costing (ABC).

This chapter deals with improvements in cost management systems, specifically with activity-based costing systems that are designed to provide more accurate costing of products and processes. The application of ABC systems to improve decision making within the organization is referred to as activity-based management (ABM). Included are decisions about product and process pricing, profitability of responsibility centers, capital equipment justification, and organizational performance. After reviewing the need for and development of ABC systems within the evolving manufacturing environment, we illustrate activity-based management with a case study from the automotive sector.

2. CONVENTIONAL COSTING SYSTEMS

The purpose of a cost management system is to provide management with information needed for external reporting and internal planning and control. A cost management system for a manufacturing firm includes cost accounting information for proper valuation of inventory and cost of goods sold; proper pricing of manufactured products, cash flow, and other information for proper justification of

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

capital expenditures; and performance measurements for proper assessment of how the business unit, plus its managers and employees, are doing relative to their expressed goals and objectives.

Johnson and Kaplan (1987) provide a useful historical review of the development of cost management systems in the United States. Conventional cost management systems were first developed with the advent of the industrial revolution in the 19th century. Early systems were designed for firms that manufactured one or two products using a fairly simple production process. These early systems and the information they provided were appropriately expanded with changes in manufacturing. By 1920, virtually all management accounting practices used today had been developed: cost accounts for material, labor, and overhead; budgets for cash, income, and capital; and flexible budgets, sales forecasts, standard costs, variance analyses, transfer prices, and divisional performance measures.

But as the scope and diversity of manufacturing continued to increase in the 1930s and 1940s, the state of the art of management accounting did not keep pace. The lack of new ideas for cost management was in part due to the enormous growth of audited financial reporting, plus the emergence of the public accounting profession. By their nature, public accountants were interested more in generating financial statements for external public consumption and less in generating financial information for internal use by management. Until recently, external financial statements also focused on financial activity in the aggregate, with little attention to the financial progress of profit centers and/or business units within the total organization. Unfortunately, there was not enough pressure by management to develop a second accounting system for internal use.

When companies did attempt to account for individual profit centers, the procedure was straightforward. Costs of material and direct labor were relatively easy to track. The difficulty was in how to assign corporate overhead costs to each profit center. Because direct labor tended to be the largest expense category for most businesses, the easiest practice was just to allocate overhead expenses based on direct labor hours. In other words, the profit center with the largest amount of direct labor would thus have the largest overhead assignment. In essence, allocation of overhead to profit centers was a single-stage process based on direct labor. And because direct labor did not necessarily correlate with the way in which corporate overhead expenses were actually incurred, the single-stage allocation system was also an *arbitrary* costing system.

3. THE EVOLVING MANUFACTURING ENVIRONMENT

During the 1980s, U.S. manufacturers faced increasing competition, largely from abroad. They allowed themselves to fall behind by failing to recognize shifts in consumer demand. They compounded the difficulty by continuing to utilize manufacturing processes and costing systems that were developed in a different time under a different competitive environment. Better news was that many U.S. manufacturers began to understand these developments and began to take steps to improve their competitive posture during the decade of the 1990s. By emphasizing product quality, building flexibility into production processes, reducing inventory, shifting factors of production from direct labor to machinery and equipment, and reorganizing product lines toward less centralized service departments, U.S. manufacturers fought back. A pivotal ingredient was the use of newer and improved information technologies. Of interest in this chapter is the role that cost accounting played in enabling improved information for improved decision making by management.

In addition to the pioneering work of Johnson and Kaplan (1987), important contributions to the literature on cost management systems were provided by Berliner and Brimson (1988), Kaplan (1990), Brimson (1991), and Hicks (1992). An earlier source on the emerging literature on manufacturing management is Hayes et al. (1988). These sources collectively acknowledged the need for improved cost management systems and provided tangible examples of how improved systems could be used to help in improved management decision making.

At the beginning of the 1990s, a comprehensive survey of cost management systems in the United States was conducted by Sullivan and Smith (1993a, b). They investigated the extent and processes by which corporate managers adapted their costing systems to the evolving manufacturing environment. They analyzed responses from 289 manufacturing firms (with a wide range of firm size) to discover linkages between changes in cost management systems and changes in production processes, competitive market conditions, and key elements of manufacturing strategies. They found that while cost management systems were changing, it was more likely to be the fine-tuning of existing systems and less likely to be bold innovation in changing costing systems that would allow management to better understand the economics of their internal managed transactions.

Sullivan and Smith concluded that the evolution in costing systems was really just beginning because plant managers were not yet satisfied that they had the proper information for effective decision making within the organization. An example of bold innovation in cost management systems at that time was activity-based costing, but only 15% of the responding firms reported that they had developed an ABC system. We turn now to a closer look at how ABC developed during the 1990s, and especially how ABC was useful in improving activity-based management.

4. ACTIVITY-BASED COSTING SYSTEMS

Broadly defined, an *activity-based accounting system* is one that attempts to identify the primary activities that allow each cost category to be allocated to all business unit products as reasonably and accurately as possible. Unfortunately, activity-based information is rarely obtainable from the financial accounts that comprise a conventional cost accounting system. Why is it important to have accurate costs? Because managers in a competitive manufacturing environment need to know the true costs in order to decide what to produce and distribute. Decisions involving product design, new product introductions, and the amount of effort expended on trying to market a new product will be influenced by anticipated costs as well as anticipated revenues. Product costs also play an important role in setting prices for products, especially those customized products where there are no comparable market prices.

As mentioned before, direct labor is the basis for overhead allocations in conventional costing systems. But in a manufacturing environment that features technological development and automation, direct labor has become less and less important. Profitability no longer results from just reducing direct labor costs. For example, quality and flexibility play important roles in determining the profitability of the manufacturing firm.

How does a manufacturing firm design and develop an activity-based costing system that provides more accurate information on product costs? As explained by Hilton (1997), management must continue to collect accurate data on direct labor and material costs. But management must also analyze the demands that various products make on the overhead resources of the firm. In so doing, it is useful to give greater attention to more expensive overhead activities, as well as those overhead activities whose consumption varies dramatically across products or product groups.

Another difference between conventional and ABC costing systems is that conventional costing is a single-stage allocation system, while ABC is a two-stage system of allocation. In conventional costing, overhead costs are allocated to profit centers or business units based on direct labor. In ABC, overhead costs are first allocated to *cost pools* of activity. Examples of cost pools in a manufacturing firm are machinery, engineering, purchasing, receiving, inspection, materials handling, quality assurance, packaging, and shipping. For each cost pool, the key activity or *cost driver* is identified. Examples of cost drivers are purchase orders, units produced, inspections, and shipments. In the second stage of allocation, costs in each overhead pool are allocated to profit centers based on the cost driver that is unique to that cost pool.

As a result of a two-stage process of cost allocation, activity-based costing is apt to reflect different costs than in a conventional costing system. For example, costs associated with ordering parts, keeping track of parts, inspecting parts, and setting up to produce parts as input to larger assemblies make up a large fraction of what normally might be classified as purchasing, as just one segment of total manufacturing overhead. However, there are two cost categories that should *not* be included in an activity-based costing system. Research and development costs for new products are more appropriately treated as a capital expenditure. And cost of excess capacity should be separated from product costs so as not to penalize products unfairly during periods of slack demand.

While there is a need for improved cost information, establishing an activity-based cost management system is not easy. And it becomes more difficult as the business firm becomes larger and more complicated. Many corporate managers have been reluctant to abandon a single conventional costing system. But the increased availability of computational power over time has enabled firms to develop and install an improved, second accounting system that provides useful information for internal decision making. Armed with more accurate product cost information, managers can contemplate improved strategic decisions on product lines, product prices, capital expenditures, and organizational performance. That is why the application of ABC is referred to as activity-based management

Activity-based costing represents a distinct improvement over conventional costing. But according to Kaplan and Cooper (1998), most U.S. firms are only at the third of four possible stages of development. Their first stage includes many firms that a decade or more ago had a conventional accounting system that simply did not provide accurate information. The second stage includes firms having an adequate accounting system for external reporting, but nothing that would help with internal management decisions. Kaplan and Cooper believe that many firms are now in the third stage because they have realized the need for improved information for managerial decision making and have developed a second and alternative accounting system for measuring and monitoring costs internally that leads toward improved decision making. The authors suggest a fourth stage in which firms develop a comprehensive accounting system that provides for both external financial reporting and internal decision making. It also should have the capacity to integrate financial reporting of the past and management decision making in the future.

5. CASE STUDY OF AN AUTOMOTIVE MANUFACTURER

To illustrate activity-based costing systems and how they can be used to make improved management decisions, we examine the case of the Titantic Auto Production Company (TAP). An earlier version

TABLE 1 Titatic Auto Production Company Balance Sheet as of December 31, 1999 (in millions)

Cash et al.	$3.1	Current liabilities	$18.1
Receivables	9.4	Long-term debt	13.3
Inventory	12.7	Common stock	7.2
Net fixed assets	30.8	Retained earnings	17.4
Total	$56.0	Total	$56.0

of the case study was presented in Smith and Leksan (1991). Henry Hankinson, president and chief executive officer of TAP has decided that his firm must do something rather drastic to improve their cost management system. Two years ago, Mr. Hankinson attended a seminar in Boston entitled ‘‘Manufacturing Strategies for the Future.’’ The seminar was attended by about 70 CEOs of medium-sized manufacturing firms. In addition to making interesting new contacts, Hankinson found the presentations excellent and the ensuing discussions both intense and useful.

Among the topics covered at the seminar, Hankinson was particularly interested in one session on cost management. The featured speaker at that session, a well-regarded professor from an eastern university, discussed management's critical need for better financial information. The thesis of his presentation was that for many manufacturing firms, today's financial information is based on antiquated cost-accounting systems that do not reflect the reality of the increasingly competitive environment for manufacturing.

That single presentation, together with the lively discussion that followed, convinced Hankinson that he and the TAP board of directors were just not getting the information needed to make important decisions at this critical time in the history of their firm. Among those questions were the pricing of one particular auto model, the relative profitability of TAP's three product lines, the possibility of further automation, and the overall ability of TAP management to control the business. As a result, Henry Hankinson has decided to look beyond the monthly reports that he now receives.

Last month at the TAP annual meeting, CEO Henry Hankinson presented the financial statements for 1999. As seen in Tables 1 and 2, TAP had total sales revenue of $66.5 million on assets of $56 million at year end. After-tax income of $4.9 million (up from $4.6 million in 1998) amounted to 7.4% of sales, 8.8% of assets, and a return on equity of 19.9%. These profitability numbers were well received at the annual meeting, but Hankinson is not sure how TAP compares with other firms in the automotive industry. He also is not sure what changes, if any, should be made in the TAP product line and marketing strategy so as to remain competitive in the years ahead. In order to answer these questions, as well as to learn more about TAP's cost management system, Mr. Hankinson requests a meeting with corporate comptroller Bradley Bartlett.

Bradley Bartlett is well prepared for his meeting with Henry Hankinson. Breakdowns of sales and profitability for each of the three automobiles produced by TAP—compact, midsize, and luxury—are shown in Tables 3 and 4, respectively. In Table 3, we see that total sales volume reached 10,000 autos in 1999, an increase of almost 650 vehicles from the prior year. Although sales volume increased for all three models, compacts accounted for most of that. Management has been pleased with the results of the vigorous advertising program for compacts that led to that increase. For purposes of this illustration, we assume that TAP is operating nearly at maximum capacity.

TABLE 2 Titantic Auto Production Company Income Statement for the Year Ended December 31, 1999 (in millions)

Gross revenues	66.5
Operating expenses	−58.3
Gross margin	8.2
Corporate overhead	− 1.2
Before-tax income	7.0
Federal taxes (30%)	− 2.1
After-tax income	$4.9

TABLE 3 Schedule of Sales Revenues by Model for the Year Ended December 1999

Model	Volume	Price	Revenue
Compact	7,000	$ 5,000	$35.0 million
Midsize	2,900	10,000	29.0 million
Luxury	100	25,000	2.5 million
Totals	10,000		$66.5 million

Profitability by model is analyzed in Table 4, and we see a quite different picture, with compacts actually losing almost $2.3 million. In Table 4, manufacturing overhead totaling $25.0 million is allocated by the conventional method on the basis of direct labor hours at the rate of $25 per hour. By far the largest contributor to TAP's gross margin in 1999 was the midsize product line. Bartlett explains that TAP's corporate overhead of $1.2 million must be subtracted from the gross margin of $8.2 million (Table 4) in order to obtain the firm's before-tax income of $7.0 million (Table 2). Profitability of an individual vehicle is presented in Table 5. Under a conventional cost management system, we see that TAP loses $325 on each compact sold. In contrast, TAP makes $3,150 on each midsize and $13,400 on each luxury automobile.

Despite the favorable aggregate results of TAP for 1999, Hankinson and Bartlett agree that they need to reconsider the emphasis on particular models within the total product line. In particular, should TAP continue its aggressive advertising for compact automobiles when that particular product apparently is losing money? Mr. Hankinson shares his notes and experiences from the seminar that he attended, and he asks Mr. Bartlett to see if he can apply some of the newer thinking to their firm—especially the concept of ''activity-based'' cost accounting.

Two weeks later, Messrs. Hankinson and Bartlett have a second meeting to continue their discussion of the relative profitability of the TAP automotive products. Because Mr. Hankinson is scheduled to attend a local meeting of financial analysts to discuss the latest earnings prognosis for TAP, he has time for only a summary of Bradley's work during the last few days. Mr. Bartlett thus decides to present only the comparison in Table 6.

The comparison between conventional and activity-based cost accounting systems is striking. We see that the results have reversed. Namely, compacts made money, while luxuries lost money. Meanwhile, midsized automobiles continued to generate the largest part of gross margin. Both Hankinson and Bartlett are surprised at these results. Henry Hankinson asks for further explanation at their next

TABLE 4 Schedule of Profitability by Model—Conventional Cost Management System—for the Year Ended December 31, 1999 (in millions)

Model	Sales Revenue	Material	Direct Labor	Overhead	Total Cost	Gross Margin
Compact	$35.0	$14.0	$6.65	$16.62	$37.27	($2.27)
Midsize	29.0	8.7	3.19	7.98	19.87	9.13
Luxury	2.5	0.6	0.16	0.40	1.16	1.34
Totals	$66.5	$23.3	$10.00	$25.00	$58.30	$8.20

TABLE 5 Schedule of Profitability by Vehicle—Conventional Cost Management System—for the Year Ended December 31, 1999 (in millions)

Model	Sales Revenue	Material	Direct Labor*	Overhead	Total Cost	Gross Margin
Compact	$5,000	$2,000	$950	$2,375	$5,325	($325)
Midsize	10,000	3,000	1,100	2,750	6,850	3,150
Luxury	25,000	6,000	1,600	4,000	11,600	13,400

*Direct labor is charged at $10/hr.

TABLE 6 Comparison of Accounting Systems on Gross Margin for the Year Ended December 31, 1999 (in millions)

Model	Sales Revenue	Gross Margin	
		Conventional-Cost	Activity-Based
Compact	$35.0	($2.27)	$2.00
Midsize	29.0	9.13	8.79
Luxury	2.5	1.34	(2.59)
Totals	$66.5	$8.20	$8.20

meeting as to why activity-based cost accounting leads to such different results. Bradley Bartlett agrees to try to provide the requested information.

6. IMPROVED PRODUCT PRICING AND PROFITABILITY

Ten days later, Bradley Bartlett provides the necessary backup information to justify the comparison in Table 6. He explains that TAP has four distinct overhead functions, and thus it is necessary to examine each function individually. The four overhead departments are purchasing, production planning and control, quality control and inspection, and inventory control. The respective analyses are presented as Tables 7–10.

For each overhead department, it is necessary to define the particular activities that drive the functions within that department. For example, in 1999 the purchasing department (see panel A of Table 7) incurred total costs of $5 million. Cost drivers were the purchasing of raw materials, the purchasing of components, and vendor relations. Appropriate allocation measures for these activities are the number of orders for purchasing, and the number of vendors for vendor relations. Levels of those activities for each auto model are indicated in panel B. In turn, these levels lead to the total allocated costs for each auto model in panel C.

TABLE 7 Activities and Costs for Purchasing for the Year Ended December 31, 1999

Panel A: Analysis

Activity	Number of Employees	Total Cost	Allocation Measure	Unit Cost
Purchasing materials	20	$2.0 million	Number of orders	$2,000
Purchasing components	5	1.0 million	Number of orders	250
Vendor relations	10	2.0 million	Number of vendors	20,000
Totals		$5.0 million		

Panel B: Activities

Model	Number of Purchase Orders		Number of Vendors
	Materials	Components	
Compact	500	2000	25
Midsize	300	1500	30
Luxury	200	500	45
Totals	1000	4000	100

Panel C: Costs (in millions)

Model	Purchasing Materials	Purchasing Components	Vendor Relations	Total Cost
Compact	$1.00	$0.50	$0.50	$2.00
Midsize	0.60	0.38	0.60	1.58
Luxury	0.40	0.13	0.90	1.43
Totals	$2.00	$1.00	$2.00	$5.00

TABLE 8 Activities and Costs for Production Planning and Control for the Year Ended December 31, 1999

Panel A: Analysis

Activity	Number of Employees	Total Cost	Allocation Measure	Unit Cost
Developing manufacturing plan	10	$1.0 million	Number of units produced	$100
Controlling manufacturing plan	10	1.0 million	Number of units produced	100
Expediting manufacturing plan	10	1.0 million	Number of units expedited	500
Totals		$3.0 million		

Panel B: Activities

Model	Number of Units Produced	Expedited
Compact	7,000	1,500
Midsize	2,900	450
Luxury	100	50
Totals	10,000	2,000

Panel C: Costs (in millions)

Model	Developing Manufacturing Plan	Controlling Manufacturing Plan	Expediting Manufacturing Plan	Total Cost
Compact	$0.70	$0.70	$0.750	$2.150
Midsize	0.29	0.29	0.225	0.805
Luxury	0.01	0.01	0.025	0.045
Totals	$1.00	$1.00	$1.000	$3.000

The other overhead departments for TAP are production planning and control ($3 million), inventory control ($10 million), and quality control and inspection ($7 million). Similar activity and cost analyses for the other overhead departments (see Tables 8–10) provide improved overhead allocations that can be used to better understand the relative profitability of each of the TAP auto models. The results are included in a revised profitability schedule in Table 11. It shows how a conventional costing system can lead to incorrect decisions by management.

For example, Simon Starling, senior manager of the TAP compact product line, recently suggested that the price of compacts should be increased by 10% in the next model year from $5000 to $5500. According to Starling, if sales remained at the current level of 7000 vehicles, revenues and costs would increase, but gross margin would improve from a loss of $2.27 million to a profit of $1.2 million. Such an improvement certainly would be welcomed. But, more realistically, a price increase of 10% may well affect sales volume for compacts, especially since the compact model competes in a price-sensitive segment of the market. Mr. Starling forecasts an *expected* sales volume of only 5500 vehicles as a result of the 10% price increase. At the same time, Starling's colleagues believe that sales of midsize and luxury automobiles are likely to remain the same during the next model year.

A comparison of direct labor utilization is shown in Table 12. Total overhead costs of $25 million remain the same, at least in the short run, and per hour costs thus increase from $25 per hour (before the price increase) to $25 million/857,500 hr = $29.15/hr (after the price increase). Recomputation of overhead allocations on model profitability via a conventional cost system is presented in Table 13.

Comptroller Bradley Bartlett notes that sales revenue for compacts would be expected to decrease from $35 million to $30.2 million as a result of the lower volume. The gross margin for compacts improves from a loss of $2.27 million (before price increase) to a loss of $1.28 million (after price increase). Gross margins of both midsize and luxury models decrease because they are penalized by having to absorb more of the total corporate overhead. Under the conventional cost management system, management might also look at the decreased gross margins for the midsize and luxury models and thus propose price increases for those products as well. In other words, conventional cost management systems give incorrect signals that are not in the best interest of the firm and its owners.

TABLE 9 Activities and Costs for Inventory Control for the Year Ended December 31, 1999

Panel A: Analysis

Activity	Number of Employees	Total Cost	Allocation Measure	Unit Cost
Receiving parts	25.0	$5.0 million	Number of shipments	$1,250
Receiving materials	12.5	2.5 million	Number of shipments	2,500
Disbursing materials	12.5	2.5 million	Number of production runs	50,000
Totals		$10.0 million		

Panel B: Activities

Model	Number of Shipments		Number of Production Runs
	Parts	Materials	
Compact	2000	500	10
Midsize	1500	300	15
Luxury	500	200	25
Totals	4000	1000	50

Panel C: Costs (in millions)

Model	Receiving Parts	Receiving Materials	Disbursing Materials	Total Costs
Compact	$2.500	$1.250	$0.500	$4.250
Midsize	1.875	0.750	0.750	3.375
Luxury	0.625	0.500	1.255	2.375
Totals	$5.000	$2.500	$2.500	$10.000

If TAP instead were to utilize an activity-based costing system, Simon Starling probably would not have proposed a price increase, since compacts already were making a positive contribution (see Table 11) to the overall profitability of TAP. Under the activity-based costing system, the difficulty caused by lower volume would be avoided by treating the cost of idle capacity as a *period cost,* rather than attributing it to individual products. This is illustrated in Tables 14 and 15. The Starling proposal would cause the gross margin of compacts to increase from $2 million (in Table 11) to $4.01 million (in Table 15). However, that improvement would be more than offset by the $2.39 million cost of idle capacity, so the project should be rejected. However, note that the activity-based costing system suggests that TAP's problem is not with the compacts, but with the luxury line of automobiles. Specifically, either the price of the luxury line should be increased substantially, and/or costs for luxury vehicles should be decreased if possible. Alternatively, TAP should consider abandoning that segment of the market.

7. IMPROVED CAPITAL INVESTMENT JUSTIFICATION

Two months ago, the TAP board of directors heard a presentation from the president of a foreign firm that manufactures high-technology industrial equipment. The firm has a new machine that performs a variety of inspection activities with great precision and considerable flexibility. The president of the foreign firm argues that their new machine is ideally suited for relatively low volume, high-quality manufacturing, such as TAP's midsize automobiles. The new machine costs $2.2 million, has an expected useful lifetime of six years and an estimated salvage value of $400,000, and is expected to reduce by two-thirds the manual inspection of midsize automobiles.

The TAP board was impressed by the presentation, and some members believe that the proposed new machine will add to the profitability of the midsize line. As a result, Henry Hankinson asks Bradley Bartlett to run the numbers, factor in other relevant considerations, and make a recommendation. Bartlett wonders if his new data, using activity-based cost analysis, will have any impact on his eventual recommendation.

Bradley Bartlett has worked hard on trying to understand the full implications of the proposed new machine for automated inspection of midsize automobiles. He wonders what incentive there

TABLE 10 Activities and Costs for Quality Control and Inspection for the Year Ended December 31, 1999

Panel A: Analysis

Activity	Number of Employees	Total Cost	Allocation Measure	Unit Cost
Inspecting materials	15	$3 million	Number of shipments	$600
Inspecting autos	20	4 million	Number of inspection points × number of autos	12.50
Totals		$7 million		

Panel B: Activities

Model	Number of Shipments	Number of Inspection Points/Auto	Total Number of Points
Compact	2500	28	196,000
Midsize	1800	41	118,900
Luxury	700	51	5,100
Totals	5000		320,000

Panel C: Costs (in millions)

Model	Inspecting Materials	Inspecting Autos	Total Cost
Compact	$1.500	$2.450	$3.950
Midsize	1.080	1.486	2.566
Luxury	0.420	0.064	0.484
Totals	$3.000	$4.000	$7.000

TABLE 11 Revised Schedule of Profitability by Model—Activity-Based Cost Management System—for the Year Ended December 31, 1999 (in millions)

Model	Sales Revenue	Material	Direct Labor	Overhead	Total Cost	Gross Margin
Compact	$35.0	$14.0	$6.65	$12.35	$33.00	$2.00
Midsize	29.0	8.7	3.19	8.32	20.21	8.79
Luxury	2.5	0.6	0.16	4.33	5.09	(2.59)
Totals	$66.5	$23.3	$10.00	$25.00	$58.30	$8.20

TABLE 12 Revised Schedule of Direct Labor Utilization—Conventional Cost Management System—for the Next Model Year

	Before Price Change		After Price Change	
Model	Volume	Labor Hours	Volume	Labor Hours
Compact	7,000	665,000 (66.5%)	5,500	522,500 (61%)
Midsize	2,900	319,000 (31.9%)	2,900	319,000 (37%)
Luxury	100	16,000 (1.6%)	100	16,000 (2%)
Totals	10,000	1,000,000	8,500	857,500

TABLE 13 Revised Schedule of Profitability by Model—Conventional Cost Management System—for the Next Model Year (in millions)

Model	Sales Revenue	Material	Direct Labor	Overhead	Total Cost	Gross Margin
Compact	$30.2	$11.0	$5.23	$15.25	$31.48	($1.28)
Midsize	29.0	8.7	3.19	9.25	21.14	7.86
Luxury	2.5	0.6	0.16	0.50	1.21	1.29
Totals	$61.7	$20.3	$8.58	$25.00	$53.88	$7.82

TABLE 14 Comparison of Direct Labor Utilization for the Next Model Year

Model	Conventional Cost	Activity-Based Cost
Compact	522,500	522,500
Midsize	319,000	319,000
Luxury	16,000	16,000
Idle capacity	0	142,500
Totals	857,500	1,000,000

TABLE 15 Revised Schedule of Profitability by Model—Activity-Based Cost Management System—for the Next Model Year (in millions)

Model	Sales Revenue	Material	Direct Labor	Overhead	Total Cost	Gross Margin
Compact	$30.2	$11.0	$5.23	$9.96	$26.19	$4.01
Midsize	29.0	8.7	3.19	8.32	20.21	8.79
Luxury	2.5	0.6	0.16	4.33	5.09	(2.59)
Subtotals	61.7	20.3	8.58	22.61	51.49	10.21
Cost of idle capacity				2.39	2.39	(2.39)
Totals	$61.7	$20.3	$8.58	$25.00	$53.88	$7.82

TABLE 16 Allocation of Overhead on All Product Lines—Proposed Inspection Equipment for Mid-Sized Automobiles—Conventional Cost Management System

Model	Before	After	Savings
Compact	$4,655,000	$3,990,000	665,000
Midsize	2,233,000	1,914,000	319,000
Luxury	112,000	96,000	16,000
Totals	$7,000,000	$6,000,000	$1,000,000

might be for management of the midsize line to adopt such a cost-savings device, since under the conventional cost management system the savings would reduce that portion of total overhead attributable to quality control and inspection, which in turn is then allocated to all three TAP product lines. The midsize product line thus would not receive the full benefit of the new inspection equipment. To understand the implication of this, he decides to prepare an analysis of the potential savings from the proposed foreign equipment under both cost management systems.

The conventional cost management system cannot even tell TAP management how many hours currently go into inspection. In a sense, this isn't important to the conventional system, since total overhead expenses are allocated on the basis of direct labor hours. Bartlett proceeds as follows for his analysis of the midsize product line:

20 employees × 32 effective hr per week × 50 weeks = 32,000 hr/yr.

320,000 inspection points/32,000 hr/yr = 10 inspection points/hr.

Each midsize auto has 41 inspection points, hence 4.1 hr of inspection time per auto.

2,900 autos × 4.1 hr = 12,000 hr of inspection.

Reduced by two-thirds from 12,000 to 4,000 hours, hence savings = 8,000 hr, or 5 employees.

5 employees × ($10 wages/hr + $2 benefits/hr) × 40 hours/wk × 52 weeks = $125,000/yr.

Since this is a permanent reduction in number of employees, the fixed overhead attributed to each employee is also eliminated, for a savings of $1 million.

Assuming overhead percentage among the three product lines remains the same, under conventional costing, the midsize line would have an overhead reduction of $1 million × 31.9% (see Table 12) = $319,000.

Total quality control and inspection overhead of $7 million is reduced by $125,000 in labor saved and $875,000 of fixed overhead to become $6 million. The savings by product line is shown in Table 16.

Bradley Bartlett is puzzled by this result. Under the conventional cost management system, the primary beneficiary of the proposed inspection equipment for midsized automobiles is the compact line, even though that part of the business has nothing to do with the proposed new equipment. Management of the midsize product line would have little incentive to even propose the project, since they would be charged the entire $2.2 million cost of the new machine and yet receive credit for only part of the potential savings. In fact, under conventional costing, the project would actually reduce the return on investment of the midsize line, even though the project would help the corporation overall.

Bartlett next proceeds to analyze of how the proposed new equipment for midsized inspection would look under an ABC system. His calculations are shown in Table 17. The results are striking. Under an activity-based system, the $1 million savings per year is fully attributed to the midsized product line. If TAP has a tax rate of 30% and an annual after-tax cost of capital (i.e., required rate of return) of 15%, then the net present value of the six year investment would be as follows:

Present value:		
Savings	(3.784)(1,000,000)	$3,784,000
Tax shield	(3.784)(2,200,000 − 400,000)/6 yr	340,560
Salvage	(0.432)(400,000)	172,800
		$4,297,360
Cost of the inspection equipment		−2,200,000
Net present value		$2,097,360

This, of course, looks very good for TAP, and it is directly a result of a proposal appropriately attributable to the midsize product line.

8. IMPROVED PERFORMANCE MEASUREMENT

Bradley Bartlett had just completed his analysis of the inspection equipment when he received another proposal from the compact division. Their top management proposes a $2 million automation project that is expected to reduce direct labor from 95 hr to 90 hr per automobile. Again, he wonders how this proposal will look under different cost management systems. He also feels that he should be careful to include all relevant considerations in his analysis.

Mr. Bartlett proceeds to analyze the new proposal from the compact division. The automation project will result in the reduction of direct labor from 95 hours to 90 hours per auto. That would represent an annual savings of:

5 hr × 7,000 autos × $12/hr = $420,000

The new automated process costs $2 million and has an expected lifetime of five years and an estimated salvage value of $300,000. The annual tax shield would be:

TABLE 17 Activities and Costs for Quality Control and Inspection for the Year Ended December 31, 1999

Panel A: Analysis

Activity	Number of Employees	Total Cost	Allocation Measure	Unit Cost
Inspecting materials	15	$3 million	Number of shipments	$600
Inspecting	15	3 million	Number of inspection points × number of autos	12.50
Totals		$6 million		

Panel B: Activities

Model	Number of Shipments	Number of Inspection Points/Auto	Total Number of Points
Compact	2500	28	196,000
Midsize	1800	⅓ × 41 = 14	40,600
Luxury	700	51	5,100
Totals	5000		241,700

Panel C: Costs (in millions)

Model	Inspecting Materials	Inspecting Autos	Total Cost
Compact	$1.500	$2.450	$3.950
Midsize	1.080	0.500	1.580
Luxury	0.420	0.064	0.484
Totals	$3.000	$3.000	$6.000

$$30\% \times (\$2 \text{ million} - \$300{,}000)/5 \text{ years} = \$102{,}000$$

The present value calculation, again using a 15% cost of capital, would be as follows:

Present value:

Savings	(3.353)(350,000)	$1,408,260
Tax shield	(3.353)(102,000)	342,000
Salvage	(0.497)(300,000)	149,100
		$1,899,360
Cost of investment		−2,000,000
Net present value		($100,640)

Because net present value is negative, the proposal would not be accepted using the conventional costing system.

Suppose, however, that the new equipment will also improve the flexibility of TAP to offer additional options of the component whose production is being automated. The machine can perform certain manufacturing steps in very short times that were economically infeasible to perform under the current production process.

This ability to offer additional variations of the component is expected to be valued by consumers, and TAP will be able to charge more for compacts that have these option variations. It is anticipated that 2000 compacts per year would be sold with the more expensive option at a price of $5,070 (i.e., $70 higher). Added material cost for the option is $10, and hence TAP would benefit each year by the after-tax amount of:

$$2{,}000 \text{ autos} \times (\$70 - \$10) \times (1 - 30\%) = \$84{,}000$$

The present value of this is:

$$\$84{,}000 \times 3.353 = \$281{,}650$$

The net present value of the proposal becomes $181,010, and the project should now be accepted. Once again, activity-based costing leads to an improved management decision.

Henry Hankinson and Bradley Bartlett are preparing for the next monthly meeting of the TAP board of directors. As part of his CEO report, Hankinson has decided to make a presentation to the board on the firm's improved cost management system. He asks Mr. Bartlett to prepare illustrations for some recent management proposals where the older cost management system would have led to incorrect or myopic decisions. More work needs to be done to improve their cost management system further, but Henry Hankinson is pleased with the progress as TAP strives to be more competitive in the automobile industry. Mr. Hankinson also decides that he must confront the issue of whether the Titantic Auto Production Company should consider moving into the next stage of development, namely an integrated accounting system that would provide accurate information both externally and internally.

9. SUMMARY

As suggested by the TAP case study, it is likely that much of the information generated by conventional costing systems is inadequate in the evolving manufacturing environment. Focusing on the key costs of a bygone competitive era, conventional costing systems cannot adequately explain the rise of key costs today, namely, overhead. Making matters worse, conventional costing systems allocate overhead on the basis of direct labor, which in many operations is a relatively minor expense. This is where product cost distortions occur. Finally, the standard approaches toward pricing and investment decision making may well not reflect all of the relevant factors that have an impact on the firm.

Managers need to know what their products cost in order to make appropriate decisions concerning the products that they manufacture. This has become critical in the evolving manufacturing environment in which U.S. firms find themselves. However, product-costing decisions cannot be made in a vacuum. In and of itself, this does not increase the competitive advantage of a company. Company management must analyze its entire manufacturing strategy and make necessary changes that will lead to a competitive advantage in its operating activities.

The cost management system of a successful firm needs to be developed to handle two functions. First, it must generate nonfinancial indicators of operating results on a timely basis so that management can control the manufacturing process. Second, it must cost products on the basis of the underlying activities and their costs, rather than on just the purely financial numbers that are utilized for product costing in their conventional systems. In sum, the focus of their cost management system much be changed from its current role of inventory costing for financial statement purposes to a role that involves the effective control of operating activities, the accurate pricing of a company's products, and correct decisions about proposed capital investments and how the organization is appraised overall. Only when the cost management system reaches the third stage can it be used as a competitive weapon in the evolving manufacturing environment of the 20th century. The fourth stage of development for cost management systems is even more promising for firms in the 20th century as they explore the possibility of developing an integration of their external and internal accounting systems.

REFERENCES

Berliner, C., and Brimson, J. A., Eds. (1988), *Cost Management for Today's Advanced Manufacturing,* Harvard Business School Press, Boston.

Brimson, J. A. (1991), *Activity Accounting: An Activity-Based Costing Approach,* John Wiley & Sons, New York.

Hayes, R. H., Wheelwright, S. C., and Clark, K. B. (1988), *Dynamic Manufacturing: Creating the Learning Organization,* Free Press, New York.

Hicks, D. T. (1992), *Activity-Based Costing for Small and Mid-Sized Businesses: An Implementation Guide,* John Wiley & Sons, New York.

Hilton, R. W. (1997), *Managerial Accounting,* 7th Ed., McGraw-Hill, New York.

Johnson, H. T., and Kaplan, R. S. (1987), *Relevance Lost: The Rise and Fall of Managerial Accounting,* Harvard Business School Press, Boston.

Kaplan, R. S., Ed. (1990), *Measures for Manufacturing Excellence,* Harvard Business School Press, Boston.

Kaplan, R. S., and Cooper, R. (1998), *Cost and Effect: Using Integrated Cost Systems to Drive Profitability and Performance,* Harvard Business School Press, Boston.

Smith, K. V., and Leksan, M. P. (1991), ''A Manufacturing Case Study on Activity-Based Costing,'' *Journal of Cost Management,* Vol. 5, No. 2, pp. 45–54.

Sullivan, A. C., and Smith, K. V. (1993a), ''What Really Is Happening to Cost Management Systems in U.S. Manufacturing,'' *Review of Business Studies,* Vol. 2, No. 1, pp. 51–68.

Sullivan, A. C., and Smith, K. V. (1993b), ''Investment Justification for U.S. Factory Automation Projects,'' *Journal of the Midwest Finance Association,* Vol. 22, pp. 24–35.

CHAPTER 90

Discounted Cash Flow Methods

RAYMOND P. LUTZ
The University of Texas at Dallas

1. **DEFINING ALTERNATIVE CAPITAL EXPENDITURE PROPOSALS 2332**
   - 1.1. Determining the Economic Life of an Asset 2332
   - 1.2. Developing Cash Flow Profiles 2332
     - 1.2.1. Traditional Engineering Economy Cash Flow Profiles 2332
     - 1.2.2. Computer Spreadsheet Cash Flow Profiles 2333
   - 1.3. Selecting the Interest Rate 2333
     - 1.3.1. Weighted Average Cost of Capital 2334
     - 1.3.2. Factors Impacting the Selection of an Interest Rate 2335
2. **USING INTEREST FACTORS TO FIND EQUIVALENT MONETARY AMOUNTS 2336**
   - 2.1. Simple Interest 2336
   - 2.2. Compound Interest 2336
   - 2.3. Nominal and Effective Interest Rates 2337
   - 2.4. Compound Interest Factors: Discrete Cash Flow, Discrete Compounding 2337
     - 2.4.1. Compound Amount Factor (Single Payment) 2338
     - 2.4.2. Present Worth Factor (Single Payment) 2338
     - 2.4.3. Compound Amount Factor (Uniform Series) 2339
     - 2.4.4. Sinking Fund Factor 2339
     - 2.4.5. Capital Recovery Factor 2340
     - 2.4.6. Present Worth Factor (Uniform Series) 2341
     - 2.4.7. Arithmetic Gradient Conversion Factor (to a Uniform Series) 2342
     - 2.4.8. Arithmetic Gradient Conversion Factor (to a Present Value) 2343
   - 2.5. Compound Interest Factors: Discrete Cash Flow, Continuous Compounding 2343
     - 2.5.1. Continuous Compounding Compound Amount Factor (single payment) 2345
     - 2.5.2. Continuous Compounding Geometric Conversion Factor (to present value) 2345
   - 2.6. Compound Interest Factors: Continuous Uniform Cash Flows, Continuous Compounding 2345
3. **COMPARING ALTERNATIVES 2346**
   - 3.1. Present Worth 2346
   - 3.2. Annual Worth 2347
     - 3.3. Future Worth 2348
     - 3.4. Rate of Return 2348
     - 3.5. Payback Period 2349
     - 3.6. Benefit–Cost Analysis 2349
4. **NONUNIFORM SERVICE LIVES 2350**
5. **PERPETUITIES AND CAPITALIZED COSTS 2350**

REFERENCES 2351

ADDITIONAL READING 2351

APPENDIX: INTEREST TABLES: SELECTED RATES 2352

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. DEFINING ALTERNATIVE CAPITAL EXPENDITURE PROPOSALS

Expenditures are made with the anticipation that more will be gained in benefits from expenditures than they will cost. If those benefits were to accrue over more than one year, then they would be called *capital expenditures*. Buildings, machinery, equipment, software, or research and development would be typical capital expenditures. When evaluating an expenditure proposal, the primary question is whether or not the money invested will generate a worthwhile annual flow of benefits.

To determine the economic feasibility of a proposed expenditure, it is necessary to define and relate four variables:

1. The economic life of the proposed alternative
2. The amount and pattern of the expenditures and the time period over which they will occur
3. The amount and pattern of the benefits and the time period over which they will occur
4. The interest rate that will appropriately represent the capital structure of the firm and the risk involved

The recognition of risk is especially important because all estimates and actual outcomes are subject to variability.

1.1. Determining the Economic Life of an Asset

The economic life of a capital expenditure is the number of years that this asset will make a positive economic contribution to the firm. The equipment used for a project will be retired when management observes (1) unsatisfactory functional characteristics, such as wear and deterioration; (2) unsatisfactory economic characteristics: it costs more than it earns; (3) termination of need: no one wants to pay for its output anymore; or (4) obsolescence due to changes in policy regulations or technology. The economic life is estimated by considering when the preceding conditions will occur. Such estimations can be facilitated by using data from historical service records of comparable assets, either within the firm or from other sources required to keep accurate records. Utilities or government agencies will often make such data available. Economic service lives can vary significantly. Telecommunications and computer equipment can be obsolete in 18 months, while aircraft and power generators can have life cycles of 30 years. Estimation of economic service lives can be complicated by the potential of legislative or regulatory action terminating or significantly modifying the use of an asset. Finally, economic service lives are not the same as tax lives, which are established for tax strategies, not operational plans.

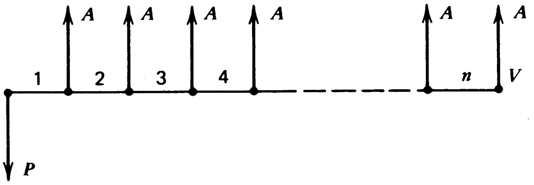
1.2. Developing Cash Flow Profiles

The cash flow profile should include all cash items anticipated to flow into or out of a project. Each cash flow item must be identified specifically according to the time at which this flow occurs. The major items included in an engineering economy study are:

1. First cost, P, is the sum of the costs of engineering, construction, purchasing, installation, and so on, to bring the asset into service. This cost is considered to occur at the installation of the project, $t = 0$. However, if the project requires a number of years to construct, then the major expenditures to be made each year during construction should be identified at the time the cash flow occurs.
2. Salvage, V, is the net sum realized from the disposal of the project or asset at the termination of its useful economic life.
3. Income (revenues), other benefits, and expenses are identified according to their type of flow over time.
   - (a) Periodic cash flow items occur at specific times, such as an overhaul of an engine every three years.
   - (b) Uniform series revenues or expenses are equal periodic amounts such as property taxes, leasing costs, interest on debt, etc.
   - (c) Continuous flow of revenues or expenses occur continuously and uniformly over the life of the project, such as the savings realized from a new assembly technique or the revenue stream from evenly divided alphabetical billing.

1.2.1. Traditional Engineering Economy Cash Flow Profiles

To ensure that all positive and negative cash flow items are included in the analysis and to better visualize these cash flows, it is useful to construct a *cash flow diagram* or *table.* A generalized illustration of a cash flow diagram for n periods of time is shown below. In the diagram, P is the present value of the first cost; A is the value of an annuity, a uniform series of equal cash savings occurring at the end of each period; and V is the salvage value received for the asset upon the termination of the project.



Convention in engineering economy studies dictates that all discrete cash flows are considered to occur at the end of the period, which is normally at the end of each year. However, the period may be quarterly, monthly, or even daily.

To describe a specific cash flow, consider the following discrete items, with the cash flow developed using a tabular format.

End of Year	Receipts	Disbursements	F_{jt}
0	\$ 0	\$−6000	\$6000
1	2000	−500	1500
2	3500	−1000	2500
3	5000	−2000	3000
4	5000	−1000	4000

In this table F_{jt} = net cash flow for investment j at time t. If $F_{jt} < 0$, then F_{jt} represents a net cash disbursement of expense. If $F_{jt} > 0$, then F_{jt} represents a net gain or revenue.

1.2.2. *Computer Spreadsheet Cash Flow Profiles*

Spreadsheets, such as Microsoft Excel, have largely replaced the traditional tabular cash flow profiles and calculators for setting up and solving cash flow problems. The spreadsheet format has become the business standard for collecting, analyzing, and displaying data. Entering the previous data into an Excel format would result in the following table.

	A	B	C	D
1 2	End of Year	Receipts \$	Disbursements \$	Fjt \$
3	0	—	(6,000)	(6,000)
4	1	2,000	(500)	1,500
5	2	3,500	(1,000)	2,500
6	3	5,000	(2,000)	3,000
7	4	5,000	(1,000)	4,000

The calculation of F_{j0} would be accomplished as before, using the Excel commands = B3 + C3. Then the remaining F_{jt} values would be found by duplicating the formula for all subsequent F_{jt} cells.

The principal value of spreadsheets, other than their universal acceptance and use over the past decade, is the ability to display efficiently and use individual estimates of costs and benefits for each period over the life of the project, reflecting anticipated variability rather than a uniform average value or an approximate mathematical series. This will be discussed as each discounted cash flow method is presented. The principal disadvantage of a spreadsheet is that computational errors are hidden, even though an "audit tool" exists. However, the problem of quality assurance is not a new one, and the engineer must always be alert to the prevention and elimination of errors.

1.3. Selecting the Interest Rate

When analyzing the feasibility of any investment, two principles must be observed. First, the value of a sum of money is a function of the time span between the base point of reference and the date an expenditure or the receipt of revenue will occur. For example, the value of $100 one year from today is not the same as that of $100 today. Consequently, money has a time value, which is measured by the interest rate. Second, because of this time value of money, all comparisons of receipts and disbursements must be made at a common specific point of time. Compound interest factors provide the means of shifting cash flows to this common time.

1.3.1. Weighted Average Cost of Capital

The minimum interest rate a company should earn on its invested capital is determined by the capital structure of the firm. The firm must earn an adequate return both to support the long-term debt and to compensate the stockholders adequately for their equity investment. This minimum interest rate, C_C, is calculated using the capital asset pricing model and is often referred to as the weighted cost of capital.

$$C_C = C_D + C_E$$

where C_C = weighted cost of capital, %
C_D = weighted cost of long-term debt, %
C_E = weighted cost of equity, %

The weighted cost of long-term debt, C_D, can be determined by

$$C_D = k_D \times W_{LTD}$$

where k_D is the return necessary to support debt, stated as a percentage and W_{LTD} is the percentage of the firm's capital structure represented by long-term debt. The values of k_D and W_{LTD} are determined by examining the financial statements of the firm. The value of k_D would represent the average interest rate charged on the firm's long-term obligations.

$$W_{\mathrm{LTD}} = \frac{\text{long-term debt}}{\text{long-term debt} + \text{equity}} = \frac{\mathrm{LTD}}{\mathrm{LTD} + E}$$

The weighted cost of equity, C_E, can be determined by

$$C_E = k_E \times W_E$$

where k_E is the return necessary to support the firm's equity stated as a percentage and W_E is the percentage of the firm's capital structure represented by equity. As would be expected, W_E is the complement of W_{LTD} and would be calculated by

$$W_E = \frac{\text{equity}}{\text{long-term debt} + \text{equity}} = \frac{E}{\mathrm{LTD} + E}$$

The before-tax rate of return required to support the firm's equity structure would reflect both the risk attributable to equity instruments and the volatility associated with the firm's stock price relative to the equity market average. The rate of return, k_E, required to support the equity portion of the firm's capital structure is

$$k_E = \frac{(\text{risk-free rate}) + (\text{risk premium})}{1 - \text{tax rate}}$$

$$\mathrm{k}_E = \frac{r^* + \beta(R_m - r^*)}{1 - t}$$

where r^* = risk-free rate of return
R_m = risk attributable to the general equity market
t = effective tax rate
β = systematic risk of a stock due to underlying movements in security prices

The value of the risk-free rate of return, r^*, can be estimated as being equal to the interest rate paid on U.S. Treasury Bills or other guaranteed savings instruments, approximately 6% in mid-2000. The interest rate equivalent to the general equity market, R_m, was found to be approximately 9% by Fisher and Lorie (1968). Over the past 30 years this figure has risen to 11%. The effective tax rate, t, can be calculated from data obtained from the annual consolidated statement of earnings in the firm's annual report. Values for the systematic risk of a stock due to underlying movements in security prices, β, can be obtained for the majority of firms traded on major stock exchanges from references such as *Value Line Investment Survey.*

As an example of determining the minimum rate of return that must be earned to maintain the capital structure of the firm, consider the following calculation for Johnson & Johnson using the data contained in its*1999 Annual Report.* The data are in $million.

$$t = \frac{\text{taxes paid}}{\text{profit before taxes}} = \frac{\$\ 1{,}586}{\$\ 5{,}753} = 27.6\%$$

$$k_E = \frac{r^* + \beta(R_M - r^*)}{1 - t} = \frac{6 + 0.8(11 - 6)}{1 - 0.276} = 13.8\%$$

$$W_E = \frac{E}{\text{LTD} + E} = \frac{\$16{,}231}{\$18{,}689} = 0.8675$$

Cost of equity

$$C_E = W_E \times k_E = .8675 \text{ x } 13.8 = 11.97\%$$

$$W_D = \frac{\text{LTD}}{\text{LTD} + E} = \frac{\$\ 2{,}476}{\$18{,}689} = 0.1325$$

Cost of long-term debt

$$C_D = W_D \times k_D = 0.1325 \times 6.42 = 0.85\%$$

Weighted cost of capital

$$C_C = C_D + C_E = 11.97\% + 0.85\% = 12.82\%$$

1.3.2. Factors Impacting the Selection of an Interest Rate

The preceding calculation will determine the *minimum* interest rate a company should earn on its invested capital to preserve its capital structure. Any lesser rate would erode the firm's reserves. Most companies use a "hurdle rate" of approximately two times the weighted cost of capital. Since research has shown that the average project returns approximately half of the original estimate, using a higher rate of return is a good policy. However, a firm may adopt differing interest rates in order to establish strategic investment guidelines for different types of business or functions within the business. For example, an energy company might require quite different interest rate hurdles between marketing and exploration and production, or a technology firm could encourage research and development expenditures by using a rate in a feasibility study which would be a fraction of that applied to a "cash cow" operation. For example, consider projects for:

1. *Safety, quality, or legal requirement:* A specific rate of return is usually not required. The immediate need for the project or the lack of an alternative solution may preclude an earnings test for such an expenditure.
2. *Increased profit:*
    - **(a)** *Cost reduction:* Projects with a rate of return of at least 25% would qualify for consideration.
    - **(b)** *Existing product:* Projects to increase the production capacity, flexibility, or delivery speed for an existing product would be considered if the return were at least 25%.
    - **(c)** *New product line:* Because of the greater risk of demand, technology, and life cycle, the rate of return should be at least 50% over the projected life of the project.
3. *Country risk:* When operating offshore, the stability of a country's government, labor force, and infrastructure operations can vary the risk to the firm. The hurdle rate should be varied accordingly.

The interest rate used to evaluate capital investment alternatives should be greater than the weighted cost of capital. However, the specific rate to be used in capital budgeting and project feasibility analysis must be a management decision, depending on the type of activity, the risk, and the opportunity costs.

2. USING INTEREST FACTORS TO FIND EQUIVALENT MONETARY AMOUNTS

Equivalence is equating monetary values occurring at different points in time. Comparisons of financial alternatives must be made among equivalent units. The Interest rate provides the mechanism for converting a cash flow at one specific time into an equivalent cash flow at another time. This conversion is facilitated by using either the Excel spreadsheet functions in Table 1 or the interest tables at the end of the chapter.

2.1. Simple Interest

The simple interest payment each year, i_n, is found by multiplying the interest rate, i, times the invested capital, or principal, P. Thus, $i_n = Pi$. After any n time periods, the accumulated value of money owed under simple interest, F_n, would be

$$F_n = P(1 + i_n)$$

For example, \$100 invested now at 9% simple interest for eight years would yield

$$F_8 = \$100[1 + 0.09(8)] = \$172$$

Simple interest forgoes the money earned by annual compounding and is rarely used in engineering economy analyses.

2.2. Compound Interest

The interest payment each year, or each period, is found by multiplying the interest rate by the accumulated value of money, both principal and interest.

TABLE 1 Compound Interest Factors: Discrete Cash Flow, Discrete Compounding

To Find	Given	Name of Factor	Algebraic	Format	
				Functional	Excel
F	P	Compound amount factor (single payment)	$(1 + i)^n$	$(F/P, i\%, N)$	FV (rate, nper, pmt, pv, type)
P	F	Present worth factor (single payment)	$(1 + i)^{-n}$	$(P/F, i\%, N)$	PV(rate, nper, pmt, fv, type)
F	A	Compound amount factor (uniform series)	$\frac{(1 + i)^n - 1}{i}$	$(F/A, i\%, N)$	FV(rate, nper, pmt, pv, type)
A	F	Sinking fund factor	$\frac{i}{(1 + i)^n - 1}$	$(A/F, i\%, N)$	PMT(rate, nper, pv, fv, type)
A	P	Capital recovery factor	$\frac{i(1 + i)^n}{(1 + i)^n - 1}$	$(A/P, i\%, N)$	PMT(rate, nper, pv, fv, type)
P	A	Present worth factor (uniform series)	$\frac{(1 + i)^n - 1}{i(1 + i)^n}$	$(P/A, i\%, N)$	PV(ratte, nper, pmt, fv, type)
A	G	Arithmetic gradient conversion factor to uniform series	$\frac{(1 + i)^n - (1 + ni)}{(1 + i)^n - 1}$	$(A/G, i\%, N)$	
P	G	Arithmetic gradient conversion factor to present value	$\frac{1 - (1 + ni)(1 + i)^{-n}}{i^2}$	$(P/G, i\%, N)$	

End of Period (EOP)	Accumulated BOP Value of Amount Owed (1)	Interest for Period (2)	Amount Owed or Value Accumulated End of Period (3) = (1) + (2)
0	0		P
1	P	Pi	$P + Pi = P(1 + i)^2$
2	$P(1 + i)^1$	$[P(1 + i)^1]i$	$P(1 + i)^1 + P(1 + i) = P(1 + i)^2$
3	$P(1 + i)^2$	$[P(1 + i)^2]i$	$P(1 + i)^2 + P(1 + i)^2\, i = P(1 + i)^3$
. . .	. . .	. . .	. . .

Consequently, the value for an amount P invested for n periods at i rate of interest would be

$$F_n = P(1 + i)^n$$

For example, \$100 invested now at 9% compound interest for eight years would yield

$$F_8 = \$100(1 + 0.09)^8 = 100(1.9926) = \$199.26$$

2.3. Nominal and Effective Interest Rates

For many financial feasibility studies it is appropriate to consider interest periods of one year. However, financial agreements may call for interest to be compounded or paid more frequently, say quarterly, monthly, or even daily. Interest rates associated with more frequent compounding, say of quarterly interest periods, are usually stated as "8% compounded quarterly."

The nominal interest rate, r, is expressed as an annual rate, without considering the impact of any compounding per period during the year. It is obtained by multiplying the periodic interest rate, i, by the number of periods per year, m.

$$r = im$$

For example, the nominal annual rate would be 8% with a 2% quarterly rate.

The effective annual interest rate is the true or actual annual interest rate, taking into account the monetary gain obtained by compounding the invested capital each period during the year. The effective interest rate per year, i_a, is

$$i_a = (1 + i)^m - 1 \text{ when } m < \infty$$

For example when the effective interest rate is 2% per month, the nominal interest rate per year is

$$r = (0.02)(12) = 24\%$$

and the effective interest rate per year is

$$i_a = (1 + 0.02)^{12} = 26.8\%$$

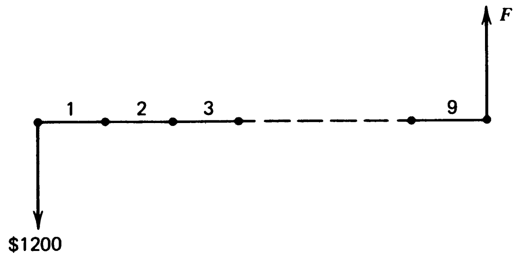
2.4. Compound Interest Factors: Discrete Cash Flow, Discrete Compounding

The compound interest factors described in this section are used for discrete cash flows compounded discretely at the end of each interest period. All of these factors can be found in Table 1, including algebraic and functional formats and the Excel functions. The numerical values for each factor for selected interest rates can be found in the tables at the end of this chapter. Complete tables can be found in the Additional Reading at the end of the chapter. The notation used in this chapter is

i = effective interest rate.
n = number of compounding periods.
A = end-of-period cash flows (or equivalent end-of-period vales) in a uniform series continuing for a specified number of periods (the letter A implies annual or an annuity).
F = future sum of money (the letter F implies future or equivalent future value).
P = present sum of money (the letter P implies present or equivalent present value).

2.4.1. Compound Amount Factor (Single Payment)

This factor finds the equivalent future worth, F, of a present investment, P, held for n periods at a rate of i interest. For example, what is the value in nine years of \$1200 invested now at 10% interest?



Algebraic format

$$\begin{aligned} F &= P(1 + i)^n \\ &= \$1200(1 + 0.10)^9 \\ &= \$1200(2.3579) \\ F &= \$2829 \end{aligned}$$

Functional format (note that throughout this chapter only the functional format will be used)

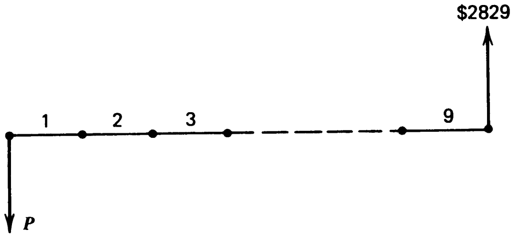
$$\begin{aligned} F &= P(F/P, i\%, N) \\ &= P(F/P, 10\%, 9) \\ &= \$1200(2.3579) \\ F &= \$2,829 \end{aligned}$$

Excel format
Highlight the location where you want the solution. Click on the f_x button on the Standard Tool Bar. Follow the Excel instructions to enter the appropriate variables.

F = FV (rate, nper, pmt, pv, type)
= FV (0.10, 9, 0, A1, 0)
pmt = no intervening payments
pv = A1, location of the present value payment of \$1,200 in the spreadsheet. Note that from the cash flow diagram, pv should be negative if F is to be positive.
Type = 0 for end of the period payments. Excel default. 0 for beginning of the period payments
F = \$2830

2.4.2. Present Worth Factor (Single Payment)

This factor finds the equivalent present value, P, of a single future cash flow, F, occurring at n periods in the future when the interest rate is $i\%$ per period. Note that this factor is the reciprocal of the compound amount factor (single payment). For example, what amount would you have to invest now to yield \$2829 in nine years if the interest rate were 10%?



$$\begin{aligned} P &= F(P/F, 10\%, 9) \\ &= \$2829\ (0.4241) \\ &= \$1200 \end{aligned}$$

Functional format

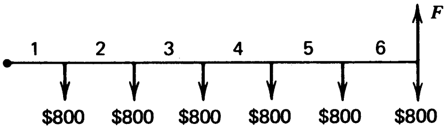
$$\begin{aligned} P &= F(P/F, \text{i}\%, N) \\ &= F(P/F, 10\%, 9) \\ &= \$2839(0.4241) \\ &= \$1200. \end{aligned}$$

Excel format

$$\begin{aligned} P &= \text{PV(rate, nper, pmt, pv, type)} \\ &= \text{FV}(0.10, 9, 0, \$1200, 0) \\ &= \$(1200) \end{aligned}$$

2.4.3. *Compound Amount Factor (Uniform Series)*

This factor finds the equivalent future value, F, of the accumulation of a uniform series of equal annual payments, A, occurring over n periods at i rate of interest per period. For example, what would be the future worth of an annual year-end cash flow of \$800 for six years at 12% interest per year?



$$\begin{aligned} \mathbf{F} &= \mathbf{A(F/A, 12\%, 6)} \\ &= \mathbf{\$800\ (12.2997)} \\ &= \mathbf{\$9840} \end{aligned}$$

Functional format

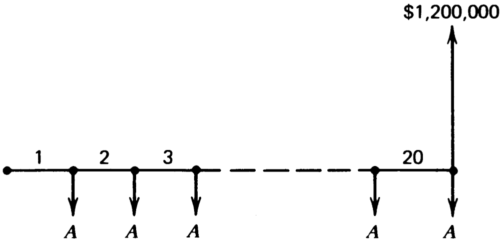
$$\begin{aligned} F &= A(F/A, \text{i}\%, N) \\ &= A(F/A, 12\%, 6) \\ &= \$800(8.1152) \\ &= \$6492 \end{aligned}$$

Excel format

$$\begin{aligned} F &= \text{FV(rate, nper, pmt, pv, type)} \\ &= \text{FV}(0.10, 6, -800, 0, 0) \\ &= \$6492 \end{aligned}$$

2.4.4. *Sinking Fund Factor*

This factor determines how much must be deposited each period in a uniform series, A, occurring over n periods at i rate of interest per period to yield *a specified* future sum, F. For example, if a \$1.2 million bond issue is to be retired at the end of 20 years, how much must be deposited annually into a sinking fund at 10% interest per year?



$$\begin{aligned} A &= F(A/F, 7\%, 20) \\ &= \$1{,}200{,}000\ (0.0244) \\ &= \$29{,}280 \end{aligned}$$

Functional format

$$\begin{aligned} A &= F(A/F, \text{i}\%, N) \\ &= F(A/F, 10\%, 20) \\ &= \$1{,}200{,}000(0.01746) \\ &= \$20{,}952 \end{aligned}$$

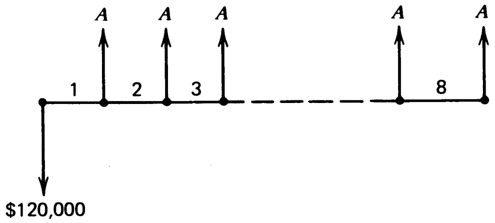
Excel format

$$\begin{aligned} A &= \text{PMT(rate, nper, pv, fv, type)} \\ &= \text{PMT}(0.10, 20, 0, \$1{,}200{,}000, 0) \\ &= \$20{,}952 \end{aligned}$$

This factor was historically used to find the required annual payments that must be made into a “sinking fund” to retire a bond issue by a particular date.

2.4.5. Capital Recovery Factor

This factor finds an annuity, or uniform series of payments, over n periods at $i\%$ interest per period that is equivalent to a present value, P. For example, what savings in annual manufacturing costs over an eight-year period would justify the purchase of a $120,000 machine if a firm’s minimum attractive rate of return (MARR) were 20%?



$$\begin{aligned} A &= P(A/P, 25\%, 8) \\ &= \$120{,}000\ (0.3004) \\ &= \$36{,}048 \end{aligned}$$

Functional format

$$\begin{aligned} A &= P(A/P,\ \text{i}\%,\ N) \\ &= P(A/P,\ 20\%,\ 8) \\ &= \$120{,}000(0.2606) \\ &= \$31{,}272 \end{aligned}$$

Excel format

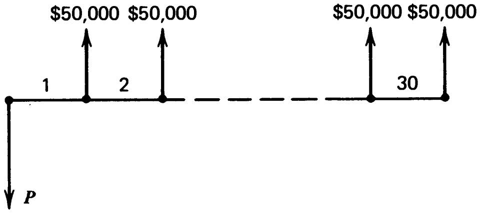
$$\begin{aligned} A &= \text{PMT(rate, nper, pv, fv, type)} \\ &= \text{PMT}(0.20,\ 8,\ \$120{,}000,\ 0) \\ &= \$31{,}273 \end{aligned}$$

This problem of finding the revenue that must be generated each period to justify a capital expenditure is one of the most common facing the engineer. The analyst should be reminded that there are two elements of the capital recovery factor. The first is the recovery of the $120,000 original investment, and the second is the necessity of earning 20% on the capital invested over the life of the project. In this case,

$$\begin{aligned} 8 \text{ payments} \times \$31{,}273 &= \$250{,}184 \text{ cash flow} \\ &= \underline{(\$120{,}000)} \text{ return of original investment} \\ &= \$130{,}184 \text{ interest paid on invested capital} \end{aligned}$$

2.4.6. Present Worth Factor (Uniform Series)

This factor finds the equivalent present value, *P*, of a series of end-of-period payments, *a*, for *n* periods at *i%* interest per period. For example, a donor has offered to give the Hospital Authority a new wing for treatment of allergies. However, since the operation and maintenance of the existing facility requires all of the funds available under the authority's taxing limits, how much additional endowment would be required to provide $50,000/year over the 30-year economic life of the structure. The endowment is expected to earn 10% on its invested capital.



$$\begin{aligned} P &= \$50{,}000\ (P/A,\ 7\%,\ 30) \\ &= \$50{,}000\ (12{,}409) \\ &= \$620{,}450 \end{aligned}$$

Functional format

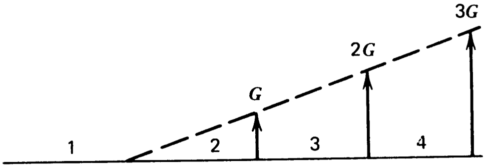
$$\begin{aligned} P &= A(P/A,\ i\%,\ N) \\ &= \$50{,}000(P/A,\ 10\%,\ 30) \\ &= \$50{,}000(9.4269) \\ &= \$471{,}345 \end{aligned}$$

Excel format

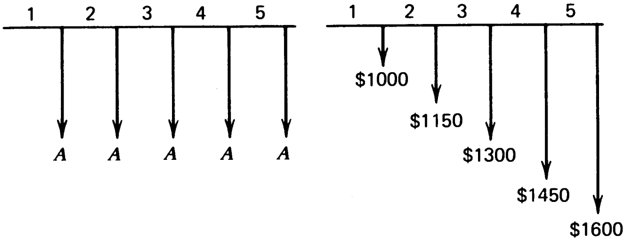
$$\begin{aligned} P &= \text{PV(rate, nper, pmt, fv, type)} \\ &= \text{PV}(0.10,\ 30,\ -50000,\ 0,\ 0) \\ &= \$471{,}345 \end{aligned}$$

2.4.7. Arithmetic Gradient Conversion Factor (to a Uniform Series)

Many times annual payments do not occur in equal amounts. Inflation causes annual increases in operating costs, and maintenance costs often increase with the age of the equipment. If a series of payments increases by an equal amount or gradient, G, each year, then a special compound interest factor can be used to reduce the gradient series to an equivalent equal-payment series. The following illustration shows a four-period gradient series that increases by G each period.



The arithmetic gradient conversion factor (to uniform series) is used when it is necessary to convert a gradient series into a uniform series of equal payments. For example, what would be the equal annual series, A, that would have the same net present value (i.e., be equivalent) at 20% interest per year to a five-year gradient series that started at \$1000 for the first year and increased \$150 every year thereafter?



$$\begin{aligned} A &= A_g + G(A/G, 20\%, 5) \\ &= \$1000 + \$150\ (1.6405) \\ &= \$1246 \end{aligned}$$

Functional format

$$\begin{aligned} A &= A_g + G(A/G, i\%, N) \\ &= A_g + G(A/G, 20\%, 5) \\ &= \$1000 + \$150(1.6405) \\ &= \$1246 \end{aligned}$$

where A_g is the uniform base value of the gradient series. In the previous four-period gradient series illustration, $A_g = 0$.

Excel format

It is necessary to solve this problem in two steps

1. Net present value of the series of cash flows:

$$\begin{aligned} \text{NPV} &= (\text{rate, value 1, value 2, value 3, value 4, value 5}) \\ &= (0.20, 1000, 1150, 1300, 1450, 1600) \\ &= \$3762 \end{aligned}$$

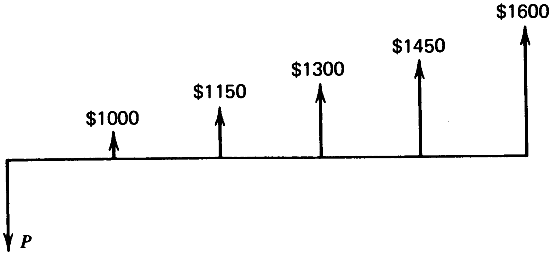
2. Equal annual cash flows equivalent to the series of cash flows:

$$\begin{aligned} A &= \text{PMT (rate, nper, pv, fv, type)} \\ &= \text{PMT } (0.20, 5, 3726, 0, 0) \\ &= \$1246 \end{aligned}$$

Excel Spreadsheet can calculate equivalent values from a series of cash flows regardless of whether they follow a particular gradient series, such as in this example, or are just best estimates of the cash flows for each period. This is a powerful feature eclipsing the requirement of uniform cash flow series necessary in traditional engineering economy models.

2.4.8. Arithmetic Gradient Conversion Factor (to a Present Value)

This factor converts a series of cash amounts increasing by a gradient value, G, each period to an equivalent present value at $i\%$ interest per period. For example a machine will require \$1000 maintenance in the first year of its five-year operating life. Further, the cost of maintenance will increase by \$150 each year. What is the present worth of this series of maintenance costs if the firm's minimum attractive rate of return (MARR) is 20%?



$$\begin{aligned} P &= A(P/A, 20\%, 5) + G(P/G, 20\%, 5) \\ &= \$1000\,(2.9906) + \$150\,(4.9061) \\ &= \$3727 \end{aligned}$$

Functional format

$$\begin{aligned} P &= A(P/A, i\%, N) + G(P/G, i\%, N) \\ &= A(P/A, 20\%, 5) + \$150(P/G, 20\%, 5) \\ &= \$1000(2.9906) + \$150(4.9061) \\ &= \$3727 \end{aligned}$$

Excel format—referring to the previous example

$$\begin{aligned} \text{NPV} &= \text{(rate, value 1, value 2, value 3, value 4, value 5)} \\ &= (0.20, 1000, 1150, 1300, 1450, 1600) \\ &= \$3726 \end{aligned}$$

2.5. Compound Interest Factors: Discrete Cash Flow, Continuous Compounding

Monetary institutions and industrial firms alike strive to keep their funds working at all times. Techniques of cash management, such as electronic funds transfer, provide the potential to put to work cash receipts immediately. This has shortened the compounding periods to the point where the use of continuous compounding is the most appropriate cash flow model. In the concept of discrete cash flows with continuous compounding, it is assumed that the cash flows occur once per year but that compounding is continuous throughout the year. Thus, if

r = nominal interest rate per year
M = number of compounding periods per year
N = number of years

then at the end of one year one unit of principal will equal

$$\left[1 + \left(\frac{r}{n}\right)\right]^{M} \tag{1}$$

Letting $k = M/r$, Eq. (1) becomes

$$\left[1 + \frac{1}{k}\right]^{rk} = \left[\left(1 + \frac{1}{k}\right]^{k}\right]^{r} \tag{2}$$

The limit of $(1 + 1/k)^k$ as k approaches infinity is e. Thus, Eq. (2) can be written as e^r, and the single-payment continuous compounding amount factor at $r\%$ nominal annual interest rate for N years is e^{rN}. Also, since e^{rN} (for continuous compounding) corresponds to $(1 + i)^N$ for discrete compounding,

$$e^r = 1 + i$$

or

$$i = e^r - 1$$

By the use of this relationship, the compound interest factors for discrete cash flows compound continuously shown in Table 1 can be derived from the discrete compounding factors in Table 2.

TABLE 2 Compound Interest Factors: Discrete Cash Flow, Continuous Compounding

To Find	Given	Name of Factor	Format	
			Algebraic	Functional
F	P	Continuous compounding Compound amount factor (single payment)	e^{m}	$(F/P, r\%, N)$
P	F	Continuous compounding Present worth factor (single payment)	e^{-rn}	$(P/F, r\%, N)$
F	A	Continuous compounding Compound amount factor (uniform series)	$\dfrac{e^{rm} - 1}{e^{r} - 1}$	$(F/A, r\%, N)$
A	F	Continuous compounding Sinking fund factor	$\dfrac{e^{r} - 1}{e^{rm} - 1}$	$(A/F, r\%, N)$
A	P	Continuous compounding Capital recovery factor	$\dfrac{e^{rm}(e^{r} - 1)}{e^{rn} - 1}$	$(A/P, r\%, N)$
P	A	Continuous compounding Present worth factor (uniform series)	$\dfrac{e^{rn} - 1}{e^{rn}\,(e^{r} - 1)}$	$(P/A, r\%, N)$
A	G	Continuous compounding Arithmetic gradient conversion factor (to uniform series)	$\dfrac{1}{e^{r} - 1} - \dfrac{n}{e^{rn} - 1}$	$(A/G, r\%, N)$
P	G	Continuous compounding Arithmetic gradient conversion factor (to present value)	$\dfrac{e^{rn} - 1 - n(e^{r} - 1)}{e^{rn}(e^{r} - 1)^2}$	$(P/G, r\%, N)$
P	$A_{1,}c$	Continuous compounding Geometric gradient Conversion fctor (to present value)	$\dfrac{1 - e^{(C-r)n}}{e^{r} - e^{c}}$	$(P/A, r\%, c\%, N)$ $r \neq c$
F	$A_{1,}c$	Continuous compounding Geometric gradient Conversion factor (to uniform series)	$\dfrac{e^{rn} - e^{cn}}{e^{r} - e^{c}}$	$(F/A, r\%, c\%, N)$ $r \neq c$

2.5.1. Continuous Compounding Compound Amount Factor (Single Payment)

This factor is used to find the equivalent future worth, F, of a present value, P, when the interest is continuously compounded at the nominal annual rate of $r\%$. For example, consider the problem of finding the future worth in six years of \$5,000 invested now at 9% nominal interest rate compounded continuously.

$$F = Pe^{rn}$$

Functional format

$$\begin{aligned} F &= P(F/P, r\%, N) \\ &= P(F/P, 9\%, 6) \\ &= \$5000(1.7160) \\ &= \$8580 \end{aligned}$$

Note that the only difference between continuous compounding and discrete compounding in finding equivalent values of F, P, A, and G is the interest factor used (r, the nominal annual interest rate). Consequently, to solve discrete cash flow continuous compounding problems, use the same procedures illustrated for discrete compounding with the functional format.

2.5.2. Continuous Compounding Geometric Conversion Factor (to Present Value)

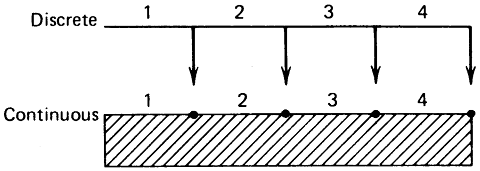
When conditions such as inflation cause a gradient that increases at a fixed percent per period, c, then geometric conversion factors should be used. For example, find the present worth of a series of costs that increase 10% per year from the initial first-year cost of \$1000 for five years when the firm's MARR is 15%.

Functional format

$$\begin{aligned} P &= A(P/A, r\%, c\%, N) \\ &= A(P/A, 15\%, 10\%, 5) \\ &= \$1000\ (3.9037) \\ &= \$3904 \end{aligned}$$

2.6. Compound Interest Factors: Continuous Uniform Cash Flows, Continuous Compounding

Many cash flows that engineers must consider can be assumed to occur continuously, such as accounts receivable, the cost savings resulting from productivity improvements, or the costs of carrying inventories. The following cash flow diagrams serve to illustrate the differences between discrete and continuous cash flows.



Again, as with discrete cash flows, to solve for one variable given another, it is only necessary to select the proper interest factor for continuous cash flow, continuous compounding. A listing of these factors can be found in Table 3.

As an example, consider the equivalent future worth of a uniform series of continuous cash flows totaling \$2000 per year for 10 years compounded continuously at 15% nominal annual rate of interest.

$$\begin{aligned} F &= \overline{\mathbf{A}}(F/\overline{\mathbf{A}}, r\%, N) \\ &= \$2000\ (F/\overline{\mathbf{A}}, 15\%, 10) \\ &= \$2000\ (23.2113) \\ &= \$46{,}423 \end{aligned}$$

TABLE 3 Compound Interest Factors: Continuous, Uniform Cash Flow, Continuous Compounding

To Find	Given	Name of Factor	Format	
			Algebraic	Functional
F	$\overline{A}$	Continuous compounding Compound amount factor (continuous, uniform payments)	$\frac{e^{rn}-1}{r}$	$(F/\overline{A}, r\%, N)$
$\overline{A}$	F	Continuous compounding Sinking fund factor (continuous, uniform payments)	$\frac{r}{e^{rn}-1}$	$(\overline{A}/F, r\%, N)$
$\overline{A}$	P	Continuous compounding Capital recovery factor (continuous, uniform payments)	$\frac{re^{rn}}{e^{rn}-1}$	$(\overline{A}/P, r\%, N)$
P	$\overline{A}$	Continuous compounding Present worth factor (continuous, uniform payments)	$\frac{e^{rn}-1}{re^{rn}}$	$(P/\overline{A}, r\%, N)$

3. COMPARING ALTERNATIVES

In comparing alternatives to meet a need or an objective, plans (1) should provide the same quality and quantity (or level) of service and (2) should provide that service over the same period of time. Competing plans should be alternative ways to accomplish the same end. Any differences in expected revenue or other benefits must be credited to the plan providing the additional services. The analysis is concerned only with the *differences* in the cash flows between the alternatives.

3.1. Present Worth

The present worth method compares all of a project's estimated expenditures to all of its estimated revenues and other benefits at a reference time called the "present" ($t = 0$). For a particular interest rate, if the present value of the revenues and other benefits exceeds the present value of the expenses, the project is considered acceptable. The present worth of alternative j with cash flows that last of n periods of time at $i\%$ interest per period is

$$\text{PW }(i)_j = \sum_{t=1}^{N} F_{jt}\,(P/F,\, i\%,\, t)$$

If two or more alternatives are being compared, the alternative with the greatest present worth or net present value is recommended. To compare alternatives fairly using the present worth method, it is necessary that all alternatives have a common retirement date.

For example, two pieces of equipment are being considered by a hospital to perform a particular service. Brand A will cost \$30,000 and will have an annual operating and maintenance cost of \$5000 over its eight-year economic life with a salvage value of \$3000. Brand B will cost \$15,000, will have an annual operating and maintenance cost of \$8000 over the first three years and \$10,000 over the last three years of its economic life, and will have a negligible salvage value. Which brand of equipment would you recommend, using the present worth comparison with an interest rate of 10%/year?

Cash Flow Item	Brand A	Brand B
First cost	\$(30,000)	\$(15,000)
Operating and maintenance		
A: = \$(5,000) (P/A,10%,6)		
= \$(5,000) (4.3553)	(21,776)	
B: = \$(8,000)(P/A,10%,3)		
= \$(8,000)(2.4869)	(19,895)	

Cash Flow Item	Brand A	Brand B
= $(10,000)[(*P*/*A*,10%,3)(*P*/*F*,10%,3)]		
= $(10,000)[(2.4869)(0.7531)]	$(18,684)	
Salvage value		
A: = $3,000(*P*/*F*,10%,6)		
= $3,000(0.5645)	1,694	
B:		0
Present worth	$(50,082)	$(53,579)

Brand A would be the recommended alternative since the present worth of its total cost is smaller or its present worth is greater.

3.2. Annual Worth

The annual worth method converts all cash flows to an equivalent uniform series of equal annual payments. As in the present worth method, if the annual worth of the revenues is greater than the annual worth of the costs for the specified interest rate, then the project is acceptable. The annual worth of alternative *j* and *i* percent rate of interest per period, which lasts for *n* periods, is

$$\mathrm{AW}(i)_j = \mathrm{PW}(i)_j(A/P,\ i\%,\ N)$$

It is usually necessary to calculate the present worth of all cash flows, PW(i), first since these cash flows are rarely a uniform series that can be summed directly to find AW(i).

If two or more alternatives are being compared, the alternative with the greatest annual worth (cash receipts are positive and disbursements are negative) is the recommended alternative.

If you must compare alternatives with differing economic lives, the annual worth method is preferred when the "repeatability assumption" is valid for the analysis. (See Section 4 for a detailed discussion of the comparison of alternatives with unequal service lives.) If this assumption is valid, then the annual worth at the time of renewal of the asset is exactly the same as before. Therefore, you are actually comparing the annual worth of two infinite series.

Using the preceding example, the following calculations are made.

Cash Flow Item	Brand A	Brand B
First cost		
A: = $(30,000) (*P*/*A*, 10%, 6)		
= $(30,000)(0.2296)	$ (6,888)	
A: = $(15,000) (*P*/*A*, 10%, 6)		
= $(15,000) (0.2296)		$ (3,444)
Operating and maintenance		
A: =	$ (5,000)	
B: = $(8,000)(*P*/*A*, 10%, 3) (*A*/*P*, 10%, 6)		
= $(8,000)(2.4869)(0.2296)		$ (4,568)
= $(10,000)[(*P*/*A*, 10%, 3)(*P*/*F*, 10%, 3) (*A*/*P*, 10%, 6)]		
= $(10,000)[(2.4869)(0.7531)(0.2296)]		$ (4,290)
Salvage value		
A: = $3,000(*A*/*F*, 10%, 6)		
= $3,000(0.1296)	389	
B:		0
Annual worth	$(11,499)	$(12,302)

Brand A is the recommended alternative since it has the greatest annual worth.

Note that either of these methods of comparison recommends the same alternative because they are dealing with equivalent monetary amounts.

A common method of finding the annual worth of an alternative is

$$\mathrm{AW}(i)_j = R_j - \mathrm{FC}(A/P,\ i\%,\ N) - (\mathrm{O\&M}) + V(A/F,\ i\%,\ N)$$

where R = revenues per year (uniform series)
FC = first cost
V = salvage value
O&M = operating and maintenance cost (uniform series)

This is simply an equation form of the preceding tabular format.

3.3. Future Worth

The future worth (FW) method is comparable to the present worth method except that the comparison between the project's estimated expenditures and benefits occurs at a reference time called the ''future'' ($t = F$). As in present worth analysis, in future worth analysis a project is acceptable at a particular interest rate if the future value of the revenues and other benefits exceeds the future value of the expenses. Likewise, the preferred alternative, given equal future benefits, would be the alternative with the lowest future costs.

For example, if the estimated future worth of a stream of revenues and other benefits from proposed materials handling equipment at the end of 10 years is \$1,200,000, should the new equipment be purchased? The firm's MARR before taxes is 25%. The initial cost would be \$125,000, and the annual maintenance cost would be \$750/year for the 10-year life.

$$\begin{aligned}\text{FW}(25\%) &= \$125{,}000\ (F/P, 25\%, 10) + \$750\ (F/A, 25\%, 10)\\ &= \$125{,}000\ (9.313) + \$750\ (33.253)\\ &= \$1{,}164{,}125 + \$25{,}940\\ &= \$1{,}189{,}065\end{aligned}$$

Since the future worth of the benefits exceed the future worth of the costs, the purchase of the equipment would be justified.

3.4. Rate of Return

The rate of return method finds the interest rate that equates the cash flows of receipts and disbursements. That is an alternative's rate of return is the interest rate at which the present worth of the cash flows is equal to 0.

$$0 = \sum_{t=0}^{N} F_{jt}\ (1 + i)^{-k}$$

Thus, for alternative j the rate of return is the break-even interest rate between incomes and expenses.

For example, what rate of return would be earned from a \$64,000 investment in a testing device if the savings were to be \$16,000/year for 8 years?

Functional format

$$-\$64{,}000 = \$16{,}000\ (P/A, i\%, 8)$$

Solving for (P/A) and interpolating from the interest tables

$$i = 18.62\%$$

Excel format

IRR calculates the rate of return for a series of cash flows. Assume the data were located in an Excel spreadsheet with −64,000 in cell A1 and 16,000 in each of cells A2 through A9. Going to the function button and then to the financial micros,

$$i = \text{IRR(values, guess)}$$

Enter the location of the data array, A1:A9, in the value command box. The ''Formula Result'' then indicates the internal rate of return,

$$i = 0.1862$$

A more complex example would consider the investment yielding the four-year stream of cash flows illustrated in the discussion on cash flow profiles. What rate of return would equate the −\$6000 disbursement at $t = 0$ with the positive cash flows of \$1500, \$2500, \$3000, and \$4000 at the end of years 1, 2, 3, and 4, respectively?

$$\begin{aligned}0 = -5000 &+ \$1500\ (\mathrm{P/F},\ \mathrm{i}\%,\ 1) + \$2500\ (\mathrm{P/F},\ \mathrm{i}\%,\ 2)\\ &+ \$3000\ (\mathrm{P/F},\ \mathrm{i}\%,\ 3) + \$4000\ (\mathrm{P/F},\ \mathrm{i}\%,\ 4)\end{aligned}$$

Functional format
The rate of return for this set of cash flows is approximately 34%/year and would be determined using an iterative trial-and-error solution method, guessing values of *i* until a zero solution is obtained.

Excel format
From an array of values of cash flow on the spreadsheet arranged from $n = 0$ to $n = 4$, enter A1: a5 into the values box.

$$i = \text{IRR(values, guess)}$$

$$i = 33.99\% \text{ by the formula}$$

The Excel solution is far faster and less error prone than an iterative method in the functional format.

3.5. Payback Period

The payback period method determines the length of time required to recover the initial investment, or first cost, at a zero rate of interest. The payback period (PP) for alternative *j* is

$$\mathrm{PP}_j = \frac{\text{first cost of the project}}{\text{uniform } net \text{ benefits per period}}$$

$$\mathrm{PP}_j = \frac{\mathrm{FC}_j}{R - D}$$

where *R* equals the equivalent uniform benefits per period and *D* equals the equivalent uniform costs per period. For example, find the payback period for a \$10,000 investment that will return net uniform benefits of \$1,250/year.

$$\mathrm{PP} = \frac{\$\ 10{,}000}{\$\ 1{,}250} = 8 \text{ years}$$

The alternative with the shortest payback period would be the preferred alternative.

The payback period method is an approximate measure of preference. First, it does not consider the timing of cash flows prior to payback, ignoring the time value of money. It weighs cash flows 10 years from now the same as cash flows occurring today. Second, it ignores the duration of the cash flows. Cash flows after the payback period, such as major overhauls, are not included in the calculation. These weaknesses in the payback period method render it less desirable than the other measures of merit presented in this section.

Payback period is still widely used for comparing alternatives. Its use precludes the necessity of specifying an interest rate or performing interest rate calculations. However, its widest use is as a surrogate measure of risk. The faster the cash is returned from the project, the less the likelihood of unforeseen risks harming the outcome.

3.6. Benefit–Cost Analysis

The benefit–cost method is often utilized to determine the feasibility of public sector expenditures. The benefit–cost criterion for the *j*th alternative, B/C_j, can be expressed as

$$B/C_j = \frac{\sum_{t=1}^{Nt} B_{jt}\,(1 + i)^{-t}}{\sum_{t=1}^{tN} C_{jt}\,(1 + i)^{-t}}$$

where B_{jt} equals the public benefits accruing to alternative project *j* during year *t* and C_{jt} equals the governmental costs associated with the alternative project *j* during year *t*. A project is deemed to be acceptable if $B/C_j \geq 1.0$, that is, if the project's benefits equal or exceed its costs.

The calculation of the benefit–cost ratio is not different in principle from the other methods of comparing alternatives. However, determining the monetary value of the costs and benefits associated with projects in the public sector is difficult. This difficulty arises because of the often subjective nature of the costs and benefits and because they often occur far into the future. A detailed discussion of the problems of defining public costs and benefits and of their use in calculating the benefit–cost ratio can be found in other sources, such as Smith (1987) or White et al. (1998).

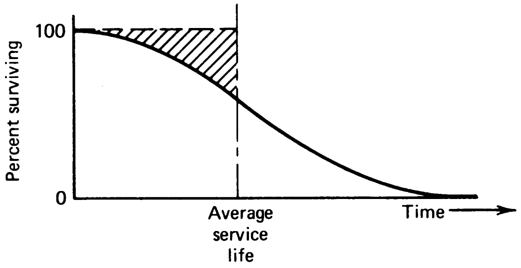
An example of the use of the benefit–cost method is as follows: A government bridge project requires an initial investment of $10 million and operation and maintenance costs of $250,000/year for the 20-year life of the project. The annual user benefits of $1,950,000/year are estimated to arise from savings in travel distance and time. At an annual interest rate of 7%, the benefit–cost ratio is

$$\begin{aligned} B/C &= \frac{\$1{,}950{,}000\ (\mathrm{P/A},7\%,20)}{\$10{,}000{,}000 + \$250{,}000\ (\mathrm{P/A},7\%,20)} \\ &= \frac{\$1{,}950{,}000\ (10.5940)}{\$10{,}000{,}000 + \$250{,}000\ (10.5940)} \\ &= \frac{\$20{,}658{,}300}{\$10{,}000{,}000 + \$2{,}648{,}500} \\ &= 1.63 \end{aligned}$$

Thus, the project benefits would exceed its costs.

4. NONUNIFORM SERVICE LIVES

Engineers are often required to purchase or install a number of items at one time in order to provide a service or manufacture a product. The group of such items is called a vintage group. For example, you may install 2000 telephone poles or 50 automobiles. Assuming that all these items must be in working order in order to provide the required level of service, whenever an item fails, it must be replaced. Thus, if 10 poles are destroyed or fail during the first year of service, they must be replace in order to maintain the desired level of service. The result of having to replace items to maintain the desired level of service means that the true cost will always be higher than the estimated by conventional interest factors. Therefore, the error is always one of *underestimation.* The following figure illustrates the effect of assuming the rectangular distribution as opposed to a survival distribution, representing the true failure rate of the items in question. The shaded portion of the distribution indicates the items that must be replaced to maintain a given level of service.



It is the cost of these replacement items that provides the error of underestimation. See Smith (1987) for details on how to correct for this error in group properties that is due to nonuniform service lives.

5. PERPETUITIES AND CAPITALIZED COSTS

In some major public works projects, such as dams, locks, or bridges, the life of the investment is considered to be infinite. In the case of an infinite life asset, the amount needed to construct or acquire that asset initially plus the amount to provide for the perpetual maintenance and replacement of that asset is referred to as the capitalized cost. A "perpetuity" is a uniform series of payments that continues indefinitely, such as one would find from the conversion of a capitalized cost to an annuity.

For example, what would be the capitalized cost of an irrigation dam at a 5% rate of interest per year if the initial construction cost were $5 million/year for three years and there was a $50,000 maintenance cost/year forever after the completion of the construction?

$$
\begin{aligned}
P &= \$5{,}000{,}000\ (P/A,\ 5\%,\ 3) + \$50{,}000\ (P/A,\ 5\%,\ \infty)\ (P/F,\ 5\%,\ 3) \\
&= \$5{,}000{,}000\ (2.7233) + \$50{,}000\ (20.00)(0.8638) \\
&= \$13{,}616{,}500 + \$863{,}800 \\
&= \$14{,}480{,}300
\end{aligned}
$$

Note that in this example the present worth factor of the annuity is the reciprocal of the interest rate.

$$(P/A,\ i\%,\ \infty) = 1/i$$

REFERENCES

Fisher, L., and Lorie, J. H. (1968), “Rates of Return on Investments in Common Stock: The Year-by-Year Recod, 1926–1965,” *Journal of Business,* Vol. 41, July, pp. 291–316.

Smith, G. W., *Engineering Economy: Analysis of Capital Expenditures,* 4th Ed., Iowa State University Press, Ames, 1987.

White, J. A., Case, K. E., Pratt, D. B., and Agee, M. H. (1998), *Principles of Engineering Economic Analysis*, 4th Ed., John Wiley & Sons, New York.

ADDITIONAL READING

Blank, L. T., and Tarquin, A. J., *Engineering Economy*, 4th Ed., McGraw-Hill, New York, 1997.

Fabrycky, W. J., and Thuesen, G. J., *Engineering Economy,* 8th Ed., Prentice Hall, Englewood Cliffs, NJ, 1993.

Park, C. S., *Contemporary Engineering Economy*, 2nd Ed., Addison-Wesley, Reading, MA, 1997.

Park, C. S. and Sharp-Bette, G. P., *Advanced Engineering Economics*, John Wiley & Sons, New York, 1990.

Schenbach, T. E., *Engineering Economy: Applying Theory to Practice*, Richard D. Irwin, Homewood, IL, 1995.

Sullivan, W. G., Ed., *Engineering Economy,* 11th Ed., Prentice Hall, Upper Saddle River, NJ, 2000.

Young, D., *Modern Engineering Economy*, John Wiley & Sons, New York, 1993.

APPENDIX
Interest Tables: Selected Rates

CONTINUOUS COMPOUNDING INTEREST FACTORS
Interest Rate 9%

	Single Payment		Uniform Series				Arithmetic Gradient		
N	Compound Amount Factor Find *F* Given *P* *F*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *F* *P*/*F*, *i*, *N*	Capital Recovery Factor Find *A* Given *P* *A*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *A* *P*/*A*, *i*, *N*	Compound Amount Factor Find *F* Given *A* *F*/*A*, *i*, *N*	Sinking Fund Factor Find *A* Given *F* *A*/*F*, *i*, *N*	Uniform Series Factor Find *A* Given *G* *A*/*G*, *i*, *N*	Present Worth Factor Find *P* Given *G* *P*/*G*, *i*, *N*	*N*
1	1.0942	0.9139	1.0942	0.9139	1.0000	1.0000	0.0000	0.0000	1
2	1.1972	0.8353	0.5717	1.7492	2.0942	0.4775	0.4775	0.8353	2
3	1.3100	0.7634	0.3980	2.5126	3.2914	0.3038	0.9401	2.3620	3
4	1.4333	0.6977	0.3115	3.2103	4.6014	0.2173	1.3878	4.4551	4
5	1.5683	0.6376	0.2599	3.8479	6.0347	0.1657	1.8206	7.0056	5
6	1.7160	0.5827	0.2257	4.4306	7.6030	0.1315	2.2388	9.9193	6
7	1.8776	0.5326	0.2015	4.9632	9.3190	0.1073	2.6424	13.1149	7
8	2.0544	0.4868	0.1835	5.4500	11.1966	0.0893	3.0316	16.5221	8
9	2.2479	0.4449	0.1696	5.8948	13.2510	0.0755	3.4065	20.0810	9
10	2.4596	0.4066	0.1587	6.3014	15.4990	0.0645	3.7674	23.7401	10
11	2.6912	0.3716	0.1499	6.6730	17.9586	0.0557	4.1145	27.4559	11
12	2.9447	0.3396	0.1426	7.0126	20.6498	0.0484	4.4479	31.1914	12
13	3.2220	0.3104	0.1366	7.3229	23.5945	0.0424	4.7680	34.9158	13
14	3.5254	0.2837	0.1315	7.6066	26.8165	0.0373	5.0750	38.6033	14
15	3.8574	0.2592	0.1271	7.8658	30.3419	0.0330	5.3691	42.2327	15
16	4.2207	0.2369	0.1234	8.1028	34.1993	0.0292	5.6507	45.7866	16
17	4.6182	0.2165	0.1202	8.3193	38.4200	0.0260	5.9201	49.2512	17
18	5.0531	0.1979	0.1174	8.5172	43.0382	0.0232	6.1776	52.6155	18
19	5.5290	0.1809	0.1150	8.6981	48.0913	0.0208	6.4234	55.8711	19
20	6.0496	0.1653	0.1128	8.8634	53.6202	0.0186	6.6579	59.0117	20
21	6.6194	0.1511	0.1109	9.0144	59.6699	0.0168	6.8815	62.0332	21
22	7.2427	0.1381	0.1093	9.1525	66.2893	0.0151	7.0945	64.9326	22
23	7.9248	0.1262	0.1078	9.2787	73.5320	0.0136	7.2972	67.7087	23
24	8.6711	0.1153	0.1065	9.3940	81.4568	0.0123	7.4900	70.3612	24
25	9.4877	0.1054	0.1053	9.4994	90.1280	0.0111	7.6732	72.8908	25
26	10.3812	0.0963	0.1042	9.5957	99.6157	0.0100	7.8471	75.2990	26
27	11.3589	0.0880	0.1033	9.6838	109.9969	0.0091	8.0122	77.5879	27
28	12.4286	0.0805	0.1024	9.7642	121.3558	0.0082	8.1686	79.7603	28
29	13.5991	0.0735	0.1016	9.8378	133.7844	0.0075	8.3168	81.8193	29
30	14.8797	0.0672	0.1010	9.9050	147.3835	0.0068	8.4572	83.7683	30
31	16.2810	0.0614	0.1003	9.9664	162.2632	0.0062	8.5899	85.6109	31
32	17.8143	0.0561	0.0998	10.0225	178.5442	0.0056	8.7155	87.3511	32
33	19.4919	0.0513	0.0993	10.0738	196.3585	0.0051	8.8340	88.9928	33
34	21.3276	0.0469	0.0988	10.1207	215.8504	0.0046	8.9460	90.5401	34
35	23.3361	0.0429	0.0984	10.1636	237.1780	0.0042	9.0516	91.9970	35
36	25.5337	0.0392	0.0980	10.2027	260.5140	0.0038	9.1512	93.3678	36
37	27.9383	0.0358	0.0977	10.2385	286.0477	0.0035	9.2451	94.6563	37
38	30.5694	0.0327	0.0974	10.2712	313.9861	0.0032	9.3335	95.8667	38
39	33.4483	0.0299	0.0971	10.3011	344.5555	0.0029	9.4167	97.0028	39
40	36.5982	0.0273	0.0968	10.3285	378.0038	0.0026	9.4950	98.0684	40
41	40.0448	0.0250	0.0966	10.3534	414.6020	0.0024	9.5685	99.0673	41
42	43.8160	0.0228	0.0964	10.3763	454.6469	0.0022	9.6377	100.0030	42
43	47.9424	0.0209	0.0962	10.3971	498.4629	0.0020	9.7026	100.8791	43
44	52.4573	0.0191	0.0960	10.4162	546.4053	0.0018	9.7635	101.6988	44
45	57.3975	0.0174	0.0958	10.4336	598.8626	0.0017	9.8207	102.4654	45
46	62.8028	0.0159	0.0957	10.4495	656.2601	0.0015	9.8743	103.1819	46
47	68.7172	0.0146	0.0956	10.4641	719.0629	0.0014	9.9245	103.8513	47
48	75.1886	0.0133	0.0954	10.4774	787.7801	0.0013	9.9716	104.4764	48
49	82.2695	0.0122	0.0953	10.4895	862.9687	0.0012	10.0157	105.0598	49
50	90.0171	0.0111	0.0952	10.5006	945.2382	0.0011	10.0569	105.6042	50

CONTINUOUS COMPOUNDING CONTINUOUS FLOW FACTORS
Interest Rate 15%

N	Capital Recovery Factor Find $\bar{A}$ Given P $A/P, i, N$	Present Worth Factor Find P Given $\bar{A}$ $P/A, i, N$	Compound Amount Factor Find F Given $\bar{A}$ $F/A, i, N$	Sinking Fund Factor Find $\bar{A}$ Given F $A/F, i, N$	N
1	1.0769	0.9286	1.0789	0.9269	1
2	0.5787	1.7279	2.3324	0.4287	2
3	0.4139	2.4158	3.7887	0.2639	3
4	0.3325	3.0079	5.4808	0.1825	4
5	0.2843	3.5176	7.4467	0.1343	5
6	0.2528	3.9562	9.7307	0.1028	6
7	0.2307	4.3337	12.3843	0.0807	7
8	0.2147	4.6587	15.4674	0.0647	8
9	0.2025	4.9384	19.0495	0.0525	9
10	0.1931	5.1791	23.2113	0.0431	10
11	0.1857	5.3863	28.0465	0.0357	11
12	0.1797	5.5647	33.6643	0.0297	12
13	0.1749	5.7182	40.1913	0.0249	13
14	0.1709	5.8503	47.7745	0.0209	14
15	0.1677	5.9640	56.5849	0.0177	15
16	0.1650	6.0619	66.8212	0.0150	16
17	0.1627	6.1461	78.7140	0.0127	17
18	0.1608	6.2186	92.5315	0.0108	18
19	0.1592	6.2810	108.5852	0.0092	19
20	0.1579	6.3348	127.2369	0.0079	20
21	0.1567	6.3810	148.9071	0.0067	21
22	0.1557	6.4208	174.0843	0.0057	22
23	0.1549	6.4550	203.3359	0.0049	23
24	0.1542	6.4845	237.3216	0.0042	24
25	0.1536	6.5099	276.8072	0.0036	25
26	0.1531	6.5317	322.6830	0.0031	26
27	0.1527	6.5505	375.9830	0.0027	27
28	0.1523	6.5667	437.9089	0.0023	28
29	0.1520	6.5806	509.8564	0.0020	29
30	0.1517	6.5926	593.4475	0.0017	30
31	0.1514	6.6029	690.5666	0.0014	31
32	0.1512	6.6118	803.4028	0.0012	32
33	0.1511	6.6194	934.4998	0.0011	33
34	0.1509	6.6260	1086.8127	0.0009	34
35	0.1508	6.6317	1263.7751	0.0008	35
36	0.1507	6.6366	1469.3761	0.0007	36
37	0.1506	6.6408	1708.2504	0.0006	37
38	0.1505	6.6444	1985.7827	0.0005	38
39	0.1504	6.6475	2308.2292	0.0004	39
40	0.1504	6.6501	2682.8586	0.0004	40
41	0.1503	6.6524	3118.1159	0.0003	41
42	0.1503	6.6544	3623.8127	0.0003	42
43	0.1502	6.6561	4211.3486	0.0002	43
44	0.1502	6.6576	4893.9679	0.0002	44
45	0.1502	6.6589	5687.0584	0.0002	45
46	0.1502	6.6599	6608.4981	0.0002	46
47	0.1501	6.6609	7679.0583	0.0001	47
48	0.1501	6.6617	8922.8718	0.0001	48
49	0.1501	6.6624	10367.9769	0.0001	49
50	0.1501	6.6630	12046.9494	0.0001	50

DISCRETE COMPOUND INTEREST FACTORS
Interest Rate 10%

	Single Payment		Uniform Series				Arithmetic Gradient		
N	Compound Amount Factor Find *F* Given *P* *F*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *F* *P*/*F*, *i*, *N*	Capital Recovery Factor Find *A* Given *P* *A*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *A* *P*/*A*, *i*, *N*	Compound Amount Factor Find *F* Given *A* *F*/*A*, *i*, *N*	Sinking Fund Factor Find *A* Given *F* *A*/*F*, *i*, *N*	Uniform Series Factor Find *A* Given *G* *A*/*G*, *i*, *N*	Present Worth Factor Find *P* Given *G* *P*/*G*, *i*, *N*	*N*
1	1.1000	0.9091	1.1000	0.909090	1.0000	1.0000	0.0000	0.0000	1
2	1.2100	0.8264	0.5762	1.735537	2.1000	0.47619	0.47619	0.8264	2
3	1.3310	0.7513	0.4021	2.486852	3.3100	0.3021	0.93656	2.3291	3
4	1.4641	0.6830	0.3155	3.1699	4.6410	0.21547	1.38117	4.3781	4
5	1.6105	0.6209	0.2638	3.790787	6.1051	0.1638	1.81013	6.8618	5
6	1.7716	0.5645	0.2296	4.355261	7.7156	0.12961	2.22356	9.6842	6
7	1.9487	0.5132	0.2054	4.868419	9.4872	0.10541	2.62162	12.7631	7
8	2.1436	0.4665	0.1874	5.334926	11.4359	0.08744	3.00448	16.0287	8
9	2.3579	0.4241	0.1736	5.7590	13.5795	0.07364	3.37235	19.4215	9
10	2.5937	0.3855	0.1627	6.144567	15.9374	0.06275	3.72546	22.8913	10
11	2.8531	0.3505	0.1540	6.495061	18.5312	0.0540	4.06405	26.3963	11
12	3.1384	0.3186	0.1468	6.813692	21.3843	0.04676	4.3884	29.9012	12
13	3.4523	0.2897	0.1408	7.103356	24.5227	0.04078	4.69879	33.3772	13
14	3.7975	0.2633	0.1357	7.366687	27.9750	0.03575	4.99553	36.8005	14
15	4.1772	0.2394	0.1315	7.60608	31.7725	0.03147	5.2789	40.1520	15
16	4.5950	0.2176	0.1278	7.823709	35.9497	0.02782	5.54934	43.4164	16
17	5.0545	0.1978	0.1247	8.021553	40.5447	0.02466	5.8071	46.5819	17
18	5.5599	0.1799	0.1219	8.2014	45.5992	0.02193	6.05256	49.63954	18
19	6.1159	0.1635	0.1195	8.36492	51.1591	0.01955	6.2861	52.58268	19
20	6.7275	0.1486	0.1175	8.513564	57.2750	0.01746	6.50808	55.40691	20
21	7.4002	0.1351	0.1156	8.648694	64.0025	0.01562	6.71888	58.10952	21
22	8.1403	0.1228	0.1140	8.77154	71.4027	0.0140	6.9189	60.6893	22
23	8.9543	0.1117	0.1126	8.883218	79.5430	0.01257	7.1085	63.14621	23
24	9.8497	0.1015	0.1113	8.984744	88.4973	0.0113	7.28805	65.4813	24
25	10.8347	0.0923	0.1102	9.0770	98.3471	0.01017	7.4580	67.6964	25
26	11.9182	0.0839	0.1092	9.160945	109.18177	0.00916	7.61865	69.7940	26
27	13.1100	0.0763	0.1083	9.237223	121.09994	0.00826	7.77044	71.77726	27
28	14.4210	0.0693	0.1075	9.306567	134.20994	0.00745	7.9137	73.64953	28
29	15.8631	0.0630	0.1067	9.369606	148.63093	0.00673	8.04886	75.41463	29
30	17.4494	0.0573	0.1061	9.4269	164.4940	0.00608	8.17632	77.07658	30
31	19.1943	0.0521	0.1055	9.479013	181.94342	0.0055	8.29617	78.63954	31
32	21.1138	0.0474	0.1050	9.526376	201.13777	0.0050	8.4091	80.10777	32
33	23.2252	0.0431	0.1045	9.569432	222.25154	0.0045	8.5152	81.48559	33
34	25.5477	0.0391	0.1041	9.608575	245.4767	0.00407	8.61494	82.77729	34
35	28.1024	0.0356	0.1037	9.644158	271.02437	0.00369	8.7086	83.98715	35
36	30.9127	0.0323	0.1033	9.676508	299.1268	0.0033	8.7965	85.11938	36
37	34.0039	0.0294	0.1030	9.705917	330.0395	0.0030	8.87892	86.17808	37
38	37.4043	0.0267	0.1027	9.732651	364.0434	0.00275	8.95617	87.16727	38
39	41.1448	0.0243	0.1025	9.7570	401.4478	0.00249	9.02852	88.09083	39
40	45.2593	0.0221	0.1023	9.779051	442.5926	0.00226	9.09623	88.9525	40
41	49.7852	0.0201	0.1020	9.7991	487.8518	0.0020	9.15958	89.7560	41
42	54.7637	0.0183	0.1019	9.817397	537.6370	0.00186	9.2188	90.50466	42
43	60.2401	0.0166	0.1017	9.8340	592.4007	0.00169	9.27414	91.20187	43
44	66.2641	0.0151	0.1015	9.849089	652.6408	0.00153	9.32582	91.85079	44
45	72.8905	0.0137	0.1014	9.862808	718.9048	0.00139	9.3740	92.45443	45
46	80.1795	0.0125	0.1013	9.87528	791.7953	0.00126	9.4190	93.01567	46
47	88.1975	0.0113	0.1011	9.886618	871.9749	0.00115	9.4610	93.53723	47
48	97.0172	0.0103	0.1010	9.896926	960.1723	0.0010	9.5001	94.02168	48
49	106.7190	0.0094	0.1009	9.906296	1057.1896	0.00095	9.53651	94.47146	49
50	117.3909	0.0085	0.1009	9.914814	1163.9085	0.00086	9.57041	94.88887	50

DISCRETE COMPOUND INTEREST FACTORS
Interest Rate 12%

	Single Payment		Uniform Series				Arithmetic Gradient		
N	Compound Amount Factor Find *F* Given *P* *F*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *F* *P*/*F*, *i*, *N*	Capital Recovery Factor Find *A* Given *P* *A*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *A* *P*/*A*, *i*, *N*	Compound Amount Factor Find *F* Given *A* *F*/*A*, *i*, *N*	Sinking Fund Factor Find *A* Given *F* *A*/*F*, *i*, *N*	Uniform Series Factor Find *A* Given *G* *A*/*G*, *i*, *N*	Present Worth Factor Find *P* Given *G* *P*/*G*, *i*, *N*	*N*
1	1.1200	0.8929	1.1200	0.89286	1.0000	1.0000	0.0000	0.0000	1
2	1.2544	0.7972	0.5917	1.69005	2.1200	0.4717	0.4717	0.7972	2
3	1.4049	0.7118	0.4163	2.40183	3.3744	0.29635	0.92461	2.2208	3
4	1.5735	0.6355	0.3292	3.03735	4.7793	0.20923	1.35885	4.1273	4
5	1.7623	0.5674	0.2774	3.60478	6.3528	0.15741	1.77459	6.3970	5
6	1.9738	0.5066	0.2432	4.11141	8.1152	0.12323	2.1720	8.9302	6
7	2.2107	0.4523	0.2191	4.56376	10.0890	0.09912	2.55147	11.6443	7
8	2.4760	0.4039	0.2013	4.96764	12.2997	0.0813	2.91314	14.4714	8
9	2.7731	0.3606	0.1877	5.3282	14.7757	0.06768	3.25742	17.3563	9
10	3.1058	0.3220	0.1770	5.65022	17.5487	0.05698	3.58465	20.2541	10
11	3.4785	0.2875	0.1684	5.9377	20.6546	0.0484	3.89525	23.1288	11
12	3.8960	0.2567	0.1614	6.19437	24.1331	0.04144	4.18965	25.9523	12
13	4.3635	0.2292	0.1557	6.42355	28.0291	0.03568	4.4683	28.7024	13
14	4.8871	0.2046	0.1509	6.62817	32.3926	0.03087	4.73169	31.3624	14
15	5.4736	0.1827	0.1468	6.81086	37.2797	0.02682	4.9803	33.9202	15
16	6.1304	0.1631	0.1434	6.97399	42.7533	0.02339	5.21466	36.3670	16
17	6.8660	0.1456	0.1405	7.11963	48.8837	0.02046	5.4353	38.6973	17
18	7.6900	0.1300	0.1379	7.24967	55.7497	0.01794	5.64274	40.9080	18
19	8.6128	0.1161	0.1358	7.36578	63.4397	0.01576	5.83752	42.9979	19
20	9.6463	0.1037	0.1339	7.46944	72.0524	0.01388	6.0202	44.96757	20
21	10.8038	0.0926	0.1322	7.5620	81.6987	0.01224	6.19132	46.81876	21
22	12.1003	0.0826	0.1308	7.64465	92.5026	0.0108	6.35141	48.55425	22
23	13.5523	0.0738	0.1296	7.71843	104.6029	0.00956	6.5010	50.17759	23
24	15.1786	0.0659	0.1285	7.78432	118.1552	0.00846	6.64064	51.69288	24
25	17.0001	0.0588	0.1275	7.8431	133.3339	0.0075	6.7708	53.10464	25
26	19.0401	0.0525	0.1267	7.89566	150.33393	0.00665	6.8921	54.4177	26
27	21.3249	0.0469	0.1259	7.94255	169.3740	0.0059	7.00491	55.63689	27
28	23.8839	0.0419	0.1252	7.98442	190.69889	0.00524	7.10976	56.76736	28
29	26.7499	0.0374	0.1247	8.02181	214.58275	0.00466	7.20712	57.81409	29
30	29.9599	0.0334	0.1241	8.05518	241.3327	0.00414	7.29742	58.78205	30
31	33.5551	0.0298	0.1237	8.0850	271.29261	0.00369	7.3811	59.6761	31
32	37.5817	0.0266	0.1233	8.11159	304.84772	0.0033	7.45858	60.5010	32
33	42.0915	0.0238	0.1229	8.13535	342.42945	0.00292	7.53025	61.26122	33
34	47.1425	0.0212	0.1226	8.15656	384.5210	0.0026	7.59649	61.96123	34
35	52.7996	0.0189	0.1223	8.1755	431.6635	0.00232	7.65765	62.60517	35
36	59.1356	0.0169	0.1221	8.19241	484.46312	0.00206	7.71409	63.1970	36
37	66.2318	0.0151	0.1218	8.20751	543.59869	0.0018	7.76613	63.74058	37
38	74.1797	0.0135	0.1216	8.2210	609.83053	0.00164	7.81406	64.23936	38
39	83.0812	0.0120	0.1215	8.2330	684.0102	0.00146	7.85819	64.69675	39
40	93.0510	0.0107	0.1213	8.24378	767.09142	0.0013	7.89879	65.11587	40
41	104.2171	0.0096	0.1212	8.25337	860.14239	0.0012	7.93611	65.4997	41
42	116.7231	0.0086	0.1210	8.26194	964.3595	0.0010	7.9704	65.85095	42
43	130.7299	0.0076	0.1209	8.2696	1081.0826	0.00092	8.00188	66.17222	43
44	146.4175	0.0068	0.1208	8.27642	1211.8125	0.00083	8.03076	66.4659	44
45	163.9876	0.0061	0.1207	8.28252	1358.2300	0.00074	8.0572	66.73421	45
46	183.6661	0.0054	0.1207	8.2880	1522.2176	0.000	8.0815	66.97922	46
47	205.7061	0.0049	0.1206	8.29282	1705.8838	0.000	8.1037	67.20284	47
48	230.3908	0.0043	0.1205	8.29716	1911.5898	0.0005	8.12408	67.40684	48
49	258.0377	0.0039	0.1205	8.3010	2141.9806	0.00047	8.1427	67.59286	49
50	289.0022	0.0035	0.1204	8.3045	2400.0182	0.00042	8.15972	67.76241	50

DISCRETE COMPOUND INTEREST FACTORS
Interest Rate 15%

	Single Payment		Uniform Series				Arithmetic Gradient		
N	Compound Amount Factor Find *F* Given *P* *F*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *F* *P*/*F*, *i*, *N*	Capital Recovery Factor Find *A* Given *P* *A*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *A* *P*/*A*, *i*, *N*	Compound Amount Factor Find *F* Given *A* *F*/*A*, *i*, *N*	Sinking Fund Factor Find *A* Given *F* *A*/*F*, *i*, *N*	Uniform Series Factor Find *A* Given *G* *A*/*G*, *i*, *N*	Present Worth Factor Find *P* Given *G* *P*/*G*, *i*, *N*	*N*
1	1.1500	0.8696	1.1500	0.869565	1.0000	1.0000	0.0000	0.0000	1
2	1.3225	0.7561	0.6151	1.625709	2.1500	0.46512	0.46512	0.7561	2
3	1.5209	0.6575	0.4380	2.283225	3.4725	0.2880	0.90713	2.0712	3
4	1.7490	0.5718	0.3503	2.8550	4.9934	0.20027	1.32626	3.7864	4
5	2.0114	0.4972	0.2983	3.352155	6.7424	0.14832	1.72281	5.7751	5
6	2.3131	0.4323	0.2642	3.784483	8.7537	0.11424	2.09719	7.9368	6
7	2.6600	0.3759	0.2404	4.16042	11.0668	0.09036	2.44985	10.1924	7
8	3.0590	0.3269	0.2229	4.487322	13.7268	0.07285	2.78133	12.4807	8
9	3.5179	0.2843	0.2096	4.7716	16.7858	0.05957	3.09223	14.7548	9
10	4.0456	0.2472	0.1993	5.018769	20.3037	0.04925	3.3832	16.9795	10
11	4.6524	0.2149	0.1911	5.233712	24.3493	0.0411	3.65494	19.1289	11
12	5.3503	0.1869	0.1845	5.420619	29.0017	0.03448	3.9082	21.1849	12
13	6.1528	0.1625	0.1791	5.583147	34.3519	0.02911	4.14376	23.1352	13
14	7.0757	0.1413	0.1747	5.724476	40.5047	0.02469	4.36241	24.9725	14
15	8.1371	0.1229	0.1710	5.84737	47.5804	0.0210	4.5650	26.6930	15
16	9.3576	0.1069	0.1679	5.954235	55.7175	0.01795	4.75225	28.2960	16
17	10.7613	0.0929	0.1654	6.047161	65.0751	0.01537	4.92509	29.7828	17
18	12.3755	0.0808	0.1632	6.1280	75.8364	0.01319	5.08431	31.15649	18
19	14.2318	0.0703	0.1613	6.198231	88.2118	0.01134	5.23073	32.42127	19
20	16.3665	0.0611	0.1598	6.259331	102.4436	0.00976	5.36514	33.58217	20
21	18.8215	0.0531	0.1584	6.312462	118.81012	0.00842	5.48832	34.64479	21
22	21.6447	0.0462	0.1573	6.358663	137.6316	0.0073	5.6010	35.6150	22
23	24.8915	0.0402	0.1563	6.398837	159.2764	0.00628	5.7040	36.49884	23
24	28.6252	0.0349	0.1554	6.433771	184.1678	0.00543	5.79789	37.30232	24
25	32.9190	0.0304	0.1547	6.4641	212.7930	0.0047	5.8834	38.03139	25
26	37.8568	0.0264	0.1541	6.490564	245.7120	0.00407	5.96123	38.6918	26
27	43.5353	0.0230	0.1535	6.513534	283.56877	0.00353	6.0319	39.2890	27
28	50.0656	0.0200	0.1531	6.533508	327.10408	0.00306	6.0960	39.82828	28
29	57.5755	0.0174	0.1527	6.550877	377.16969	0.00265	6.15408	40.3146	29
30	66.2118	0.0151	0.1523	6.5660	434.7451	0.0023	6.20663	40.75259	30
31	76.1435	0.0131	0.1520	6.579113	500.95692	0.002	6.25412	41.14658	31
32	87.5651	0.0114	0.1517	6.590533	577.10046	0.0017	6.2970	41.50060	32
33	100.6998	0.0099	0.1515	6.600463	664.66552	0.0015	6.33567	41.81838	33
34	115.8048	0.0086	0.1513	6.609099	765.36535	0.00131	6.37051	42.10334	34
35	133.1755	0.0075	0.1511	6.616607	881.17016	0.00113	6.40187	42.35864	35
36	153.1519	0.0065	0.1510	6.623137	1014.3457	0.0010	6.43006	42.58717	36
37	176.1246	0.0057	0.1509	6.628815	1167.4975	0.0009	6.45539	42.79157	37
38	202.5433	0.0049	0.1507	6.633752	1343.6222	0.00074	6.47812	42.97425	38
39	232.9248	0.0043	0.1506	6.6380	1546.1655	0.00065	6.49851	43.13739	39
40	267.8635	0.0037	0.1506	6.641778	1779.0903	0.00056	6.51678	43.2830	40
41	308.0431	0.0032	0.1505	6.6450	2046.9539	0.0005	6.53313	43.4128	41
42	354.2495	0.0028	0.1504	6.647848	2354.9969	0.00042	6.54777	43.52858	42
43	407.3870	0.0025	0.1504	6.6503	2709.2465	0.00037	6.56086	43.63168	43
44	468.4950	0.0021	0.1503	6.652437	3116.6334	0.00032	6.57255	43.72346	44
45	538.7693	0.0019	0.1503	6.654293	3585.1285	0.00028	6.5830	43.80513	45
46	619.5847	0.0016	0.1502	6.655907	4123.8977	0.00024	6.5923	43.87776	46
47	712.5224	0.0014	0.1502	6.65731	4743.4824	0.00021	6.6006	43.94232	47
48	819.4007	0.0012	0.1502	6.658531	5456.0047	0.0002	6.6080	43.99967	48
49	942.3108	0.0011	0.1502	6.659592	6275.4055	0.00016	6.61461	44.05061	49
50	1083.6574	0.0009	0.1501	6.660515	7217.7163	0.00014	6.62048	44.09583	50

DISCRETE COMPOUND INTEREST FACTORS
Interest Rate 20%

	Single Payment		Uniform Series				Arithmetic Gradient		
N	Compound Amount Factor Find *F* Given *P* *F*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *F* *P*/*F*, *i*, *N*	Capital Recovery Factor Find *A* Given *P* *A*/*P*, *i*, *N*	Present Worth Factor Find *P* Given *A* *P*/*A*, *i*, *N*	Compound Amount Factor Find *F* Given *A* *F*/*A*, *i*, *N*	Sinking Fund Factor Find *A* Given *F* *A*/*F*, *i*, *N*	Uniform Series Factor Find *A* Given *G* *A*/*G*, *i*, *N*	Present Worth Factor Find *P* Given *G* *P*/*G*, *i*, *N*	*N*
1	1.2000	0.8333	1.2000	0.83333	1.0000	1.0000	0.0000	0.0000	1
2	1.4400	0.6944	0.6545	1.52778	2.2000	0.45455	0.45455	0.6944	2
3	1.7280	0.5787	0.4747	2.10648	3.6400	0.2747	0.87912	1.8519	3
4	2.0736	0.4823	0.3863	2.5887	5.3680	0.18629	1.27422	3.2986	4
5	2.4883	0.4019	0.3344	2.99061	7.4416	0.13438	1.64051	4.9061	5
6	2.9860	0.3349	0.3007	3.32551	9.9299	0.10071	1.97883	6.5806	6
7	3.5832	0.2791	0.2774	3.60459	12.9159	0.07742	2.29016	8.2551	7
8	4.2998	0.2326	0.2606	3.83716	16.4991	0.06061	2.57562	9.8831	8
9	5.1598	0.1938	0.2481	4.0310	20.7989	0.04808	2.83642	11.4335	9
10	6.1917	0.1615	0.2385	4.19247	25.9587	0.03852	3.07386	12.8871	10
11	7.4301	0.1346	0.2311	4.32706	32.1504	0.0311	3.28929	14.2330	11
12	8.9161	0.1122	0.2253	4.43922	39.5805	0.02526	3.4841	15.4667	12
13	10.6993	0.0935	0.2206	4.53268	48.4966	0.02062	3.6597	16.5883	13
14	12.8392	0.0779	0.2169	4.61057	59.1959	0.01689	3.81749	17.6008	14
15	15.4070	0.0649	0.2139	4.67547	72.0351	0.01388	3.9588	18.5095	15
16	18.4884	0.0541	0.2114	4.72956	87.4421	0.01144	4.08511	19.3208	16
17	22.1861	0.0451	0.2094	4.77463	105.9306	0.00944	4.19759	20.0419	17
18	26.6233	0.0376	0.2078	4.8122	128.1167	0.00781	4.29752	20.68048	18
19	31.9480	0.0313	0.2065	4.8435	154.7400	0.00646	4.38607	21.2439	19
20	38.3376	0.0261	0.2054	4.86958	186.6880	0.00536	4.46435	21.73949	20
21	46.0051	0.0217	0.2044	4.89132	225.0256	0.00444	4.53339	22.17423	21
22	55.2061	0.0181	0.2037	4.90943	271.0307	0.0037	4.5941	22.5546	22
23	66.2474	0.0151	0.2031	4.92453	326.2369	0.00307	4.6475	22.88671	23
24	79.4968	0.0126	0.2025	4.9371	392.4842	0.00255	4.69426	23.1760	24
25	95.3962	0.0105	0.2021	4.9476	471.9811	0.00212	4.7352	23.42761	25
26	114.4755	0.0087	0.2018	4.95632	567.3773	0.00176	4.77088	23.6460	26
27	137.3706	0.0073	0.2015	4.9636	681.8528	0.00147	4.8020	23.83527	27
28	164.8447	0.0061	0.2012	4.96967	819.2233	0.00122	4.8291	23.99906	28
29	197.8136	0.0051	0.2010	4.97472	984.0680	0.0010	4.85265	24.14061	29
30	237.3763	0.0042	0.2008	4.9789	1181.8816	0.00085	4.87308	24.26277	30
31	284.8516	0.0035	0.2007	4.98245	1419.2579	0.0007	4.89079	24.36809	31
32	341.8219	0.0029	0.2006	4.98537	1704.1095	0.0006	4.9061	24.45878	32
33	410.1863	0.0024	0.2005	4.98781	2045.9314	0.00049	4.91935	24.5368	33
34	492.2235	0.0020	0.2004	4.98984	2456.1176	0.00041	4.93079	24.60384	34
35	590.6682	0.0017	0.2003	4.99154	2948.3411	0.00034	4.94064	24.6614	35
36	708.8019	0.0014	0.2003	4.99295	3539.0094	0.0003	4.94914	24.71078	36
37	850.5622	0.0012	0.2002	4.99412	4247.8112	0.0002	4.95645	24.7531	37
38	1020.6747	0.0010	0.2002	4.9951	5098.3735	0.0002	4.96273	24.78936	38
39	1224.8096	0.0008	0.2002	4.9959	6119.0482	0.00016	4.96813	24.82038	39
40	1469.7716	0.0007	0.2001	4.9966	7343.8578	0.00014	4.97277	24.8469	40
41	1763.7259	0.0006	0.2001	4.9972	8813.6294	0.0001	4.97674	24.8696	41
42	2116.4711	0.0005	0.2001	4.99764	10577.3553	0.0001	4.98015	24.88897	42
43	2539.7653	0.0004	0.2001	4.9980	12693.8263	0.0001	4.98306	24.9055	43
44	3047.7183	0.0003	0.2001	4.99836	15233.5916	0.0001	4.98556	24.91961	44
45	3657.2620	0.0003	0.2001	4.99863	18281.3099	0.0001	4.9877	24.93164	45
46	4388.7144	0.0002	0.2000	4.99886	21938.5719	0.0000	4.9895	24.94189	46
47	5266.4573	0.0002	0.2000	4.99905	26327.2863	0.0000	4.9911	24.95063	47
48	6319.7487	0.0002	0.2000	4.99921	31593.7436	0.0000	4.9924	24.95807	48
49	7583.6985	0.0001	0.2000	4.99934	37913.4923	0.0000	4.99354	24.9644	49
50	9100.4382	0.0001	0.2000	4.99945	45497.1908	0.0000	4.99451	24.96978	50

GEOMETRIC SERIES FACTORS: DISCRETE COMPOUNDING
FUTURE WORTH FACTOR F/A
Interest Rate 15%

N	$C = 4$	$C = 5$	$C = 6$	$C = 8$	$C = 10$	$C = 12$	$C = 20$	N
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2
3	4	4	4	4	4	4	4	3
4	5	5	6	6	6	6	7	4
5	7	8	8	8	8	9	10	5
6	10	10	10	11	11	12	14	6
7	13	13	13	14	15	16	20	7
8	16	17	17	18	19	21	27	8
9	20	21	21	23	25	27	37	9
10	25	26	27	29	31	34	49	10
11	30	31	33	36	39	43	64	11
12	37	38	40	44	48	53	83	12
13	44	46	48	53	59	66	108	13
14	53	56	58	65	73	82	139	14
15	63	67	70	79	88	100	178	15
16	75	80	84	95	107	122	227	16
17	90	95	100	113	129	149	288	17
18	106	112	119	136	156	181	365	18
19	125	133	142	162	187	219	460	19
20	148	157	168	193	224	264	579	20
21	174	185	198	229	268	318	728	21
22	204	218	234	271	319	381	912	22
23	240	256	275	321	380	457	1141	23
24	281	301	324	379	451	547	1425	24
25	329	353	380	447	535	653	1778	25
26	385	414	446	527	634	779	2214	26
27	450	484	523	620	750	928	2753	27
28	526	566	613	729	887	1104	3420	28
29	614	662	718	857	1047	1311	4244	29
30	716	774	840	1006	1234	1556	5261	30
31	836	903	982	1179	1454	1844	6516	31
32	974	1054	1147	1382	1711	2184	8064	32
33	1136	1230	1339	1619	2013	2584	9970	33
34	1323	1434	1563	1895	2366	3054	12319	34
35	1541	1672	1824	2217	2779	3608	15210	35
36	1795	1948	2127	2592	3262	4258	18769	36
37	2089	2269	2480	3029	3826	5023	23146	37
38	2432	2643	2891	3539	4486	5920	28527	38
39	2830	3077	3369	4132	5256	6974	35142	39
40	3293	3582	3924	4824	6156	8210	43270	40
41	3830	4169	4570	5629	7207	9660	53254	41
42	4455	4852	5322	6567	8434	11361	65513	42
43	5182	5645	6195	7658	9865	13354	80562	43
44	6026	6567	7211	8929	11536	15689	99031	44
45	7007	7639	8392	10407	13484	18425	121692	45
46	8147	8885	9765	12128	15756	21628	149489	46
47	9472	10332	11361	14130	18405	25377	183579	47
48	11011	12015	13216	16460	21494	29766	225377	48
49	12800	13970	15373	19171	25094	34900	276615	49
50	14879	16243	17880	22323	29289	40906	339415	50

GEOMETRIC SERIES FACTORS: DISCRETE COMPOUNDING
PRESENT VALUE FACTOR: P/A
Interest Rate 15%

N	$C = 4$	$C = 5$	$C = 6$	$C = 8$	$C = 10$	$C = 12$	$C = 20$	N
1	0.8607	0.8607	0.8607	0.8607	0.8607	0.8607	0.8607	1
2	1.6318	1.6395	1.6473	1.6632	1.6794	1.6960	1.7655	2
3	2.3225	2.3442	2.3663	2.4115	2.4582	2.5066	2.7168	3
4	2.9413	2.9818	3.0233	3.1092	3.1991	3.2932	3.7168	4
5	3.4956	3.5588	3.6238	3.7597	3.9037	4.0566	4.7680	5
6	3.9922	4.0808	4.1726	4.3662	4.5741	4.7974	5.8732	6
7	4.4370	4.5532	4.6742	4.9317	5.2117	5.5163	7.0351	7
8	4.8356	4.9806	5.1326	5.4590	5.8182	6.2140	8.2565	8
9	5.1926	5.3673	5.5515	5.9507	6.3952	6.8910	9.5405	9
10	5.5124	5.7173	5.9344	6.4091	6.9440	7.5481	10.8903	10
11	5.7989	6.0339	6.2844	6.8365	7.4660	8.1857	12.3094	11
12	6.0556	6.3204	6.6042	7.2350	7.9626	8.8045	13.8012	12
13	6.2855	6.5797	6.8965	7.6066	8.4350	9.4050	15.3695	13
14	6.4915	6.8142	7.1636	7.9530	8.8843	9.9877	17.0183	14
15	6.6760	7.0265	7.4078	8.2761	9.3117	10.5533	18.7515	15
16	6.8413	7.2185	7.6309	8.5773	9.7183	11.1021	20.5736	16
17	6.9894	7.3923	7.8348	8.8581	10.1050	11.6347	22.4892	17
18	7.1220	7.5495	8.0212	9.1199	10.4729	12.1515	24.5029	18
19	7.2409	7.6918	8.1915	9.3641	10.8229	12.6531	26.6199	19
20	7.3473	7.8206	8.3472	9.5917	11.1557	13.1399	28.8455	20
21	7.4427	7.9370	8.4895	9.8040	11.4724	13.6122	31.1851	21
22	7.5281	8.0424	8.6195	10.0019	11.7736	14.0706	33.6447	22
23	7.6046	8.1378	8.7383	10.1864	12.0601	14.5155	36.2304	23
24	7.6732	8.2241	8.8470	10.3584	12.3326	14.9472	38.9487	24
25	7.7346	8.3022	8.9462	10.5189	12.5918	15.3661	41.8064	25
26	7.7897	8.3728	9.0369	10.6684	12.8384	15.7727	44.8105	26
27	7.8389	8.4368	9.1198	10.8079	13.0730	16.1673	47.9687	27
28	7.8831	8.4946	9.1956	10.9379	13.2961	16.5502	51.2888	28
29	7.9227	8.5469	9.2649	11.0591	13.5084	16.9217	54.7792	29
30	7.9581	8.5943	9.3282	11.1722	13.7103	17.2823	58.4485	30
31	7.9898	8.6372	9.3860	11.2776	13.9023	17.6323	62.3059	31
32	8.0183	8.6759	9.4389	11.3759	14.0850	17.9719	66.3611	32
33	8.0438	8.7110	9.4872	11.4675	14.2588	18.3014	70.6242	33
34	8.0666	8.7428	9.5313	11.5529	14.4241	18.6212	75.1059	34
35	8.0870	8.7715	9.5717	11.6326	14.5813	18.9316	79.8174	35
36	8.1053	8.7975	9.6086	11.7069	14.7309	19.2328	84.7704	36
37	8.1218	8.8210	9.6423	11.7761	14.8732	19.5251	89.9774	37
38	8.1365	8.8423	9.6731	11.8407	15.0085	19.8088	95.4513	38
39	8.1496	8.8615	9.7013	11.9009	15.1372	20.0840	101.2059	39
40	8.1614	8.8789	9.7270	11.9570	15.2597	20.3512	107.2556	40
41	8.1720	8.8947	9.7505	12.0094	15.3762	20.6104	113.6154	41
42	8.1814	8.9090	9.7720	12.0582	15.4870	20.8620	120.3013	42
43	8.1899	8.9219	9.7916	12.1037	15.5924	21.1061	127.3300	43
44	8.1975	8.9336	9.8096	12.1461	15.6926	21.3431	134.7190	44
45	8.2043	8.9441	9.8260	12.1856	15.7880	21.5730	142.4869	45
46	8.2104	8.9537	9.8410	12.2225	15.8787	21.7961	150.6531	46
47	8.2159	8.9623	9.8547	12.2569	15.9650	22.0126	159.2380	47
48	8.2208	8.9702	9.8672	12.2890	16.0471	22.2228	168.2630	48
49	8.2252	8.9773	9.8787	12.3189	16.1252	22.4267	177.7507	49
50	8.2291	8.9837	9.8891	12.3468	16.1995	22.6246	187.7249	50

CHAPTER 91

Economic Risk Analysis*

G. A. FLEISCHER
University of Southern California

1. **INTRODUCTION** **2361**

2. **SENSITIVITY ANALYSIS** **2361**
   - 2.1. Numerical Example: Certainty Analysis 2361
   - 2.2. Classical Sensitivity Analysis: Single Variable 2362
     - 2.2.1. Algebraic Solution 2362
     - 2.2.2. Graphical Presentation 2362
     - 2.2.3. Percent Deviation Graph 2363
   - 2.3. Sensitivity to Two Parameters Considered Simultaneously 2364
   - 2.4. Sensitivity to More Than Two Parameters 2366

3. **RISK ANALYSIS** **2367**
   - 3.1. Alternative Risk Measures 2367
   - 3.2. Determining the Probability Distribution for Present Worth 2367
     - 3.2.1. Expected Present Worth 2367
     - 3.2.2. Variance of Present Worth 2368
   - 3.3. Cash Flows with Uncertain Timing 2369
     - 3.3.1. Single Cash Flow 2369
     - 3.3.2. Uncertain Initiation and Duration 2370
   - 3.4. Uncertain Project Life and Uncertain Cash Flow 2371
   - 3.5. Other Models 2371
   - 3.6. Analysis Based on the Probability Distribution for Present Worth 2371
     - 3.6.1. Discrete Distribution for Present Worth 2372
     - 3.6.2. Using Only the Mean and Variance of the PW Distribution 2373
     - 3.6.3. When the Normal Distribution Can be Assumed 2374
   - 3.7. Comparing Risky Proposals 2376

4. **DECISION THEORY APPLICATIONS** **2376**
   - 4.1. Problem Statement 2376
   - 4.2. Dominance 2377
   - 4.3. Principles for Decisions under Risk 2377
     - 4.3.1. The Principle of Expectation 2377
     - 4.3.2. The Principle of Most Probable Future 2378
     - 4.3.3. The Aspiration Level Principle 2378
   - 4.4. Principles for Decisions under Uncertainty 2378
     - 4.4.1. The Minimax (or Maximin) Principle 2378
     - 4.4.2. The Minimin (or Maximax) Principle 2379
     - 4.4.3. The Hurwicz Principle 2379

---

*Portions of this chapter are adopted substantially unchanged from the *Handbook of Industrial Engineering,* 2nd ed., Chapter 52. That material, in turn, was based in large part on the *Handbook of Industrial Engineering,* 1st ed., Chapter 9.5, by James R. Buck and Jose M. Tanchoco, as well as other material adopted, by permission of the publisher, from G. A. Fleischer, *Engineering Economy,* Chapter 8, PWS-Kent Publishing Company, Boston, 1984.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

4.4.4. The Laplace Principle (Insufficient Reason) 2380
4.4.5. The Savage Principle (Minimax Regret) 2381
4.5. Summary of Results 2382
5. DECISION TREES 2382
5.1. Deterministic Decision Trees 2382
5.2. Stochastic Decision Trees 2385
6. DIGITAL COMPUTER (MONTE CARLO) SIMULATION 2385
6.1. Sampling from a Discrete Distribution 2385
6.2. Sampling from a Normal Distribution 2386
6.3. General Framework 2386
6.4. Numerical Example 2388
7. OTHER APPROACHES FOR DEALING WITH THE UNCERTAIN/RISKY FUTURE 2391
7.1. Increasing the Minimum Attractive Rate of Return 2391
7.2. Differentiating Rates of Return by Risk Class 2391
7.3. Decreasing the Expected Project Life 2392
7.4. Utility Models 2392
REFERENCES 2392
ADDITIONAL READING 2392

1. INTRODUCTION

Various methods of analysis for economic justification are shown in Chapter 90 based on the assumption that all of the component cash flows for the proposed investment are known and certain. However, in most cases the amount and timing of these cash flows are estimated, and uncertainties exist in the estimation process. Furthermore, there is usually more uncertainty with some component cash flows than others, and some of these component flows affect the economic criteria more than others. Thus, additional methodologies and concepts are needed for economic analysis when explicit information on the effects of uncertainties in the timing and amounts of the cash flows is important. These methodologies and concepts are the focus of this chapter.

Numerous factors contribute to the uncertainties in the estimates of the amount and timing of component cash flows. Delivery or construction delays, unexpected bottlenecks in new projects, inflationary or recessionary pressures, labor negotiations, and problems in R&D are but a few examples of changes that can and do occur to alter the amounts and timing of disbursements and receipts of monies. Although these possibilities are usually recognized during the early planning phases of a project, the actual cash flows are uncertain, and there is a risk associated with the resulting project's present worth, benefit–cost ratio, or other measure of economic merit being used. Since this economic risk is as important to the decision maker as the other aspects of economic analysis, explicit information regarding the risk should be developed as part of the analysis. Approaches to this form of analysis and some of the relevant techniques are described in this chapter.

A variety of measures have been proposed for dealing with a noncertain operating environment, that is, where the relevant parameters of the analytical model cannot be assumed with certainty. The relevant literature is very extensive, and an encyclopedic treatment is beyond the scope of this chapter. Our discussion will be limited, therefore, to a limited number of concepts related for their popularity among practitioners and because they are representative of the spectrum of possible approaches to this issue. We begin with *sensitivity analysis,* that technique which, surveys show, appears to be most commonly used in industry.

2. SENSITIVITY ANALYSIS

Sensitivity analysis is the process whereby one or more system input variables are changed and corresponding changes in the system output, or figure of merit, are observed. If a decision is changed as a certain input is varied over a reasonable range of possible values, the decision is said to be *sensitive* to that input; otherwise it is *insensitive.*

The term *break-even analysis* is often used to express the same concept for a single input variable. Here, the value of the input variable at which the decision is changed is determined. If the break-even point lies within the range of expected values, the decision is said to be sensitive to that point. Thus, sensitivity and the break-even point are directly related.

2.1. Numerical Example: Certainty Analysis

A manufacturing firm is considering the introduction of a new product to be produced and sold over a 15-year period. The initial cost of capital facilities is $100,000; the anticipated net salvage value at the end of 15 years is $20,000. It is expected that 7000 units will be produced each year at a cost of \$10 per unit and sold at \$12 per unit. The firm's minimum attractive rate of return (MARR) is 10% per year.

The anticipated "profitability" of this proposed investment can be measured by present worth (PW) as follows.

$$\text{PW} = Q(r - c)(P/A, i, N) - P + S(P/F, i, N) \tag{1}$$

where Q = quantity sold per year
r = revenue per unit
c = cost per unit
P = initial cost of capital facilities
S = net salvage value of capital facilities
N = project life, in years
i = MARR, the discount rate per year

[Note the factor (P/A, i, N) is the *functional* form of the uniform series present worth factor, the *algebraic* form of which is $((1 + i)^N - 1)/(i(1 + i)^N)$. Similarly, the functional form of the single payment present worth factor, (P/F, i, N), represents the algebraic form $(1 + i)^{-N}$. See Chapter 90 for additional discussion.]

Assuming the "certainty estimates" for these seven parameters as described in the preceding paragraph, the solution is

$$\begin{aligned}\text{PW} &= 7000(\$12 - \$10)(P/A, 10\%, 15) - \$100{,}000 + \$20{,}000(P/F, 10\%, 15)\\ &= \$14{,}000(7.606) - \$100{,}000 + \$20{,}000(0.2394)\\ &= \$11{,}273\end{aligned}$$

Since the PW is positive, we conclude that the proposal appears to be economically attractive. This result, of course, is based on the presumption that all of the parameter values assumed for the analysis will in fact occur as anticipated.

2.2. Classical Sensitivity Analysis: Single Variable

2.2.1. Algebraic Solution

Suppose there is some reason to question the validity of the assumption concerning the number of units produced and sold annually. Additional investigation, for example, may suggest that the "certainty estimate" of 7,000 units per year is questionable; it now appears that this parameter value could occur anywhere over the range of 6,000 to 7,500 units. With this new information the resulting range of values for the present worth is

$$\begin{aligned}\text{Min PW} &= 6000(\$2)(7.606) - \$95{,}212 = -\$3{,}940\\ \text{Max PW} &= 7500(\$2)(7.606) - \$95{,}212 = \$18{,}878\end{aligned}$$

The break-even point can be determined by determining that value of $Q = Q_0$ such that PW = 0:

$$\text{PW} \equiv 0 = Q_0(\$2)(7.606) - \$95{,}212$$

Solving,

$$Q_0 = \$95{,}212/\$15.212 = 6259 \text{ units}$$

Since the break-even point lies within the range ($6000 < 6259 < 7500$), the decision *is* sensitive to the estimate for Q.

2.2.2. Graphical Presentation

Sensitivity analyses are usually presented in graphical format. Indeed, it is this "power of pictures" that probably accounts for its widespread popularity. The graphical portrayal of sensitivity of PW to the variable Q in our example is illustrated in Figure 1. The linear function in the figure is the graph of

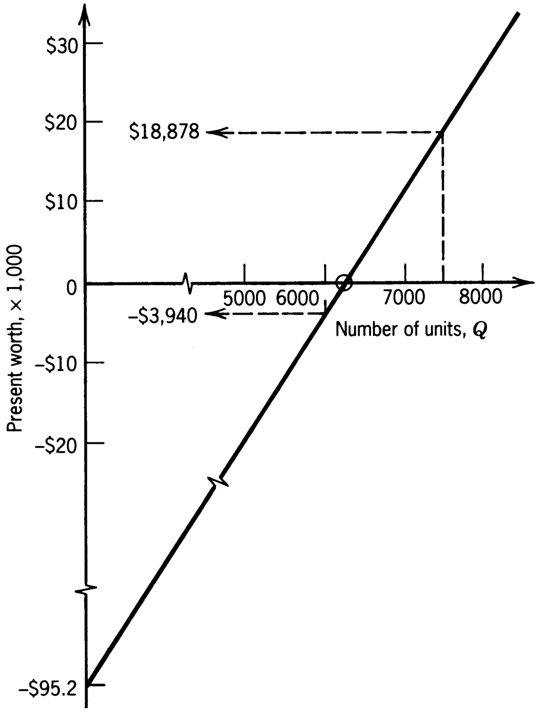


Figure 1 Present Worth as a Function of Number of Units Produced and Sold Annually. (Break-even = 6259 units)

$$PW = \$15.212Q - \$95,212 \qquad 0 \leq Q \leq 9,000$$

Note the break-even point at $Q_0 = 6259$. Also note that the range for Q is highlighted at $Q(\text{min}) = 6000$ and $Q(\text{max}) = 7500$.

2.2.3. Percent Deviation Graph

An alternative approach is a plot of the figure of merit—here, the present worth (PW)—as a function of the *percent deviation* of the variable of interest In the example let p_Q = the percent deviation of Q such that

$$\begin{aligned} PW &= \$15.212(7000)(1 - p_Q) - \$95,212 \\ &= \$11,272 - \$106,485p_Q \end{aligned} \tag{2}$$

The function is graphed in Figure 2. Also shown in the figure are similar graphs for percent deviation for revenue per unit (p_r), cost per unit (p_c), and the number of units produced and sold annually (p_Q).

Although percent deviation graphs for one or more variables may be shown in a single illustration, it should be emphasized that sensitivity to only one variable at a time is being examined. The graph of PW as a function of p_Q, for example, is based on the assumptions that *all* other variables (r, c, P, S, N, i) are held constant at their "certainty estimates." When sensitivity to p_r is being examined, we set $Q = 7000$. And so on.

One notable advantage of the percent deviation graph is that it makes apparent the relative degree of sensitivity for the various parameters. The greater the slope (steepness of the function) the more likely is the decision to be sensitive to that parameter, that is, the break-even point for percent deviation will be relatively small. In Figure 2 it is apparent that the decision is somewhat more sensitive to per unit revenue (r) and cost (c) and is relatively insensitive to number of years of service (N). This conclusion may be misleading, however, because it is based on the presumption of equal

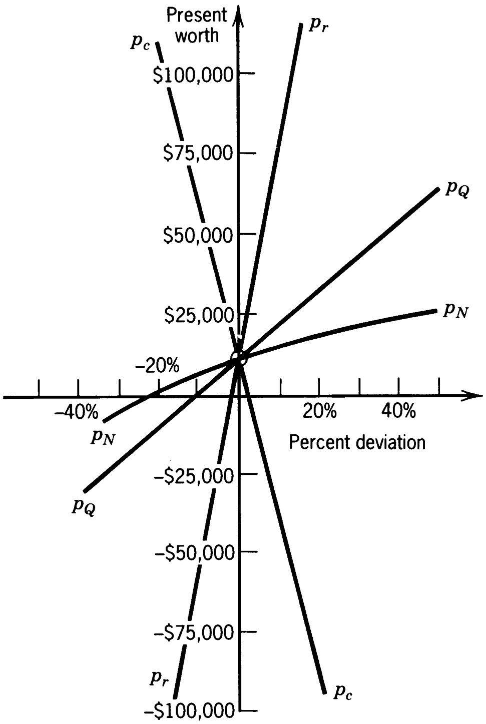


Figure 2 Present Worth as Function of Percent Deviation in Estimates for r, c, Q, and N.

likelihoods of deviation for the various parameters. To illustrate, we found that the break-even percent deviations are about -23% for p_N and -11% for p_Q. But suppose that there is evidence to suggest that:

Parameter	Certainty Estimate	Range	Deviation
Quantity	7000 units	6000 to 7500	-14 to $+7\%$
Life	15 years	10 to 18	-33 to $+20\%$

Thus, it would appear that the decision maker would be well advised to give careful attention to the assumption concerning service life (N) as well as quantity produced (Q). The point here is that the *range* of interest for percent deviation may be different for different parameters.

2.3. Sensitivity to Two Parameters Considered Simultaneously

Suppose that our decision maker in this example is concerned about the sensitivity to the revenue per unit (r) as well as the number of units produced and sold (Q). Considering these two parameters, now variables, simultaneously,

$$\text{PW} = Q(r - \$10)(7.606) - \$95{,}212 \tag{3}$$

As before, assume $6000 \leq Q \leq 7500$, and assume further that $\$11.25 \leq r \leq \$12.50$.

One approach to sensitivity analysis for two variables considered simultaneously is to construct a three-dimensional graph with the x and y axes representing the two variables and the z axis serving as the figure of merit. The combined function is now a *surface* and we now have a break-even *line.* But three-dimensional graphs are difficult to construct and generally harder to interpret. A useful alternative is a variant of the two-dimensional graph as illustrated in Figure 3. One of the two variables is represented along the x axis. The second variable is reflected by a family of curves, specifically, curves based on the maximum and minimum values of the variable.

The two functions plotted in Figure 3 are

$$\mathrm{PW} = Q(\$12.50 - \$10)(7.606) - \$95{,}212$$

and

$$\mathrm{PW} = Q(\$11.25 - \$10)(7.606) - \$95{,}212$$

These represent the upper and lower bounds of the r variable, respectively. Two additional vertical lines are drawn at the lower and upper bounds of the Q variable, at 6000 and 7500 units. The polygon thus formed contains all possible combinations of r and Q, and the maximum and minimum values of the figure of merit (PW) can be readily determined. The decision is insensitive if the polygon lies either wholly above the x axis (PW = \$0) or wholly below the x axis.

One problem in the interpretation of sensitivity graphs can be illustrated by this numerical example. It would appear from Figure 3 that, since the area of the polygon lying above the x axis is roughly the same as the area lying below the line, the likelihood of making money on this project (PW > \$0) is about the same as the likelihood of losing money. Implicit in this conclusion is the assumption that all points in the polygon are equally probable. But this is not necessarily the case. Indeed, it would be reasonable to assume that there is an inverse relationship between price per unit and quantity sold, so that Q would decrease as r increases. This dependency is not reflected in the graph.

The simultaneous consideration of sensitivity to two variables can also be displayed in a percent deviation format. In Figure 4, the percent deviations for each of the variables are shown on the x and y axes. The function plotted is

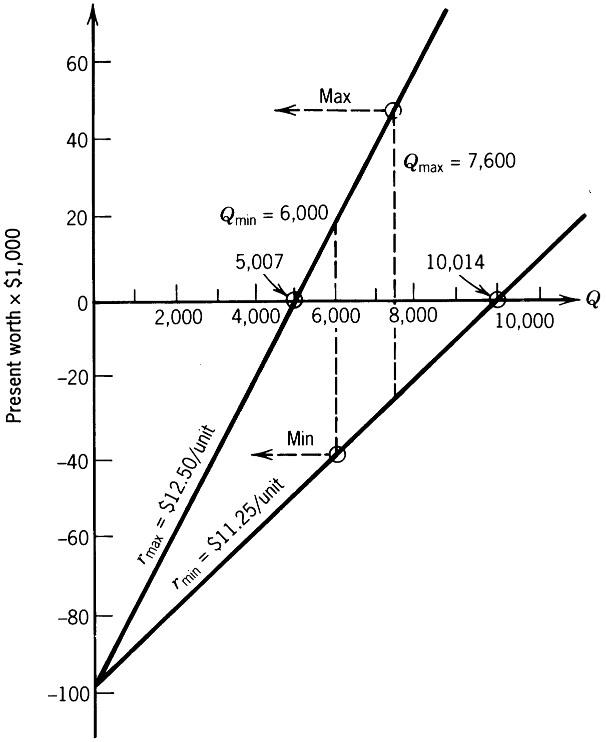


Figure 3 Present Worth as a Function of Units Produced (Q) and Revenue per Unit (r).

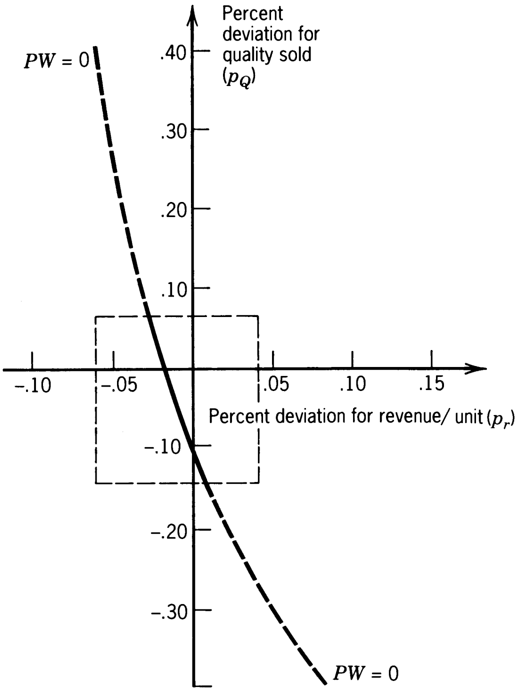


Figure 4 Percent Deviation Graph for Two Variables: Quantity Sold and Revenue per Unit.

$$\mathrm{PW} = 7000(1 + p_Q)[\$12(1 + p_r) - \$10](7.606) - \$95{,}212 = 0$$

This is an *indifference curve,* the locus of all points (p_r, p_Q) such that PW = \$0, which is in fact the *break-even line.* The solid portion of the line represents the set of possible outcomes: ($-6\% \leq p_r \leq +4\%$) and ($-14\% \leq p_Q \leq +7\%$).

2.4. Sensitivity to More Than Two Parameters

Using the previous example, suppose that, in addition to uncertainty about quantity sold (Q) and revenue per unit (r), there is also uncertainty as to the cost per unit (c). Suppose that the following range of values is possible:

Parameter	Minimum	Most likely	Maximum
Quantity	6000	7000	7500
Revenue/unit	\$11.25	\$12.00	\$12.50
Cost/unit	\$ 9.00	\$10.00	\$11.00

As mentioned previously, a percent deviation graph, as in Figure 2, permits the plotting of the figure of merit (PW) as a function of the percent deviation from the most likely value for any number of parameters. However, the *interactive* effects of the parameter are ignored.

It is possible, of course, to reduce the original problem to a series for two-dimensional graphs. Here, for example, consider: (1) PW as a function of quantity, assuming C = \$9.00, and a family of curves for r = \$11.25 and \$12.50; and (2) PW as a function of quantity assuming C = \$11.00, and a family of curves for r = \$11.25 and \$12.50. This approach suffers from two defects. First, although a series of smaller problems is solved, we are not testing for the sensitivity of *all* parameters *simultaneously.* Second, the number of graphs required grows exponentially as the number of uncertain parameters increases arithmetically.

A second approach is based on the *a fortiori* (“strength of the argument”) principle. If it can be shown that a certain course of action is indicated regardless of the input assumptions, then it has been proven, *a fortiori,* that there can be no other possible outcome. To illustrate, the following is a computation of both the minimum and maximum possible values for PW, given the ranges for the input assumptions:

$$\text{Min PW} = 6000(\$11.25 - \$11.00)(7.606) - \$95{,}212 = -\$83{,}803$$
$$\text{Max PW} = 7500(\$12.50 - \$9.00)(7.606) - \$95{,}212 = \$104{,}446$$

If both present worths had been negative, we would have proven, *a fortiori,* that the proposal should be rejected on economic grounds. Conversely, if both PW values had been positive, an “accept” decision would have been indicated.

Unfortunately, this test of extreme values rarely yields a clear result, and the *a fortiori* argument cannot be used. Nevertheless, analysts would be well advised to try this approach before proceeding further. The calculations can be completed relatively easily, and the few cases for which a clear signal *is* indicated more than justify the time involved.

3. RISK ANALYSIS

3.1. Alternative Risk Measures

A number of different statistics have been proposed for the measure of “riskiness” of proposed plans, programs, and projects. Perhaps the most widely used measure is the *variance* (*or standard deviation*) *of the prospective return,* where return is generally the present worth, internal rate of return, and so forth. The variance (σ^2) of the distribution for a continuous random variable x is given by

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \qquad (4)$$

Large variance signifies large risk; relatively small variance indicates relatively small risk. In general, everything else being equal, risk is to be minimized.

An alternative view is that the *semivariance* is a preferred statistic as it focuses on the variability in negative return, that is, on the reduction of losses. The semivariance (S_h) of a distribution for a random variable X is given by

$$S_h = \int_{-\infty}^{h} (h - x)^2 f(x)dx \qquad (5)$$

Still another measure of risk is the *probability of loss,* a statistic that measures the probability that the return will lie below some predetermined critical level, h. The probability of loss (L) for a continuous random variable is given by

$$L = \int_{-\infty}^{h} f(x)dx \qquad (6)$$

Limited space precludes a full discussion of these (and other) risk measures. Therefore, in the remainder of this section we will limit our remarks to the variance, a statistic that has proven most popular in use as well as in the literature of engineering economy.

3.2. Determining the Probability Distribution for Present Worth

3.2.1. Expected Present Worth

Consider an uncertain stream of cash flows, A_j, occurring at the end of periods 1, 2, . . . j, . . . , N. If the project life, N, the discount rate, i, and the amounts and timing of the cash flows are known with certainty, then

$$\text{PW} = \sum_{j=0}^{N} A_j(1 + i)^{-j} \qquad (7)$$

Now suppose that the cash flows are random variables with associated probability or density functions $f(A_j)$. The PW is a function of random variables, so it is itself a random variable with mean, μ_p,

$$\mu_p = \text{Exp[PW]} = \sum_{j=0}^{N} \mu_j(1 + i)^{-j} \tag{8}$$

where $\mu_j = \text{Exp}[A_j]$ for $j = 0, 1, \ldots, N$.

3.2.2. *Variance of Present Worth*

The variance of the PW distribution depends upon the degree of correlation between the individual cash flows. In general

$$\sigma_p^2 = \text{Var[PW]} = \sum_{j=0}^{N} \sigma_j^2(1 + i)^{-2j} + 2 \sum_{j=0}^{N-1} \sum_{k=j+1}^{N} \rho_{jk}\sigma_j\sigma_k(1 + i)^{j+k} \tag{9}$$

where ρ_{jk} is the correlation coefficient between cash flows, A_j and A_k and σ_j and σ_k are the standard deviations of the distribution of A_j and A_k, respectively. This formulation is intractable in practice because of the difficulty, if not impossibility, in estimating the correlation coefficients. However, formulations of the variance under the two extreme cases—independent cash flows ($\rho_{jk} = 0$) and perfectly correlated cash flows ($\rho_{jk} = 1$)—is helpful, as will be shown.

3.2.2.1. Independent Cash Flows If there is no causative or consequential relationship between the cash flows, they are said to be independent and

$$\sigma_p^2 = \sum_{j=0}^{N} \sigma_j^2(1 + i)^{-2j} \tag{10}$$

A numerical example, summarized in Table 1, illustrates Eqs. (9) and (10). This is a five-period project life with means (μ_j) and variances (σ_j^2) of the cash flows as shown. The results: $\mu_p = \$69.44$ and $\sigma_p = \sqrt{\$802.9325} = \$28.34$.

3.2.2.2. Perfectly Correlated Cash Flows Cash flows in any two periods, x and y, are perfectly correlated if, given that A_x is the actual value of $\mu_x + d\sigma_x$, then

$$A_y = \mu_y + d\sigma_y$$

In words, if random factors cause A_x to deviate from its mean value by d standard deviations, the same factors will cause A_y to deviate from its mean in the same direction by d standard deviations. Under these conditions

$$\sigma_p = \sum_{j=0}^{N} \sigma_j(1 + i)^{-j} \tag{11}$$

To illustrate, consider the example summarized in Table 2. Assuming a 10% discount rate, the expected value of the PW of the five cash flows is \$625.92, and the standard deviation of the PW is \$75.82. Note that the expected value of PW is given by Eq. (8) and is independent of the degree of correlation.

TABLE 1 Numerical Example: Determining the Mean and Variance of PW Given Probabilistic Cash Flows and 10% Discount Rate—Independent Cash Flows

End of Period	Cash Flow Estimates			Present Worth at 10%	
j	Mean μ_j	Variance σ_j^2		$\mu^j(1.10)^{-j}$	$\sigma_j^2(1.10)^{-2j}$
0	−\$400	\$\$ 0^2		−\$400.00	\$\$ 0.0000
1	100	10^2		90.91	82.6446
2	130	15^2		107.44	153.6780
3	160	20^2		120.21	225.7896
4	130	20^2		88.79	186.6030
5	100	20^2		62.09	154.2173
			Totals	\$ 69.44	\$\$802.9325

TABLE 2 Numerical Example: Determining the Mean and Standard Deviation of PW Given Probabilistic Cash Flows and 10% Discount Rate—Perfectly Correlated Cash Flows

End of Period	Cash Flows			Present Worth at 10%	
	Mean	Standard Deviation			
j	μ_j	σ_j		$\mu_j(1.0)^{-j}$	$\sigma_j(1.10)^{-j}$
1	$100	20		$ 90.91	$18.18
2	150	20		123.97	16.53
3	200	20		150.26	15.03
4	200	20		136.60	13.66
5	200	20		124.18	12.42
			Totals	$625.92	$75.82

3.2.2.3. Combining Independent and Perfectly Correlated Cash Flows Suppose that it is feasible, in a given problem situation, to identify two types of cash flows: those that are statistically independent and those that are perfectly correlated. In this case the variance of the PW distribution is the sum of (a) the sum of the variances of the independent cash flows, discounted, and (b) the sum of the variances of each of the subsets of perfectly correlated cash flows, where the variance of each subset is the square of the sum of the standard deviations of the cash flows in that subset. That is,

$$\sigma_p^2 = \sum_{j=0}^{N} \sigma_j^2(1 + i)^{-2j} + \sum_{k-1}^{M} \left\{ \sum_{j=0}^{N} [\sigma_{jk}(1 + i)^{-j}] \right\}^2 \tag{12}$$

where σ_j^2 = variance of the distribution of the independent A_j's
σ_{jk} = standard deviation of the distribution of the perfect correlated cash flows in subset k, $k = 1, 2, \ldots, M$

Returning to the previous example (Table 2), suppose that there is a cash flow A_0 such that $\mu_0 = -\$500$ and $\sigma_0 = \$10$, and A_0 is independent of the positive cash flows in periods 1 through 5. All cash flows for the proposal are now completely specified, and

$$\mu_p = \sum_{j=0}^{5} \mu_j(1.10)^{-j} = -\$500 + \$625.93 = \$125.93$$

$$\sigma_p = \$10 + \$75.82 = \$85.82$$

Note in this example that $M = 1$; there is only one subset of perfectly correlated cash flows.

3.3. Cash Flows with Uncertain Timing

3.3.1. Single Cash Flow

Consider a single (impulse) cash flow, F, occurring at time t. If both t and F are deterministic, and assuming that interest is compounded/discounted continuously at nominal rate r per period, then the present worth is given by

$$\text{PW} = Fe^{-n} \tag{13}$$

If the timing, t, is a random variable with probability density function $f(t)$, then

$$\mu_p = \text{Exp[PW]} = F \int_0^\infty f(t)e^{-n}\, dt \tag{14}$$

and

$$\sigma_p^2 = \text{Var[PW]} = F^2 \int_0^\infty f(t)e^{-n}\, dt - \{F[\text{Exp}(e^{-n})]\}^2 \tag{15}$$

When the cash flow, F, is also a random variable with known μ_F and σ_F^2, then the PW is the product of two random variables. Determination of μ_p and σ_p^2 results from a straightforward application of probability theory with respect to products of random variables.

3.3.2. Uncertain Initiation and Duration

Consider a uniform continuous cash flow, A, which begins at time m and continues for an uncertain duration t. Assume that m and t are statistically independent random variables with known probability functions $f(m)$ and $f(t)$. It may be shown that

$$\mu_p = (\overline{A}/r)[\mathrm{Exp}(e^{-rm})][1 - \mathrm{Exp}(e^{-rt})]$$

and

$$\sigma_p^2 = (\overline{A}/r)^2[\mathrm{Var}(e^{-rm}) + \mathrm{Var}(e^{-r(m+t)})] \tag{16}$$

To illustrate, consider a uniform cash flow of \$1000 per year beginning at some uncertain time m and continuing for a duration of t years. The delay to initiation is *uniformly* distributed between 6 months and 1 year. The project duration is *gamma* distributed with mean of 3 years and standard deviation of 1 year; the parameters of the gamma distribution yielding these statistics are $a = 3$ and $b = 9$. The nominal interest rate is 10% compounded continuously. It is assumed that the initiation time and project duration are independent random variables. Our problem is to determine the equivalent present value of these cash flows. (This problem is taken from Park and Sharp-Bette (1990, p. 411)

This problem may be solved by use of integral calculus in connection with Eq. (16). However, it may be instructive to use *Laplace transform* methodology to evaluate μ_p and σ_p^2. If a function $f(x)$ is considered to be piecewise continuous, then the Laplace transform of the function, written $\mathcal{L}\{f(x)\}$, is defined as a function $F(s)$ of the variable r by the integral

$$\mathcal{L}\{f(x)\} = F(r) = \int_0^\infty f(x)e^{-rx}\,dx = \mathrm{Exp}(e^{-rx})$$

over the range of values of r for which the integral exists. For the *uniform* distribution

$$F(x) = \frac{e^{-ra} - e^{-rb}}{r(b - a)}$$

and for the *gamma* distribution

$$F(r) = \left[1 + \left(\frac{r}{a}\right)\right]^{-b}$$

Returning to our example, for the uniformly distributed delay time, m,

$$\mathrm{Exp}(e^{-rm}) = \mathcal{L}(r)_m = \frac{e^{-0.10(0.5)} - e^{-0.10(1.0)}}{0.10(1.0 - 0.5)} = 0.92784$$

and

$$\begin{aligned}\mathrm{Var}(e^{-rm}) &= \mathcal{L}(2r) - \mathcal{L}(r)^2 \\ &= \frac{e^{-0.2(0.5)} - e^{-0.2(1.0)}}{0.2(1.0 - 0.5)} - (0.92784)^2 \\ &= 0.86107 - 0.86089 = 0.00018\end{aligned}$$

Similarly, considering the gamma-distributed random variable, t,

$$\mathrm{Exp}(e^{-rt}) = \mathcal{L}(r)_t = [1 + (0.10/3)]^{-9} = 0.74445$$

and

$$\begin{aligned}\mathrm{Var}(e^{-rt}) &= [1 + (0.20/3)]^{-9} - (0.74445)^2 \\ &= 0.559425 - 0.554206 = 0.00522\end{aligned}$$

Next, we must determine

$$
\begin{aligned}
\mathrm{Var}(e^{-r(m+t)}) &= \mathrm{Var}(e^{-rm}e^{-rt}) \\
&= \mathcal{L}(r)_m^2[\mathrm{Var}(e^{-n}) + \mathcal{L}(r)_t^2[\mathrm{Var}(e^{-rm})] + [\mathrm{Var}(e^{-rm})][\mathrm{Var}(e^{-n})] \\
&= (0.92784)^2(0.00522) + (0.74445)^2(0.00018) + (0.0018)(0.00522) \\
&= 0.00459
\end{aligned}
$$

Finally, using Eq. (16),

$$
\begin{aligned}
\mu_p &= (\$1{,}000/0.10)(0.92784)(1 - 0.74455) \\
&= \$2{,}371
\end{aligned}
$$

and

$$
\begin{aligned}
\sigma_p^2 &= (\$1{,}000/0.10)^2(0.00018 + 0.00459) \\
&= \$477{,}000
\end{aligned}
$$

or

$$\sigma_p = \$691$$

3.4. Uncertain Project Life and Uncertain Cash Flow

Consider the case in which the amounts and timing of cash flows are random variables with known means and variances, and the project life, also, is a random variable, N. Here, N must be integer valued and, of course, must be positive. If cash flows are statistically independent,

$$\mu_p = \mathrm{Exp}[\mathrm{PW}] = \sum_{N=1}^{\infty} \left[\sum_{j=1}^{N} \mu_j (1 + i)^{-j} \right] p_N \tag{17}$$

where p_N is the probability mass function for N. Moreover,

$$\sigma_p^2 = \mathrm{Var}[\mathrm{PW}] = \sum_{k=1}^{N} \left\{ \sum_{j=0}^{k} \mathrm{Var}(X_j) + \left[\sum_{j=0}^{k} \mathrm{Exp}(x_j) \right]^2 \right\} P_k - \mu_p^2 \tag{18}$$

where

$$\mathrm{Exp}(X_j) = \mu_j(1 + i)^{-j} \tag{19a}$$

$$\mathrm{Var}(X_j) = \sigma_j^2(1 + i)^{-2j} \tag{19b}$$

To illustrate, consider the problem summarized in Table 3. There are three risky, independent, end-of-year cash flows; the means and variances of their respective probability functions are given in columns (2) and (3) of Table 3. Project life, N, is also a random variable, with probability mass function as shown in columns (6) and (7) of the table. A 10% discount rate is assumed. Determination of the expected present worth, μ_p, based on Eq. 17, is summarized in the table. Here, $\mu_p = -\$82.64$. The variance of present worth, σ_p^2, based on Eqs. 18 and 19, may be shown to be

$$\sigma_p^2 = \$121{,}687 \qquad \text{or} \qquad \sigma_p = \$349$$

3.5. Other Models

There are a variety of other analytical models for assessing risky investments. The randomness ("riskiness") of cash flow amounts, timing, project life, and discount rate are considered singly and/or in combination. The complexity of the analytical procedure is roughly a function of the number of variables considered as well as the assumptions concerning mutual independence between random variables. In almost all cases the *mean* and *variance* of the distribution of the figure of merit are of primary concern. In some instances it is also possible to approximate the statistical *distribution* as well. Space limitations preclude an exhaustive review of the extant literature. For further readings consult the bibliography at the end of this chapter.

3.6. Analysis Based on the Probability Distribution for Present Worth

As before, the mean and variance of the probability distribution for the present worth statistic (PW) are denoted by μ_p and σ_p^2, respectively. These are measures of central tendency and variability, or

TABLE 3 Numerical Example: Both Cash Flows (A_j) and Project Life (N) Are Random Variables

End of Year j (1)	Mean μ_j (2)	Variance σ_j^2 (3)	Present Worth at 10%	
			$\mu_j(1.10)^{-j}$ (4)	$\sigma_j^2(1.10)^{-2j}$ (5)
0	−\$1,000	\$ 50^2	−\$1,000.00	\$ 2,500.00
1	500	100^2	454.55	8,264.46
2	800	200^2	661.16	27,320.54

Project Life N (6)	Probability p_N (7)	$\sum_{j=0}^{N}\mu_j(1.10)^{-j}$ (8)	$p_N \sum_{j=0}^{N} \mu_j(1.10)^{-j}$ (9)
0	0.00	−\$1,000.00	\$ 0
1	0.30	− 545.45	− 163.64
2	0.70	115.71	81.00
Totals	1.00		−\$ 82.64

From Park and Sharp-Bette 1990, p. 416.

dispersion, of the PW distribution. Under certain conditions the underlying probability distribution may be fully or partially characterized. When such is the case, it may be useful to describe the riskiness of the figure of merit in terms other than the variance of the distribution; for example, the probability that the PW will exceed some specified critical level.

3.6.1. Discrete Distribution for Present Worth

Consider a two-period problem as summarized in Figure 5. A cash outlay of \$100 occurs at the start of period 1 ($j = 0$). There are two possible discrete cash flows at end of period 1: $A_1 = \$50$ with probability 0.5 or $A_1 = \$70$ with probability 0.5. *If* $A_1 = \$50$, there are two possibilities for the cash flow at end of period 2: either $A_2 = \$60$ with probability 0.3 or $A_2 = \$80$ with probability 0.7. *If* $A_1 = \$70$, then either $A_2 = \$44$ with probability 0.4 or $A_2 = \$90$ with probability 0.6. The diagram of the possible outcomes shown in Figure 5 is sometimes known as a *probability tree.*

There are four possible present worths (outcomes), each with an associated joint probability. Assuming a 10% discount rate:

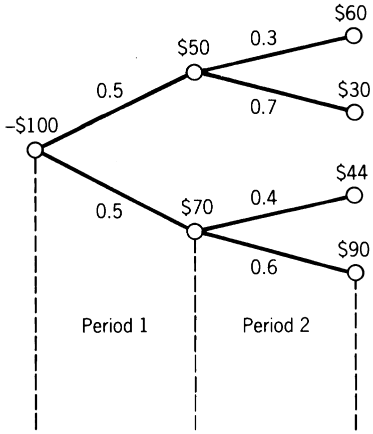


Figure 5 Cash Rows and Their Probabilities. Example problem from Park and Sharpe-Bette 1990, p. 419.

Outcome	A_0	A_1	A_2	$\sum_{j=0}^{2} A_j(1.10)^{-j}$	Joint Probability
1	−\$100	\$50	\$60	−\$4.96	$0.5 \times 0.3 = 0.15$
2	−100	50	80	11.57	$0.5 \times 0.7 = 0.35$
3	−100	70	44	0	$0.5 \times 0.4 = 0.20$
4	−100	70	90	38.02	$0.5 \times 0.6 = 0.30$

The remainder of the analysis is summarized in Table 4. Note that column (4) reflects the calculation of Exp[PW] and column (5) reflects the calculation of Exp[(PW)^2]. Moreover

$$\sigma_p^2 = \text{Exp}[(\text{PW})^2] - [\text{Exp}(\text{PW})]^2 \tag{20}$$

as discussed previously.

Now, suppose that it is of interest to determine the probability that this investment will be profitable, that is, PW > \$0. Only two of the possible outcomes, 2 and 4, meet this requirement, and, as they are independent events, the sum of their probabilities is

$$\begin{aligned}\text{Prob}[\text{PW} > \$0] &= \text{Prob}[\text{PW} = \$11.57] + \text{Prob}[\text{PW} = \$38.02] \\ &= 0.35 + 0.30 = 0.65\end{aligned}$$

3.6.2. Using Only the Mean and Variance of the PW Distribution

Tchebycheff's (sometimes written Chebyshev's) *inequality* states that

$$\text{Prob}[\mu - k\sigma < X < \mu + k\sigma] \geq 1 - 1/k^2 \tag{21}$$

where X is any random variable having mean μ and variance σ^2 and k is a positive constant. This is a useful relationship when only the mean and variance of the distribution are known. In terms of an unknown PW distribution with known mean μ_p and variance σ_p^2,

$$\text{Prob}[\mu_p - k\sigma_p < \text{PW} < \mu_p + k\sigma_p] \geq 1 - 1/k^2 \tag{22}$$

To illustrate, suppose that the mean and variance of the PW distribution have been determined to be \$800 and (\$50)^2, respectively. The analyst has been asked to determine the probability that the PW lies between two values, say, between \$600 and \$1000. Note here that

$$\mu_p - k\sigma_p = \$800 - k(\$50) = \$600$$

and

TABLE 4 Determining the Expected Present Worth

Outcome (1)	PW at 10% (2)	Joint Probability (3)	(4) = (2) × (3)	(5) = (2)^2 × (3)
1	−\$ 4.96	0.15	−\$ 0.744	\$\$ 3.690
2	11.57	0.35	4.050	46.853
3	0	0.20	0	0
4	38.02	0.30	11.406	433.656
	Totals	1.00	\$14.712	\$\$484.199

$$\begin{aligned}\text{Exp[PW]} &= \underline{\$14.712} \\ \text{Var[PW]} &= \$\$484.199 - (\$14.712)^2 = \$\$267.756 \\ \sigma_P = \sqrt{\text{Var[PW]}} &= \underline{\$16.363}\end{aligned}$$

$$\mu_p + k\sigma_p = \$800 + k(\$50) = \$1000$$

from which it is apparent that $k = 4$. Thus

$$\text{Prob}[\$600 < \text{PW} < \$1000] \geq 1 - 1/16 \quad \text{or} \quad 0.9375$$

Put somewhat differently, in the absence of any knowledge as to the *shape* of the distribution, the probability is *at least* 0.9375 that the random variable lies with $\pm 2\sigma$ of the mean.

3.6.3. When the Normal Distribution Can Be Assumed

Consider a stream of risky cash flows A_j occurring at the ends of periods 1, 2, . . . j, . . . N. The project life N and the discount rate i are known with certainty. The only stochastic variable here is the amount of the cash flow. The resulting PW is a random variable with mean given by Eq. (8) and, assuming independent cash flows, with variance given by Eq. (10). Under some general conditions application of the central limit theorem leads to the result that

$$Z_N = \frac{\text{PW} - \Sigma_{j=0}^{N} \mu_j(1 + i)^{-j}}{\sqrt{\Sigma_{j=0}^{N} \sigma_j^2}} \tag{23}$$

is approximately normally distributed, with $\mu = 0$ and $\sigma = 1$, as N approaches infinity. The "general condition" may be summarized as follows: the terms A_j, taken individually, contribute a negligible amount to the variance of the sum, and it is unlikely that any single A_j makes a relatively large contribution to the sum.

The terms A_j may have essentially any distribution. As a general rule of thumb, if the A_j's are approximately normally distributed, then the central limit theorem is a very good approximation when $N \geq 4$. If the distribution of the A_j's has no prominent mode(s), that is, approximately uniformly distributed, then $N \geq 12$ is a reasonable rule of thumb for applicability of the central limit theorem.

To illustrate the application of Eq. (23), consider the numerical example given in Table 1. It was determined that $\mu_p = \$69.44$ and $\sigma_p = \sqrt{\$\$802.9325} = \$28.34$. The probability distribution for PW is shown in part (*a*) of Figure 6; the equivalent standardized normal distribution is shown in part (*b*).

Consider the question: What is the probability that this proposal will result in a present worth greater than $50?

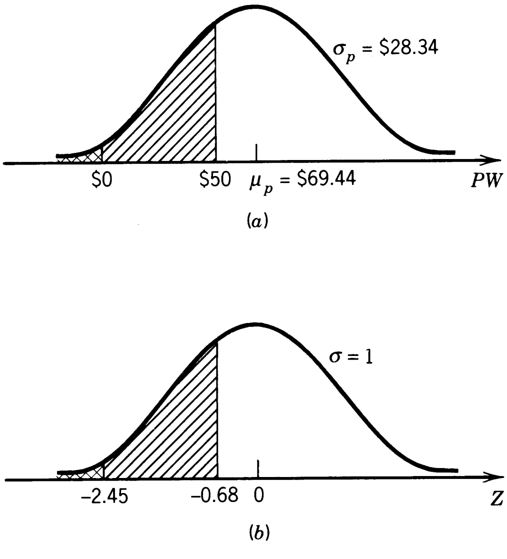


Figure 6 Probability Distribution for Present Worth.

TABLE 5 Cumulative Standard Normal Distribution

$$\phi(z) = \int_{-\infty}^{2} \frac{1}{\sqrt{2\pi}}\, e^{-\mu^2}/2\; du$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.0	0.500 00	0.503 99	0.507 98	0.511 97	0.515 95	0.519 94	0.523 92	0.527 90	0.531 88	0.535 86	0.0
0.1	0.539 83	0.543 79	0.547 76	0.551 72	0.555 67	0.555 62	0.563 56	0.567 49	0.571 42	0.575 34	0.1
0.2	0.579 26	0.583 17	0.587 06	0.590 95	0.594 83	0.598 71	0.602 57	0.606 42	0.610 26	0.614 09	0.2
0.3	0.617 91	0.621 72	0.625 51	0.629 30	0.633 07	0.636 83	0.640 58	0.644 31	0.648 03	0.651 73	0.3
0.4	0.655 42	0.659 10	0.662 76	0.666 40	0.670 03	0.673 64	0.677 24	0.680 82	0.684 38	0.687 93	0.4
0.5	0.691 46	0.694 97	0.698 47	0.701 94	0.705 40	0.708 84	0.712 26	0.715 66	0.719 04	0.722 40	0.5
0.6	0.725 75	0.729 07	0.732 37	0.735 65	0.738 91	0.742 15	0.745 37	0.748 57	0.751 75	0.754 90	0.6
0.7	0.758 03	0.761 15	0.764 24	0.767 30	0.770 35	0.773 37	0.776 37	0.779 35	0.782 30	0.785 23	0.7
0.8	0.778 14	0.791 03	0.793 89	0.796 73	0.799 54	0.802 34	0.805 10	0.807 85	0.810 57	0.813 27	0.8
0.9	0.815 94	0.818 94	0.821 21	0.823 81	0.826 39	0.828 94	0.831 47	0.833 97	0.836 46	0.838 91	0.9
1.0	0.841 34	0.843 75	0.846 13	0.848 49	0.850 83	0.853 14	0.855 43	0.857 69	0.859 93	0.862 14	1.0
1.1	0.864 33	0.866 50	0.868 64	0.870 76	0.872 85	0.874 93	0.876 97	0.879 00	0.881 00	0.882 97	1.1
1.2	0.884 93	0.886 86	0.888 77	0.890 65	0.892 51	0.894 35	0.896 16	0.897 96	0.899 73	0.901 47	1.2
1.3	0.903 20	0.904 90	0.906 58	0.908 24	0.909 88	0.911 49	0.913 08	0.914 65	0.916 21	0.917 73	1.3
1.4	0.919 24	0.920 73	0.922 19	0.923 64	0.925 06	0.926 47	0.927 85	0.929 22	0.930 56	0.931 89	1.4
1.5	0.933 19	0.934 48	0.935 74	0.936 99	0.938 22	0.939 43	0.940 62	0.941 79	0.942 95	0.944 08	1.5
1.6	0.945 20	0.946 30	0.947 38	0.948 45	0.949 50	0.950 53	0.951 54	0.952 54	0.953 52	0.954 48	1.6
1.7	0.955 43	0.956 37	0.957 28	0.958 18	0.959 07	0.959 94	0.960 80	0.961 64	0.962 46	0.963 27	1.7
1.8	0.964 07	0.964 85	0.965 62	0.966 37	0.967 11	0.967 84	0.968 56	0.969 26	0.969 95	0.970 62	1.8
1.9	0.971 28	0.971 93	0.972 57	0.973 20	0.973 81	0.974 41	0.975 00	0.975 58	0.976 15	0.976 70	1.9
2.0	0.977 25	0.977 78	0.978 31	0.978 82	0.979 32	0.979 82	0.980 30	0.980 77	0.981 24	0.981 69	2.0
2.1	0.982 14	0.982 57	0.983 00	0.983 41	0.983 82	0.984 22	0.984 61	0.985 00	0.985 37	0.985 74	2.1
2.2	0.986 10	0.986 45	0.986 79	0.987 13	0.987 45	0.987 78	0.988 09	0.988 40	0.988 70	0.988 99	2.2
2.3	0.989 28	0.989 56	0.989 83	0.990 10	0.990 36	0.990 61	0.990 86	0.991 11	0.991 34	0.991 58	2.3
2.4	0.991 80	0.992 02	0.992 24	0.992 45	0.992 66	0.992 86	0.993 05	0.993 24	0.993 43	0.993 61	2.4
2.5	0.993 79	0.993 96	0.994 13	0.994 30	0.994 46	0.994 61	0.994 77	0.994 92	0.995 06	0.995 20	2.5
2.6	0.995 34	0.995 47	0.995 60	0.995 73	0.995 85	0.995 98	0.996 09	0.996 21	0.996 32	0.996 43	2.6
2.7	0.996 53	0.996 64	0.996 74	0.996 83	0.996 93	0.997 02	0.997 11	0.997 20	0.997 28	0.997 36	2.7
2.8	0.997 44	0.997 52	0.997 60	0.997 67	0.997 74	0.997 81	0.997 88	0.997 95	0.998 01	0.998 07	2.8
2.9	0.998 13	0.998 19	0.998 25	0.998 31	0.998 36	0.998 41	0.998 46	0.998 51	0.998 56	0.998 61	2.9
3.0	0.998 65	0.998 69	0.998 74	0.998 78	0.998 82	0.998 86	0.998 89	0.998 93	0.998 97	0.999 00	3.0
3.1	0.999 03	0.999 06	0.999 10	0.999 13	0.999 16	0.999 18	0.999 21	0.999 24	0.999 26	0.999 29	3.1
3.2	0.999 31	0.999 34	0.999 36	0.999 38	0.999 40	0.999 42	0.999 44	0.999 46	0.999 48	0.999 50	3.2
3.3	0.999 52	0.999 53	0.999 55	0.999 57	0.999 58	0.999 60	0.999 61	0.999 62	0.999 64	0.999 65	3.3
3.4	0.999 66	0.999 68	0.999 69	0.999 70	0.999 71	0.999 72	0.999 73	0.999 74	0.999 75	0.999 76	3.4
3.5	0.999 77	0.999 78	0.999 78	0.999 79	0.999 80	0.999 81	0.999 81	0.999 82	0.999 83	0.999 83	3.5
3.6	0.999 84	0.999 85	0.999 85	0.999 86	0.999 86	0.999 87	0.999 87	0.999 88	0.999 88	0.999 89	3.6
3.7	0.999 89	0.999 90	0.999 90	0.999 90	0.999 91	0.999 91	0.999 92	0.999 92	0.999 92	0.999 92	3.7
3.8	0.999 93	0.999 93	0.999 93	0.999 94	0.999 94	0.999 94	0.999 94	0.999 95	0.999 95	0.999 95	3.8
3.9	0.999 95	0.999 95	0.999 96	0.999 96	0.999 96	0.999 96	0.999 96	0.999 96	0.999 97	0.999 97	3.9

Source: Adapted from W. W. Hines and D.C. Montgomery, *Probability and Statistics in Engineering and Management Science,* 2nd Ed., John Wiley & Sons, New York, 1980; from Park and Sharp-Bette 1990.

$$\begin{aligned}\text{Prob}[\text{PW} > \$50] &= 1 - \text{Prob}[\text{PW} < \$50] \\ &= 1 - \text{Prob}\left[Z < \frac{\$50.00 - \$69.44}{\$28.34}\right] \\ &= 1 - \text{Prob}[Z < -0.686] \simeq 1 - 0.75 = 0.25\end{aligned}$$

An abbreviated version is included here in Table 5.

Consider a second question: What is the probability that this proposal will result in a loss?

$$\begin{aligned}\text{Prob}[\text{PW} < \$0] &= \text{Prob}\left[Z < \frac{\$0 - \$69.44}{\$28.34}\right] \\ &= \text{Prob}[Z < -2.45] \simeq 0.01\end{aligned}$$

Note that the probability of a loss is identical here to the probability that the proposal's internal rate of return will be less than the minimum attractive rate of return.

3.7. Comparing Risky Proposals

As indicated previously, decision makers are generally risk avoiders. Additional risk, as measured by the variance of the figure of merit, is to be avoided whenever possible. Thus there are two criteria to be considered simultaneously: the figure of merit, for example, present worth, as measured by the expected value (mean, μ) of the distribution; and the riskiness of the outcome as measured by the variance (σ^2) of the distribution. The former is to be maximized, and the latter is to be minimized.

Consider two mutually exclusive alternations. Let (μ_1, σ_1^2) and (μ_2, σ_2^2) represent the mean and variance of alternatives I and II, respectively. The appropriate decision rules are as follows:

$$\begin{array}{ll} \text{Case A:} & \left.\begin{array}{l}\text{If } \mu_1 = \mu_2 \text{ and } \sigma_1 < \sigma_2 \\ \text{or } \mu_1 > \mu_2 \text{ and } \sigma_1 \leq \sigma_2\end{array}\right\} \text{Choose I over II} \\ \text{Case B:} & \left.\begin{array}{l}\text{If } \mu_1 = \mu_2 \text{ and } \sigma_1 > \sigma_2 \\ \text{or } \mu_1 < \mu_2 \text{ and } \sigma_1 \geq \sigma_2\end{array}\right\} \text{Choose II over I} \\ \text{Case C:} & \left.\begin{array}{l}\text{If } \mu_1 < \mu_2 \text{ and } \sigma_1 < \sigma_2 \\ \text{or } \mu_1 > \mu_2 \text{ and } \sigma_1 > \sigma_2\end{array}\right\} \text{Conclusion ambiguous} \end{array}$$

There is no clear conclusion in case C because the riskier alternative is also the one with the larger expected return. When this situation arises, trade-offs must be made between risk and return. There is little theoretical guidance short of converting to utility theory (discussed later). Additional discussion is beyond the scope of this chapter.

4. DECISION THEORY APPLICATIONS

The approach to risk analysis outlined in the previous section is based on the premise that the decision maker desires to (a) maximize expected return and (b) minimize risk. This section presents some additional principles of choice that may be appealing under certain conditions. A simple numerical example is used as a basis for the discussion.

4.1. Problem Statement

The International Manufacturing Company (IMC) is considering five mutually exclusive alternatives for constructing a new manufacturing plant in a certain Asian country. The cost of each alternative, stated in terms of present worth (or net present value) of cost, depend on the outcome of negotiations that are currently under way between IMC, lending agencies, and the government of the host country. IMC analysts have concluded that four specific mutually exclusive outcomes are possible, and they have computed the present worth (cost) for each alternative—outcome combination. These are shown as cell values in the *cost matrix* in Table 6.

If the future is known with certainty, then the least costly alternative may be selected by any of the methods presented in Chapter 91. For example, if it is known that outcome s_3 will definitely occur, then a_3 should be selected because it will result in the lowest present worth (cost). On the other hand, if a_3 is selected and s_4 perversely occurs, choosing a_3 will have resulted in the most costly event.

Assume that sufficient information exists to warrant statements about the relative probabilities of the possible future outcomes. Specifically, these probabilities (expected relative frequencies) are

TABLE 6 Cost Matrix for Illustrative Problem (Cell entries are multiples of $1,000,000)

		Possible Outcomes			
		s_1	s_2	s_3	s_4
Alternatives	a_1	18	11	10	10
	a_2	16	16	16	16
	a_3	14	14	8	20
	a_4	9	12	17	16
	a_5	10	13	17	18

$$P[s_1] = 0.3 \qquad P[s_3] = 0.2$$

$$P[s_2] = 0.4 \qquad P[s_4] = 0.1$$

Given this additional information, which alternative should be selected? A number of principles that may be applied in this situation are discussed later.

A problem statement of this type is known as a *decision under risk* because the underlying probability distribution for the future scenarios, or *states of nature,* is known or can be assumed. ‘‘Risk,’’ in the previous section, was used in a more general sense to characterize the absence of certainty. The term was used analogously to ‘‘randomness’’ or uncertainty. Here, in a more limited sense, a problem statement in which the underlying distribution for the s_j's is *not* known or assumed is a *decision under uncertainty*.

4.2. Dominance

Before applying *any* of the principles of choice, it is first desirable (although not absolutely necessary) to apply the *dominance principle* to determine which alternatives, if any, are dominated. If, of two alternatives, one would never be preferred no matter what future occurs, it is said to be dominated and may be removed from any further consideration. From the example, consider a_4 and a_5:

	s_1	s_2	s_3	s_4	
a_4	9	12	17	16	(in $ million)
a_5	10	13	17	18	

Since a_5 is always at least as costly as a_4, irrespective of which future outcome occurs, a_5 may be ignored in the remaining discussion.

If one alternative dominates all others, it is said to be *globally dominant,* and the decision maker need look no further; the optimal solution has been found. Unfortunately, globally dominant alternatives are rare. But in any event, the dominance principle is frequently effective in reducing the number of alternatives to be considered.

4.3. Principles for Decisions under Risk

4.3.1. The Principle of Expectation

The *principle of expectation* states that the alternative to be selected is the one that has the minimum expected cost (or maximum expected profit or revenue). In general,

$$\text{Min } E[C(a_i)] = \sum_j C(a_i|s_j)p_j \tag{24a}$$

or

$$\text{Max } E[R(a_i)] = \sum_j R(a_i|s_j)p_j \tag{24b}$$

where $C(a_i|s_j)$ = total cost of alternative a_i given that states of nature s_j occurs
$R(a_i|s_j)$ = total net return of alternative a_i given that state of nature s_j occurs
p_j = probability that state s_j will occur

From the example it may be shown that

$$E[C(a_1)] = \$12{,}800{,}000 \qquad E(C(a_3)] = \$13{,}400{,}000$$
$$E[C(a_2)] = \$16{,}000{,}000 \qquad E[C(a_4)] = \$12{,}500{,}000$$

Here, a_4 should be selected because it yields the minimum expected cost.

Principles that depend on determination of expected values by the mathematics of probability theory are frequently criticized on the grounds that the theory holds only when trials are repeated many times. It is argued that, for certain types of decisions—for example, whether to finance a major expansion—expectation is meaningless since this type of decision is not made very often. According to the counterargument, even if the firm is not faced with a large number of repetitive decisions, it should apply the principle to many different decisions and thus realize the long-run effects. Moreover, even if the decision is unique, the only way to approach decisions for which probabilities are known is to behave as if the decision were a repetitive one and thus minimize expected cost or maximize expected revenue or profit.

4.3.2. The Principle of Most Probable Future

Assume that the future event to expect is the most likely event. Thus, observing that s_2 has the highest probability of occurring, assume that it will in fact occur. In this case a_1 (with present worth (cost) = \$11,000,000 is the least costly of the four available alternatives.

This principle is particularly appealing in cases in which one future is significantly more probable than all other possibilities.

4.3.3. The Aspiration Level Principle

The *aspiration level principle* requires the establishment of a goal, or "level of aspiration." Thus the alternative that maximizes the probability that the goal will be met or exceeded should be selected. To illustrate, suppose that the management of IMC wishes to minimize the probability that present worth (cost) will exceed \$15,000,000. (This is identical to the requirement that it maximize the probability that costs will *not* exceed \$15,000,000.) The probabilities are

$$\text{Prob}[C(a_1) > \$15{,}000{,}000] = 0.3$$
$$\text{Prob}[C(a_2) > \$15{,}000{,}000] = 0.3 + 0.4 + 0.2 + 0.1 = 1.0$$
$$\text{Prob}[C(a_3) > \$15{,}000{,}000] = 0.1$$
$$\text{Prob}[C(a_4) > \$15{,}000{,}000] = +0.2 + 0.1 = 0.3$$

Thus, the aspiration level will be met if a_1 is selected.

Clearly, the selection from mutually exclusive alternatives is a matter of which principle is used to guide the decision.

4.4. Principles for Decisions under Uncertainty

This section examines a number of principles of choice that may be used when the relative likelihoods of future states of nature *cannot* be estimated. These principles will be demonstrated by using the example problem introduced earlier.

4.4.1. The Minimax (or Maximin) Principle

The *minimax principle* is pessimistic in the extreme. It assumes that, if any alternative is selected, the worst possible outcome will occur. The maximum cost associated with each alternative is examined, and the alternative that *minimizes* the *maximum* cost is selected. In general, the mathematical formulation of the minimax principle is

$$\underset{i}{\text{Min}}[\underset{j}{\text{Max}}(C_{ij})] \tag{25a}$$

where C_{ij} is the cost that results when alternative i is selected and state of nature j occurs. From the example

Alternative (a_j)	$\underset{j}{\text{Max}}\ C_{ij}$
a_1	18 (in $ million)
a_2	16 (in $ million)
a_3	20 (in $ million)
a_4	17 (in $ million)

If the minimax principle is adopted, a_2 is indicated because it results in minimum costs, assuming the worst possible conditions.

The mirror image of the minimax principle, the *maximin principle,* may be applied when the matrix contains *profits* or *revenue* measures. In this case the most pessimistic view suggests that the alternative to select is the one that *maximizes* the *minimum* profit or revenue associated with each alternative. The mathematical formulation of the maximin principle is

$$\underset{i}{\text{Max}}[\underset{j}{\text{Min}}(R_{ij})] \tag{25b}$$

where R_{ij} is the revenue or profit resulting from the combination of a_i and s_j.

4.4.2. The Minimin (or Maximax) Principle

The *minimin principle* is based on the view that the best possible outcome will occur when a given alternative is selected. It is optimistic in the extreme. The minimum cost associated with each alternative is examined, and the alternative that *minimizes* the *minimum* cost is selected. The mathematical formulation is

$$\underset{i}{\text{Min}}[\underset{j}{\text{Min}}(C_{ij})] \tag{26a}$$

From the example

Alternative (a_1)	$\underset{j}{\text{Min}}\ C_{ij}$
a_1	10 (in $ million)
a_2	16 (in $ million)
a_3	8 (in $ million)
a_4	9 (in $ million)

Alternative a_3 minimizes the minimum cost.

As a corollary to the minimin principle, the *maximax principle* is appropriate when the decision maker is extremely optimistic and the matrix contains measures of profit or revenue. The maximum profit (or revenue) associated with each alternative is examined, and the alternative that *maximizes* the *maximum* profit (or revenue) is selected. The mathematical formulation is

$$\underset{i}{\text{Max}}[\underset{j}{\text{Max}}(R_{ij})] \tag{26b}$$

4.4.3. The Hurwicz Principle

It may be argued that decision makers need not be either completely optimistic or pessimistic, in which case the *Hurwicz principle* permits selection of a position between the two extremes. When evaluating costs, C_{ij} the *Hurwicz criterion* for alternative a_i is given by

$$\text{Min } H(a_j) = \alpha[\underset{j}{\text{Min}}(C_{ij})] + (1 - \alpha)[\underset{j}{\text{Max}}(C_{ij})] \tag{27a}$$

where α is the "index of optimism" such that $0 \leq \alpha \leq 1$. Extreme pessimism is defined by $\alpha = 0$; extreme optimism is defined by $\alpha = 1$. The value of α used in any particular analysis is selected by the decision maker based on subjective judgment. The alternative that *minimizes* the quantity $H(a_i)$ is the alternative to select.

When evaluating profits or revenues, R_{ij}, the expression for the Hurwicz criterion is

$$\text{Max } H(a_i) = \alpha[\underset{j}{\text{Max}}(R_{ij})] + (1 - \alpha)[\underset{j}{\text{Min}}(R_{ij})] \qquad (27b)$$

The values of $H(\alpha_j)$ are plotted in Figure 7 for the sample problem. We may determine, either graphically or algebraically, that a_2 will be chosen for $0 \leq \alpha \leq 0.125$, a_4 will be selected for $0.125 \leq \alpha \leq 0.75$, and a_3 is least costly for $0.75 \leq \alpha \leq 1.00$.

4.4.4. The Laplace Principle (Insufficient Reason)

The *Laplace principle,* sometimes known as the *principle of insufficient reason,* assumes that the probabilities of future events occurring are equal. That is, in the absence of any information to the contrary, it is assumed that all future outcomes are equally likely to occur. The expected cost (or profit/revenue) of each alternative is then computed, and the alternative that yields the minimum expected cost (or maximum expected profit/revenue) is selected. The mathematical expression for this principle is

$$\underset{i}{\text{Min}} \left\{ \left(\frac{1}{k}\right) \sum_{j=1}^{k} C_{ij} \right\} \qquad (28a)$$

when the figure of merit is expressed as a cost or as

$$\underset{i}{\text{Max}} \left\{ \left(\frac{1}{k}\right) \sum_{j=1}^{k} R_{ij} \right\} \qquad (28b)$$

when the figure of merit is expressed as revenue or profit.

Returning to our example, the insufficient reason assumption yields $p_1 = p_2 = p_3 = p_4 = 0.25$. With these probabilities:

$$E[C(a_1)] = \$12{,}250{,}000 \qquad E[C(a_3)] = \$14{,}000{,}000$$

$$E[C(a_2)] = \$16{,}000{,}000 \qquad E[C(a_4)] = \$13{,}500{,}000$$

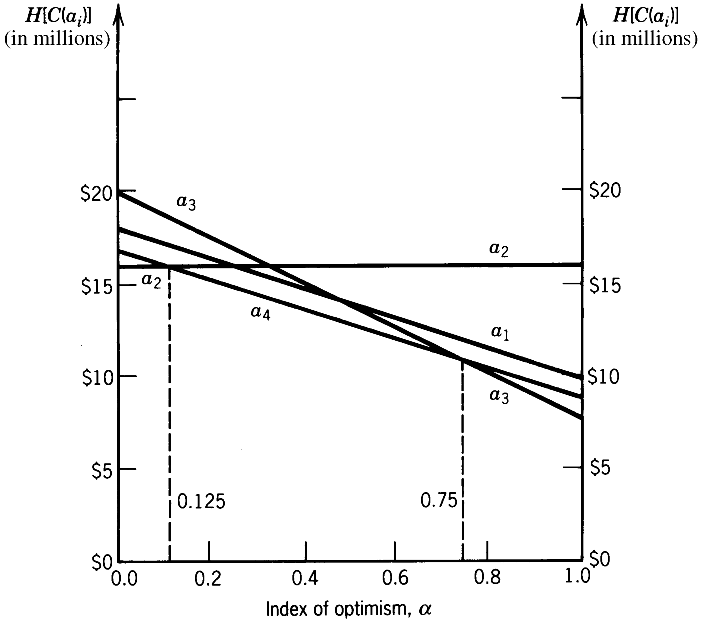


Figure 7 Sample Problem—Hurwicz Criterion as Function of Index of Optimism.

Alternative a_1 should therefore be selected because it results in the minimum expected present worth (cost).

4.4.5. The Savage Principle (Minimax Regret)

The *Savage principle,* or *principle of minimax regret,* is based on the assumption that the decision maker's primary interest is the *difference* between the actual outcome and the outcome that would have occurred had it been possible to accurately predict the future. Given these difference, or *regrets,* the decision maker then adopts a conservative position and selects the alternative that minimizes the maximum potential regret for each alternative.

A *regret matrix* is constructed, having for its cell values either

$$C_{ij} - [\underset{i}{\text{Min}}(C_{ij})] \tag{29a}$$

for cost data or

$$[\underset{i}{\text{Max}}(R_{ij})] - R_{ij} \tag{29b}$$

for revenue or profit data. In either case these cell values, or regrets, represent the differences between (a) the outcome if alternative a_j is selected and state of nature s_j subsequently occurs and (b) the outcome that would have been achieved had it been known in advance which state of nature would occur, so that the best alternative could have been selected. To illustrate, consider alternative a_1 and state of nature s_1: C_{11} = \$18,000,000. However, if we had known *a priori* that state s_1 would in fact occur, we would have selected a_4, incurring a cost of only \$9,000,000. The difference (\$18,000,000 − \$9,000,000) is a measure of "regret" about selecting a_1 when we could have selected a_4 (had we known the state of nature in advance). The complete regret matrix for the example is given in Table 7.

The alternative that *minimizes* the *maximum* regret is preferred. That is, for cost data,

$$\underset{i}{\text{Min}}\ \underset{j}{\text{Max}}\{C_{ij} - [\text{Min}(C_{ij})]\} \tag{30a}$$

or, when the cell values are based on revenue or profit data,

$$\underset{i}{\text{Min}}\ \underset{j}{\text{Max}}\{[\underset{i}{\text{Max}}(R_{ij})] - R_{ij}\} \tag{30b}$$

Equation (30a) is applicable for the sample problem, yielding the following:

Alternative	Maximum Regret
a_1	9 (in \$ millions)
a_2	8 (in \$ millions)
a_3	10 (in \$ millions)
a_4	9 (in \$ million)

Thus, according to this principle of choice, a_2 should be preferred.

TABLE 7 Regret Matrix for Sample Problem (Cell values are multiples of \$1,000,000)

		Possible Outcomes			
		s_1	s_2	s_3	s_4
Alternatives	a_1	18 − 9 = 9	11 − 11 = 0	10 − 8 = 2	10 − 10 = 0
	a_2	16 − 9 = 7	16 − 11 = 5	16 − 8 = 8	16 − 10 = 6
	a_3	14 − 9 = 5	14 − 11 = 3	8 − 8 = 0	20 − 10 = 10
	a_4	9 − 9 = 0	12 − 11 = 1	17 − 8 = 9	16 − 10 = 6

4.5. Summary of Results

There is no special reason why the principles of choice discussed in the preceding sections should yield the same solution. Indeed, each of the alternatives in this example problem were selected at least once.

Decision Under Risk

Principle	Solution
Expectation	a_4
Most probable future	a_1
Aspiration level	a_3

Decisions Under Uncertainty

Principle	Solution
Minimax	a_2
Minimin	a_3
Hurwicz $(0.1215 < \alpha < 0.75)$	a_4
Laplace (insufficient reason)	a_1
Savage (minimax regret)	a_2

Is one principle more "correct" than any other? There is no simple answer to this question—the choice of principle largely depends on the predisposition of the decision maker and the specific decision situation. Each principle has certain obvious advantages, and each is deficient in one or more desirable characteristics. Nevertheless, the principles in this section are useful because they shed some light on the subjective decision process and make the available information explicit to the decision process.

5. DECISION TREES

Decision tree methodology is useful for the evaluation of problems characterized by *sequential* decisions, each of which involves a variety of outcomes. The pictorial representation of this problem is suggestive of a tree lying on its side, with the branches in the tree representing successions of outcomes. The graphic portrayal of the problem structure is both its primary asset as well as its principal disadvantage. The ability to communicate complex dependencies is of great value, of course. However, the number of sequential decisions and outcomes (branches) is necessarily limited by the graphic medium (CRT screen, 8½ × 11 in. paper, etc.). These features will be apparent from the following discussion of deterministic and stochastic decision trees.

5.1. Deterministic Decision Trees

Consider a problem of retirement and replacement over a three-year planning horizon. The existing equipment, the defender, is now two years old. Replacement decisions are to be made now, one year hence, and two years hence. Whichever equipment is in service three years from now will be removed from service and sold in any event. Initial costs, salvage values, and operating costs for each of the 3 years are summarized in Table 8. Note that the first row of the table represents the defender: It has a current salvage value of \$50, \$40 after one year, \$30 after two years, and \$20 if sold at the end of the third year.

TABLE 8 Input Data for Deterministic Example (All cash flows have been discounted—they are shown as their PW equivalents)

Year *j*	Initial Cost if Purchased at Start of Year *j*	Salvage Value if Sold at End of Year			Operating Cost in Year		
		1	2	3	1	2	3
1	\$50[a]	\$40	\$ 30	\$ 20	\$90	\$95	\$100
1	100[b]	80	65	55	50	60	70
2	120[b]		100	85		45	55
3	130[b]			110			40

[a] Current salvage value of defender.

[b] Challengers at start of years 1, 2, and 3.

The decision tree for this problem is shown in Figure 8. The three decision points are represented by squares, the branches are the decisions, and the economic consequence of each branch is shown at that line. The solution begins at the end of the tree, that is, at the latest decision point, which is decision 3 in this case.

Decision Point	Alternative	Monetary Outcome, 3rd Year	Choice
3*a*	Keep	−$100 + $20 = −$80	
	Replace	$30 − $130 − $40 + $110 = −$30	Replace
3*b*	Keep	−$55 + $85 = $30	
	Replace	$100 − $130 − $40 + $110 = $40	Replace
3*c*	Keep	−$70 + $55 = −$15	
	Replace	$65 − $130 − $40 + $110 = $5	Replace
3*d*	Keep	−$55 + $85 = $30	
	Replace	$100 − $130 − $40 + $110 = $40	Replace

Next, the procedure rolls back to the preceding decision point, the beginning of the second year. The monetary outcomes are cumulative, that is, the economic consequences in year 2 are added to those of the optimal decisions at the beginning of year 3.

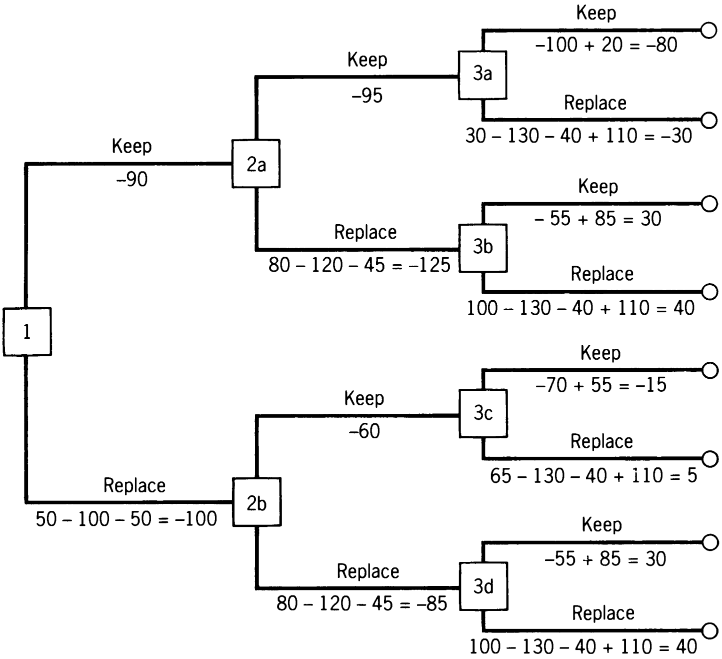

Keep
-100 + 20 = -80
Keep
-95
3a
Replace
30 - 130 - 40 + 110 = -30
Keep
-90
2a
Keep
- 55 + 85 = 30
Replace
3b
80 - 120 - 45 = -125
Replace
100 - 130 - 40 + 110 = 40
1
Keep
-70 + 55 = -15
Keep
3c
-60
Replace
65 - 130 - 40 + 110 = 5
Replace
2b
50 - 100 - 50 = -100
Keep
-55 + 85 = 30
Replace
3d
80 - 120 - 45 = -85
Replace
100 - 130 - 40 + 110 = 40

Figure 8 Decision Tree for Deterministic Example.

Decision Point	Alternative	Monetary Outcome, 2nd + 3rd Year	Choice
2*a*	Keep	(−$95) + (−$30) = −$125	
	Replace	($40 − $120 − $45) + ($40) = −$85	Replace
2*b*	Keep	(−$60) + ($5) = −$55	
	Replace	($80 − $120 − $45) + ($40) = −$45	Replace

The process continues until the first decision point. Here:

Decision Point	Alternative	Monetary Outcome, 1st, 2nd, 3rd Year	Choice
1	Keep Replace	(−$90) + (−$85) = −$175	
		($50 − $100 − $50) + (−$45) = −$145	Replace

The optimal solution is now complete. The optimal path through the tree is determined by beginning at the first decision point and continuing through subsequent decisions in accordance with the indicated decisions. In this example the optimal solution is *replace, replace,* and *replace* at the start of years 1, 2, and 3, respectively.

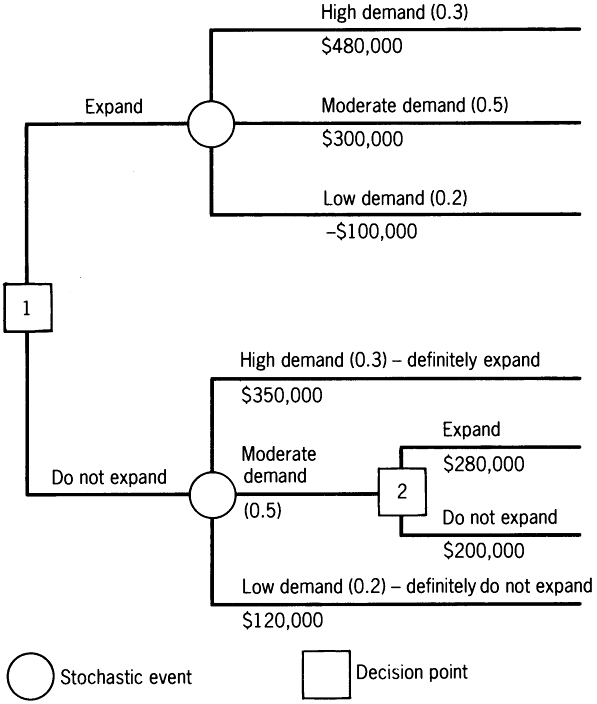

High demand (0.3)
$480,000
Expand
Moderate demand (0.5)
$300,000
Low demand (0.2)
-$100,000
1
High demand (0.3) – definitely expand
$350,000
Expand
Moderate demand
(0.5)
$280,000
2
Do not expand
Do not expand
$200,000
Low demand (0.2) – definitely do not expand
$120,000
Stochastic event
Decision point

Figure 9 Decision Tree for Stochastic Example. (All cash flows have been discounted. They are shown here in their PW equivalents.)

5.2. Stochastic Decision Trees

One of the most useful features of decision trees is their ability to illustrate variable outcomes when their probabilities of occurrence can be estimated. To be effective in a graphic format, the outcomes must be characterized as discrete random variables with a relatively small number of possibilities.

A very simple example of a stochastic decision tree is illustrated in Figure 9. The firm is considering expansion of a production line for the manufacture of a certain product. If the decision is made to forgo expansion at the present time, a subsequent opportunity will arise in about a year from now. The delay, perhaps, might be warranted based on the demand experienced over the next year. All dollar values shown in the tree are present value equivalents. There are three possible demand outcomes as shown—high, moderate, and low, with probabilities 0.3, 0.5, and 0.2, respectively.

At decision point 2 it is clear that plant expansion is preferred as PW(expand) = $280,000 and PW(do not expand) = $200,000. Rolling back to decision point 1, it is now necessary to evaluate the *expected* present worth through each of the branches emanating from that point.

$$\begin{aligned}\text{Exp[PW, expand now]} &= 0.3(\$480{,}000) + 0.5(\$300{,}000) + 0.2(-\$100{,}000)\\ &= \$274{,}000\\ \text{Exp[PW, do not expand]} &= 0.3(\$230{,}000) + 0.5(\$280{,}000) + 0.2(\$120{,}000)\\ &= \$233{,}000\end{aligned}$$

Thus, based on this analysis, it would appear that expansion at the present time is warranted.

The solution can be effected, of course, without reference to the decision tree. It is the graphic character of the tree that permits the analyst to articulate the sequential and stochastic nature of events and outcomes and to communicate these interrelationships to decision makers.

6. DIGITAL COMPUTER (MONTE CARLO) SIMULATION

The statistical procedures related to risk analysis suffer from at least one important drawback: The analytical techniques necessary to derive the mean, variance, and possibly the probability distribution of the figure of merit may be extremely difficult to implement. Indeed, the complexity of many real-world problems precludes the use of these computational techniques altogether; computations may be intractable, or the necessary underlying assumptions may not be met. Under these conditions analysts may find *digital computer* (*Monte Carlo*) *simulation* especially useful. (Strictly speaking, *Monte Carlo* simulation and *digital computer* simulation are not synonymous. Monte Carlo simulation is a sampling technique used in the digital computer simulation of systems behavior. However, in recent years, practitioners have tended to blur this semantic distinction, using the terms interchangeably.)

The objective of digital computer simulation is to generate a probability distribution for the figure of merit, generally present worth or rate of return, given the probability distributions for the various components of the analysis. The decision maker can thus compare expected returns as well as the variability of returns for two or more alternatives. Moreover, probability statements can be made, in this form: The probability is x that project y will result in a profit in excess of z.

6.1. Sampling from a Discrete Distribution

Suppose that ‘‘annual operating savings,’’ A, is a discrete random variable with probabilities as given in Table 9. The associated *cumulative distribution function* (CDF), also given in the table, represents the probability that the annual operating savings will be less than or equal to some given value.

Our problem now is one of sampling from this distribution, using either the probability function or its associated CDF in order to preserve precisely all the characteristics of the original distribution. We can do this by obtaining, say, 100 perfectly matched balls, numbered from 00 to 99. We want to label four of the balls ‘‘$2400,’’ eight of the balls ‘‘$2500,’’ and so on. The number of balls labeled with a particular amount is proportional to their relative probability in the original distribution:

Ball Numbers	Number of Balls	Labels	Ball Numbers	Number of Balls	Label
01–04	4	$2400	49–64	16	$2900
05–12	8	2500	65–80	16	3000
13–22	10	2600	81–90	10	3100
23–34	12	2700	91–98	8	3200
35–48	14	2800	99–00	2	3300

TABLE 9 Probabilities and Cumulative Distribution Function for Sample Problem

Event i	Annual Savings $\bar{A}_i$	Probability of Occurrence[a] $P(\bar{A}_i)$	Cumulative Distribution Function (CDF) $\underline{P}$(Ann. Savings $\leq \bar{A}_i$)
1	\$2400	0.04	0.04
2	2500	0.08	0.12
3	2600	0.10	0.22
4	2700	0.12	0.34
5	2800	0.14	0.48
6	2900	0.16	0.64
7	3000	0.16	0.80
8	3100	0.10	0.90
9	3200	0.08	0.98
10	3300	0.02	1.00

[a] The probability function for a *discrete* random variable is sometimes known as a *probability mass function.* The equivalent function for a *continuous* random variable is a *probability density function* (pdf).

Now we can put all the balls into a large jar, shake it thoroughly so that the balls are completely mixed, and then draw out a single ball. Then, the result of this "random sample" is recorded and the ball is placed back in the jar to select our next sample. As we continue this process through a large number of samples, or trials, we can expect the resulting frequency distribution to approximate that of the original population.

Of course, in practical applications, the sampling process does not consist of drawing balls from a jar. There are a variety of more elegant procedures, generally based on successive iterations of a predetermined formula. An alternative approach that is useful when the number of samples to be drawn is relatively small is to reference a *table of random numbers*. Such tables have been developed and the results recorded in tabular format. (A table of three-digit random numbers appears in Table 10.) Inasmuch as the numbers in the table are randomly generated, users may enter the table at any point and proceed in any direction.

6.2. Sampling from a Normal Distribution

The *normal distribution* is frequently used to describe the probabilities of certain continuous random variables. The probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \left\{ \exp\left[\frac{-1}{2} \left(\frac{x - \mu}{\sigma} \right) \right] \right\} \tag{31}$$

where μ = mean and σ = standard deviation of the distribution and x is the particular value of the random variable. A particular normal distribution is fully described by the parameter μ and σ, where μ is a measure of central tendency and σ is a measure of dispersion.

The *standard normal distribution* results from the special case wherein $\mu = 0$ and $\sigma = 1$. The area under the curve from $-\infty$ to $+\infty$ is exactly 1.0. If one can develop a table of random numbers for a uniform distribution over the interval 0–1, it is possible to map a set of equivalent values for the standard normal distribution, as in Figure 10. The value along the ordinate represents the probability that the random variable X lies in the interval $-\infty$ to x. For any random number we can compute the equivalent value x. This latter value is called the *random normal deviate*.

Tables of random normal deviates exist that contain values from which one may generate a random sample for *any* normal distribution with known parameters μ and σ. (See Table 11.) The simulated event, then, is given by

$$\{\text{Simulated event (Sample)}\} = \mu + \{\text{Random normal deviate}\}\sigma \tag{32}$$

6.3. General Framework

The previous sections discussed the process whereby random samples are drawn from an underlying probability distribution. The general procedure can be described in four steps:

Step 1: Determine the probability distribution(s) for the significant factors, as illustrated in Figure 11.

TABLE 10 Random Numbers[a]

139	407	027	030	530	687	694	017	943	787
073	886	255	332	037	264	341	948	462	774
075	259	224	042	332	890	196	693	988	467
254	352	917	614	273	643	994	956	128	193
096	119	694	625	095	727	846	565	868	405
459	637	289	778	407	468	234	472	567	681
577	111	813	903	194	321	019	757	959	726
062	868	748	951	815	863	435	621	154	365
895	362	955	001	004	798	091	394	637	554
438	170	667	256	871	953	972	528	265	370
424	995	495	044	900	283	436	601	275	016
963	666	423	819	951	864	219	317	274	820
539	136	809	158	257	900	430	504	249	235
011	483	389	765	429	720	553	115	557	840
615	910	272	467	450	776	447	227	934	337
958	745	941	218	680	646	347	045	488	555
026	442	257	096	854	034	862	896	705	447
178	578	454	305	080	768	977	233	443	091
149	856	142	171	844	800	051	635	937	689
047	106	304	149	003	210	819	804	796	572
357	279	299	816	794	199	389	569	005	190
939	454	864	876	825	097	246	882	922	123
027	834	106	157	081	356	250	823	284	073
230	747	510	611	920	554	634	594	197	869
532	647	935	317	078	396	009	523	148	464
294	111	617	479	664	707	358	063	996	936
248	843	163	423	162	443	042	793	974	488
506	670	559	604	431	680	793	415	692	449
551	546	165	599	706	623	723	758	136	270
242	550	713	112	597	599	314	775	663	531
814	883	315	971	087	061	427	544	008	935
876	874	453	128	536	588	296	268	281	309
413	977	988	663	678	882	530	275	967	607
784	769	154	777	623	772	114	018	923	907
723	954	560	800	855	210	407	076	386	412
340	360	190	184	234	276	143	151	964	450
119	939	405	508	993	172	432	073	641	475
920	770	938	474	743	226	758	792	778	064
976	057	899	910	468	891	980	389	108	921
898	126	771	771	526	746	333	066	740	873
669	432	416	134	653	493	427	152	160	875
649	553	066	201	957	961	245	098	226	003
573	190	331	302	924	103	147	484	173	461
549	174	196	889	412	997	868	013	610	577
062	457	020	541	656	846	516	512	522	805

[a] This table was prepared by Mr. Ken Molay using a PDP 22 computer with a TOPS 10 operating system. To produce these random numbers, a congruential multiplicative generator was used, based on a seed of system time in milliseconds.

Step 2: Using Monte Carlo simulation, select random samples from these factors according to their relative probabilities of occurring in the future. (See Figure 12.) Note that the selection of one factor (price, for example) may determine the probability distribution of another factor (total amount demanded, for example).

Step 3: Determine the figure of merit (rate of return or present worth, for example) for each combination of factors. One trial consists of one calculation of the figure of merit.

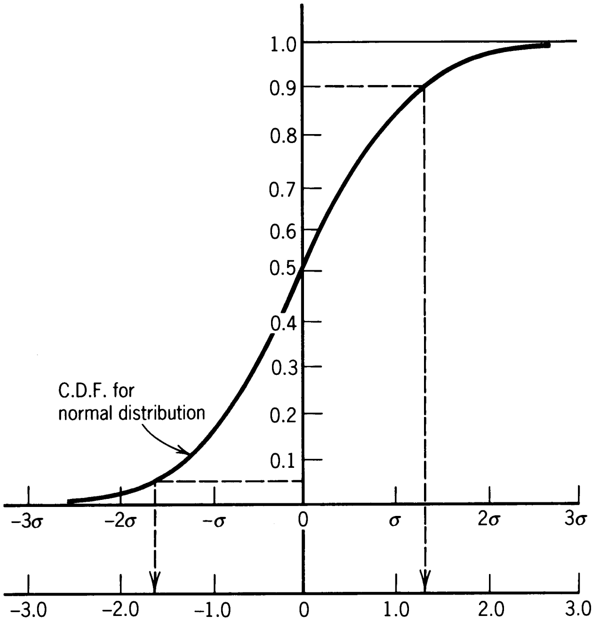


Figure 10 Sampling from a Normal Distribution.

Step 4: Repeat the process, that is, conduct a series of trials, building a frequency histogram with the results, as in Figure 13. Continue until you are reasonably satisfied that the histogram yields a clear portrayal of the investment risk. (There is no universally accepted rule for determining the optimum number of trials. It is clearly less expensive to produce a smaller number of trials, yet a larger number of trials yields more information. Substantial literature is addressed to this interesting problem, but additional discussion is not warranted here.)

6.4. Numerical Example

A certain investment, if purchased, will result in annual operating savings described by the probability distribution shown in column (b) of Table 12. The project life is described by the probability distribution shown in column (e), and the initial cost is a normally distributed random variable with $\mu = \$25,000$ and $\sigma = \$1000$. The discount rate to be used in the analysis, a certainty estimate, is 0.10. It is assumed that the annual savings, project life, and initial cost are independent random variables. If there are no other relevant consequences of the proposed investment, the expected net present value is given by the equation

$$\text{PW} = \bar{A}(\text{P}/\bar{A},\ 10\%,\ N) - P \tag{33}$$

where $\bar{A}$ = annual operating savings, N = project life, and P = initial cost.

Columns (c) and (f) of Table 12 contain the random numbers corresponding to the relative probabilities in columns (b) and (e), respectively. Note that two-digit random numbers are used. Inasmuch as the specified accuracy of the probability distributions is two significant digits, the corresponding random numbers must be specified by *at least* two digits. (A three-digit random number, say 843, could be rounded to 84 or simply truncated after the first two digits.) For the first variable, annual operating savings ($\bar{A}$), there are 100 random numbers, the first four of which correspond to the event $\bar{A} = \$2400$. The next eight numbers, 05 through 12, correspond to the event $\bar{A} = \$2500$, and so on.

The results of 10 simulated trials are shown in Table 13. Consider the first trial, for example: a random number, 09, is drawn from the table of random numbers in Table 10. As shown in Table 12, this corresponds to the event $\bar{A} = \$2500$. Next, a new random number, 52, is drawn, which corresponds to the event $N = 30$. Note that *the same random number cannot be used for both random variables because they are independent.* The third random variable, P, is normally distributed, so a random normal deviate (RND) is drawn from Table 11. This number, 0.464, indicates a simulated value for

TABLE 11 **Random Normal Deviates**[a]

0.199	−0.066	−0.205	0.455	−2.023	−0.131	−0.032	1.050
1.344	0.421	−0.599	−0.575	0.231	−0.455	1.977	2.029
−0.362	1.112	−0.200	0.072	1.044	1.399	0.910	−1.630
−0.451	−0.413	−0.159	1.421	0.286	0.499	1.402	0.750
−1.477	−0.149	−1.234	−0.644	−1.753	−0.895	1.393	0.853
−0.392	0.977	0.603	0.851	−1.161	0.206	0.294	−0.270
1.341	0.009	−1.489	0.499	0.695	−1.284	−0.542	0.682
−0.993	1.078	0.194	0.231	0.615	−1.436	−0.019	0.928
−0.708	−0.134	−0.308	1.797	−0.354	−0.445	0.019	1.355
−0.336	2.044	0.199	−0.401	−0.929	−1.964	−0.746	−0.229
0.307	−0.998	−1.083	0.104	−1.385	−1.224	0.428	0.607
−1.361	−0.203	0.675	−0.761	−0.092	−1.309	−0.966	−0.335
0.467	−0.256	0.788	0.72	−0.349	−1.401	0.205	1.043
0.373	−1.472	0.334	−0.361	−2.519	−0.658	−0.249	−1.017
1.517	0.615	−1.414	−0.665	−0.701	−0.105	−0.78	−0.266
−1.659	−0.902	−0.883	−1.679	−0.197	−1.329	0.596	−0.419
1.078	0.274	0	0.926	−1.557	−0.610	1.554	−0.139
−0.388	−1.048	−1.135	−0.878	−1.705	0.275	0.535	−0.488
0.008	0.184	−0.208	0.236	−0.134	−0.705	0.202	0.354
−2.998	−0.165	−0.295	−0.282	−0.709	1.024	0.029	1.179
0.051	−1.229	−1.265	0.440	0.593	0.276	1.053	−0.125
0.536	−0.367	2.430	0.312	0.431	0.987	0.335	0.505
1.761	0.349	−1.039	−0.814	0.299	−0.057	0.970	1.705
0.365	0.250	1.426	−1.042	−0.822	−1.065	0.708	−0.144
0.921	0.190	0.385	1.674	0.483	−0.863	−0.743	2.513
−1.308	−0.892	1.333	−0.127	−0.590	−1.590	−0.470	0.159
0.647	0.879	0.094	−0.464	0.093	0	0.614	0.393
−0.603	−0.333	−0.373	−0.523	−0.058	−1.294	0.321	−1.855
−0.214	−0.699	−0.292	0.928	0.363	0.035	0.645	−1.243
1.223	−0.868	−0.397	−0.047	0.870	−0.613	0.174	1.602
−0.649	−0.244	0.008	−0.611	0.958	−0.940	2.080	0.964
−2.215	1.712	0.941	0.537	−1.221	0.263	0.893	1.171
0.630	−0.602	−0.401	0.922	−0.734	1.992	−0.310	−1.030
−0.516	0.539	1.148	−0.373	−0.805	1.855	−0.115	−0.773
0.764	−1.190	−0.150	0.396	1.620	0.575	−0.049	−0.279
−0.519	0.772	0.817	1.003	0.306	−1.761	−0.841	−1.099
−0.144	1.254	−0.661	0.890	0.645	1.618	−1.800	−0.297
0.469	0.514	−0.304	−0.166	1.145	1.018	−0.080	0.030
1.871	0.048	−0.075	0.105	−0.617	−1.945	1.378	0.782
−2.306	−1.901	1.636	−0.725	0.264	0.169	−0.337	−0.208

[a] This table was prepared by Mr. Shay Bao Lai using an Apple II computer. The algorithm for producing these random normal deviates is from A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis,* McGraw-Hill, New York, 1982, pp. 258–259.

$$
\begin{aligned}
P &= \mu + (\text{RND})\sigma \\
&= \$25{,}000 + (0.464)(\$1000) = \$25{,}464
\end{aligned}
$$

The present worth for the first trial can now be computed:

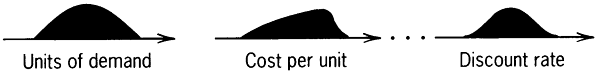


Figure 11 Probability Functions for Inputs.

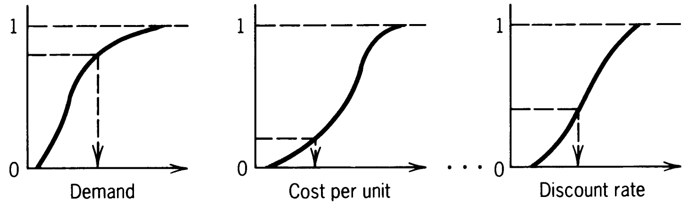


Figure 12 Cumulative Distribution Functions.

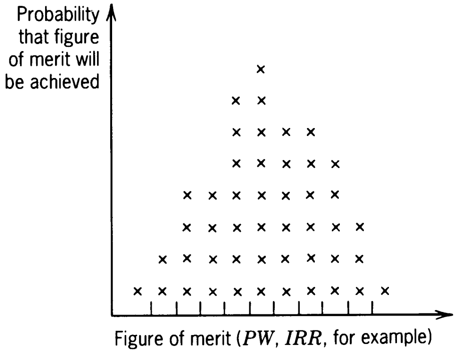


Figure 13 Frequency Histogram for Figure of Merit.

$$\text{PW} = \$2500(P/\bar{A}, 10\%, 30) - \$25{,}464 = -\$737$$

A frequency distribution can be developed from the resulting PW values, and relevant statistics can be computed. In this example the cumulative average PW after 10 trials is $2023; 40% of the trials result in a negative PW. The minimum value simulated was −$737; the maximum value simulated was $6761. Of course, if this were an actual application, the number of trials would be much larger, perhaps several thousand or more, and we would have considerably greater confidence in the resulting statistics.

In this particular example there are relatively few random variables, the relationships are not complex, and the variables are independent. Hence it is possible to compute the *theoretical expected value* for the net present value. That is

$$E[\text{PW}] = \text{E}[\bar{A}](P/\bar{A}, 10\%, E[N]) - E[P] \tag{34}$$

where

$$\begin{aligned} E[\bar{A} &= 0.04(\$2400) + 0.08(\$2500) + \ldots + 0.02(\$3300) \\ &= \$2848 \\ E[N] &= 0.05(25) + 0.10(26) + \ldots + 0.05(35) \\ &= 30 \\ E[P] &= \$25{,}000 \end{aligned}$$

Thus,

$$E[\text{PW}] = \$2848(P/A, 10\%, 30) - \$25{,}000 = \$3170$$

We can expect, therefore, that the cumulative average PW will approach $3170 as the number of trials increases. Although we can compute the theoretical *mean,* it is not possible in this case to compute the theoretical *distribution* for the PW. However, an approximate distribution can be developed through simulation.

TABLE 12 Example Simulation Problem: Input Data

Annual operating savings (a)	Probability (b)	Corresponding random numbers (c)	Project life (d)	Probability (e)	Corresponding random numbers (f)
$2400	0.04	01–04	25	0.05	01–05
2500	0.08	05–12	26	0.10	06–15
2600	0.10	13–22	27	0.10	16–25
2700	0.12	23–34	28	0.10	26–35
2800	0.14	35–48	29	0.10	36–45
2900	0.16	49–64	30	0.10	46–55
3000	0.16	65–80	31	0.10	56–65
3100	0.10	81–90	32	0.10	66–75
3200	0.08	91–98	33	0.10	76–85
3300	0.02	99–00	34	0.10	86–95
			35	0.05	96–00
	1.00			1.00	

TABLE 13 Example Simulation Problem: Simulated Trials

Trial	Random Number	Operating Savings	Random Number	Project Life (years)	Random Number Deviate	Initial Cost	Present Worth	Cumulative Average PW
1	09	$2500	52	30	0.464	$25,464	−$737	−$737
2	54	2900	80	33	0.137	25,137	3979	1621
3	42	2800	45	29	2.455	27,455	72	1105
4	01	2400	68	32	−0.323	24,677	−689	656
5	80	3000	59	31	−0.068	24,932	4904	1506
6	06	2500	48	30	0.296	25,269	−569	1160
7	06	2500	12	26	−0.288	24,712	−682	897
8	26	2700	35	28	0.060	24,940	1,426	963
9	57	2900	91	34	−2.526	22,474	6761	1607
10	79	3000	89	34	−0.531	24,469	5774	2023

7. OTHER APPROACHES FOR DEALING WITH THE UNCERTAIN/RISKY FUTURE

As indicated at the beginning of this chapter, risk and uncertainty are inherent in the general problem of resource allocation because all decisions depend on estimates about the noncertain future. Thus, risk and uncertainty have occupied the attention of a great many theoreticians and practitioners. A substantial number of approaches have been proposed, several of which were summarized earlier. Now four additional approaches are briefly identified. The first three are widely used in industry, despite certain important shortcomings; the fourth requires detailed discussion beyond the scope of this text.

7.1. Increasing the Minimum Attractive Rate of Return

Some analysts advocate adjusting the minimum attractive rate of return to compensate for risky investments, suggesting that, since the future is uncertain, stipulation of a minimum attractive rate of return of, say, $i + \Delta i$ will ensure that i will be earned in the long run. Since some investments will not turn out as well as expected, they will be compensated for by the incremental ‘‘safety margin,’’ Δi. This approach, however, fails to come to grips with the risk or uncertainty associated with estimates for specific alternatives, and thus an element Δi in the minimum attractive rate of return penalizes all alternatives equally.

7.2. Differentiating Rates of Return by Risk Class

Rather than building a safety margin into a single minimum attractive rate of return, some firms establish several risk classes with separate standards for each class. For example, a firm may require low-risk investments to yield at least 15% and medium-risk investments to yield at least 20%, and it may define a minimum attractive rate of return of 25% for high-risk proposals. The analyst then judges which class a specific proposal belongs in, and the relevant minimum attractive rate of return is used in the analysis. Although this approach is a step away from treating all alternatives equally, it is less than satisfactory in that it fails to focus attention on the uncertainty associated with the individual proposals. No two proposals have precisely the same degree of risk, and grouping alternatives by class obscures this point. Moreover, the attention of the decision maker should be directed to the causes of uncertainty, that is, to the individual estimates.

7.3. Decreasing the Expected Project Life

Still another procedure frequently employed to compensate for uncertainty is to decrease the expected project life. It is argued that estimates become less and less reliable as they occur further and further into the future; thus shortening project life is equivalent to ignoring those distant, unreliable estimates. Furthermore, distant consequences are more likely to be favorable than unfavorable; that is, distant estimated cash flows are generally positive (resulting from net revenues) and estimated cash flows near date zero are more likely to be negative (resulting from startup costs). Reducing expected project life, however, has the effect of penalizing the proposal by precluding possible future benefits, thereby allowing for risk in much the same way that increasing the minimum attractive rate of return penalizes marginally attractive proposals. Again, this procedure is to be criticized on the basis that it obscures uncertain estimates.

7.4. Utility Models

In essence, utility is a single metric on the unit interval denoting the degree of desirability of an item or a quantity of items with respect to a completely defined collection of such items. Thus an item or group of items with the greatest desirability would have a utility of, say, 100, and at a least desirable item, a zero utility. All items and groups within the collection range between these extremes in an ordered fashion. Amounts of monetary receipts and disbursements would provide a utility function from 0 to 100. A monetary gamble would be reviewed in this theory as a linear combination of the amount won and lost in the gamble, with the expected utility associated with winning. Once a person's utility function is derived, the theory of utility denotes how one should act in order to remain consistent with his or her denoted goals. Accordingly, utility theory is a description of normative economic behavior based on several stated axioms.

A number of advocates of this theory have therefore recommended that utility functions be established and economic risk analysis conducted with respect to this theory. That is, projects with the greatest expected utility should be selected by rational economic decision makers. There are many compelling features to this approach. However, it also involves the required development of the utility function, which is not a simple task; the question of whose utility function should represent the firm; and other perplexing problems. Also, it has been shown that current methods of risk cash flow analysis do represent a reasonable and rational approximation of the utility theory approach. There are also challenges to the axioms of existing theories of utility. Because of these and other detractions, the utility theory approach has not enjoyed popularity among many practitioners.

REFERENCES

Hines, W. W., and Montgomery, D. C. (1980), *Probability and Statistics in Engineering and Management Science,* 2nd Ed., John Wiley & Sons, New York.

Law, A. M., and Kelton, W. D. (1982), Simulation Modeling and Analysis, McGraw Hill, New York, 1982.

Park, C. S., and Sharp-Bette, G. P. (1990), *Advanced Engineering Economics,* John Wiley & Sons, New York, pp. 353–576.

ADDITIONAL READING

There exists a very substantial literature relevant to the topics discussed in this chapter. An exhaustive compilation is neither feasible nor, in this context, particularly useful. The references included here have been selected because of their historical importance and/or their value as additional material to augment the rather limited discussion in this *Handbook.* See especially Buck (1989) for an exceptionally comprehensive bibliography.

Buck, J. R., *Economic Risk Decisions in Engineering and Management,* Iowa State University Press, Ames, 1989.

Eschenbach, T. D., and Gimpel, R. J., "Stochastic Sensitivity Analysis," *Engineering Economist,* Vol. 35, No. 4, Summer 1990, pp. 305–321.

Estes, J. H., Moor, W. C., and Rollier, D. A., "Stochastic Cash Flow Evaluation under Conditions of Uncertain Timing," *Engineering Costs and Production Economics,* Vol. 18, 1989, pp. 65–70.

Fabrycky, W. J., Thusen, G. J., and Verma, D., *Economic Decision Analysis,* 3rd Ed., Prentice Hall, Upper Saddle River, NJ, 1998, pp. 233–295.

Fleischer, G. A., Ed., *Risk and Uncertainty: Non-Deterministic Decision Making in Engineering Economy,* Monograph Series No. 2, American Institute of Industrial Engineers, Norcross, GA, 1975.

Geweke, J., Ed., *Decision Making under Risk and Uncertainty: New Models and Empirical Findings,* Kluwer, Boston, 1992.

Goyal, A. K., Tien, J. M., and Voss, P. A., ''Integrating Uncertainty Situations in Learning Engineering Economy,'' *Engineering Economist,* Vol. 42, No. 3, Spring 1997, pp. 249–257.

Hertz, D. B., ''Risk Analysis in Capital Investment,'' *Harvard Business Review,* Vol. 42, 1964, pp. 95–106.

Hertz, D. B., and Thomas, H., *Risk Analysis and Its Applications,* John Wiley & Sons, New York, 1983.

Hillier, F. S., *The Evaluation of Risky Interrelated Investments,* North-Holland, Amsterdam, 1969.

Howard, R. A., ''Decision Analysis: Practice and Promise,'' *Management Science,* Vol. 34, No. 6, June 1988, pp. 679–695.

Ouederni, B. N., and Sullivan, W. G., ''A Semi-Variance Model for Incorporating Risk into Capital Investment Analysis,'' *Engineering Economist,* Vol. 36, No. 2, Winter 1991, pp. 83–106.

Perrakis, S., and Henin, C., ''The Evaluation of Risky Investments with Random Timing of Cash Returns,'' *Management Science,* Vol. 21, No. 1, 1974, pp. 79–86.

Young, D., and Contreras, L. E., ''Expected Present Worths of Cash Flows under Uncertain Timing,'' *Engineering Economist,* Vol. 20, No. 4, Summer 1975, pp. 257–268.

Zinn, C. D., Lesso, W. G., and Motazed, R., ''A Probabilistic Approach to Risk Analysis in Capital Investment Projects,'' *Engineering Economist,* Vol. 22, No. 4, Summer 1977, pp. 239–260.

CHAPTER 92

Inflation and Price Change in Economic Analysis

Joseph C. Hartman
Lehigh University

1. INTRODUCTION TO INFLATION AND DEFLATION 2394
   1.1. Inflation, Deflation, and Purchasing Power 2394
   1.2. Measures of Inflation 2395
   1.3. Computing Periodic and Average Inflation 2395
2. INCORPORATING INFLATION INTO ECONOMIC ANALYSIS 2396
   2.1. Inflation-Free and Market Interest Rates 2396
   2.2. Actual and Constant Dollar Cash Flows 2397
   2.3. Economic Equivalence Calculations with Inflation 2398
   2.4. Differing Inflation Rates for Component Cash Flows 2400
   2.5. Differing Inflation Rates per Time Period 2400
   2.6. Relationship between Inflation and Exchange Rates 2401
3. THE EFFECTS OF INFLATION IN ECONOMIC ANALYSIS 2401
   3.1. Before-Tax Cash Flow Analysis 2401
   3.2. After-Tax Cash Flow Analysis 2403
REFERENCES 2405
ADDITIONAL READING 2405

1. INTRODUCTION TO INFLATION AND DEFLATION

The economic analysis of investment alternatives generally entails the estimation of cash flows and the application of some measure of worth, such as net present value or the internal rate of return, in order to make a decision. The estimation of these cash flows requires the estimation of prices, whether they be the price of goods sold to forecast revenues or the estimation of wages to forecast labor costs. Over time these prices change. An increase in price is known as inflation, while a decrease in price is termed deflation. These concepts and their measurement are explained in this chapter. Cash flow analysis methods are revisited under the assumption of price changes, as their effects can be significant (Fleischer 1994). This is especially true when one considers after-tax cash flow analysis, as the effects of depreciation and taxes represent one of the most important aspects of investment analysis (Park and Sharp-Bette 1990).

1.1. Inflation, Deflation, and Purchasing Power

Consider the price of something as simple as a stamp. In 1967, a typical first class postage stamp cost just 5 cents in the United States (Park 1997). In 1999, that same stamp (in that it provided the same service as in 1967) cost 33 cents. So if you had tucked away a nickel in 1967 to buy a stamp in 1999, you would be out of luck. This is not the case with all commodities, as competition and improved manufacturing efficiencies may actually lead to decreases in the price of a good or commodity over time. Consider the average price of a one-minute long distance (interstate) call. In 1970, it cost roughly $0.25 cents per minute to place a call between states in the United States. Now, prices average in the neighborhood of $0.08 per minute (Festa 1999). This effect is known as deflation.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

Although it is rare, it has happened with numerous commodities in the last decade, mainly due to deregulation and increased competition.

Changes in price affect one's purchasing power. When inflation occurs, the worth or value of money decreases in that one cannot buy as much commodity with the same amount of money as earlier. This is known as a decrease in purchasing power. In the case of deflation, the same amount of money purchases more commodity than previously possible and thus results in an increase in purchasing power.

Because investment alternatives are generally modeled as cash flows over some time horizon, it is important to include the effects of inflation in economic equivalence calculations (Thuesen and Fabrycky 1994). This chapter illustrates measures of inflation and how to include it in cash flow analysis calculations. For the remainder of this chapter, the term *inflation* is utilized exclusively and *deflation* refers to negative inflation.

1.2. Measures of Inflation

The movement of prices is tracked in a variety of ways for a variety of products, generally with the use of a price index. A price index is merely a measure of the change in price of a commodity relative to some baseline. This is generally taken as a ratio of the price at some point in time to a price at some earlier point in time. In the United States, the Consumer Price Index (CPI) is a popular method to measure inflation associated with the cost of living. The CPI tracks the movement in price of a variety of commodities and services, including food and beverages, housing, apparel, transportation, medical care, recreation, education and communication, and energy (Bureau of Labor Statistics 1999b). While the CPI tracks this bundle of commodities and services, indexes also exist for the individual commodities and services that comprise the CPI. For example, there is a conglomerate index for transportation that is composed of indexes for private transportation and public transportation. Private transportation is composed of indexes for new and used motor vehicles, gasoline, motor vehicle parts and equipment, and motor vehicle maintenance and repair. These indexes are generally available on a regional basis and/or seasonally adjusted.

For the CPI, the baseline is given at 100 for the year 1967. The CPI for the end of 1998 was 491.0, and at the end of November 1999 the index stood at 504.1 (Bureau of Labor Statistics 1999a).

While the CPI is very popular, it is not the only price index available to those in need of estimating inflation for a specific service or commodity. The Bureau of Labor Statistics publishes a number of price indexes that may or may not influence the CPI (Standard & Poor's Statistical Service 1999). To provide some semblance of the range of available data, the Bureau of Labor Statistics provides producer price indexes for selected lumber, including hardwoods, southern pine, oak, softwoods, and Douglas fir. Breakdowns to such fine detail are available for a variety of commodities. Indexes are also available for industry trade groups to provide information for specific estimates, such as construction costs and machinery.

1.3. Computing Periodic and Average Inflation

Price indexes provide a measure of the relative change in prices from year to year. Thus, the inflation rate for one period (generally a month or a year) may be computed from the relative change in the index over the corresponding period. We will use the CPI in the following calculations, but the relationships hold for any price index. We follow notation similar to (Thuesen and Fabrycky 1994).

Define the inflation for period n as f_n. This value is calculated as:

$$f_n = \frac{CPI_n - CPI_{n-1}}{CPI_{n-1}}$$

In the one-period case, the previous period ($n - 1$) represents the base period. This calculation is valid only for single period price changes. For changes over multiple periods, an average inflation rate may be calculated.

Define the average inflation rate over n periods (period t to period $t + n$), $\bar{f}$, as:

$$CPI_t(1 + \bar{f})^n = CPI_{t+n}$$

such that:

$$\bar{f} = \left[\frac{CPI_{t+n}}{CPI_t}\right]^{1/n} - 1$$

In this calculation, time t refers to the base period for calculations. Generally, the bar is dropped and this rate is referred to as f. Note that this interest rate includes the effects of compounding and is not merely an arithmetic average of CPI values.

Consider the change in the CPI from 1967. Given an index value of 100 at the end of 1967, rising to 491.0 at the end of 1998, the average annual inflation rate over those 31 years is calculated as:

$$f = \left[\frac{491.0}{100.0}\right]^{1/31} - 1 = 0.0527 = 5.27\%$$

Again, this is an average annual rate over the respective 31 years. While this may seem like a high annual rate when one considers the recent years of low inflation, one must consider the double-digit percentage inflation rates of the 1970s in the United States.

Consider the postage stamp example presented earlier. The two prices from 1967 and 1999 can be used to determine the average inflation rate for the postage stamp over the past 31 years as follows:

$$f = \left[\frac{0.33}{0.05}\right]^{1/31} - 1 = .0628 = 6.28\%$$

Thus, it can be concluded that the average annual increases in price for a postage stamp have outpaced the average increases in the cost of living over the past 31 years in the United States. This further illustrates the need to utilize specialized indexes to achieve more accurate inflation estimates.

2. INCORPORATING INFLATION INTO ECONOMIC ANALYSIS

The role of economic analysis in engineering is to determine whether engineering projects are economically viable. This generally entails determining whether a single project should be accepted or rejected, or choosing the best project from a set of feasible projects. The analysis may take on the form of cost minimization or profit maximization. Regardless of the application or situation, the procedure generally requires the estimation of relevant cash flows and their conversion to some common denominator, such as the net present value, in order to make a decision. In this section, we consider this decision process under the assumption that the cash flows are subject to inflation. After terminology is introduced, an analytic approach is outlined to handle these types of analyses, with or without inflation. Generalizations are made for cases where the inflation rate varies according to different cash flows and over time.

2.1. Inflation-Free and Market Interest Rates

The inflation rate has been defined which allows one to predict price changes. However, economic analysis requires the use of an interest rate for discounting or compounding procedures in order to reduce a set of cash flows to a common measure for analysis, such as net present value. Because cash flows may or may not be inflated, two interest rates must be defined for use in analysis.

Define the market interest rate i as the rate of interest that can be expected to be earned on investments in the marketplace. It is a function of the investments available on the market and their associated risks. Additionally, this rate includes the effect of inflation such that if there is an upward movement in prices (inflation), there is an upward movement in the market interest rate. Other terms that may be commonly used to refer to the market rate include the minimum attractive rate of return (MARR), nominal MARR, actual interest rate, effective rate, and inflated interest rate.

Define the inflation-free interest rate i' as the market interest rate with the effects of inflation removed. While this rate is fictitious on the market, because any interest rate available on the market includes the effects of inflation, it is useful in economic analysis. This rate is also termed the real interest rate.

In the previous section, we defined the inflation rate f as the periodic increase in the price of some good or service. This rate provides the link between inflation-free and market rates because the first rate excludes inflation while the latter rate includes the effects of inflation.

To convert from an inflation-free interest rate to a market interest rate, one must add inflation. This is accomplished as follows:

$$(1 + i) = (1 + i')(1 + f)$$

such that:

$$i = i' + f + i'f$$

Similarly, to convert a market interest rate to an inflation-free rate, one must remove inflation, as:

$$i' = \frac{(1 + i)}{(1 + f)} - 1$$

Thus, the periodic inflation rate provides the link between the inflation-free interest rate and the market rate.

2.2. Actual and Constant Dollar Cash Flows

As noted earlier, economic analysis may be performed on cash flows that are either inflated or not. Thus, two interest rates were defined. The market rate includes the effects of inflation, while these effects are removed from the real interest rate. Here, the corresponding cash flows are defined.

Define actual dollars as cash flows that incorporate the effects of inflation. These may be viewed as out-of-pocket dollars because they are the true expenses paid or revenues received in business transactions at any point in time. Unfortunately, a common terminology does not exist for differentiating dollars, as actual dollars are often referred to as current, nominal, future, or inflated dollars.

Constant dollar cash flows do not include the effects of inflation. As with the inflation-free interest rate, constant dollars are a fictitious concept that represents the change in purchasing power through inflation according to some baseline in time. Constant dollars are often referred to as real, deflated, or today's dollars.

The relationship between constant dollars and actual dollars lies in the inflation rate, as with the relationship between the market and inflation-free interest rates. Because constant dollars do not include the effects of inflation, they may be converted to actual dollars by adding inflation. To show this mathematically, define the actual dollar cash flow at time n as F_n and the constant dollar cash flow at time n as F'_n. To convert constant to actual dollars at the same time period, one must incorporate the effects of inflation, or:

$$F_n = F'_n(1 + f)^n$$

Similarly, to convert actual dollars to constant dollars, one must remove inflation, as follows:

$$F'_n = \frac{F_n}{(1 + f)^n}$$

Figure 1 illustrates the two realms of cash flows: actual dollars and constant dollars. As illustrated in the figure, discounting with actual dollars requires use of the market interest rate, while the

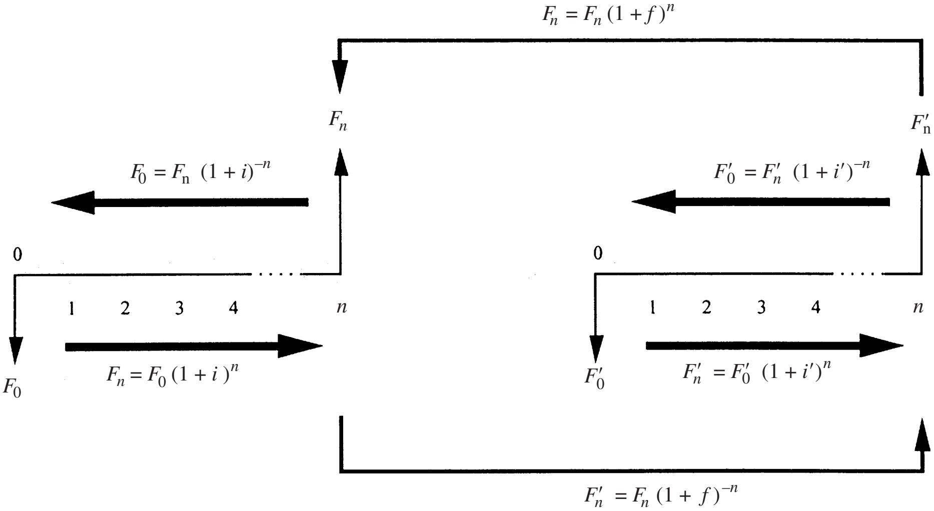

$F_n = F_n (1 + f)^n$
F_n
F'_n
$F_0 = F_\mathrm{n}\ (1 + i)^{-n}$
$F'_0 = F'_n\ (1 + i')^{-n}$
0
1
2
3
4
n
0
1
2
3
4
n
F_0
$F_n = F_0 (1 + i)^n$
F'_0
$F'_n = F'_0\ (1 + i')^n$
$F'_n = F_n (1 + f)^{-n}$

Figure 1 Time Value of Money Calculations for Actual and Constant Dollar Cash Flows and Conversions between the Two with the Inflation Rate.

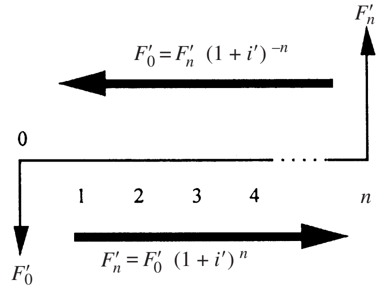


Figure 2 Time Value of Money Calculations for Constant Dolllar Cash Flows with the Inflation-free Interest Rate.

inflation-free rate is utilized with constant dollars. To convert between the interest rates or the dollar domains, the inflation rate is utilized. This is shown with the arrows that cross the domain boundaries.

2.3. Economic Equivalence Calculations with Inflation

Because there are two types of cash flows and two types of interest rates, one may perform economic analyses in one of two domains: (1) constant dollar cash flows with the inflation free interest rate or (2) actual dollar cash flows with the market interest rate. Figure 1 can be broken into the two different domains of cash flows for analysis. Figure 2 illustrates the computations in the constant-dollar domain. These calculations require the use of i'.

Similarly, Figure 3 breaks out the calculations in the actual dollar domain. As noted earlier and illustrated in Figure 1, the inflation rate provides the mathematical link between these environments because it allows one to transform constant-dollar cash flows to actual dollar-cash flows and inflation-free interest rates to market interest rates.

For a decision maker, it is critical that analysis with actual dollars be conducted with the market interest rate and analysis with constant dollars be evaluated at the inflation-free interest rate. If one is provided with constant dollars and a market interest rate or actual dollars and an inflation-free interest rate, the inflation rate must be used to convert either the dollars (from actual to constant or vice versa) or the interest rate (from market to inflation free or vice versa) to the appropriate combination for the ensuing analysis.

Example

Due to deterioration, an industrial firm replaces a piece of its machinery every year. The time zero price of the asset is \$25,000. The purchase price is expected to rise at a rate of 4.35% each year. What is the present value of the capital costs incurred by the firm over the next three years? Assume purchases occur at time zero and at the end of each of the three years and a market interest rate of 20%.

Constant Dollar Analysis Because inflationary effects on the cash flows can be ignored, the constant dollar cash flow diagram consists of four expenditures of \$25,000 each. The one caveat in

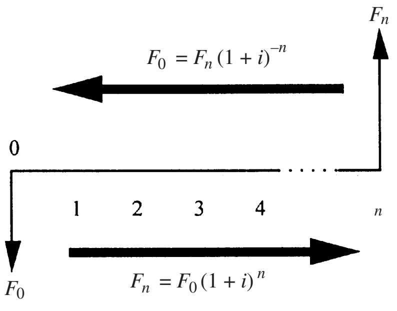


Figure 3 Time Value of Money Calculations for Actual Cash Flows with the Market Interest Rate.

this problem is that the discount rate is the market rate, not the inflation-free rate. Convert the rate as follows:

$$i' = \frac{(1 + i)}{(1 + f)} - 1 = \frac{(1 + 0.20)}{(1 + 0.0435)} - 1 = 0.14997 \simeq 15\%$$

With the inflation-free rate, the net present value (NPV) of the three expenditures in periods one through three can be brought back to time zero using the equal-payment series present-worth factor (Thuesen and Fabrycky 1994) and added to the time zero expenditure, as follows:

$$\begin{aligned} \text{NPV} &= \$25{,}000 + \$25{,}000(P/A, 15\%, 3) \\ &= \$25{,}000 + \$25{,}000 \left[\frac{(1 + 0.15)^3 - 1}{0.15(1 + 0.15)^3}\right] = \$82{,}080.63 \end{aligned}$$

An equivalent analysis can be conducted with actual dollar cash flows.

Actual Dollar Analysis The actual cash flows must be calculated in each period. To convert the constant dollar flows, the inflation rate must be utilized as follows:

$$F_n = F'_n(1 + f)^n$$

For time zero, the cash flows are the same, but for time period one:

$$F_1 = \$25{,}000(1 + .0435) = \$26{,}087.50$$

The cash flows have been calculated for each period and are listed in Table 1.

The NPV of the actual cash flows is found using the market interest rate of 20%. Because there is no pattern, each cash flow must be brought back to time zero individually as follows:

$$\text{NPV} = \$25{,}000 + \frac{\$26{,}087.50}{(1 + 0.20)} + \frac{\$27{,}222.31}{(1 + 0.20)^2} + \frac{\$28{,}406.48}{(1 + 0.20)^3} = \$82{,}082.90$$

Note that the constant dollar and actual dollar analysis lead to the same net present value of costs (within an acceptable rounding error of $2.27). To illustrate that this holds in general, the NPV value for the actual dollar cash flow can be rewritten as follows:

$$\begin{aligned} \text{NPV} &= F_0 + \frac{F_1}{(1 + i)} + \frac{F_2}{(1 + i)^2} + \frac{F_3}{(1 + i)^3} \\ &= \$25{,}000 + \frac{\$26{,}087.50}{(1 + 0.20)} + \frac{\$27{,}222.31}{(1 + 0.20)^2} + \frac{\$28{,}406.48}{(1 + 0.20)^3} \\ &= \$25{,}000 + \frac{\$25{,}000(1 + 0.0435)}{(1 + 0.20)} + \frac{\$25{,}000(1 + 0.0435)^2}{(1 + 0.20)^2} + \frac{\$25{,}000(1 + 0.0435)^3}{(1 + 0.20)^3} \\ &= F'_0 + \frac{F'_1(1 + f)}{(1 + i)} + \frac{F'_2(1 + f)^2}{(1 + i)^2} + \frac{F'_3(1 + f)^3}{(1 + i)^3} \\ &= F'_0 + \frac{F'_1}{(1 + i')} + \frac{F'_2}{(1 + i')^2} + \frac{F'_3}{(1 + i')^3} \end{aligned}$$

Therefore, if one inflation rate is assumed for all cash flows, then actual and constant dollar analyses are equivalent. This allows one to choose the more straightforward analysis for calculations. In the

TABLE 1 Constant and Actual Dollar Cash Flows for Example

Year	Constant-Dollar Cash Flow	Actual Dollar Cash Flow
0	$25,000.00	$25,000.00
1	$25,000.00	$26,087.50
2	$25,000.00	$27,222.31
3	$25,000.00	$28,406.48

previous example, the constant dollar cash flow was straightforward because the cash flows resulted in an equal payment series while the actual dollar cash flows had no pattern. Unfortunately, most applications do not allow for the decision maker to select the analysis, as inflation rates may differ by the cash flow or time period or inflation may not affect certain cash flows, such as loan payments and the cash flow effects of depreciation (in the United States). These cases are examined in the following sections.

2.4. Differing Inflation Rates for Component Cash Flows

As noted earlier, numerous price indexes exist for estimating inflation rates for various commodities. Thus, for a given analysis, one may compute a variety of inflation rates for use in a single analysis. Unless a single inflation rate affects all the cash flows involved in calculations, each component cash flow must be converted individually at the appropriate rate for proper analysis (Oakford and Salazar 1982).

Define f_k as the inflation rate for cash flow component k. If the net cash flow at time n, F_n, is made up of a number (K) of component cash flows $F_{k,n}$, then the actual net cash flow is equivalent to the sum of the actual component cash flows:

$$F_n = \sum_{k=1}^{K} F_{k,n}$$

To convert constant dollar component cash flows to a net actual dollar cash flow, one sums the individually converted cash flows as follows:

$$F_n = \sum_{k=1}^{K} F'_{k,n}\,(1 + f_k)^n$$

This is illustrated in the following example.

Example. Labor costs are estimated at \$10,000 at time zero and are assumed to increase at a rate of 3.2% per year. Time zero material costs are estimated at \$20,000 and are assumed to increase at a rate of 5%. Finally, fuel and energy costs are expected to increase at a rate of 4.5% from a time zero value of \$12,000. What are the total expected costs at the end of period 3?

Constant Dollar Analysis As all costs are given in time zero dollars, a straightforward solution would be that the total costs are:

$$\$10{,}000 + \$20{,}000 + \$12{,}000 = \$42{,}000$$

in constant dollars (with time zero serving as the baseline).

Actual Dollar Analysis The individual component inflation rates must be used to convert this to an actual dollar cash flow at the end of period 3. This is accomplished as follows:

$$\$10{,}000(1 + 0.032)^3 + \$20{,}000(1 + 0.05)^3 + \$12{,}000(1 + 0.045)^3 = \$47{,}837.54$$

This represents the expected net expenses at the end of time period 3 in actual dollars.

2.5. Differing Inflation Rates per Time Period

An additional complication is an inflation rate that changes over time. This is generally more realistic, as inflation is not a constant, which is assumed when using the average rate f defined earlier. Examining various price indexes, such as the CPI, it is clear that the inflation rate changes periodically. To perform precise economic analysis, the inflation rate must also change periodically. This is not difficult, but it complicates the analysis as calculations must be repeated.

Consider the conversion of a constant dollar cash flow at time n, F'_n, to an actual dollar cash flow F_n at the same time period:

$$F_n = F'_n(1 + f)^n$$

Because there are n periods of inflation, we may generalize this to n different periodic inflation rates. Define f_t as the inflation rate for period t. The conversion can now be written as:

$$f_n = F'_n(1 + f_1)(1 + f_2) \ldots (1 + f_n)$$

We revisit the example of the previous section to illustrate the calculation.

Example. Labor costs are estimated at \$10,000 at time zero and are assumed to increase at a rate of 2.0% the first year, 2.4% the second year, and 3.5% the third year. Time zero material costs are estimated at \$20,000 and are assumed to increase at rates of 3.0, 4.0, and 5.5% for each of the three years, respectively. Finally, fuel and energy costs are expected to increase at rates of 4.5, 4.1, and 5.2%, respectively, from a time zero value of \$12,000. What are the total expected costs at the end of period 3?

Constant Dollar Analysis As before, the constant dollar solution is as follows:

$$\$10{,}000 + \$20{,}000 + \$12{,}000 = \$42{,}000$$

in constant dollars (with time zero serving as the baseline).

Actual Dollar Analysis Again, the component inflation rates, with considerations for the change over time, must be used to convert this to an actual dollar cash flow at the end of period 3. This is accomplished as follows:

$$\$10{,}000(1 + 0.02)(1 + 0.024)(1 + 0.035) + \$20{,}000(1 + 0.03)(1 + 0.04)(1 + 0.055) \\ + \$12{,}000(1 + 0.045)(1 + 0.041)(1 + 0.052) = \$47{,}145.64$$

As before, this represents the expected net expenses at the end of time period three in actual dollars.

2.6. Relationship between Inflation and Exchange Rates

Inflation is a general term that refers to changes in price. As illustrated in this chapter, it allows one to convert a constant dollar cash flow, which does not incorporate the effects of inflation, into an actual dollar cash flow. An exchange rate, which allows for the conversion from one form of currency to another, acts in a similar manner. Exchange rates allow one to convert dollars in a given currency to another currency. Furthermore, changes in the exchange rate between two currencies over time are analogous to changes in the general inflation rate because the relative purchasing power between the two currencies changes with the exchange rate (DeGarmo et al. 1997). We examine this with the following investment example.

Example. A multinational firm based in the United States is considering an investment in Mexico. The exchange rate is currently 10 Mexican pesos to 1 U.S. dollar. (There are a variety of sources available with exchange rate data, e.g., Tukiainen 1999.) Assume an investment cost of 10 million pesos with an expected annual return of 500,000 pesos per year over a five-year horizon.

Assuming a constant exchange rate over the five-year horizon, the analysis is straightforward. In pesos, a 10 million-peso investment at time zero results in five equal annual payments of 500,000 pesos. The exchange rate converts this to U.S. dollars, in which it is equivalent to a 1 million-dollar investment at time zero with annual profits of \$50,000 per year. Because these are assumed to be constant dollars, the net present value can be computed given the appropriate inflation-free interest rate.

A more interesting question that is more directly related to the issue of inflation is whether the currency exchange rate is expected to change over time. For example, assume that the exchange rate between the U.S. and Mexico is expected to decline at 1% per year over the next five years. As the exchange rate is merely an index to convert currencies, the change in the rate corresponds to inflation. See DeGarmo et al. (1997) and Lee and Sullivan (1995) for more on this matter.

3. THE EFFECTS OF INFLATION IN ECONOMIC ANALYSIS

This section examines more complicated investment scenarios from both before-tax and after-tax cash flow perspectives. It is important to analyze these cases because inflation has a much more drastic impact on after-tax analysis due to the necessary inclusion of depreciation charges. Specifically, examples that include taxes and loans are examined to illustrate the different effects of inflation.

3.1. Before-Tax Cash Flow Analysis

The following example illustrates the effects of inflation on borrowing funds. This analysis is similar to any contract signed for future services because the cash flows are agreed upon and therefore represent out-of-pocket expenses. Thus, by definition, these expenses are actual dollar cash flows. The following example illustrates the required calculations to determine the net present value or rate of return for a project.

Example. A firm purchases a piece of machinery for \$50,000. Half of the purchase price is borrowed at an interest rate of 7% per year. Two equal payments are made at the end of years 1 and 2 to pay off the loan. The asset is expected to generate revenues of \$75,000 per year for three years,

TABLE 2 Actual Cash Flows for Investment Example

Year	0	1	2	3
Revenues		\$78,750.00	\$82,687.50	\$86,821.88
O&M expenses		(\$15,750.00)	(\$16,537.50)	(\$17,364.38)
Interest expenses		(\$1,750.00)	(\$904.59)	
Capital cost	(\$50,000)			\$11,576.25
Loan principal	\$25,000	(\$12,077.30)	(\$12,922.70)	
Before-tax cash flow	(\$25,000)	\$49,172.70	\$52,322.71	\$81,033.75

at which time it is sold for a salvage value of \$10,000. Operating and maintenance (O&M) expenses are assumed to be \$15,000 per year. All costs here are given in time zero dollars. Determine the net present value of the transaction assuming a market interest rate of 20%.

Because loan interest and principal payments are defined by a contractual agreement, they are actual dollar cash flows. Thus, we first analyze this problem from the perspective of the actual dollar domain.

Actual Dollar Analysis (*1*) In this first actual dollar analysis, we assume one (general) annual inflation rate of 5% for the revenues and expenses and the salvage value of the asset at the end of period 3. All of the actual dollar cash flows are given in Table 2. Negative cash flows are denoted in parentheses. Note that the interest and principal payments are unaffected by the inflation rate becasue they are dependent on a prior agreement.

The present value of the actual dollar before-tax cash flow is calculated using the market interest rate, as follows:

$$\text{NPV} = -\$25{,}000 + \frac{\$49{,}172.70}{(1 + 0.20)} + \frac{\$52{,}322.71}{(1 + 0.20)^2} + \frac{\$81{,}033.75}{(1 + 0.20)^3} = \$99{,}207.00$$

Constant Dollar Analysis In the original problem statement, all cash flows were presented in constant dollars. Thus, revenues, operating and maintenance expenses, and capital costs are straightforward. However, the loan principal and interest expenses are actual dollar cash flows, which means they include the effects of inflation. Thus, these cash flows must be deflated with the inflation rate of 5%. These calculated cash flows, along with the other constant cash flows, are provided in Table 3.

The present value of the constant dollar before-tax cash flow is calculated using the inflation free interest rate. The inflation-free rate is calculated from the market and inflation rates, as follows:

$$i' = \frac{1 + 0.20}{1 + 0.05} - 1 = 0.1429 = 14.29\%$$

This rate is then used to calculate the net present value with the constant dollar before-tax cash flow:

$$\text{NPV} = -\$25{,}000 + \frac{\$46{,}831.14}{(1 + 0.1429)} + \frac{\$47{,}458.24}{(1 + 0.1429)^2} + \frac{\$70{,}000.00}{(1 + 0.1429)^3} = \$99{,}197.46$$

As expected, the net present value is equivalent (within rounding error) of the actual dollar analysis.

TABLE 3 Constant Cash Flows for Investment Example

Year	0	1	2	3
Revenues		\$75,000.00	\$75,000.00	\$75,000.00
O&M expenses		(\$15,000.00)	(\$15,000.00)	(\$15,000.00)
Interest expenses		(\$1,666.67)	(\$820.49)	
Capital costs	(\$50,000)			\$10,000.00
Loan principal	\$25,000	(\$11,502.19)	(\$11,721.27)	
Before-tax cash flow	(\$25,000)	\$46,831.14	\$47,458.24	\$70,000.00

TABLE 4 Actual Cash Flows with Varying Inflation Rates for Investment Example

Year	0	1	2	3
Revenues		$78,750.00	$82,687.50	$86,821.88
O&M expenses		($15,450.00)	($15,913.50)	($16,390.91)
Interest expenses		($1,750.00)	($904.59)	
Capital costs	($50,000)			$12,597.12
Loan principal	$25,000	($12,077.30)	($12,922.70)	
Before-tax cash flow	($25,000)	$49,472.70	$52,946.71	$83,028.09

Actual Dollar Analysis (*2*) Because it is more realistic to assume inflation rates that vary for each cash flow component, we revisit the actual dollar analysis. Assume that the revenues grow at a rate of 5% per period, O&M expenses increase at a rate of 3% and the salvage value grows at a rate of 8% per period. The revised actual dollar cash flows for this example are given in Table 4. As before, interest expenses, the initial purchase cost, and the loan principal payments are unaffected.

Using the market interest rate of 20%, the net present value of the actual dollar before-tax cash flows is $101,044.46. Because O&M costs were projected lower and the salvage value higher than in the previous actual dollar cash flow analysis, this higher net present value was expected.

The calculations are not repeated here for the corresponding constant dollar analysis, as the cash flows are the same as before. However, there is one final point that must be addressed. In this problem instance, the market rate was provided. This allows for the discounting of actual dollar cash flows. Generally, with the use of an inflation rate, the market rate can be converted to an inflation-free rate. However, in this final example with component cash flows having different inflation rates, this analysis is not straightforward. In this case, a *general* inflation rate (Park 1997) is generally specified such that the inflation free rate can be calculated. This rate would generally correspond to a broader index, such as the CPI, whereas the component inflation rates would be derived from more specific indexes. If an inflation-free rate had been specified in the problem, then a general inflation rate would have also had to be specified to determine a single market interest rate for discounting.

3.2. After-Tax Cash Flow Analysis

As shown in the previous section, loans and agreed-upon prices can complicate analyses because not all cash flows are defined as either constant or actual dollars, but rather a mix. Performing after-tax analysis presents further complicating issues. However, it is generally recognized that after-tax analyses should be performed when economically validating engineering projects (Park and Sharp-Bette 1990).

When considering after-tax cash flows, some form of actual dollar analysis must be performed because taxes are paid on profits from actual revenues. This is not to say that constant dollar analyses cannot be performed, but the actual dollars must generally be estimated before conversion to constant dollars can take place.

The following after-tax analyses revisit the before-tax transaction from the previous section. It should be noted that the goal of this section is not to present current tax law, which is forever changing and in which rules are different according to the location where the asset is in use and the time the asset is both placed in and removed from service. Rather, the point here is to illustrate the effect of inflation on differing components of cash flows common to after-tax analysis.

Example. The example analyzed as a before-tax transaction is again examined here. It is assumed that the asset is depreciated using the straight-line method over three years. For simplicity, no adjustments to the amount of depreciation are made at the time of sale.

Because taxes are paid on actual revenues and expenses, the actual dollar cash flow is analyzed first. A market interest rate of 15% is assumed.

Actual Dollar Analysis Assume a single annual inflation rate of 5% for the revenues, expenses, and salvage value of the asset at the end of period 3. As before, the interest and principal payments are not affected by inflation. Table 5 provides the actual cash flows for this example. Note that depreciation expenses are provided in order to calculate taxes although they are not cash flows. For the sake of simplicity, we assume a marginal tax rate of 40%, which is also applied to capital gains.

The after-tax cash flow, which is generally found on a cash flow statement, adds the after-tax profit, purchase costs, and loan principal payments and also adds back in the depreciation expenses because depreciation is not a cash flow. The net present value of the actual dollar after-tax cash flow is then computed at the market rate of 15%. The NPV for this example is $63,474.15.

TABLE 5 Actual Cash Flows for After-Tax Investment Example

Year	0	1	2	3
Revenues		$78,750.00	$82,687.50	$86,821.88
O&M expenses		($15,750.00)	($16,537.50)	($17,364.38)
Interest expenses		($1,750.00)	($904.59)	
Depreciation expenses		($16,666.67)	($16,666.67)	($16,666.67)
Before-tax operating profit		$44,583.33	$48,578.74	$52,790.83
Salvage values				$11,576.25
Tax on capital gain				($4,630.50)
Total taxes		($17,833.33)	($19,431.50)	($25,746.83)
After-tax profit		$26,750.00	$29,147.24	$38,620.25
Purchase cost	($50,000)			
Loan principal	$25,000	($12,077.30)	($12,922.70)	
After-tax cash flow	($25,000)	$31,339.37	$32,891.21	$55,286.92

Constant Dollar Analysis Unlike previous examples presented in this chapter, the constant dollar analysis for after-tax cash flows requires the computation of actual dollar cash flows because taxes are actual dollar cash flows. Although the net actual dollar after-tax cash flow can be converted to a net constant dollar cash flow for analysis, we explicitly show each of the component cash flows in Table 6 because if the component inflation rates differ, each component cash flow must be analyzed with the appropriate rate. The cash flows given in Table 5 as depreciation expenses are no longer needed as the taxes have been calculated. The general 5% inflation rate is used here to deflate all actual dollar cash flows.

With a market interest rate of 15% and a general inflation rate of 5%, the inflation-free rate is 9.52%. This leads to a net present value of $63,480.63, which is equivalent to the actual dollar analysis (within rounding error). This is expected because all of the component cash flows were inflated at the same rate. As noted in the before-tax analysis, if the component inflation rates differ, then there may be a discrepancy in the final net present value analysis because the conversion from a market rate to an inflation-free rate (or vice versa) requires a single inflation rate.

This may lead one to ask the question of which analysis should be performed. There is no correct answer. However, as these examples illustrated, more complicated cash flow scenarios generally lead to actual dollar cash flow analysis. This is because the computation of one's tax liability is based on actual dollar cash flows. Also, contracted prices, such as loans and lease agreements, are based on actual dollar cash flows. In these situations, actual dollar cash flow analysis may be more straightforward because most of the data are also in the correct form.

Regardless of the method chosen, it is important to include inflation in analysis because it may have drastic effects. Consider the after-tax example analyzed in this section. Table 7 provides the cash flows ignoring inflation. Note that these flows are not equivalent to constant dollar cash flows, as the interest and principal payments remain as actual dollar flows. The present value, at the 15% market rate, is $54,765.84. In this example, ignoring inflation leads to a lower net present value and thus may lead to a different decision. Inflation can have various effects on project cash flows and thus greatly influence decisions. In general, inflation results in overstated income and higher taxes. This is sometimes referred to as tax bracket creep because inflation may push a company's revenues, and thus profits, into a higher tax bracket. Similarly, inflated salvage values result in larger capital gains and resulting taxes. As shown in the above example, inflation can actually reduce the cost of

TABLE 6 Constant Cash Flows for After-Tax Investment Example

Year	0	1	2	3
Revenues		$75,000.00	$75,000.00	$75,000.00
O&M expenses		($15,000.00)	($15,000.00)	($15,000.00)
Interest expenses		($1,666.67)	($820.49)	
Salvage values				$10,000.00
Total taxes		($16,984.12)	($17,624.94)	($22,241.08)
Purchase cost	($50,000)			
Loan principal	$25,000	($11,502.19)	($11,721.27)	
After-tax cash flow	($25,000)	$29,847.02	$29,833.30	$47,758.92

TABLE 7 Cash Flows Ignoring Inflation for After-Tax Investment Example

Year	0	1	2	3
Revenues		$75,000.00	$75,000.00	$75,000.00
O&M expenses		($15,000.00)	($15,000.00)	($15,000.00)
Interest expenses		($1,750.00)	($904.59)	
Salvage values				$10,000.00
Total taxes		($16,633.33)	($16,971.50)	($21,333.33)
Purchase cost	($50,000)			
Loan principal	$25,000	($12,077.30)	($12,922.70)	
After-tax cash flow	($25,000)	$29,539.37	$29,201.21	$48,666.67

financing because that cost is fixed at the agreed upon rate. In summary, the effects of inflation can alter cash flows significantly and thus strongly influence the decision to accept or reject a project. Therefore, they should not be ignored in any economic analysis.

REFERENCES

Bontadelli, J. A. and W. G. Sullivan (1980), ‘‘How an IE Can Account for Inflation in Decision-Making,’’ *Industrial Engineering,* Vol. 12, No. 3, pp. 24–33.

Estes, C. B., Turner, W. C., and Case, K. E. (1980), ‘‘Inflation—Its Roles in Engineering-Economic Analysis,’’ *Industrial Engineering,* Vol. 12, No. 3, pp. 18–22.

Freidenfelds, J., and Kennedy, M., (1979), ‘‘Price Inflation and Long-Term Present-Worth Studies,’’ *Engineering Economist,* Vol. 24, No. 3, pp. 143–160.

Freidenfelds, J., and Kennedy, M. (1982), ‘‘The Arithmetic of Inflation Corrections . . .—Comment,’’ *Engineering Economist,* Vol. 27, No. 2, pp. 144–146.

Jones, B. W., *Inflation in Engineering Economic Analysis,* John Wiley & Sons, New York, 1982.

Standard & Poor’s Statistical Service (1999), ‘‘Current Statistics,’’ *Basic Statistics,* updated monthly.

White, J. A., Case, K. E., Pratt, D. B., and Agee, M. H., *Principles of Engineering Economic Analysis,* 4th Ed., John Wiley and Sons, New York.

ADDITIONAL READING

Bureau of Labor Statistics (1999a), ‘‘Consumer Price Index Summary,’’ stats.bls.gov/cpihome.htm, U.S. Department of Labor, Washington, DC.

Bureau of Labor Statistics (1999b), ‘‘Understanding the Consumer Price Index: Answers to Some Questions,’’ stats.bls.gov/cpihome.htm, United States Department of Labor, Washington, DC.

DeGarmo, E. P., Sullivan, W. G., Bontadelli, J. A., and Wicks, E. M. (1997), *Engineering Economy,* 10th Ed., Prentice Hall, Upper Saddle River, NJ.

Festa, M. (1999), ‘‘Net Phone Market Heats up,’’ CNET News.com, March 11.

Fleischer, G. A. (1994), *Introduction to Engineering Economy,* PWS, Boston.

Lee, P. M., and Sullivan, W. G. (1995), ‘‘Considering Exchange Rate Movements in Economic Evaluation of Foreign Direct Investments,’’ *Engineering Economist,* Vol. 40, No. 2, pp. 171–191.

Oakford, R. V., and Salazar, A. (1982), ‘‘The Arithmetic of Inflation Corrections in Evaluating ‘Real’ Present Worths,’’ *Engineering Economist,* Vol. 27, No. 2, pp. 127–143.

Park, C. S. (1997), *Contemporary Engineering Economics,* 2nd Ed., Addison-Wesley, Menlo Park, CA.

Park, C. S., and Sharp-Bette, G. P. (1990), *Advanced Engineering Economics,* John Wiley & Sons, New York.

Thuesen, G. J., and Fabrycky, W. J. (1994), *Engineering Economy,* 8th Ed., Prentice Hall, Upper Saddle River, NJ.

Tukiainen, M. (1999), ‘‘Rates: Currency Resources on the Net,’’ www.uta.fi/ktmatu/rate-curres.html.

V.C

Computer Simulation

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 93

Modeling Human Performance in Complex Systems

K. Ronald Laughery, Jr.
Susan Archer
Micro Analysis and Design, Inc.

Kevin Corker
San José State University

1. INTRODUCTION **2410**
2. OBJECTIVES OF THIS CHAPTER **2411**
3. THE TYPES OF QUESTIONS THAT ARE BEING ADDRESSED BY HUMAN PERFORMANCE MODELS **2411**
4. THE TWO CLASSES OF SIMULATION MODELS OF HUMAN–SYSTEM PERFORMANCE **2412**
5. AN EXAMPLE OF A REDUCTIONIST APPROACH: TASK NETWORK MODELING **2414**
5.1. What Goes into a Task Network Model? 2414
5.2. An Example of a Task Network Model of a Process Control Operator 2419
5.3. Case Studies in the Use of Task Network Modeling to Address Specific Design Issues 2420
5.4. Using Task Network Modeling to Evaluate Crew Workload 2420
5.4.1. Modeling the Workload of a Future Command and Control Process 2421
5.4.2. Extensions of This Approach to Simulating Crew Mental Workload to Other Environments 2424
5.5. Using Task Network Modeling as a Means of Extending Research Findings on Human Performance under Stress to New Task Environments 2427
5.5.1. The Taxonomy 2427
5.5.2. Degradation Functions 2427
5.6. Summary of Examples of Task Network Modeling of Human–System Performance 2429
6. AN EXAMPLE OF A FIRST PRINCIPLED APPROACH TO HUMAN/SYSTEM PERFORMANCE MODELING: THE MAN—MACHINE INTEGRATED DESIGN SYSTEM (MIDAS) **2429**
6.1. Background 2430
6.2. System Architecture 2431
6.2.1. MIDAS Functional Architecture 2431
6.3. MIDAS Structural Architecture 2434
6.3.1. Working and Long-Term Memory Stores 2434
6.3.2. Attentional Control 2435
6.3.3. Activity Representation 2435
6.3.4. Task Agenda 2435
6.3.5. Decision Making 2435
6.3.6. Higher-Level Functions 2436
6.4. Case Studies in MIDAS Applications to Aviation 2436
6.4.1. MIDAS Case Study 1: Predicting Flight Crew Performance in the Advanced Air Traffic Management System 2436
6.4.2. Extension of Model to Air Traffic Control 2439
6.5. Other Modeling Strategies That Have Demonstrated Utility in Modeling Human Performance in Systems 2440

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

6.5.1. Sample Applications 2441
7. SUMMARY 2441
REFERENCES 2441
ADDITIONAL READING 2444

1. INTRODUCTION

Over the past few decades, human factors and ergonomics practitioners have increasingly been called upon early in the system design and development process. Early input from all disciplines results in better and more integrated designs as well as lower costs than if one or more disciplines finds that changes are required later. Our goal as human factors and ergonomics practitioners should be to provide substantive and well-supported input regarding the human(s), his or her interaction(s) with the system, and the resulting total performance. Furthermore, we should be prepared to provide this input from the earliest stages of system concept development and then throughout the entire system or product life cycle.

To meet this challenge, many human factors and ergonomics tools and technologies have evolved over the years to support early analysis and design. Two specific types of technologies are design guidance (e.g., O'Hara et al. 1995; Boff et al. 1986) and high-fidelity rapid prototyping of user interfaces (e.g., Dahl et al. 1995). Design guidance technologies, in the form of either handbooks or computerized decision support systems, put selected portions of the human factors and ergonomics knowledge base at the fingertips of the designer, often in a form tailored to a particular problem such as nuclear power plant design or UNIX computer interface design. However, design guides have the shortcoming that they do not often provide methods for making quantitative trade-offs in *system* performance as a function of design. For example, design guides may tell us that a high-resolution color display will be better than a black-and-white display, and they may even tell us the value in terms of increased response time and reduced error rates. However, this type of guidance will rarely provide good insight into the value of this improved element of the human's performance to the *overall system's* performance. As such, design guidance has limited value for providing concrete input to system-level performance prediction.

Rapid prototyping, on the other hand, supports analysis of how a specific design and task allocation will affect human and system-level performance. The disadvantage of prototyping, as with all human-subjects experimentation, is that it can be costly. In particular, prototypes of hardware-based systems, such as aircraft and machinery, are very expensive to develop, particularly at early design stages when there are many widely divergent design concepts. In spite of the expense, hardware and software prototyping are important tools for the human factors practitioner, and their use is growing in virtually every application area.

While these technologies are valuable to the human factors practitioner, what is often needed is an integrating methodology that can extrapolate from the base of human factors and ergonomics data, as reflected in design guides and the literature, in order to support system-level performance predictions as a function of design alternatives. This methodology should also bind with rapid prototyping and experimentation in a mutually supportive and iterative way. As has become the case in many engineering disciplines, a prime candidate for this integrating methodology is computer modeling and simulation.

Computer modeling of human behavior and performance is not a new endeavor. Computer models of complex cognitive behavior have been around for over 20 years (e.g., Newell and Simon 1972) and tools for computer modeling of task-level performance have been available since the 1970s (e.g., Wortman et al. 1978). However, two things have changed appreciably in the past decade that promote the use of computer modeling and simulation of human performance as a standard tool for the practitioner. First is the rapid increase in computer power and the associated development of easier-to-use modeling tools. Individuals with an interest in predicting human performance through simulation can select from a variety of computer-based tools (for a comprehensive list of these tools, see McMillan et al. 1989). Second is the increased focus by the research community on the development of *predictive* models of human performance rather than simply descriptive models. For example, the GOMS model (Gray et al. 1993) represents the integration of research into a model for making predictions of how humans will perform in a realistic task environment. Another example is the research in cognitive workload that has been represented as computer algorithms (e.g., McCracken and Aldrich 1984; Farmer et al. 1995). Given a description of the tasks and equipment with which humans are engaged, these algorithms support assessment of when workload-related performance problems are likely to occur and often include identification of the quantitative impact of those problems on overall system performance (Hahler et al. 1991). These algorithms are particularly useful when embedded as key components in computer simulation models of the tasks and the environment.

Perhaps the most powerful aspect of computer modeling and simulation is that it provides a method through which the human factors and ergonomics team can step up to the table with the other engineering disciplines that also rely increasingly on quantitative computer models. What we will discuss in this chapter are the methods through which the human factors and ergonomics community can contribute early to system design tradeoff decisions.

2. OBJECTIVES OF THIS CHAPTER

This chapter will discuss some existing computer tools for modeling and simulating human/system performance. It is intended to provide the reader with:

1. An understanding of the types of human factors and ergonomics issues that can be addressed with modeling and simulation
2. An understanding of some of the tools that are now available to assist the human factors and ergonomics specialist in conducting model-based analyses
3. An appreciation of the level of expertise and effort that will be required to use these technologies

We begin this chapter with two caveats. First, we are not yet at a point where computer modeling of human behavior allows sufficiently accurate predictions that no other analysis method (e.g., prototyping) is likely to be needed. In early stages of system concept development, high-level modeling of human–system interaction may be all that is possible. As the system moves through the design process, human factors and ergonomics designers will often want to augment modeling and simulation predictions with prototyping and experimentation. In addition to providing high-fidelity system performance data, these data can be used to enhance and refine the models. This concept of human performance modeling supporting and being supported by experimentation with human subjects is represented in Figure 1. In essence, simulation provides the human factors and ergonomics practitioner with a means of extending the knowledge base of human factors and of amplifying the effectiveness of limited experimentation.

Second, the technologies discussed here are evolving rapidly. We can be certain that (1) every tool discussed is undergoing constant change and (2) new modeling tools are being developed. We are discussing computer-based tools, and we expect the pace of change in these tools to mirror the pace in other computer tools such as word processors and spreadsheets. These detailed discussions of several of the modeling tools are included to facilitate a better understanding of human performance modeling tools. We encourage the reader to contact citations in this chapter to assess the state of any tool at that time.

3. THE TYPES OF QUESTIONS THAT ARE BEING ADDRESSED BY HUMAN PERFORMANCE MODELS

Below are a few classes of problems to which human/system modeling has been applied:

- How long will it take a human or team of humans to perform a set of tasks as a function of system design, task allocation, and individual capabilities?
- What are the trade-offs in performance for different combinations of design, task allocation, and individual capability selections?
- What are the workload demands on the human as a function of system design and automation? How will human performance and resulting system performance change as the demands of the environment change? How many individuals are required on a team to ensure safe, successful performance? How should tasks be allocated to optimize performance?
- How will environmental stressors such as heat, cold, and the use of drugs affect human–system performance?

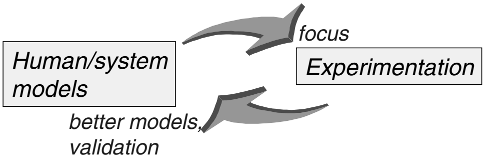


Figure 1 The Synergy between Modeling and Experimentation.

The above is a sampling rather than an exhaustive list. The tools we discuss in this chapter are inherently flexible, and we consistently discover that these tools can be used to solve problems that the tool developers never conceived.

To assess the potential of simulation to answer questions, every potential human performance modeling project should first determine the specific questions that the project is trying to answer. Then conduct a critical assessment of what is important in the human–machine system being modeled. This will define the required content and fidelity of the model. The questions that should be considered about the system include:

- *Human performance representation:* What time or duration of performance is important? How is human performance initiated and what resolution of behavior is required? What aspects of human performance, including task management, load management, and goal management, are expected?
- *Equipment representation:* What equipment is used to accomplish the tasks? To what level of functional and physical description can and should the equipment be represented? Is it operable by more than one human or system component?
- *Interface requirements:* What information needs to be conveyed to the humans and when? Is transformation of information required?
- *Control requirements:* What processes need to be controlled by the human and to what level of resolution?
- *Logical and physical constraints:* How is performance supported through equipment operability and procedural sequences? What alarms and alerts should be represented?
- *Simulation driver:* What makes the system function—the occurrence of well-defined events (e.g., a procedure), the movement of time (e.g., the control of a vehicle), or a hybrid of both?

By defining the purpose of the model and then answering the above questions, the human factors practitioner will get a sense of what is important in the system and therefore what may need to be represented in a model. In using human performance models, perhaps the most significant task of the human factors practitioner will be to determine what aspects of the human/machine system to include in the model and what to leave out. Many modeling studies have failed because of the inclusion of too many factors that, while a part of human–system performance, were not system performance drivers. Consequently, the models become overly complex and expensive to develop. In our experience, it is better to begin with a model with too few aspects of the system represented and then add to it than to begin a modeling project by trying to model everything. The first approach may succeed, while the second is often doomed.

Second, the human factors practitioner should consider the measures of effectiveness of the system that the model should be designed to predict. In building the model, it is important to remember that the goal of the model will be to predict measures of human performance that will impact system performance. Therefore, a clear definition of what is important to performance is necessary. The following aspects of performance measures should be considered:

- *Success criteria:* What operational success measures are important to the system? Can these be stated in relative terms, or must they be measured in absolute terms?
- *Range of performance to be studied:* What are the experimental variables that are to be explored by the model? How important is it to establish a range of performance for each experimental condition as a function of the stochastic (i.e., random) behavior of the system?

By asking the above questions prior to beginning a modeling project, the human factors practitioner can develop a better sense of what is important in the system in terms of both aspects that drive system performance as well as the measures of effectiveness that are truly of interest. Then and only then can a human performance modeling project begin with a reasonable hope of success.

The remainder of this chapter will discuss two classes of modeling tools for human performance simulation. After discussing each class of modeling tool, we will provide specific examples of a modeling tool and then provide case studies about how these tools have been used in answering real human performance questions.

4. THE TWO CLASSES OF SIMULATION MODELS OF HUMAN–SYSTEM PERFORMANCE

Human performance can be highly complex and involve many types of processes and behavior. Over the years, many models have been developed that predict sensory processes (e.g., Gawron et al. 1983), aspects of human cognition (e.g., Newell 1990), and human motor response (e.g., Fitts's law).

The current literature in the areas of cognitive engineering, error analysis, and human–computer interaction contains many models, descriptions, methodologies, metaphors, and functional analogies. However, in this chapter we are not focusing on the models of these individual elements of human behavior but rather on models that can be used to describe human performance in systems. These human/system performance models typically include some of these elemental behavioral models as components but provide a structural framework that allows them to be put in the context of human performance of tasks in systems.

We separate the world of human–system performance models into two general categories: *reductionist* models and *first principle* models. Reductionist models use human/system task sequences as the primary organizing structure, as shown in Figure 2. The individual models of human behavior for each task or task element are connected to this task-sequencing structure. We refer to it as reductionist because the process of modeling human behavior involves taking the larger aspects of human/system behavior (e.g., ‘‘perform the mission’’) and then successively reducing them to smaller elements of behavior (e.g., ‘‘perform the function,’’ ‘‘perform the tasks’’). This continues until a level of decomposition is reached at which reasonable estimates of human performance for the task elements can be made. One can also think of this as a top-down approach to modeling human/system performance. The example of this type of modeling that we will use in this chapter will be task network modeling, where the basis of the human–system model is a task analysis.

First-principle models of human behavior are structured around an organizing framework that represents the underlying goals and principles of human performance. Tools that support first-principle modeling of human behavior have structures embedded in them that represent elemental aspects of the human. For example, these models might directly represent processes such as goal-seeking behavior, task scheduling, sensation and perception, cognition, and motor output. To use tools that support first-principle modeling, one must describe how the system and environment interact with the modeled human processes. An example of a very simple structure that supports this type of modeling environment is presented in Figure 3. The example we will use of a tool designed to support this type of modeling is the Man–Machine Integrated Design and Analysis System (MIDAS).

It is worth noting that these two modeling strategies are not mutually exclusive and, in fact, can be mutually supportive in any given modeling project. Often, when one is modeling using a reductionist approach, one needs models of basic human behavior to represent behavioral phenomena accurately and therefore must draw on elements of first-principle models. Alternatively, when one is modeling human–system performance using a first-principled approach, some aspects of human–system performance and interrelationships between tasks may be more easily defined using a reductionist approach. Either class of model has been used to model individual and team performance. It is also worth noting that recent advances in human performance modeling tool development are blurring the distinctions between these two classes (e.g., Hoagland et al. 2001; LaVine 2000). Increased emphasis on interoperability between models has caused researchers and developers to focus on integrating reductionist and first-principle models.

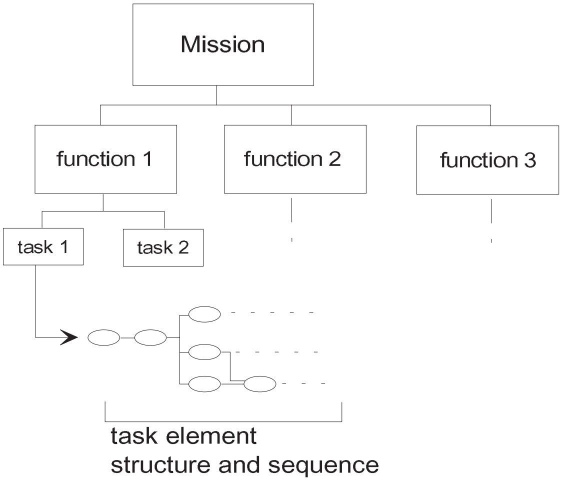


Figure 2 The Concept of Reductionist Models of Human Performance.

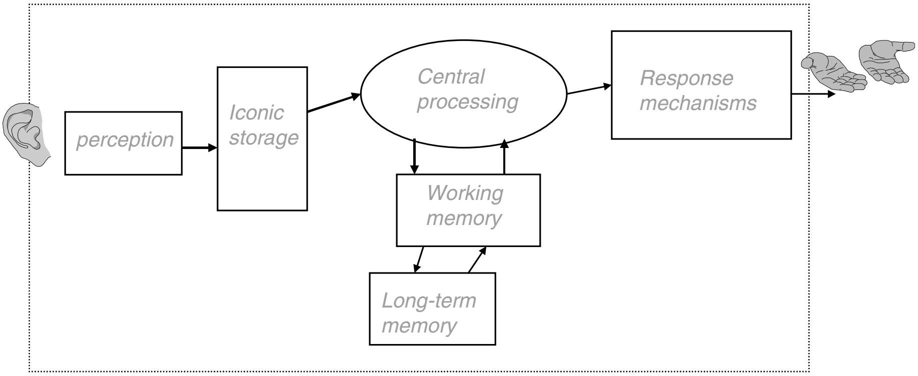


Figure 3 An Example of the Concept of First-Principle Models of Human Performance.

5. AN EXAMPLE OF A REDUCTIONIST APPROACH: TASK NETWORK MODELING

One technology that has proven useful for predicting human-system performance is task network modeling. In a task network model, human performance is decomposed into tasks. The fidelity of this decomposition can be selective, with some functions being decomposed several levels and others just one or two. This is, in human factors engineering terms, the task analysis. The sequence of tasks is defined by constructing a task network. This concept is illustrated in Figure 4, which presents a sample task network for dialing a telephone.

Task network modeling is an approach to modeling human performance in complex systems that has evolved for several reasons. First, it is a reasonable means for extending the human factors staple—the task analysis. Task analyses organized by task sequence are the basis for the task network model. Second, task network models can include sophisticated submodels of the system hardware and software to create a closed-loop representation of relevant aspects of the human/machine system. Third, task network modeling is relatively easy to use and understand. Recent advancements in task network modeling technology have made this technology more accessible to human factors practitioners. Finally, task network modeling can provide efficient, valid, and useful input to many types of issues. With a task network model, the human factors engineer can examine a design (e.g., control panel redesign) and address questions such as "How much longer will it take to perform this procedure?" and "Will there be an increase in the error rate?" Generally, task network models can be developed in less time and with substantially less effort than would be required if a prototype were developed and human subjects used. However, as stated before, for revolutionary designs, modeling may not alleviate the need for empirical data collection.

Task network models of human performance have been subjected to validation studies with favorable results (e.g., Lawless et al. 1995; However, as with any modeling approach, the real level at which validation must be considered is with respect to a particular model, not with respect to the general approach.

5.1. What Goes into a Task Network Model?

To represent complex, dynamic human/system behavior, many aspects of the system may need to be modeled in addition to simply task lists and sequence. In this subsection, we will use the task network modeling tool Micro Saint as an example.

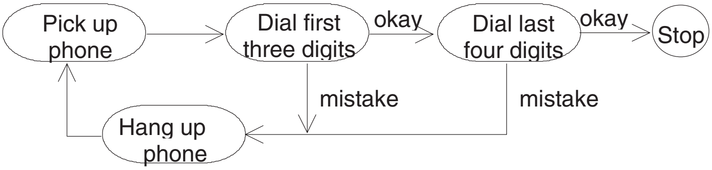


Figure 4 An Example of a Task Network Model Representing a Human Dialing a Telephone.

The basic ingredient of a Micro Saint task network model is the task analysis as represented by a network or series of networks. The level of system decomposition (i.e., how finely we decompose the tasks) and the amount of the system that is simulated depends on the particular problem. For example, in a power plant model, one can create separate networks for each of the operators and one for the power plant itself. While the networks may be independent, performance of the tasks can be interrelated through shared variables. The relationships among different components of the system, represented by different segments of the network, can then communicate through changes in these shared variables. For example, when an operator manipulates a control, this may initiate an open valve task in a network representing the plant. This could ripple through to a network representing other operators and subsystems and their response to the open valve.

This basic task network is built in Micro Saint via a point-and-click drawing palette. Through this environment, the user creates a network as shown in Figure 5. Networks can be embedded within networks, allowing for hierarchical construction. In addition, the shape of the nodes on the diagram can be chosen in order to represent specific types of activity.

To reflect complex task behavior and interrelationships, more detailed characteristics of the tasks need to be defined. By double clicking on a task, the user opens up the task description window, as shown in Figure 6. Below are descriptions of each of the items on this window.

Task number: An arbitrary number for task referencing.

Task name: any name used to identify the task.

Time distribution: Micro Saint will conduct Monte Carlo simulations with task performance times sampled from a distribution as defined by this option (e.g., normal, beta, exponential).

Mean time: This parameter defines average task performance time for this task. This can be a number, equation, or algorithm, as can all values in the fields described below.

Standard deviation: Standard deviation of task performance time.

Release condition: Data in this field will determine when a task begins executing. For example, a condition stating that this task will not start before an operator is available might be represented by a release condition such as the following:

$$\text{operator} >= 1$$

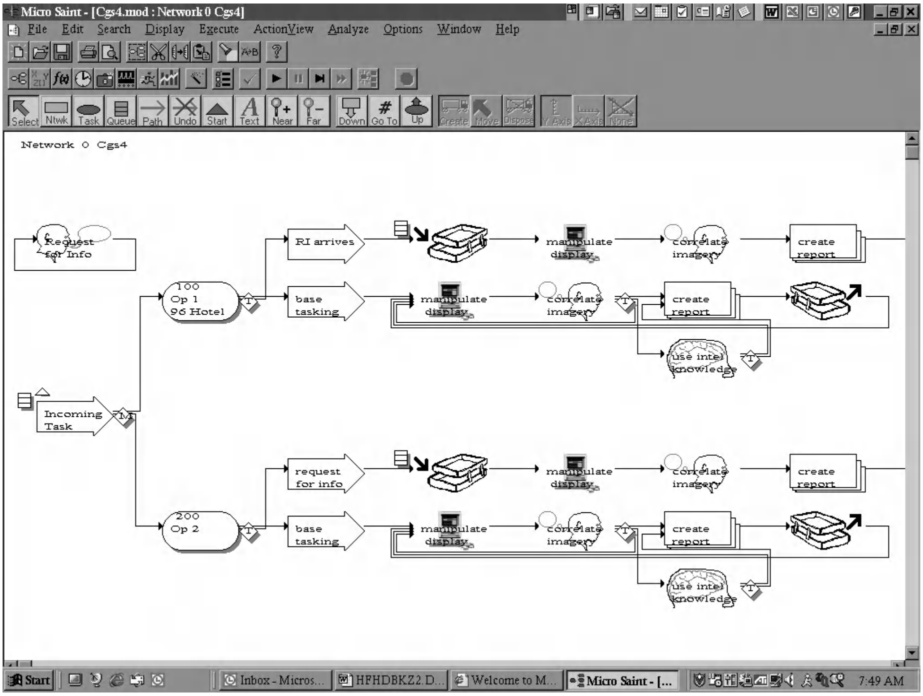
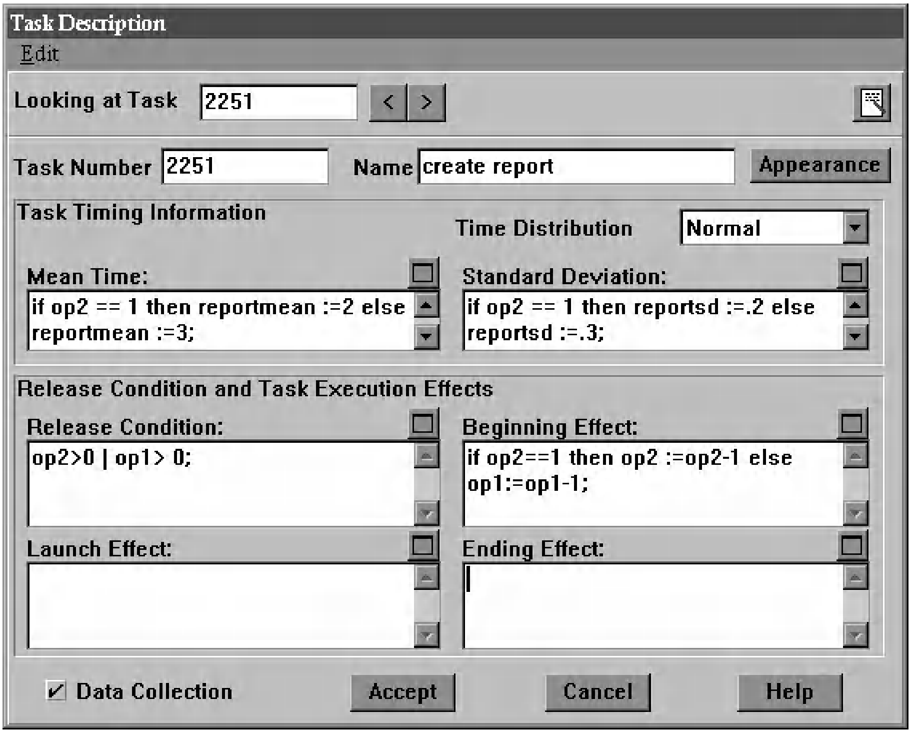

Micro Saint - [Cgs4.mod : Network 0 Cgs4]
File Edit Search Display Execute ActionView Analyze Options Window Help
Select Ntwk Task Queue Path Undo Start Text Near Far Down Go To Up Create Move Dispose Y Axis X Axis None
Network 0 Cgs4
Request for Info
RI arrives
manipulate display
correlate imagery
create report
100 Op 1 96 Hotel
base tasking
use intel knowledge
Incoming Task
M
T
request for info
200 Op 2
Start
Inbox - Micros...
HFHDBKZ2 D...
Welcome to M...
Micro Saint - [...
7:49 AM

Figure 5 The Main Window in Micro Saint for Task Network Construction and Viewing.



Looking at Task 2251

Task Number 2251

Name create report

Task Timing Information

Time Distribution Normal

Mean Time:

```
if op2 == 1 then reportmean :=2 else
reportmean :=3;
```

Standard Deviation:

```
if op2 == 1 then reportsd :=.2 else
reportsd :=.3;
```

Release Condition and Task Execution Effects

Release Condition:

```
op2>0 | op1> 0;
```

Beginning Effect:

```
if op2==1 then op2 :=op2-1 else
op1:=op1-1;
```

Launch Effect:

Ending Effect:

☑ Data Collection

Figure 6 The User Interface in Micro Saint for Providing Input on a Task.

In other words, there must be at least one operator available for the task to commence. If all operators were busy, the value of the variable operator would equal zero until a task is completed, at which time an operator becomes available. This task would wait until the condition was true before beginning execution, which would probably occur as a result of the operator completing the task he or she is currently performing.

Beginning effect: This field permits the user to define how the system will change as a result of the commencement of this task. For example, if this task used an operator that other tasks might need, we could set the following condition to show that the operator is unavailable while he performed this task:

$$\text{operator} := \text{operator} - 1$$

Assignment and modification of variables in beginning effects are one key way in which tasks are interrelated.

Launch effect: Similar to a task beginning effect but used to launch high-resolution animation of the task.

Ending effect: This field permits the definition of how the system will change as a result of the completion of this task. From the previous example, when this task was complete and the operator became available, we could set the ending effect as follows

$$\text{operator} := \text{operator} + 1$$

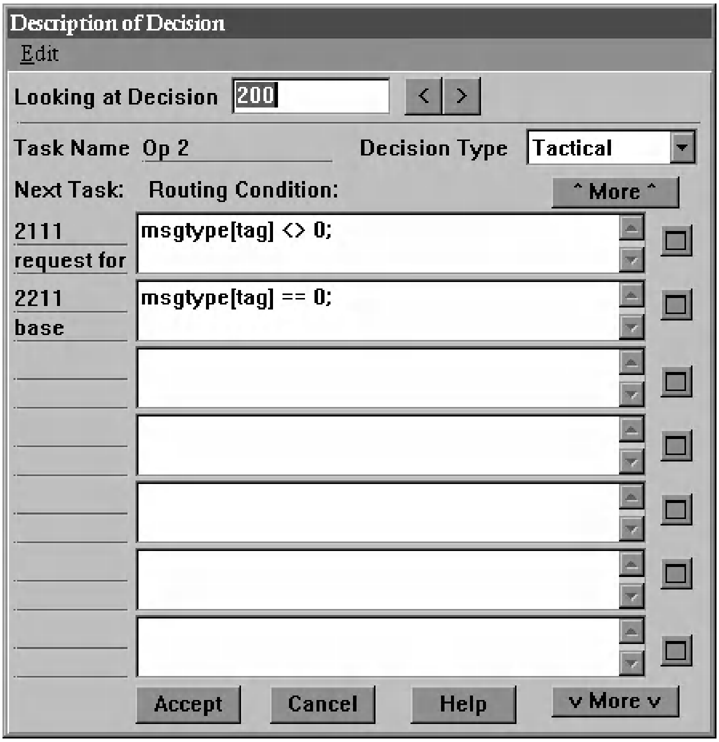
at which point another task waiting for an operator to become available could begin. Ending effects are another key way in which tasks can be interrelated through the assignment and modification of variables.

Another notable aspect of the task network diagram window shown in Figure 5 is the diamond-shaped icons that follow some tasks. These are present every time more than one path out of a task is drawn. In a task network model, this means that several tasks might commence at the completion of this task. Often this represents a human decision-making process. In that case, the branches align to potential courses of action that the modeled human could select. To define the decision logic, the user of Micro Saint would double-click on the diamond to open up a window, as shown in Figure 7.

There are only three general types of decisions to model:

- *Probabilistic:* In probabilistic decisions, the human will begin one of several tasks based on a random draw weighted by the probabilistic branch value. These weightings can be dynamically calculated to represent the current context of the decision. For example, this decision type might be used to represent human error likelihoods and would be connected to the subsequent tasks that would be performed.
- *Tactical:* In tactical decisions, the human will begin one of several tasks based on the branch with the highest "value." This could be used to model the many types of rule-based decisions that humans make, as illustrated in Figure 7.
- *Multiple:* This would be used to begin several tasks at the completion of this task, such as when one human issues a command that begins other crew members' activities.

The fields in Figure 7 labeled "Routing Condition" represent the values associated with each branch. The values can be numbers, expressions, or complicated algorithms defining the probability (for probabilistic branches) or the desirability (for tactical and multiple branches) of taking a particular branch in the network. Again, any value on this screen can not simply be numbers but also include variables, algebraic expressions, logical expressions, or groups of algebraic and logical expressions that would essentially form a subroutine. As the model executes, Micro Saint includes a parser that evaluates the expressions included in the branching logic when it is encountered in the task network flow. This results in a dynamic network in which the flow through the tasks can be controlled with variables that represent equipment state, scenario context, or the task loading of the humans in the



Looking at Decision 200

Task Name Op 2

Decision Type Tactical

Next Task:	Routing Condition:
2111 request for	msgtype[tag] <> 0;
2211 base	msgtype[tag] == 0;

Figure 7 The User Interface in Micro Saint for Defining Task Branching Decision Logic.

system, to name a few examples. It is the power of this parser that provides many task network models the ability to address complex problems.

There are other aspects of task network model development. Some items are defining a simulation scenario, defining continuous processes within the model, and defining queues in front of tasks. Further details of these features can be obtained from the Micro Saint User's Guide (Micro Analysis and Design 1999).

As a model is being developed and debugged, the user can execute the model to test it and collect data. There are several display modes, reflecting differing levels of information provided to the user during execution. In the most detailed mode, the simulation pauses after every simulated task. Another mode shows the user nothing about the simulation except when it is completed. There is also a model animation mode in which the task network is drawn on the screen and tasks that are currently executing are highlighted. In the model animation mode, the analyst can get a very clear picture of what events are occurring in what sequence in the model. Figure 8 presents a sample display during model animation.

Once a model is executed and data are collected, the analyst has a number of alternatives for data analysis. The data created during a model execution can be reviewed in the model analysis environment or exported to statistical and graphics packages.

The above discussion should indicate that task network modeling is a relatively straightforward concept that is a logical extension of task and systems analysis. Task network modeling is an evolution, not a revolution, to the human factors practitioner. As stated before, the basis for task network models of human performance is the mainstay of human engineering analysis—the task analysis. Much of the information discussed is generally included in the task analysis. Task network modeling, however, greatly increases the power of task analysis since the ability to simulate a task network with a computer permits prediction of human performance, rather than simply the description of human performance that a task analysis provides. What may not be as apparent, however, is the power of task network modeling as a means for modeling human performance in systems. Simply describing the systems activities in this step-by-step manner allows complex models of the system to be developed where the human's interaction with the system can be represented in a closed-loop manner.

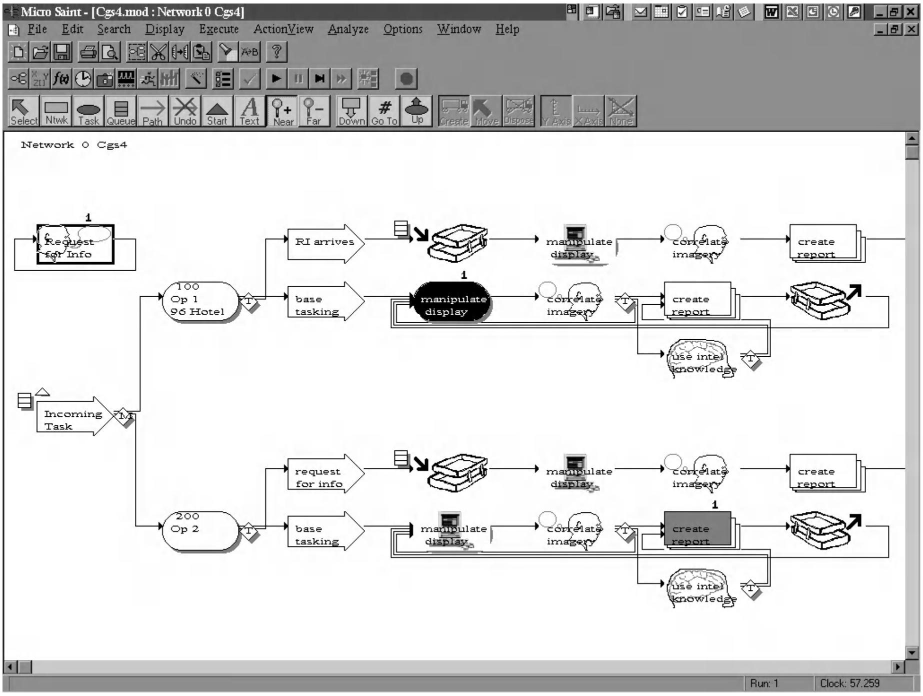

Micro Saint - [Cgs4.mod : Network 0 Cgs4]
File Edit Search Display Execute ActionView Analyze Options Window Help
Select Ntwk Task Queue Path Undo Start Text Near Far Down Go To Up Create Move Dispose Y Axis X Axis None
Network 0 Cgs4
1
Request for Info
RI arrives
manipulate display
correlate imagery
create report
100 Op 1 96 Hotel
T
base tasking
1
manipulate display
correlate imagery
T
create report
use intel knowledge
T
Incoming Task
M
request for info
manipulate display
correlate imagery
create report
200 Op 2
T
base tasking
manipulate display
correlate imagery
T
1
create report
use intel knowledge
T
Run: 1
Clock: 57.259

Figure 8 An Example of a Task Network Animation during Model Execution in Micro Saint.

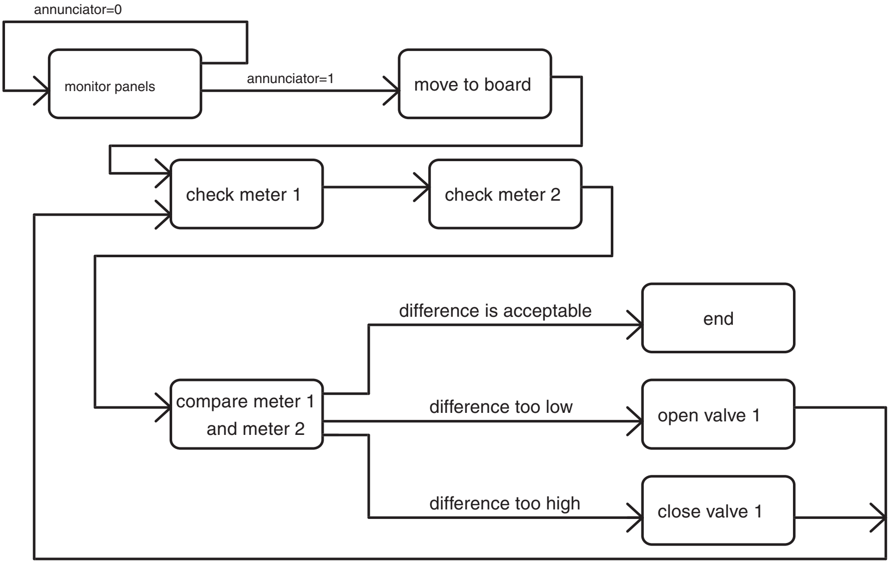

annunciator=0
monitor panels
annunciator=1
move to board
check meter 1
check meter 2
difference is acceptable
end
compare meter 1 and meter 2
difference too low
open valve 1
difference too high
close valve 1

Figure 9 Sample Task Network Model of a Process Control Operator Responding to an Annunciator.

The above discussion, in addition to being an introduction to the concepts, is also intended to support the argument that task network modeling is a mature technology ready for application in a wide range of problem domains.

5.2. An Example of a Task Network Model of a Process Control Operator

This simple hypothetical example illustrates how many of the basic concepts of task network modeling can be applied to studying human performance in a process control environment. It is intended to illustrate many of the concepts described above.

The simple human task that we want to model is of an operator responding to an annunciator. The procedure requires that the operator compare readings on two meters. Based on the relative values of these readings, the operator must either open or close a valve until the values on the two meters are nearly the same. The task network in Figure 9 represents the operator activities for this model. Also, to allow the study of the effects of different plant dynamics (e.g., control lags), a simple one-node model of the line in which the valve is being opened is included in Figure 10.

The operator portion of the model will run the monitor panels task until the values of the variables meter1 and meter2 are different. The simulation could begin with these values being equal and then precipitate a change in values based on what is referred to as a scenario event (e.g., an event representing the effects of a line break on plant state). This event could be as simple as:

$$\text{meter1} := \text{meter1} + 2.0$$

or as complex as an expression defining the change in the meter as a function of line break size,

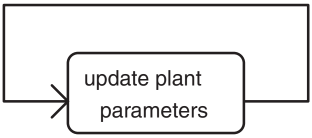

update plant parameters

Figure 10 A Simple One-Node Model of the Plant That Is Integrated with the Detailed Operator Model.

flow rates, and so on. An issue that consistently arises in model construction is how complex the plant system model should be. If the problem under study is purely operator performance, simple models will usually suffice. However, if overall plant behavior is of interest, then the models of plant dynamics, such as meter values, are more important. Again, we recommend the ‘‘start simple’’ approach whenever possible.

When the transient occurs and the values of meter1 and meter2 start to diverge, the annunciator signal will trigger. This annunciator would be triggered in the plant portion of the model by a task-ending effect such as:

if meter1 <> meter2 then annunciator := 1

Once the plant model sets the value of the variable annunciator to 1, the operator will begin to move to the appropriate board. Then the operator will continue through a loop to check the values for meter1 and meter2 and either open valve1, close valve1, or make no change. The determination of whether to make a control input is determined by the difference in values between the two meters. If the value is less than the acceptable threshold, then the operator would open the valve further. If the value is greater than the threshold, then the operator would close the valve. This opening and closing of the valve would be represented by changes in the value of the variable valve1 as a task-ending effect of the tasks open valve1 and close valve1. In this simple model, operators do not consider rates of change in values for meter1 and therefore would get into an operator-induced oscillation if there were any response lag. A more sophisticated operator model could use rates of change in the value for meter1 in deciding whether to open or close valves.

Again, this is a very small model reflecting simple operator activity on one control via a review of two displays. However, it illustrates how large models of operator teams looking at numerous controls and manipulating many displays could be built via the same building blocks used in this model. The central concepts of a task network and shared variable reflecting human–system dynamics remain the same.

Given a task network model of a process control operator in a current control room, how might the model be modified to address human-centered design questions? Some examples are:

1. Modifying task times based on changes in the time required to access a new display
2. Modifying task times and accuracies based upon changes in the content and format of displays
3. Changing task sequence, eliminating tasks, and/or adding tasks based upon changes in plant procedures
4. Changing allocation of tasks and ensuing task sequence based upon reallocation of tasks among operators
5. Changing task times and accuracies based upon stressors such as sleep loss or the effects of circadian rhythm

The above is not intended as a definitive list of all the ways that these models may be used to study design or operations concepts, but it should illustrate how these models can be used to address design and operational issues.

5.3. Case Studies in the Use of Task Network Modeling to Address Specific Design Issues

In this section, we will examine two case studies in the use of task network simulation for studying human performance issues. The first case study explores how task network modeling can be used to assess task allocation issues in a cognitively demanding environment. The second example explores how task network modeling has been used to extend laboratory and field research on human performance under stress to new task environments.

We should state clearly that these examples are intended to be representative of the types of issues that task network modeling can address, as well as approaches to modeling human performance with respect to these issues. *They are not intended to be comprehensive with respect to either the issues that might be addressed or the possible techniques that the human factors practitioner might apply.* Simulation modeling is a technology whose application leaves much room for creativity on the part of the human factors practitioner with respect to application areas and methods. These two case studies are representative.

5.4. Using Task Network Modeling to Evaluate Crew Workload

Perhaps the greatest contributor to human error in many systems is the extensive workload placed upon the human operator. The inability of the operator to cope effectively with all of his or her information and responsibilities contributes to many accidents and inefficiencies. In recognition of this problem, new automation technologies have been introduced to reduce workload during periods of high stress. Some of these technologies are in the form of enhanced controls and displays, some are in the form of tools that push information to the operator and alert the operator in order to focus attention, and still others consist of adaptive tools that take over tasks when they sense that the operator is overloaded. Unfortunately, these technical solutions often introduce new tasks to be performed that affect the visual, auditory, and/or psychomotor workload of the operators.

Recently, new concepts in crew coordination have focused on better management of human workload. This area shows tremendous promise and is benefiting from efforts of human factors researchers. However, their efforts are hindered because there are limited opportunities to examine empirically the performance of different combinations of equipment and crew composition in a realistic scenario or context. Additionally, high workload is not typically caused by a single task but by situations in which multiple tasks must be performed or managed simultaneously. It is not simply the quantity of tasks that can lead to overload, but also the composition of those tasks. For example, two cognitive tasks being performed in parallel are much more effortful than a simple motor task and an oral communication task being performed together. The occurrence of these situations will not typically be discovered through normal human engineering task analysis or subjective workload analysis until there is a system to be tested. That is often too late to influence design.

To rectify this problem, there has been a significant amount of recent research and development aimed at human workload *prediction* models. Predictive models allow the designers of a system to estimate operator workload *without human subjects experimentation*. From this and other research, a solid theoretical basis for human workload prediction has evolved, as is described in Wickens (1989).

This section discusses a study using task network modeling to predict the impact of task allocation on human workload. While these examples are posed in the context of the design of a military system, the same techniques have been used in nonmilitary applications such as process control and user–computer interface design.

5.4.1. Modeling the Workload of a Future Command and Control Process

The Army command and control (C2) community is concerned with how new information, technology, and organizational changes projected for tomorrow's battlefield will impact soldier tasks and workload. To address this concern, an initiative was taken to model soldier performance under current and future operational conditions. In this way, the impact of performance differences could be quantitatively assessed so that equipment and doctrine design could be influenced in a timely and effective manner.

In one C2 project, the primary concern was to determine how tasks should be allocated and automated such that a C2 team could evaluate all the relevant data and make decisions within an environment with particularly high time pressure. Specifically, the effort was to address the following key questions:

- How many crew members do you need?
- How do you divide tasks among jobs?
- How does decision authority flow?
- Can the crew meet decision time line requirements?
- Is needed information usable and accessible?

We used task network modeling to study crew member, task and scenario combinations in order to examine these questions.

Figure 11 shows the top level diagram of the task network. Essentially, the crew members receive and monitor information about the system and the environment until an event occurs that pushes them out of the 10000 and 20000 networks into either a series of planning tasks, and/or a series of evaluation, decision, direction, and execution tasks. The purpose of the planning task is to update tactical battle plans based on new information received from the system or the environment. Receipt of new intelligence data about the enemy's intention or capability is an example of an event that would cause the crew members to undertake planning tasks. Similarly, receipt of information from the system about resource limitations might trigger the crew members to proceed down the alternative path (through evaluate to execute). Specifically, limited resources might cause the crew members to evaluate whether the engagement is proceeding appropriately (30000), decide how to adjust system parameters (40000), direct the appropriate response to the correct level of command (50000), and then execute the order (60000). Upon completion, the crew members would return to monitoring the system and situation.

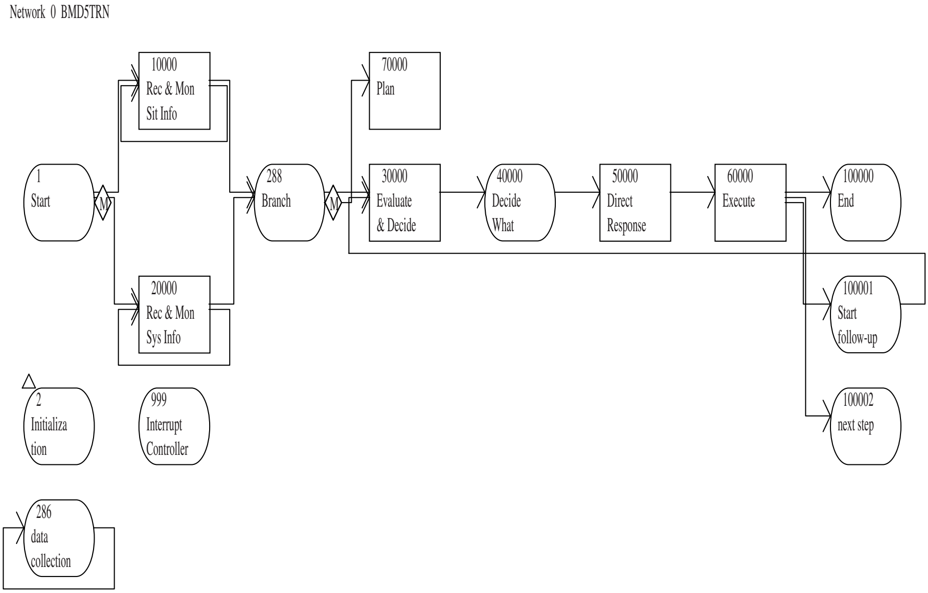

Network 0 BMD5TRN
10000 Rec & Mon Sit Info
70000 Plan
1 Start
M
288 Branch
M
30000 Evaluate & Decide
40000 Decide What
50000 Direct Response
60000 Execute
100000 End
20000 Rec & Mon Sys Info
100001 Start follow-up
2 Initializa tion
999 Interrupt Controller
100002 next step
286 data collection

Figure 11 Upper-Level Task Network.

Each of the rectangles in the task network shown in Figure 11 actually consists of a network of tasks. An example of the tasks that belong to Network 10000, receive and monitor situation information, is shown in Figure 12. As described under Figure 7 of this chapter, the tasks in Network 10000 are linked by probabilistic and tactical decisions.

Each of the tasks in the C2 task network is associated with several items of human performance data. These include:

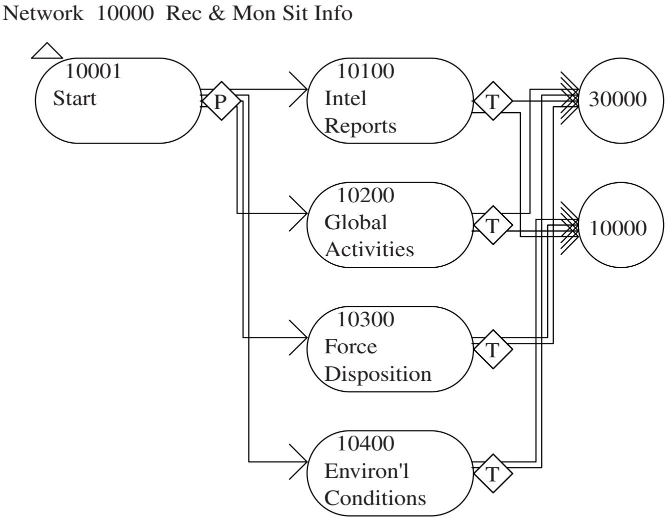

Network 10000 Rec & Mon Sit Info
10001 Start
P
10100 Intel Reports
T
30000
10200 Global Activities
T
10000
10300 Force Disposition
T
10400 Environ'l Conditions
T

Figure 12 Second- Level Task Network.

- *Task performance time:* These data consist of a mean, standard deviation, and distribution. The data were collected from a combination of three sources: (1) human factors literature (e.g., Fitts's Law); (2) empirical studies during operator-in-the-loop simulator exercises; and (3) subject matter experts.
- *Branching logic:* While the task network indicates a general process flow, this particular model was designed to respond to scenario events. Because of that design decision, each task includes logic to determine the following task. For example, if the scenario is very intense and multiple target tracks are available, the crew members will follow a different task flow than if they were performing routine system checks.
- *Release rules:* Logic controlling the number and types of parallel tasks each crew member can perform are contained in each task's release condition.

Since one purpose of the model was to examine various task-allocation strategies, the model was designed to incorporate several measures of crew member workload. The basis of this technique is an assumption that excessive human workload is not usually caused by one particular task required of the operator. Rather, the human having to perform several tasks simultaneously leads to overload. Since the factors that cause this type of workload are intricately linked to these dynamic aspects of the human's task requirements, task network modeling provides a good basis for studying how task allocation and sequencing can affect operator workload.

However, task network modeling is not inherently a model of human workload. The only relevant output common to all task network models is the time required to perform a set of tasks and the sequence in which the tasks are performed. Time information alone would suffice for some workload-evaluation techniques, such as Siegel and Wolf (1969), whereby workload is estimated by comparing the time available to perform a group of tasks to the time required to perform the tasks. Time available is driven by system performance needs, and time required can be computed with a task network model. However, it has long been recognized that this simplistic analysis misses many aspects of the human's tasks that influence both perceived workload as well as ensuing performance. At the very least, this approach misses the fact that some pairs of tasks can be performed in combination better than other pairs of tasks.

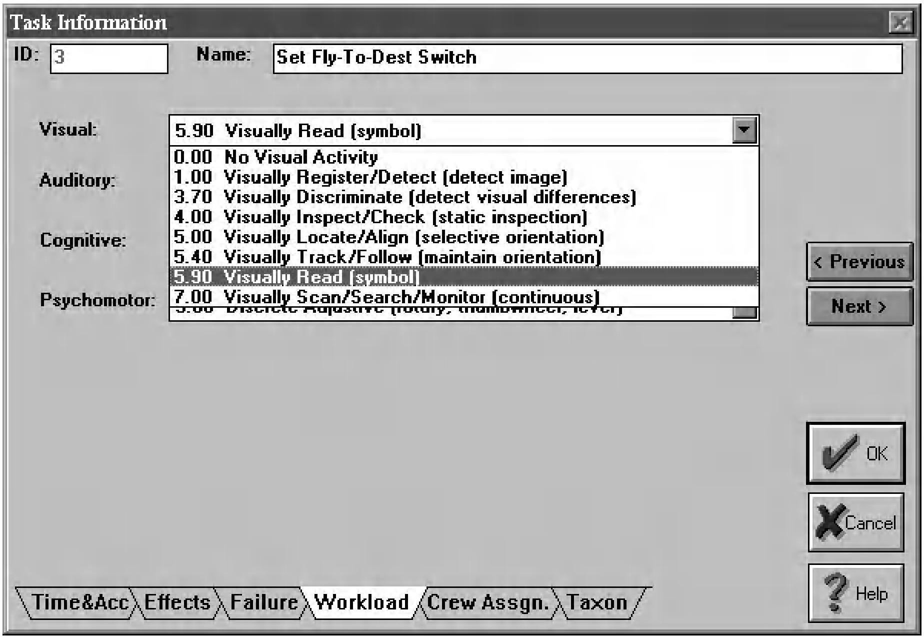
One of the most promising theories of operator workload, which is consistent with task network modeling, is the multiple resource theory proposed by Wickens (e.g., Wickens et al. 1983). Simply stated, the multiple-resource theory suggests that humans have several different resources that can be tapped simultaneously. Depending upon the nature of the information-processing tasks required of a human, these resources would have to process information sequentially (if different tasks require the same types of resources) or possibly in parallel (if different tasks required different types of resources). There are many versions of this multiple-resource theory in workload literature (e.g., McCracken and Aldrich: Bierbaum et al. 1989; North and Riley 1989; Little et al. 1993; Knapp et al. 1999). In this chapter, we will provide a discussion of the underlying methodology of the basic theory.

Multiple-resource workload theory is implemented in a task model in a fairly straightforward manner. First, each task in the task network is characterized by the workload demand required in each human resource, often referred to as a workload channel. Examples of commonly used channels include auditory, visual, cognitive, and psychomotor. Particular implementations of the theory vary in the channels that are included and the fidelity with which each channel is measured (high, medium, low vs. seven-point scale). In fact, Bierbaum et al. 1989 present reliable benchmark scales for determining demand for each channel. As an example, the scale for visual demand is presented in Figure 13.

Similar scales have been developed for the auditory, cognitive, and psychomotor channels. With this approach, each operator task can be characterized as requiring some amount of each of the four kinds of resources, as represented by a value between one and seven. All operator tasks can be analyzed with respect to these demands and values assigned accordingly.

In performing a set of tasks pursuant to a common goal (e.g., engage an enemy target), crew members frequently must perform several tasks simultaneously, or at least nearly so. For example, they may be required to monitor a communication network while visually searching a display for target track. Given this, the workload literature indicates that the crew member may either accept the increased workload (with some risk of performance degrading) or begin dumping tasks perceived as less important. To factor these two issues into task network simulations, two approaches can be incorporated: (1) evaluate combined operator workload demands for tasks that are being performed concurrently, and/or (2) determine when the operator would begin dumping tasks due to overload.

During a task network simulation, the model of the crew may indicate they are required to perform several tasks simultaneously. The task network model evaluates total attentional demands for each human resource (e.g., visual, auditory, psychomotor, and cognitive) by combining the attentional demands across all tasks that are being performed simultaneously. This combination leads to an overall workload demand score for each crewmember.



Task Information

ID: 3 Name: Set Fly-To-Dest Switch

Visual: 5.90 Visually Read (symbol)

0.00 No Visual Activity
1.00 Visually Register/Detect (detect image)
3.70 Visually Discriminate (detect visual differences)
4.00 Visually Inspect/Check (static inspection)
5.00 Visually Locate/Align (selective orientation)
5.40 Visually Track/Follow (maintain orientation)
5.90 Visually Read (symbol)
7.00 Visually Scan/Search/Monitor (continuous)

Figure 13 An Example of a Visual Workload Scale.

To implement this approach in Micro Saint, the task beginning effect can be used to increment variables that represent the current workload score in each resource. Then, while the tasks are being performed, these variables track attentional demands. When the tasks are completed, the task ending effects can decrement the values of these variables accordingly. Therefore, if these workload variables were recorded and then plotted as the model runs, the output would look something like what is shown in Figure 14. This result can be used to identify points of high workload throughout the scenario being modeled. The human factors practitioner can then review the tasks that led to the points of high workload and determine whether they should be reallocated or redesigned in order to alleviate the peak. This is a common approach to modeling workload.

Once the task networks were verified with knowledgeable crewmembers, they became part of the human factors team's analytical test bed. Figure 15 shows the overall method that we used to examine aspects of crew member performance across a wide variety of operational scenarios and crew configuration concepts.

The center of this diagram, labeled task network, represents the tasks that the crew performs as we just described them. The network itself, representing the flow of the tasks, does not change between model runs. Rather, the model has been parameterized so that an event scenario stimulates the network. The left side of the diagram illustrates the types of data that are used to drive the task network model. In this case, those data include crew configurations, or allocations of tasks to different crew members and automation devices, as well as scenario events. The scenario events represent an externally generated time-ordered list of the events that trigger the crew members to perform tasks in the task network. The right side of Figure 15 represents the types of outputs that can be produced from this task network model. One of the primary outputs is a crew member workload graph, such as that shown in Figure 14.

5.4.2. Extensions of This Approach to Simulating Crew Mental Workload to Other Environments

The workload-analysis methodology described above has recently been developed into a stand-alone task network modeling tool by the Army Research Laboratory (ARL) Human Research and Engineering Directorate (HRED), as part of the Improved Performance Research Integration Tool (IMPRINT) (Allender et al. 1994; Archer and Adkins 1999). IMPRINT integrates task network modeling software with features to support specifically the multiple-resource theory of workload discussed

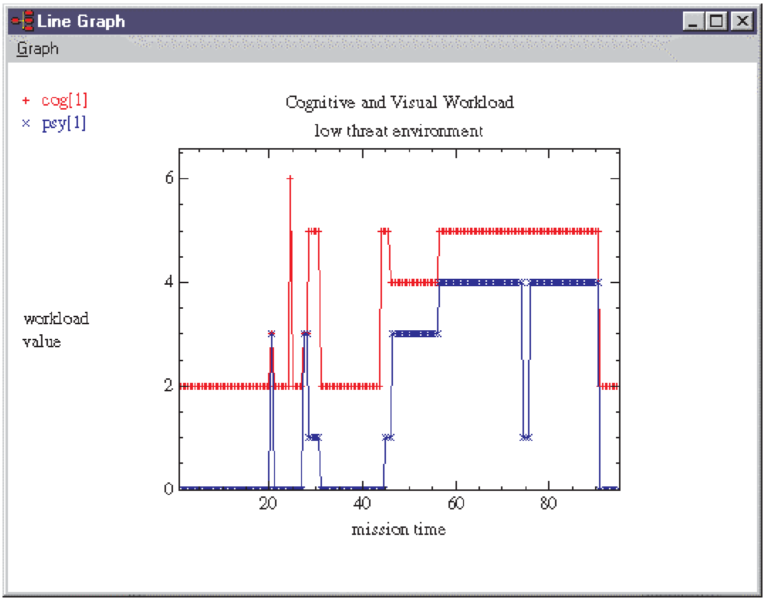


Figure 14 An Example Workload Output from a Task Network Model.

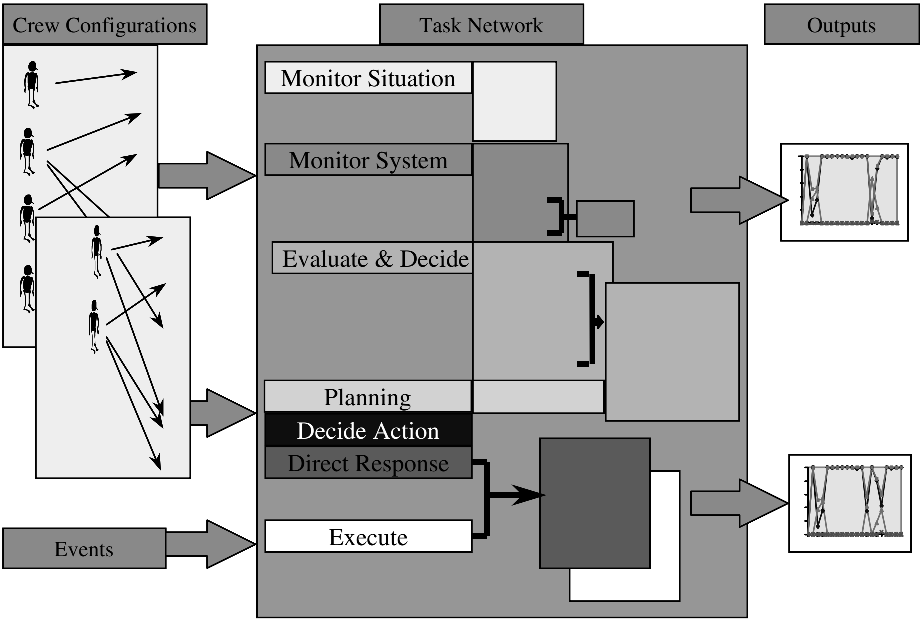

Crew Configurations
Task Network
Outputs
Monitor Situation
Monitor System
Evaluate & Decide
Planning
Decide Action
Direct Response
Execute
Events

Figure 15 Overall Method for Examining Workload in a Complex System.

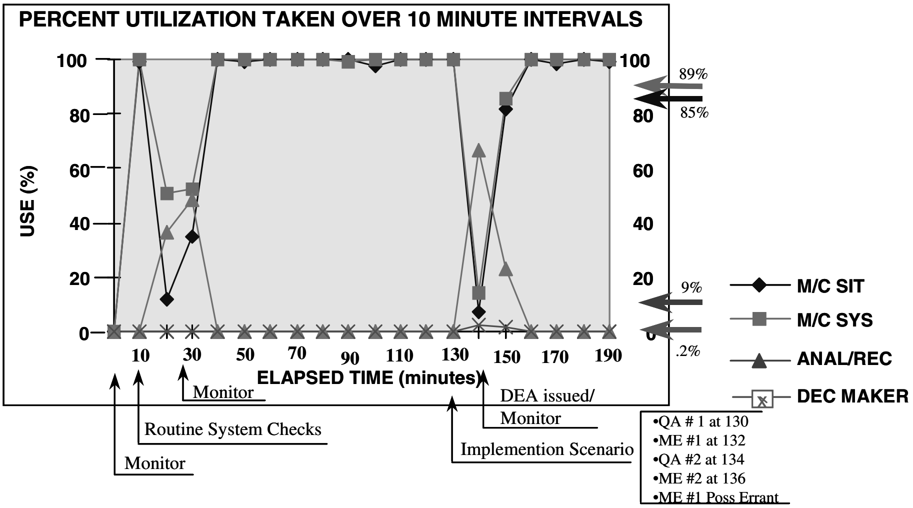


Figure 16 Example of Model Predictions of Operator Utilization over Time.

above. It provides the human factors practitioner with an environment that supports the analysis of task assignment to crewmembers based on four factors:

1. *Workload of crew members:* Tasks should be assigned to minimize the amount of time crew members will spend in situations of excessive workload.
2. *Time performance requirements:* Tasks must be assigned and/or sequenced so that they are completed within the available time. This consideration is essential because time constraints often will drive the need to perform several tasks simultaneously.
3. *Likelihood of successful performance and consequences of failure:* Tasks must be assigned and/or sequenced so that they can be completed within a specified accuracy measure.
4. *Access to controls and displays:* Tasks cannot be assigned to crew members who do not have access to the necessary controls and displays.

Of course, there are numerous theoretical questions regarding this simplistic approach to assessing workload in an operational environment. However, even the use of this simple approach has been shown to provide useful insight during design. For example, in a study conducted by the Army (Allender 1995), a three-man crew design was evaluated using a task network model. The three-man model was constructed using data from a prototype four-man system. From this model-based analysis, the three-man design was found to be unworkable . Later, human subjects experimentation verified that the model's workload predictions were sufficiently accurate to point the design team in a valid direction.

Finally, ARL HRED has developed another custom software package named WinCrew (Plott 1995). It includes capabilities, also available in an advanced version of IMPRINT, to implement more refined methods of predicting workload. WinCrew overlays the W/INDEX manifestation of the multiple resource theory of workload (Boettcher et al. 1989) into a task network-based environment. In addition to a better estimate of workload, WinCrew is unique in that it has built-in constructs for simulating workload management strategies that operators would employ to accommodate points of high operator workload. The ultimate result of simulating the workload management strategies is that the operator task network being modeled is dynamic. In other words, the task sequence, assignments to operators, and individual task performance may change in response to excessive operator workload as the task network model executes. These changes may be as simple as one operator handing tasks off to another operator to reduce workload to an acceptable level or as complex as the operator beginning to time share tasks in order to complete all the assigned tasks, potentially task time and/or accuracy. Ultimately, the tool provides an estimate of system-level performance as a result of these realistic workload management strategies. This innovation in modeling provides greater fidelity in efforts that model human behavior in the context of system performance, particularly in high-workload environments such as complex system control and management.

5.5. Using Task Network Modeling as a Means of Extending Research Findings on Human Performance under Stress to New Task Environments

Task network modeling was used by LaVine et al. (1995) to extend laboratory data and field data collected on one set of human tasks to predicting performance on similar tasks. The problem of extending laboratory or field human performance data to other tasks has plagued the human engineering community for years. We intuitively know that human performance data can be used to predict performance for similar tasks. However, often the task whose performance we want to predict is similar in some ways but different in others. The approach described below uses a skill taxonomy to quantify task similarity and therefore provides a means for determining how other tasks will be affected when exposed to a common stressor on human performance. Once functional relationships are defined between a skill type and a stressor, task network modeling is used to determine the effect of the stressor on performance of a complex task that simultaneously uses many of these skills.

The specific approach below is being used by the U.S. Army to predict crew performance degradation as a function of a variety of stressors. It not intended to represent a universally acceptable taxonomy for simulating human response to stress. The selection of the best taxonomy would depend upon the particular tasks and stressors being studied. What this example is intended to illustrate is another way that task network modeling can be used to predict human performance by making a series of reasonable assumptions that can be played together in a model for the purpose of making predictions that would be impossible to make otherwise.

The methodology for predicting human performance degradation as a function of stressors consists of three parts: (1) a *taxonomy* for classifying tasks according to basic human skills, (2) *degradation functions* for each skill type for each stressor, and (3) *task network models* for the human-based system whose performance is being predicted. Conceptually, either laboratory or field data can be used to develop links between a human performance stressor (e.g., heat, fatigue) and basic human skills. By selecting a skills taxonomy that is sufficiently discriminating to make this assumption reasonable, one can assume that the effects of the stressor on all tasks involving the skill will be approximately the same. The links between the level of a stressor (e.g., fatigue) and resulting skill performance (e.g., the expected task time increase from fatigue) are defined mathematically as the degradation function. The task network model is the means for linking these back to complex human/system performance.

5.5.1. The Taxonomy

The basic premise behind the taxonomy is that the tasks that humans perform can be broken down into basic human skills or atomic tasks (Roth 1992). The taxonomy used by Roth consists of five skill types: attention, perception, psychomotor, physical, and cognitive skills. These taxonomic skills are described by Roth as follows:

1. *Attention:* The ability to attend actively to a stimulus complex for extended periods of time in order to detect specified changes or classes of changes that indicate the occurrence of some phenomenon that is critical to task performance.
2. *Perception:* The ability to detect and categorize specific stimulus patterns embedded in a stimulus complex.
3. *Psychomotor:* The ability to maintain one or more characteristics of a situation within a set of defined conditions over a period of time, either by direct manipulation or by manipulation of controls that cause changes in the characteristics.
4. *Physical:* The ability to accomplish sustained, effortful muscular work.
5. *Cognitive:* The ability to apply concepts and rules to information from the environment and from memory in order to select or generate a course of action or a plan. This includes communicating the course of action or plan to others.

These five skills covered most of the tasks that were of interest to the Army for this study and still provided a manageable number of categories for an analyst to use.

5.5.2. Degradation Functions

The degradation functions quantitatively link skill performance to the level of a stressor. The degradation functions can be developed from any data source, including standard test batteries or actual human tasks. Through statistical analysis, one can build skill-degradation functions for each taxon. These functions map the performance decrement expected on a skill based on the parameters of the

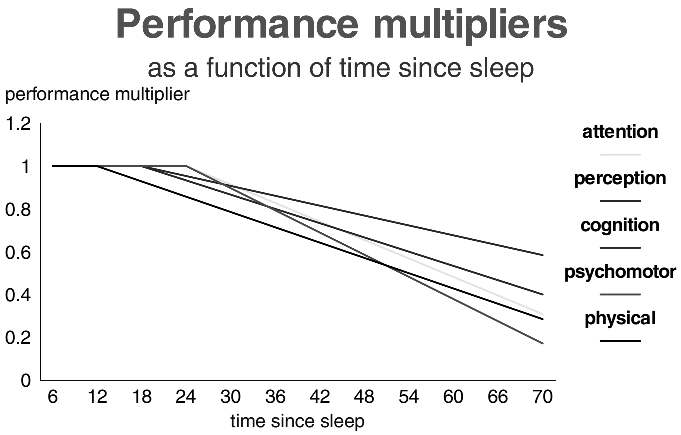


Figure 17 An Example of the Performance Degradation Functions Associated with each of the Human Skills from the Taxonomy.

performance-shaping factor (e.g., time since sleep). An example of these functions is presented in Figure 17.

5.5.2.1. Incorporating the Degradation Functions into Task Network Models to Predict Overall Human/System Performance Degradation The key to making this approach useful to predicting complex human performance is the task network model of the new task. In the task network model of the human's activities, all tasks are defined with respect to the percentage of each skill required from the taxonomy. For example, the following are ratings for tasks faced by a console operator responding to telephone contacts:

- Detect ring—50% attention, 50% perception
- Select menu item using a mouse—40% attention, 60% psychomotor
- Interpret customer's request for information—100% cognitive

In building the task network model, one can build functions to degrade a specific task's performance through an arithmetic weighting of skill-degradation multipliers that are derived from the degradation functions. For example, if the fatigue parameter was time since sleep and the value of that parameter was 36 hours since sleep, the task time performance multipliers would be as follows in the example above:

- Attention performance multiplier = 0.82
- Perception performance multiplier = 0.808
- Cognition performance multiplier = 0.856
- Psychomotor performance multiplier = 0.784
- Physical performance multiplier = 0.727

Based upon these multipliers and the above task weightings, the specific task effects would be:

- Detect ring—50% attention, 50% perception

$$\text{Task multiplier} = 0.5 \times 0.82 + 0.5 \times 0.808 = 0.814$$

- Select menu item using a mouse—40% attention, 60% psychomotor

$$\text{Task multiplier} = 0.4 \times 0.82 + 0.6 \times 0.784 = 0.7984$$

- Interpret customer's request for information—100% cognitive

$$\text{Task multiplier} = 0.856$$

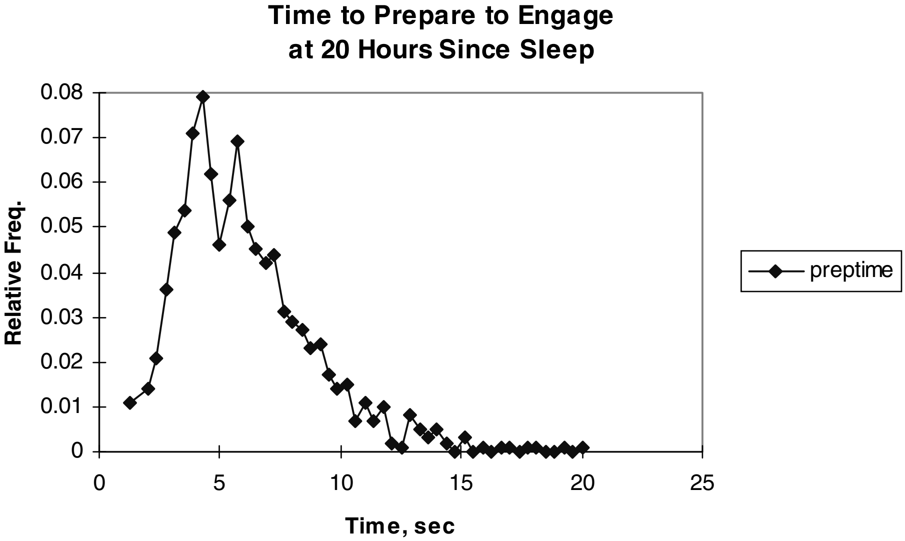


Figure 18 Frequency Distribution of Expected Human Performance as a Function of Time Since Sleep That Was Derived Using Task Network Modeling.

In a model of the complex tasks examined by LaVine et al. (1995), the task networks consisted of several dozen or even several hundred tasks. Through the approach described above, each task in a model exhibited a unique response to a stressor depending upon the particular skills that it required. The task network model then provided the means for relating the individual task performance to overall human/system performance as a function of stressor level (e.g., the time to perform a complex series of tasks involving decision making and error correction). Through this type of analysis, LaVine et al. were able to develop curves such as that shown in Figure 18 relating human performance to a stressor. These relationships would have been virtually impossible to develop experimentally.

Again, there were a number of simplifying assumptions that were made in this research. However, by being willing to accept these assumptions, LaVine et al. were able to characterize how complex human/system performance would be affected by a variety of stressors over a wide range in a relatively short time. They were thus able to estimate the effects of stressors that would have otherwise been pure guesswork.

5.6. Summary of Examples of Task Network Modeling of Human–System Performance

Once again, the above are intended to serve as examples, not as a catalogue of problems or approaches that are appropriate for task network modeling. Task network modeling is an approach to extend task and systems analysis to make predictions of human system performance. The creative human factors and ergonomics practitioner will find many other useful applications and approaches.

6. AN EXAMPLE OF A FIRST-PRINCIPLED APPROACH TO HUMAN/SYSTEM PERFORMANCE MODELING: THE MAN MACHINE INTEGRATED DESIGN SYSTEM (MIDAS)

The other fundamental approach to modeling human performance is based upon the mechanisms that underlie and cause human behavior. Since this approach is based on fundamental principles of the human and his or her interaction with the system and environment, we have designated them first-principle models. By integrating these models with models of the system and environment, the human factors specialist can predict the full behavior of large-scale interactive human–machine systems. The Man–Machine Integrated Design and Analysis System (MIDAS) follows in the tradition of integrated, first-principled models of human performance such as PROCRU (Baron et al. (1980) in that the modeling framework provides models of emergent human behavior based on elementary models of human behaviors such as perception, attention, working memory, and decision making. In the operation of these elementary models, MIDAS shares some of the characteristics of the task network approach. However, MIDAS is focused around an integrated architecture where micromodels of

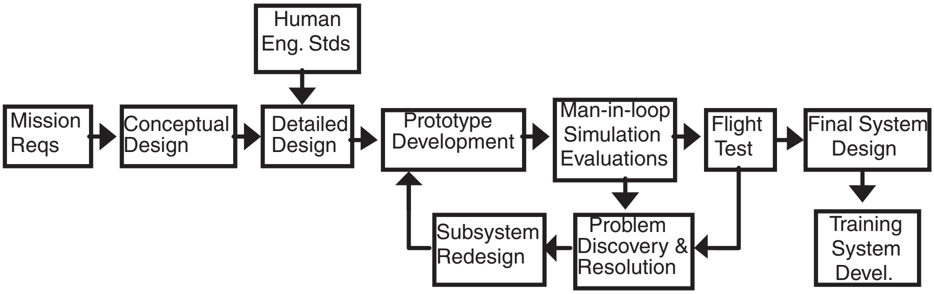

Human Eng. Stds
Mission Reqs
Conceptual Design
Detailed Design
Prototype Development
Man-in-loop Simulation Evaluations
Flight Test
Final System Design
Subsystem Redesign
Problem Discovery & Resolution
Training System Devel.

Figure 19 Current Crew-Station and Human–Computer Interaction Development Process.

human performance feed forward and feedback to the other constituent models in the human system rather than being linked primarily to the human's activities as in task network models.

6.1. Background

The A^{3}I Program was initiated in 1985 to support exploration of computational representations of human–machine performance to aid designers of crew systems. The major product of this effort was a human factors computer-aided engineering system called MIDAS (Man–Machine Integration Design and Analysis System). MIDAS is intended to revise the system design process in order to place more accurate information into the hands of the designers early in the process of human engineering design so that the impact and cost of changes are minimal. It is also intended to identify and model human–automation interactions with flexible representations of human–machine function. The crew station development process, as it is currently undertaken, is illustrated in Figure 19. The design proceeds from requirements and capabilities in conceptual design through increasing specification to hardware and software prototypes and simulation tests. Human performance evaluation occurs after prototype design and development. Results from testing the prototype are then used to guide prototype redesign.

MIDAS integrates the design process by using human performance models in the conceptual design phases of system development. Human–system integration and development enabled by the computational human performance models methodology are illustrated in Figure 20.

In this revised process, human performance considerations are accounted for early in the designs and played out for evaluation in the simulation mode. Iteration in this mode is flexible and timely. The flow then proceeds with a refined design into the standard design and prototype development phases. MIDAS provides a prototyping test bench, based on human performance models. Designers can work with computational representations of the crew station and human operators rather than relying solely on hardware simulators and man-in-the-loop studies, to discover problems and ask "what-if" questions regarding the projected mission, equipment, and environment.

In addition to its use in development and design, MIDAS offers a structure or framework in which to test and implement models of human cognition. The MIDAS framework systematizes and unifies the interaction of human performance representations in a common structure and with a common

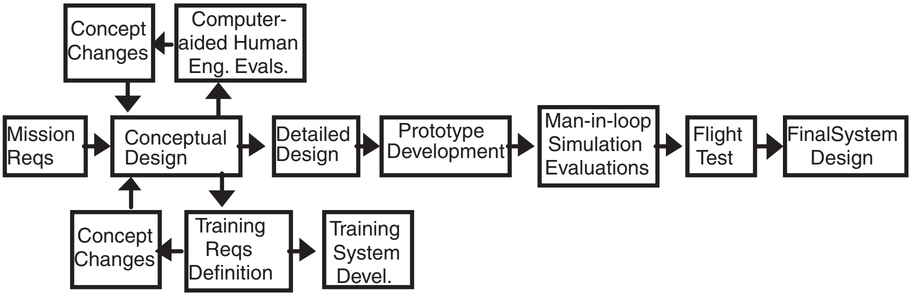

Concept Changes
Computer-aided Human Eng. Evals.
Mission Reqs
Conceptual Design
Detailed Design
Prototype Development
Man-in-loop Simulation Evaluations
Flight Test
FinalSystem Design
Concept Changes
Training Reqs Definition
Training System Devel.

Figure 20 Design Methodology Made Possible through Modeling and MIDAS.

language for interaction. The representation is a tightly linked set of computational descriptions of the elemental aspects of human performance. Models of human performance from perception through cognition and action are implemented within this framework. The interplay of the models produces simulations of behavior.

In developing MIDAS, three challenges were considered. First, the level of representation of human behavior has to be sufficiently detailed to predict individual performance and guide design for individual aiding and support systems. At the same time, that behavioral representation must be able to provide input into large-scale analyses to predict global consequences of system modification. Second, the models of human–system performance have to be sufficiently computational to support design specification in control theoretic terms but also sufficiently flexible and robust to account for a range of human behavior influenced by cultural (corporate, professional, and national) and environmental context. Third, the human performance model has to represent individuals and teams of human operators. This requires control of not only single cognitive behavior but also the cognitive activities associated with group and organizational behavior. Such applications require representation of many intelligent agents sharing world models, and coordinating action/intention with cooperative scheduling of goals and actions in a potentially unpredictable world of operations. The simulated operators' activity structures must provide for anticipation (knowledge of the intention and action of remote operators) and respond to failures of the system and other operators in the system in context-sensitive processes.

MIDAS is intended to meet these challenges by including the following characteristics:

- *Modifiability and manipulability:* The basic mode of operation for MIDAS users is to explore the impact of changes to the baseline design. Thus, the capability for systematic change is critical. Of equal importance is system extensibility. To be generally useful, the modeling environment should be applicable to many types of design changes, and to many operational domains. The MIDAS architecture is designed to allow extensions of this type with minimal disruption to the existing core MIDAS system.
- *Transparency:* The analysis system must provide designers with explicit and transparent reference to the rules, decision-making strategies, heuristics, and assumptions under which the human–machine system is assumed to be operating, as well as predicted performance. For example, at any point in the simulation, a designer should be able to examine the cognitive state of the human operators, the rules that are being used to guide their behavior, and their nominal workload. The designer should also be able to perform sensitivity analyses on critical parameters of the human–machine system. Similarly, the state of equipment or mission progress should be able to be probed in order to relate the system state to the operator's performance.
- *Dynamic analysis capability:* The simulation system must produce a stream of behavior in the form of dynamic timelines describing not only its state and structure, but also sequences of action over time and contingent responses of the human/system behavior. The system must support testable hypotheses. Designers must be able to analyze the events occurring in a simulation scenario and relate this performance to man-in-the-loop simulation data. In MIDAS, each action taken, decision made, and communication event is logged by the analysis system.

6.2. System Architecture

There are two perspectives on the MIDAS system architecture that describe the system to support these modes of analysis: the functional architecture and structural bases of the system. Each is discussed below.

6.2.1. MIDAS Functional Architecture

The MIDAS system has evolved over a period of 15 years of development (Corker and Smith 1992). The basic structure of the core system presented here is based on the work of Tyler et al. (1998). This architectural version of MIDAS has throughout its development been used to evaluate helicopter crew stations, short-haul civil tiltrotor emergency handling operations, and the impact of MOPP flight gear on crew performance (Atencio et al. 1996, 1998; Shively et al. 1995). The specific development for analysis of air traffic management systems will be provided below.

The user enters the system through the graphical user interface (GUI), which provides the main interaction between the designer and the MIDAS system. The user selects among four functions in the system:

1. Create and/or edit a domain model that includes establishment and selection of the parameters of performance for the human operator model(s) in the simulation
2. Select the graphical animation or view to support that simulation or a set of simulations

3. Specify in the simulation module the parameters of execution and display for a given simulation set
4. Specify in the results analysis system the data to be collected and analyzed as a result of running the simulation.

The results-analysis system also provides for archival processes for various simulation sessions. The overall functional architecture is provided in Figure 21.

The user would typically use all of the top-level features to support a new simulation. If a user were exploring, for instance, the assignment of function between a human operator and an automated assistant, the user could maintain the majority of the extant domain, graphical, and analytic models and make modification through the domain model to the human operator model, the equipment model, and the simulation scenario.

6.2.1.1. Domain Model The domain model consists of descriptors and libraries supporting the creation of:

- *Vehicle characteristics:* Location space, aerodynamic models of arbitrarily detailed fidelity, and guidance models for vehicle (automatic) control.
- *Environment characteristics:* This provides the external interactions, including terrain from selected databases at varied levels of resolution, weather features insofar as they affect vehicle performance or operator sensory performance, and cultural features (towns, towers, wires, etc.). In short, the analyst here specifies the world of action of the experiment/simulation.
- *Crew Station/equipment characteristics:* The crew station design module and library is a critical component in the MIDAS operation. Descriptions of discrete and continuous control operation of the equipment simulations are provided at several levels of functional detail. The system can provide discrete equipment operation in a stimulus–response (black box) format, a time-scripted/event driven format, or a full discrete-space model of the transition among equipment states. Similarly, the simulated operator's knowledge of the system can be at the same varied levels of representation or can be systematically modified to simulate various states of misunderstanding the equipment function.
- *The human operator model* (*HO*): The human performance model in MIDAS allows for the production of behavior and response for single and multiple operators in the scenarios. The

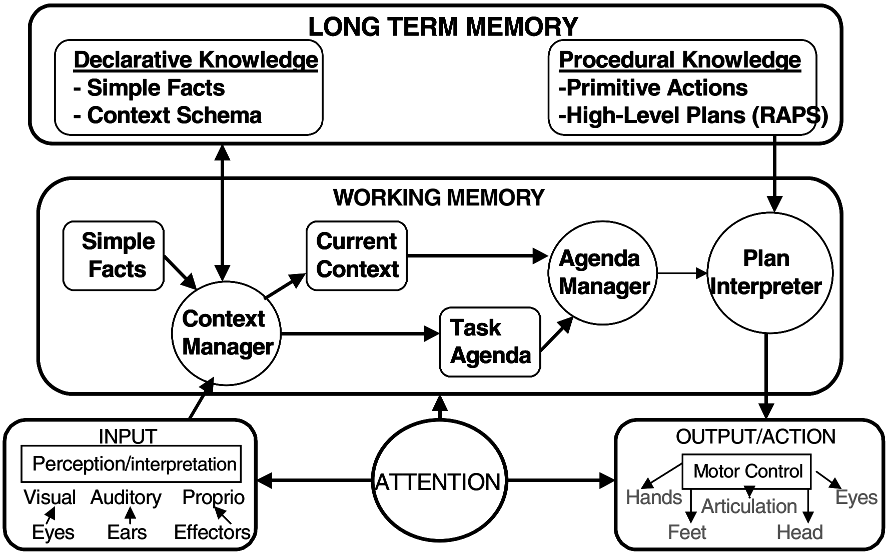

LONG TERM MEMORY
Declarative Knowledge
- Simple Facts
- Context Schema
Procedural Knowledge
-Primitive Actions
-High-Level Plans (RAPS)
WORKING MEMORY
Simple Facts
Current Context
Agenda Manager
Plan Interpreter
Context Manager
Task Agenda
INPUT
Perception/interpretation
Visual
Auditory
Proprio
Eyes
Ears
Effectors
ATTENTION
OUTPUT/ACTION
Motor Control
Hands
Articulation
Eyes
Feet
Head

Figure 21 Illustration of the Overall Functional Architecture of the Human Operator Model in MIDAS. The functional architecture is repeated because multiple operators need to be represented in larger team simulations.

human operator model is the key to the MIDAS function as a predictive design aid. The human operator performance model is a combination of a series of functionally integrated micromodels of specific cognitive capabilities within a human operator. The human operator model functions as a closed-loop control model with inputs coming from the world and action being taken in the world. The model provides psychological plausibility through the explicit representation of cognitive constructs (detailed in Section 6.3).

- *Mission and activity models:* describe in a hierarchic structure the goals and the available recovery activities from missions-not-as-planned that make up the human operators' high-level behavioral repertoire in the mission. The next level of decomposition of the action of the mission is a set of high-level procedures (that can be stored as a fairly generic set of routines, e.g., look at or fixate). Finally, there are the specific activities in "active action packets" RAPS, which are the process by which the human operator affects the simulation.

6.2.1.2. Graphical Simulation Model MIDAS provides the user with a set of CAD development tools for both the design and modification of the simulation domain elements. The system also provides for the selection of a set of views or graphical simulation windows. These windows can provide various perspectives into the ongoing simulation, from that of a God's-eye view of the vehicles, environment and operators to views from the eyes of a particular human agent in the simulation.

In addition to the standard physical views, MIDAS supports a number of analytic views. The user can select views into the activity performance over time or views into the cognitive activities of any of the operators of the simulation. In addition, these views can be linked to the overall simulation view so that the user can view the cognitive activities and their effect in the real world at the same time. The time synching also allows for retrospective reply after a particular simulation has been performed. Figure 22 illustrates a standard set of views that might be selected during a simulation.

6.2.1.3. Analytic Tools After the simulation trials have been run, MIDAS provides the user with an editor to aid in the specification of analyses to be performed on the data generated. A set of analyses can be undertaken in the MIDAS model environment, which will allow the analyst views

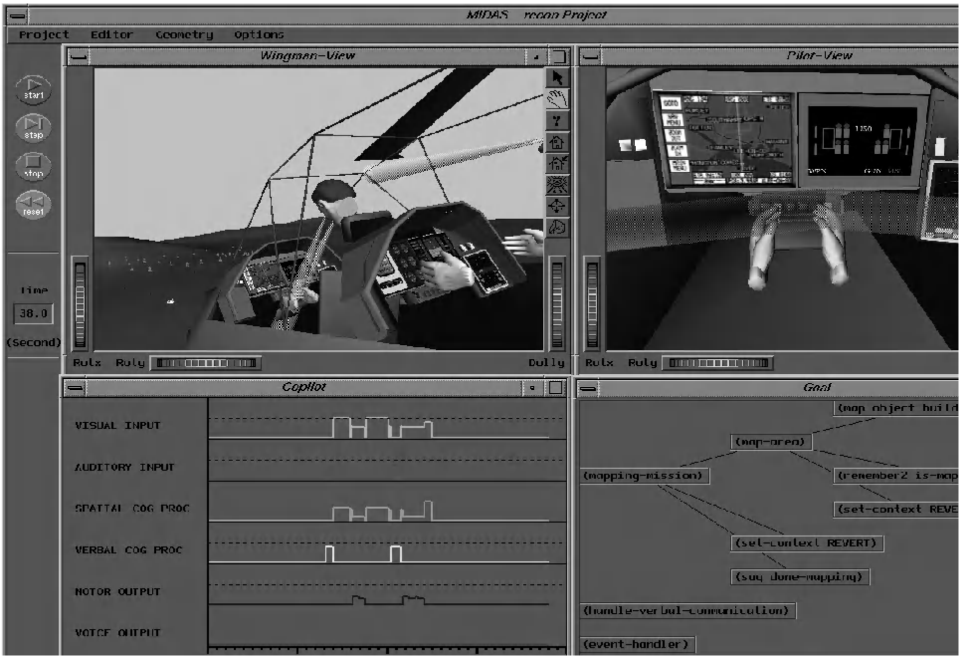

MIDAS recon Project
Project
Editor
Geometry
Options
Wingman-View
Pilot-View
start
step
stop
reset
Time
38.0
(Second)
Rulx
Ruly
Dully
Rulx
Ruly
Copilot
VISUAL INPUT
AUDITORY INPUT
SPATIAL COG PROC
VERBAL COG PROC
MOTOR OUTPUT
VOICE OUTPUT
Goal
(map object build
(map-area)
(mapping-mission)
(remember2 is-map
(set-context REVE
(set-context REVERT)
(say done-mapping)
(handle-verbal-communication)
(event-handler)

Figure 22 Illustration of a Run-Time View of the MIDAS System with Cockpit, Pilot, and Copilot, and an Activity and Load Time as Well as the Procedures and Goals That Are Currently Active in the System.

TABLE 1 Representational Models Within MIDAS.

Micromodels	Empirical research
Visual processing (field of view)	Arditi and Azueta 1992; Lubin and Bergen 1992
Visual perception	Remington et al. 1992
Auditory processing	Card et al. 1983
Central processing and memory	Baddeley and Hitch 1974
Effectors/output behavior (35 primitive tasks)	Hamilton et al. 1990
Attention—multiple-resource theory	Wickens 1984
Anthropometric models	Badler et al. 1993 (28)

Source: Gore and Corker 1999.

into the task timeline and load levels for the operators simulated in the scenario. MIDAS supports packaging and exportation of specific data associated with the simulation entities to external statistical packages as well.

6.3. MIDAS Structural Architecture

The structural architecture of MIDAS is that of a federated set of models organized into groupings that represent the various agents in the simulation. We will concentrate here on the structural integration of the models that compose the human operator(s) in the MIDAS modeling system. These models have been developed in a structure that represents an empirically based human information-processing model. This structural integration has been termed a first-principles model, based on its intentional integration of cognitive models that represent separable elements of the cognitive process.

The first-principles models of human performance are based on the mechanisms that underlie and cause human behavior. First-principles models integrate human perceptual and cognitive systems and human motor systems, thus incorporating the higher-level behaviors that are characteristic of human performance. This incorporation supports emergent human behavior based on elementary model function. The representational models and the research upon which these models have been based can be seen in Table 1.

Figure 23 presents a schematic of the way that the models can be structured for many human performance-modeling applications.

The cognitive submodels function as follows.

6.3.1. Working and Long-Term Memory Stores

Working memory is the store that is susceptible to interference and loss in the ongoing task context. Long-term memory loss would represent, for instance, a loss of skills or deep procedural memory

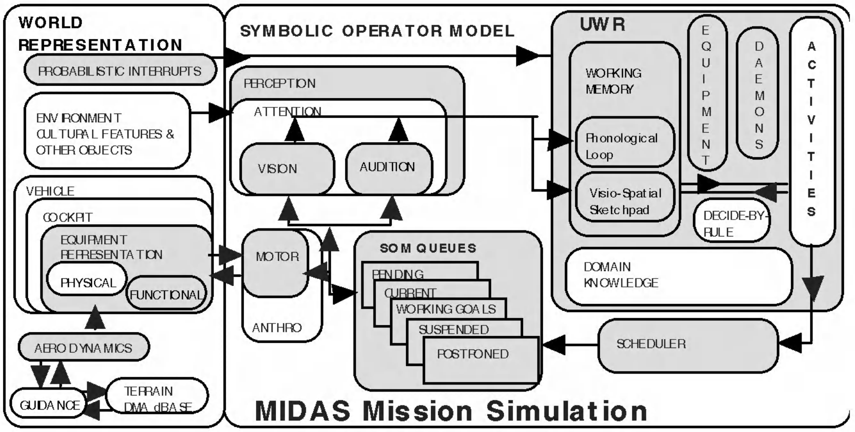

WORLD REPRESENTATION
PROBABILISTIC INTERRUPTS
ENVIRONMENT CULTURAL FEATURES & OTHER OBJECTS
VEHICLE
COCKPIT
EQUIPMENT REPRESENTATION
PHYSICAL
FUNCTIONAL
AERO DYNAMICS
GUIDANCE
TERRAIN DMA dBASE
SYMBOLIC OPERATOR MODEL
PERCEPTION
ATTENTION
VISION
AUDITION
MOTOR
ANTHRO
SOM QUEUES
PENDING
CURRENT
WORKING GOALS
SUSPENDED
POSTPONED
UWR
WORKING MEMORY
Phonological Loop
Visio-Spatial Sketchpad
EQUIPMENT
DAEMONS
ACTIVITIES
DECIDE-BY-RULE
DOMAIN KNOWLEDGE
SCHEDULER
MIDAS Mission Simulation

Figure 23 Structure of Cogntive Models.

of how to perform tasks. Neither the process of loss nor development of long-term memory, learning, is explicitly represented as a MIDAS function. We have modeled human memory structures as divided into long-term (knowledge) and working memory (short-term store). We have implemented working memory, described by Baddeley and Hitch (1974) as composed of a central control processor (of some limited capacity), an ‘‘articulatory loop’’ (temporary storage of speech-based information), and a ‘‘visuo-spatial scratch pad’’ (temporary storage of spatial information). The point of transference of information from the flight deck and ATC displays to the operators’ working memory is the critical juncture in the subsequent use of the information that is exchanged.

Long-term memory structure is provided via a semantic net. The interaction of procedure with memory is provided by a goal decomposition method implemented as a form of cognitive schema. In order to capture the central role of schema and internal representation, we have an elaborate representation of both declarative and procedural information in the MIDAS model. In MIDAS, the internal updatable world representation (UWR) provides a structure whereby simulated operators access their own tailored or personalized information about the operational world. The structure and use of the UWR are akin to human long-term memory and are one of the aspects of MIDAS unique from most human–system modeling tools. UWR contents are defined by presimulation loading of required mission, procedural, and equipment information. Data are then updated in each operator’s UWR as a function of the mediating perceptual and attentional mechanisms previously described. These mechanisms function as activation filters, allowing more or less of the stimuli in the modeled environment to enter the simulated operator’s memory. Knowledge of what is on each operator’s mind is a key modeling feature that allows MIDAS to examine decision making and the information exchange that is critical to decision making.

6.3.2. *Attentional Control*

Representation of human–automation integration requires functions of attentional control and concurrent task performance. Distributed attention and attention switching refer to an operator’s ability to perform multiple tasks simultaneously. In many cases, a second task can be added to the performance of a primary task with little or no impact to the performance of the first task. In other cases, the performance of two tasks simultaneously has a disastrous interaction. Such context-and order-sensitive effects are determined in the scheduling and agenda management function provided in the MIDAS model. Attention capture functions are represented through a preattentive filter mechanism that responds to physical characteristics of environmental stimuli (e.g., color, blinking, auditory characteristics).

6.3.3. *Activity Representation*

Tasks or activities available to an operator are contained in that operator’s UWR and generate a majority of the simulation behavior. Within MIDAS, a hierarchical representation is used (similar to, but more flexible than, the mission–phase–segment–function–task decomposition employed by many task-analysis systems). Each activity contains slots for attribute values, describing, for example, preconditions, temporal or logical execution constraints, satisfaction conditions, estimated duration, priority, and resource requirements. A continuum of contingent or decision-making behavior is also represented in MIDAS, following the skill, rule, knowledge-based distinction reported by Rasmussen (1983). The activity structures in MIDAS are currently being implemented as sketchy plans in the reactive action packets paradigm (RAPS) of Firby (1989). This structure of activities will interact with resource and context mangers to structure an agenda.

6.3.4. *Task Agenda*

The agenda structure stores instantiated RAPS as goals with subnetworks and logical control flags, object bindings, and history of state and completion. This network represents the current set of tasks to be performed by the operators of the simulation given the current goals and context. The network can complete successfully, be interrupted by other task networks, or be aborted. The relationship among the actions in terms of logic of performance (e.g. sequential or concurrent tasks) is also specified in the agenda structure. Whether, in fact, tasks can be performed concurrently is a function of resource relations in the cognitive model (sensation/reception, central/attentional/effectors). Work is currently underway to unify the representation of action and resources in the various version of the MIDAS system.

6.3.5. *Decision Making*

Quick, skill-based, low-effort responses to changes in values of information held in the UWR are captured by ‘‘daemons’’ when a triggering state or threshold value, sensed by perception, is reached. Daemons represent well-trained behaviors such as picking up a ringing phone or extinguishing a caution light. Classic production rule-based behavior is also available and is used when conditions in the simulation world match user-defined rule antecedent clauses active for the scenario modeled. Finally, more complex or optimization-oriented decision making is represented via a set of six prescriptive algorithms (e.g., weighted additive, elimination by aspect, etc.), as reported by Payne et al. (1988). Each of these algorithms uses a different combination of attribute values, weights, and cutoff values for calculating the ‘‘goodness’’ of the options.

6.3.6. Higher-Level Functions

The cognitive submodel architecture allows for the development of higher-order functions of cognition. For example, Shiveley et al. 1995 has developed and demonstrated a ‘‘situation awareness’’ function that combines characteristics of working memory, long-term memory search, and perceptual models to develop an abstraction termed situation, which can then be used to guide behavior or to serve as a measure of adequacy of information in the environment and in the crew’s knowledge store to meet task demands.

Context is a combination of declarative memory structures and incoming world information that is mapped to the agenda manager who is taking the plan (overall mission). This, combined with the plan interpreter, provides a series of actions to be performed in order to meet mission goals and handle contingent activities (such as interruption or plan repair). Verma (2000) has explored the extension of situation to ‘‘contextual control’’ (Hollnagel 1993) by using the elements of rule-based behavior, number of goals in working memory, and decision horizon in the MIDAS planning module.

6.4. Case Studies in MIDAS Applications to Aviation

The world community of aviation operations is engaged in a vast, system-wide evolution in human–system integration. The nature of this change is to relax restrictions in air transport operations wherever it is feasible. The relaxation includes schedule control, route control, and, potentially, separation authority in some phases of flight, such as aircraft self-separation in en route and oceanic operations. The consistent result of the relaxation of system constraints is to change and challenge human performance in that system in two dimensions. First, the decision-making process becomes distributed. This distributed decision differs from current operation and has a direct impact on crew and team resource management processes. Second, the dynamic concept of operations provides a new challenge to the human operators of that system. The human operators (pilots, air traffic controllers, and airline operations personnel) must monitor and predict any change in the distribution of authority and control that might result as a function of the airspace configuration, aircraft state or equipage, and other operational constraints. The operators are making decisions and sharing information not only about the management of the airspace, but also about the operating state of that airspace. In order to support collaborative and distributed control between air and ground for separation assurance, modifications to the roles and decision authority must be explored. The evolution of the air transportation system, therefore, profoundly challenges human performance prediction and the cognitive sciences. Previous models of human performance linked to machine performance have had distinct boundaries in the human–machine elements to be modeled. Current system design requires models of human, machine/automation, aircraft, airline operations, air traffic management, and National Airspace (NAS) management to be tightly coupled in order to guide design, evaluate the effectiveness, and ensure the safe operation of the system.

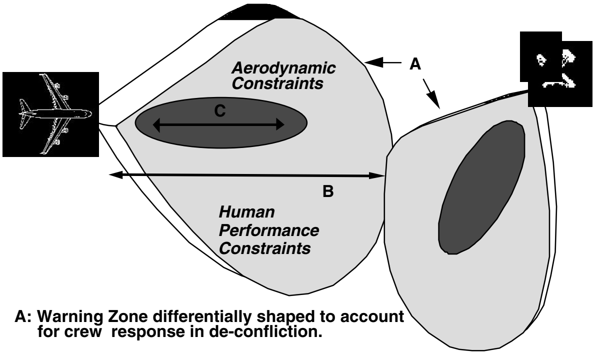
6.4.1. MIDAS Case Study 1: Predicting Flight Crew Performance in the Advanced Air Traffic Management System

We have focused our early investigation on critical issues in air ground coordination in relation to aircraft self-separation. The interaction among aircraft and controllers is proposed to occur at points in space around each aircraft called *alert* and *protected zones*.

These zones are to be used by an alerting system to monitor and advise the flight crew on conflicting traffic flying within these areas. In a cockpit-based system, the alerting system would warn the flight crew of any aircraft entering the alert zone. The crew could evaluate the situation and choose or negotiate a preferred deviation. If the intruding aircraft continued into the smaller warning zone, the crew would be advised to take evasive action.

Much discussion and debate has gone into the further definition of the warning and alert zones, including their description as complex surfaces that take into account the speed, performance, and turning radius of the aircraft. Up to this time, the process of definition lacked data for inclusion of human performance in the size and shape of these areas. Figure 6 proposes a redefinition of these zones based on a human–machine system performance. Built upon the well-defined physical aerodynamic response of the aircraft are the more varying machine (sensing, communication, computation) and human (perception, communication, decision, action) responses to any alert. These zones might also differ, depending on the speed of the aircraft, configuration of the conflict, and procedures used to process the conflict.

6.4.1.1. Study 1: Model Analysis The goal of this study was to develop a better understanding of the impact of joint and distributed decision making among flight crews and air traffic control on



A: Warning Zone differentially shaped to account for crew response in de-confliction.

B: Crew response time (RT) determines perimeters of warning/alert zones:

RT= (Perception t) (Decision t) (Communication t) (Neuromotor Response t)
/ modulation function of intent (expected (+) unexpected (-))

C: Defined by minimum reaction time, similar to TCAS Resolution Alert

Figure 24 Alert and Protected Zones Calibrated to Human Performance Parameters, Aircraft Performance Parameters, and Communication Systems Parameters.

the size and shape of the alert zones. This was accomplished by first analyzing and modeling the cognitive and procedural requirements of several candidate encounter scenarios. These models were then populated with performance data derived from human-in-the-loop experiments. The specified scenarios were then represented within the MIDAS computational modeling and simulation system.

With Monte Carlo simulation techniques, each scenario could be exercised many times, eventually establishing a statistical distribution for the human–machine performance of that configuration. Combining this with the aerodynamic performance of the system (in this case the closing speed of conflicting aircraft at differing encounter angles) meant that the differences in warning requirements between the different scenarios should have emerged. All encounters were assumed to be two-ship interactions.

- *Scenario 1:* In this scenario, both aircraft are equipped with some type of CDTI detection equipment. Here a single aircraft can detect and avoid the conflicting aircraft by acting on its own (no communications are necessary). This might describe a situation where one aircraft is slowly closing on another from behind.
- *Scenario 2:* Both aircraft are again equipped. However, because of the geometry of the encounter and conflicting goals, both aircraft must be involved and negotiate to resolve the problem. The solution would be arrived at though communication and negotiation between the two flight crews.
- *Scenario 3:* This scenario describes an encounter where communications with ATC are required to resolve the problem. This is necessary because one aircraft is equipped with the required suite of equipment while the other is not. Such encounters might be common early in the implementation of free flight or when encountering older, nonupgraded aircraft.

6.4.1.2. High-Level Activity Definition in the MIDAS Model An initial cognitive and physical task analysis was performed for each of the three scenario cases. The result was a sequential model identifying the high-level processes (or activities) performed by the operators. In scenarios 2 and 3, the activities that were to be performed in parallel by the other flight crew and ATC were also defined. Falling out of this analysis was a recognizable cycle of alert, recognition, communication, decision, then communication and action by the crews. This process was replicated throughout the scenarios for each flight crew interaction.

6.4.1.3. Lower-Level Activity Specification Using these sequences as a guide, the lower, or leaf-level, activities (corresponding to the physical or cognitive tasks actually performed by the operators) were defined for each high level task. Columns 2 and 3 of Figure 7 show the high- and lower (leaf)-level activities defined for scenario 1. The remaining columns show the interrupt recovery, duration, and VACM (visual, auditory, cognitive, and motor channel capacity requirements) specifications assigned to those activities. Where possible, the activities were chosen to correspond to research that had been performed in previous studies (Corker and Pisanich 1995). This provided access to fully defined activity specifications. New activities, along with their specifications, were developed by interpolating prior results.

The MIDAS model can contain activities that may interrupt the flight crew from the normal activities (for example, a question in the cockpit may interrupt a flight crew member from a CDU entry task). The interrupt resumption specifications define how an activity is resumed after being suspended. Resumption methods are individually defined based on the characteristics of the activity and the sequence in which it operates. The resumption methods used on this simulation include not-interruptible (cannot be interrupted); resume (resume activity where interrupted); and restart (restart the activity from its beginning). Interruptions and the way that an activity is resumed directly affect the duration of the activity sequence.

6.4.1.4. Experimental Runs After specification and testing, each scenario was loaded into Air-MIDAS and 50 Monte Carlo runs were gathered for that simulation. The data recorded for each run included the activity sequence along with the individual activities and their duration for that sequence (including any interrupt activities). These data were written to a file for analysis in Microsoft Excel format. This data was postprocessed using the rules described earlier to extract a proper time for the parallel activity sets. This allowed the establishment of a total duration (time required for all operators to complete their tasks) for each scenario run. This was the dependent variable in this study.

6.4.1.5. Results A standard set of descriptive statistics was generated for each scenario based on the set of 50 Monte Carlo runs. The temporal performance data were also plotted as a histogram using a bin size of 10 seconds, illustrated in Figure 25.

The performance observed in each scenario above was applied to a 90° crossing conflict. In this geometry, the initial traffic alert was proposed to be signaled at 40 miles from the crossing point and assumed a typical commercial aircraft cruise speed (Mach 0.82). The measure in this case was the closing distance (straight-line distance between the aircraft). The minimum, maximum, and average human–machine performance times are illustrated in Figure 25. This calculation also allowed the determination of a closing distance for each performance time (potentially from 56 down to 0 miles). Using this geometry, an idea of the initial warning distance could be inferred using the worst-case performance criteria. Although the average clearing distances in scenarios 1 and 3 differ significantly, given a warning at 40 miles, the worst-case time in both scenarios would allow an avoidance maneuver to begin well before the aircraft were 20 miles from each other. In scenario 3, however, the

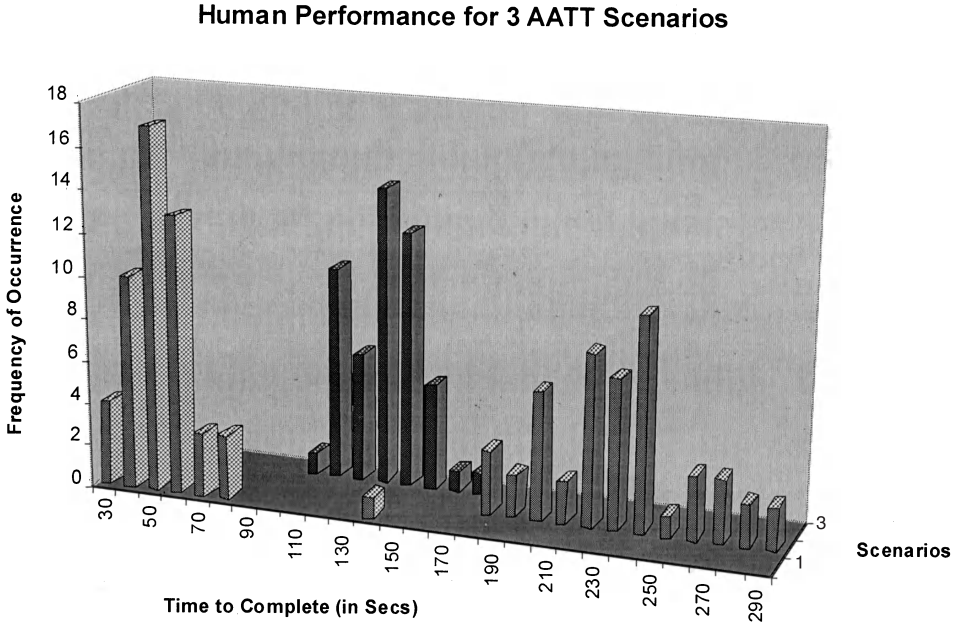


Figure 25 Response Times for an Air-to-Air Encounter at 90° Intercept.

worst, and even average, clearing distances observed would indicate that the alert point for that type of interaction should be initiated well beyond 40 miles.

To investigate this idea further, a second application of the human performance data was performed. Calculations were again made with both aircraft maintaining a speed of Mach 0.82. For each 15° angle around the aircraft, the resulting closing speed was calculated. Combining that speed and the performance distribution of each scenario resulted in a distance traveled for that angle. In this case, two standard deviations above and below the average were used as the minimum and maximum points respectively. Five miles were added to these distances to account for the warning zone. When plotted, these points create the heart-shaped rosettes shown in Figure 26.

In addition to showing the difference in warning distance needed to maintain the same performance at differing closing angles, these plots are interesting because they also illustrate a difference in performance area (size of the area between minimum and maximum performance) between the three scenarios. Given the performance observed, the higher closing speeds actually exacerbate the differences between the scenarios. Although scenarios 1 and 3 looked comparable in the 90° closure shown earlier, at shallower angles scenario 3 actually requires a significantly earlier warning point to maintain the five mile alert zone.

6.4.2. Extension of Model to Air Traffic Control

The second example of MIDAS use concentrates on the air traffic controller operations in response to free-flight self-maneuvering aircraft. In this study, the human operator's performance measures in the distributed air/ground air traffic management (ATM) that characterizes the multiple-controller, multiple-aircraft system include visual monitoring, perception, spatial reasoning, planning, decision making, communication, procedure selection, and execution. Two scenarios will be created in the current modeling effort. The first scenario will be operated consistent with the air traffic control rules of flight. The second scenario will be operated consistent with free flight rules of operation established by RTCA. Each scenario involves a response to a scripted conflict situation in a number of conditions.

The MIDAS model of the controller was developed using the rules of current and free-flight operation. Estimates and measured times for controllers performance on these tasks was input into the task descriptions. Here the interest was in how stable the system was to emergency conditions under current and free-flight conditions. The control scenario was focused on the handoff between

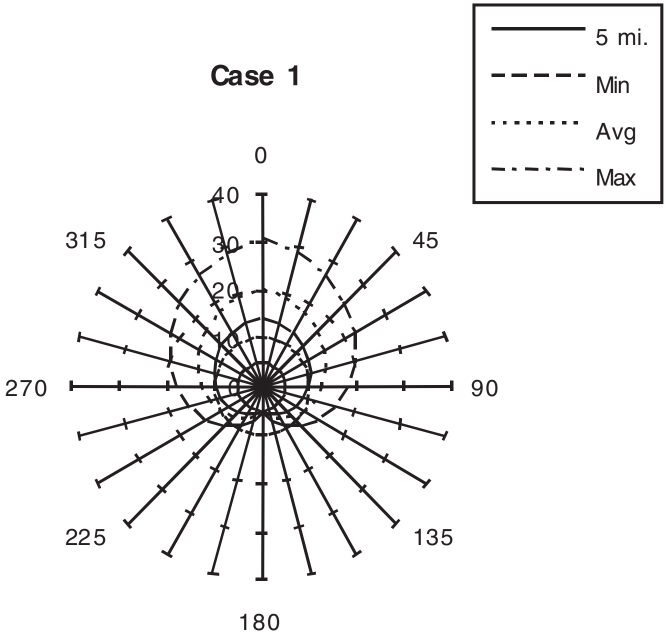


Figure 26 Cartoid Shape of Minimum Safe Distance to Alert Calculated as a Function of Model-Generated Crew Response Times.

sectors in an enroute segment of flight. This scenario then required the development and interaction of several MIDAS human performance models. There were two controllers modeled that interacted with a single aircraft flight crew. The scenario is illustrated in Figure 27.

Results were compiled for various conditions of weather and emergency management under the two flight regimes and controllers' performance profiles for workload were developed. In addition, operational measures, such as aircraft maneuvers and points of closest approach, were also calculated. The details of these analyses can be found in Gore (1999).

6.5. Other Modeling Strategies That Have Demonstrated Utility in Modeling Human Performance in Systems

There has been a flurry of interest over the past decade in the use of object-oriented tools that focus on the representation of human knowledge and how humans use that knowledge in real situations to make decisions and act. An example of this approach is the Distributed Operator Model Architecture (DOMAR), developed by Dr. Michael Young of the Wright Patterson Air Force Research Laboratories and Mr. Stephen Deutsch of BBN Inc. (Young and Deutsch 1999). The system is composed of a framework of software languages and model-development tools. As with the other models discussed in this section, the intention is to provide the analyst with the necessary components to develop models that simulate both the human operators and the systems with which they interact.

DOMAR is unique in its intention to allow these models to interact not only to allow with other entities in the model environment, that is, with simulations, but to allow the human performance models developed in OMAR to interact with live human operators and real systems in a hybrid human–simulation modeling capability.

The component software tools that are part of DOMAR allow the developers to construct their own unique cognitive architecture for the operators simulated in DOMAR.

DOMAR is made up of tools to support the development of appropriate cognitive models of human operators interacting with systems. The following description is provided by Young and Deutsch (1999):

- The Simple Frame Language (SFL) is used to represent knowledge about objects, their attributes, and the relationships among objects. It provides an object-oriented substrate for the entire simulation environment. One of the most important classes of objects is an agent—the special type of object that can execute procedures and thereby run in the DOMAR simulator. The Rule Definition Language (RDL) provides computational primitives for representing rules (i.e., condition action pairs of the form if-then). Rules are developed as rule packets, which specify when and how the rules are to be applied to a specific situation. The Simulation Core Language

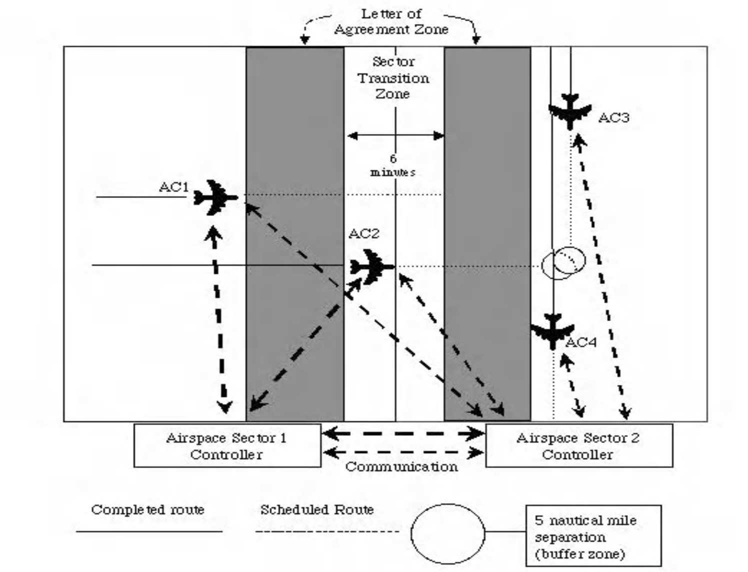


Figure 27 Cross-Sector Handoff Scenario

(SCORE) is a procedure definition language that provides a set of computational primitives for modeling the priority-based serial or parallel execution of goals, plans, and tasks.

- The DOMAR publish–subscribe protocol, implemented in a set of SCORE forms, plays a central role in several important aspects of model building. Within the human performance models themselves, it is used to coordinate the concurrent operation of related procedures. It is frequently employed in the modeling of person-to-person communication, and it is the basis for the distributed execution, in real time or fast time, that DOMAR supports. Very large agent-based simulations may be developed in a distributed computing environment.
- Knowledge engineering: DOMAR provides graphical tools to support knowledge acquisition, the development and management of large bodies of code, and data-analysis tools to support model development and experiment evaluation. SFL objects and agents are defined using the graphical concept editor, which provides an interactive graphical display of the hierarchical relationship among object definitions as a directed network of connected nodes. It provides a table window for displaying and editing slots associated with objects. SCORE procedures are created with a text editor (e.g., EMACS). They can then be viewed with the procedure browser, which provides several different graphical views of the programmatic interrelationships among the procedures.
- Analysis capabilities: DOMAR provides tools for monitoring events as reported by the simulator during a model run. Events that can be monitored include goal events, procedure events, stimulus-and-response events, and user-defined events such as communication events. Using a menu, the DOMAR user chooses the events to be displayed before the start of a simulation or at any point during a run. Finally, DOMAR has two post-run analysis displays. The event timeline window displays a listing of time-tagged events for one or more simulation agents. Individual agents have their own timelines within the display. The task timeline window displays the set of concurrently running tasks for one or more agents over a specific period of time. It provides the start and end time for a task and the task's status and priority. All of the displays are written in the Java language and can be connected to any DOMAR image that may be running on a local or remote computer.

6.5.1. Sample Applications

DOMAR was created to support the development of human performance models. DOMAR's sophisticated agent framework has recently been used by other researchers to develop a variety of agent-based systems. DOMAR is being used by AFRL in its Human Interaction with Software Agents (HISA) project to create human–computer interfaces incorporating intelligent agents; by the Defense Advanced Research Project Agency (DARPA) in its Collaborating Agent Based Systems (CoABS) program to demonstrate a collaborating agent communication framework and in its Joint Forces Air Component Commander (JFACC) program to create a business enterprise model of the JFACC process; and by the Navy in its Conning Officer Virtual Environment (COVE) program to create an intelligent training system.

7. SUMMARY

This chapter has provided the need for simulating performance of complex human-based systems as an integral part of system design, development, testing, and life-cycle support. It has also defined two fundamentally different approaches to modeling human performance: a reductionist approach and a first-principled approach. Additionally, we have provided detailed examples of two modeling environments that typify these two approaches, along with representative case studies.

As we have stated and demonstrated repeatedly throughout this chapter, the technology for modeling human performance in systems is evolving rapidly. Furthermore, the breadth of questions being addressed by models is constantly expanding. We encourage the human factors practitioner with a little creativity and computer savvy to consider how computer simulation can provide a better and more cost-effective basis for human factors analysis.

REFERENCES

Allender, L. (1995), personal communication, December.

Allender, L., Lockett, J., Headley, D., Promisel, D., Kelley, T., Salvi, L., Richer, C., Mitchell, D., and Feng, T. (1994), "HARDMAN III and IMPRINT Verification, Validation, and Accreditation Report," Prepared for the US Army Research Laboratory, Human Research and Engineering Directorate, December.

Archer, S. G. ,and Adkins, R. (1999), "Improved Performance Research Integration Tool (IMPRINT) Version 5 User's Manual," Prepared for the the US Army Research Laboratory, Human Research and Engineering Directorate, December.

Arditi, A., and Azueta, S. (1992), "Visualization of 2-D and 3-D Aspects of Human Binocular Vision," Presented at the Society for Information Display International Symposium (available from the Lighthouse, Inc.).

Atencio, A., Shively, R. J., and Shankar, R. (1996), "Evaluation of Air Warrior Baselines in a Longbow Apache Helicopter Crewstation in a MIDAS Simulation," American Helicopter Society 52nd Annual Forum, Washington, DC.

Atencio, A., Banda, C., and Tamais, G. (1998), "Evaluation of a Short Haul Civil Tiltrotor in Emergency Go-Around—A MIDAS Simulation," American Helicopter Society 54th Annual Forum, Washington, DC.

Baddeley, A. D., and Hitch, G. J. (1974), "Working Memory," in *Advances in Research and Theory,* Vol. 8 of The Psychology of Learning and Motivation, G. H. Bower, Ed., Academic Press, New York, pp. 47–90.

Badler, N., Phillips, C., and Weber, B. (1993), *Simulating Humans: Computer Graphics, Animation and Control,* Oxford University Press, New York.

Baron, S., Zacharias, G., Muralidharan, R., and Lancraft, R. (1980), "PROCRU: A Model for Analyzing Flight Crew Procedures in Approach to Landing," in *Proceedings of the Eighth IFAC World Congress* (Tokyo).

Bierbaum, C., Szabo, S., and Aldrich, T. (1989), *Task Analysis of the UH-60 Mission and Decision Rules for Developing a UH-60 Workload Prediction Model.* US Army Research Institute Aviation R&D Activity, Fort Rucker, AL.

Boettcher, K., Riley, V., and Collins, C. (1989), "A Modeling Approach for Analyzing Human-Machine Interaction Dynamics in Computer Systems," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (Cambridge, MA).

Boff, K. R., Kaufman, L., and Thomas, J.P. (1986), *Handbook of Perception and Cognition,* John Wiley & Sons, New York.

Card, S. K., Moran, T. P., and Newell, A. (1983), *The Psychology of Human–Computer Interaction,* Erlbaum, Hillsdale, NJ.

Corker, K. M., and Pisanich, G. M. (1995), "Analysis and Modeling of Flight Crew Performance in Automated Air Traffic Management Systems," Presented at: 6th IFAC/IFIP/IFORS/IEA Symposium: Analysis, Design, and Evaluation of Man–Machine Systems, Boston.

Corker, K. M., and Smith, B. (1992), "An Architecture and Model for Cognitive Engineering Simulation Analysis: Application to Advanced Aviation Analysis," Presented at AIAA Conference on Computing in Aerospace, San Diego.

Dahl, S. G., Allender, L., Kelley, T., and Adkins, R. (1995), Transitioning Software to the Windows Environment—Challenges and Innovations," in *Proceedings of the 1995 Human Factors and Ergonomics Society Meeting, Human Factors and Ergonomics Society* (Santa Monica, CA, October).

Farmer, E. W., Belyavin, A. J., Jordan, C. S., Bunting, A. J., Tattershall, A. J., and Jones, D. M. (1995), "Predictive Workload Assessment: Final Report," Report No. DRA/AS/MMI/CR95100/, Defence Research Agency, Farnborough, UK, March.

Firby, R. J. (1989), "Adaptive Execution in Complex Dynamic Worlds," Ph.D. Thesis, Yale University, Technical Report YALEU/CSD/RR #672.

Gawron, V. J., Laughery, K. R., Jorgensen, C. C., and Polito, J. (1983), "A Computer Simulation of Visual Detection Performance Derived from Published Data," in *Proceedings of the Ohio State University Aviation Psychology Symposium* (Columbus, April).

Gore, B., F., and Corker, K. M. (1999), "System Interaction in 'Free Flight': A Modeling Tool Cross Comparison," in *Proceedings of the Digital Human Modeling Conference and Exposition,* SAE International Paper # 199-01-1987, Warrendale, PA, June.

Gray, W. D., John, B. E., and Atwood, M. E. (1993), "Project Ernestine: Validating a GOMS Analysis for Predicting and Explaining Real-World Task Performance," *Human–Computer Interaction,* Vol. 8, pp. 237–309.

Hahler, B., Dahl, S., Laughery, R., Lockett, J., and Thein, B. (1991), "CREWCUT—A Tool for Modeling the Effects of High Workload on Human Performance," in *Preceedings of the 35th Annual Human Factors Society Conference,* (San Francisco).

Hamilton, D. B., Bierbaum, C. R., and Fulford, L. A. (1990), "Task Analysis/Workload (TAWL) User's Guide, Version 4," Research Project 91-11, U.S. Army Research Institute for the Behavioral and Social Sciences (AD A241 861), Alexandria, VA.

Hoagland, D., Martin, E., Anesgart, M , Brett, B., LaVine, N., and Archer, S. (2001), "Representing Goal-Oriented Human Performance in Constructive Simulations: Validation of a Model Performing Complex Time-Critical-Target Missions," in *Proceedings of the Simulation Interoperability Workshop* (Orlando, FL, April).

Hollnagel, E. (1993), *Human Reliability Analysis: Context and Control,* Academic Press, London.

Knapp, B., Archer, S. G., Archer, R. D., and Walters, B. (1999), "Innovative Approaches to Modeling – An Application for National Missile Defense," in *Proceedings of the Society for Computer Simulation Conference* (Chicago).

LaVine, N. D., Peters, S. D., and Laughery, K. R. (1995), "A Methodology for Predicting and Applying Human Response to Environmental Stressors," Micro Analysis & Design, Inc., Boulder, CO, December.

Lawless, M. L., Laughery, K. R., and Persensky, J. J. (1995), "Micro Saint to Predict Performance in a Nuclear Power Plant Control Room: A Test of Validity and Feasibility," NUREG/CR-6159, Nuclear Regulatory Commission, Washington, DC, August.

Little, R., Dahl, S. G., Plott, B., Wickens, C., Powers, J., Tillman, B., Davilla, D., and Hutchins, C. (1993), "Crew Reduction in Armored Vehicles Ergonomic Study (CRAVES)," Report No. ARL-CR-80, prepared for the Army Research Laboratory, July.

Lubin, J., and Bergen, J. (1992), "NASA Cockpit Display Visibility Modeling Project," Final Report NAS2-12852, SRI/David Sarnoff Research Center, Moffett Field, CA.

McCracken, J. H., and Aldrich, T. B. (1984), "Analysis of Selected LHX Mission Functions: Implications for Operator Workload and System Automation Goals," Technical Note ASI 479-024-84(B), prepared by Anacapa Sciences, Inc., June.

McMillan, G. R., Beevis, D., Salas, E., Strub, M. H., Sutton, R., and Van Breda, L. (1989), *Applications of Human Performance Models to System Design,* Plenum Press, New York.

Micro Analysis and Design (1999), *Micro Saint, Version 3.0: User's Guide, Micro Analysis and Design,* Boulder, CO, May.

Newell, A. (1990), *Unified Theories of Cognition,* Harvard University Press, 1990.

Newell, A., and Simon, H. A. (1972), *Human Problem Solving,* Prentice Hall, Englewood Cliffs, NJ.

North, R. A., and Riley, V. (1989), "W/INDEX: A Predictive Model of Operator Workload," in *Applications of Human Performance Models to System Design,* G. McMillan et al., Eds., Plenum Press, New York.

O'Hara, J. M., Brown, W. S., Stubler, W. F., Wachtel, J. A., and Persensky, J. J. (1995), "Human–System Interface Design Review Guideline: Draft Report for Comment," NUREG-0700 Rev.1, U.S. Nuclear Regulatory Commission, Washington, DC.

Payne, W., Bettman, J. R., and Johnson, E. J. (1988), "Adaptive Strategy in Decision Making," *Journal of Experimental Psychology: Learning, Memory and Cognition,* Vol. 14, No. 3, pp. 534–552.

Plott, B. (1995), "Software User's Manual for WinCrew, the Windows-Based Workload and Task Analysis Tool," U.S. Army Research Laboratory, Aberdeen Proving Ground, MD.

Rasmussen, J. (1983), "Skills, Rules, and Knowledge; Signals, Signs and Symbols, and Other Distinctions in Human Performance Models," *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-13, No. 3, pp. 257–266.

Remington, R. W., Johnston, J. C., and Yantis, S. (1992), "Involuntary Attentional Capture by Abrupt Onsets," *Perception and Psychophysics,* Vol. 51, No. 3, pp. 279–290.

Roth, J. T. (1992), "Reliability and Validity Assessment of a Taxonomy for Predicting Relative Stressor Effects on Human Task Performance," Technical Report 5060-1, prepared under contract DNA001-90-C-0139, Micro Analysis and Design, Inc., Boulder, CO, July.

Shively, R. J. et al. (1995), "MIDAS Evaluation of AH-64D Longbow Crew Procedures in a Air-Ground Flight Segment: MOPP versus Unencumbered," Sterling Federal Systems, NASA Ames Research Center.

Siegel, A. I., and Wolf, J. A. (1969), *Man–Machine Simulation Models,* Wiley-Interscience, New York.

Tyler, S., Neukom, C., Logan, M., and Shively, J. (1998), "The MIDAS Human Performance Model," in *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (Chicago), pp. 320–325.

Verma, S. (2000), "Introduction of Context in Human Performance Models as Applied to Dynamic Resectorization," Master's thesis, San José State University, Human Factors and Ergonomics Program, San José, CA.

Wickens, C. D. (1984), *Engineering Psychology and Human Performance,* Merrill, Columbus, OH.

Wickens, C. D. (1989), "Models of Multitask Situations," in *Applications of Models to System Design*, G. McMillan, Ed., Plenum Press, New York, pp. 259-273.

Wickens, C. D., Sandry, D. L., and Vidulich, M. (1983), "Compatibility and Resource Competition Between Modalities of Input, Central Processing, and Output," *Human Factors,* Vol. 25, pp. 227–248.

Wortman, D. B., Duket, S. D., Seifert, D. J., Hann, R. l., and Chubb, A. P. (1978), *Simulation Using SAINT: A User-Oriented Instruction Manual,* Aerospace Medical Research Laboratory, AMRL-TR-77-61. Wright-Patterson Air Force Base, OH, July.

Young, M. and Deutsch, S. (1999), www.he.afrl.af.mil/hes/hess/programs/omar/2omarnext.html.

ADDITIONAL READING

Agha, G., "Actors: A Model of Concurrent Computation in Distributed Systems," Technical Report 844, MIT Artificial Intelligence Laboratory, Cambridge, MA, 1985.

Allender, L., Kelley, T., Salvi, L., Headley, D. B., Promisel, D., Mitchell, D., Richer, C., and Feng, T., "Verification, Validation, and Accreditation of a Soldier-System Modeling Tool," *Proceedings of the 39th Human Factors and Ergonomics Society Meeting* (San Diego, October 9–13, 1995) (available from the Human Factors and Ergonomics Society, Santa Monica, CA).

Archer, R., Drews, C. W., Laughery, K. R., and Dahl, S. G., "Data on the Usability of Micro Saint," in *Proceedings of NAECON Meeting* (Dayton, OH, May 1986).

Baron, S., and Corker, K., "Engineering-Based Approaches to Human Performance Modeling," in *Applications of Human Performance Models to System Design,* Grant McMillan, Ed., Plenum Press, New York, 1989.

Corker, K. M., and Pisanich, G. M., "When Reasonable Expert Systems Disagree," Presented at the Topical Meeting of the American Nuclear Society, Computer-Based Human Support Systems: Technology, Methods, and Future, Philadelphia, 1995.

Corker, K. M., Lozito, S., and Pisanich, G., "Flight Crew Performance in Automated Air Traffic Management," in *Human Factors in Aviation Operations,* R. Fuller, N. Johnston, and N. McDonald, Eds., Vol. 3, Avebury Aviation, Hants, UK.

Corker, K. M., and Smith, B., "An Architecture and Model for Cognitive Engineering Simulation Analysis: Application to Advanced Aviation Analysis," Presented at AIAA Conference on Computing in Aerospace, San Diego, 1993.

Craik, K. J. W., "Theory of the Human Operator in Control Systems I. The Operator as an Engineering System," *British Journal of Psychology,* Vol. 38, 1947, pp. 56–61.

Drews, C., and Laughery, K. R., "A Modeling System Designed Around the User Interface," Presented at the Summer Computer Simulation Conference, Chicago, July 1985.

Drews, C., Laughery, R. R., Kramme, K., and Archer, R., "LHX Cockpits: Micro SAINT Simulation Study and Results," Report prepared for Texas Instruments—Equipment Group, Dallas, June 1985.

Gore, B., "A Comparison of Human Performance Models Applied to Advanced Air Traffic Management Operations," Master's thesis, San José State University, San José, CA.

Reason, J. T., "The Chernobyl Errors," *Bulletin of the British Psychological Society,* Vol. 40, 1987, pp. 201–206.

Reason, J. T., *Human Error,* Cambridge University Press, Cambridge, 1990.

Remington, R. W., Johnston, J. C., and Yantis, S., "Involuntary Attentional Capture by Abrupt Onsets," *Perception and Psychophysics,* Vol. 51, No. 3, pp. 279–290.

Shankar, R., "Z-Scheduler: Integrating Theories of Scheduling Behavior into a Computational Model," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* pp. 1219–1223.

Tustin, A., "The Nature of the Operator's Response in Manual Control and Its Implication for Controller Design," *Journal of the IEE,* Vol. 94 (Part IIA, No. 2).

Tversky, A., and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science,* Vol. 185, 1974, pp. 1124–1131.

CHAPTER 94

Simulation Packages

ABE NISANCI
Bradley University

ROBERT E. SCHWAB
Caterpillar Inc.

- **1. INTRODUCTION 2445**
  - 1.1. Simulation Languages 2446
  - 1.2. Simulation Packages 2446
  - 1.3. Complementary Products 2446
- **2. SELECTION STRATEGY 2447**
  - 2.1. Establish the Uses of the Tool 2447
  - 2.2. Select the Users of the Tool 2447
  - 2.3. Design the Training Program 2448
  - 2.4. Assess Support Needs 2448
- **3. EVALUATING SIMULATION TOOL CHARACTERISTICS 2449**
  - 3.1. Language Characteristics 2449
  - 3.2. Operational Characteristics 2450
  - 3.3. Input Characteristics 2450
  - 3.4. Output Characteristics 2450
  - 3.5. Hardware Characteristics 2451
  - 3.6. Vendor Characteristics 2451
  - 3.7. Evaluation Methodology 2452
- **4. SIMULATION TOOLS 2454**
  - 4.1. Languages 2454
  - 4.2. Packages 2456
  - 4.3. Applications 2461
- **5. PITFALLS INHERENT IN THE USE OF SIMULATION TOOLS 2461**
  - 5.1. General Pitfalls 2461
    - 5.1.1. Having the Tool and Not Using It 2461
    - 5.1.2. Failure to Establish and Adhere to Project Objectives 2462
    - 5.1.3. No Review of Inadequate Review 2462
    - 5.1.4. Not Having the Proper View of the Tool 2463
    - 5.1.5. Not Viewing Simulation as a Statistical Tool 2463
    - 5.1.6. Thinking of Simulation as an Optimization Tool 2463
  - 5.2. Pitfalls Related to Methodology 2463
    - 5.2.1. Failure to Establish Appropriate Model Boundaries and Parameters 2463
    - 5.2.2. Improper Data Collection 2463
    - 5.2.3. Not Verifying Data 2463
    - 5.2.4. Not Collecting the Proper Data 2463
    - 5.2.5. Avoidance of the Proper Level of Analysis 2463
    - 5.2.6. Validation Problems 2463
- **6. CURRENT STATUS AND TRENDS 2463**
- **REFERENCES 2465**

1. INTRODUCTION

The increased complexity and cost of modern service and manufacturing systems, coupled with the advances in computer and simulation technologies, have led to an extensive use of simulation. Parallel

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

to this development, simulation languages, and products have also proliferated. In the early 1970s, there were only a few special-purpose languages such as SIMSCRIPT, Q-GERT, GASP, GPSS, and DYNAMO, and general-purpose languages such as ALGOL, PL/1, and FORTRAN (Nance 1996; Schwab and Nisanci 1992). Initially, the most significant use of simulation was by the military. Currently, simulation plays a significant role in almost every major service and manufacturing industry throughout the world. The *Directory of Simulation Software* published by the Society for Computer Simulation (Rodrigues 1994), lists 174 simulation products that are available in the marketplace. The annual Buyer's Guide of simulation software (Institute of Industrial Engineers 1999) lists 27 companies providing 40 simulation software products. Twenty-eight simulation products are listed under the categories of general purpose simulation software, manufacturing-oriented software, business process engineering, simulation-based scheduling, animators, simulation support software, and optimization (Banks 1998a). The simulation software market is very dynamic, resulting in many mergers and new companies replacing the existing ones. There are many more products that are in various states of development and use at companies and universities. Most major companies employ at least one simulation specialist on staff or have developed a close relationship with a consulting firm that offers simulation service. Major manufacturers started requiring simulation analysis prior to approving significant capital expenditures as early as the 1980s (Krepchin 1988; Schwab 1987). This soaring interest in simulation and the simultaneous availability of PCs have produced a corresponding growth in simulation companies and products. Simulation products can be divided into two distinct categories: simulation tools and associated/complementary products. Traditionally, simulation tools include languages and simulation packages (sometimes called simulators). However, the difference between languages and packages seems to disappear over time. A simulation package may have all the power of a simulation language. Complementary products are used to support simulation serving as front and back ends.

1.1. Simulation Languages

Simulation languages allow the user to mathematically describe a process using constructs and syntax specifically designed for the analysis of dynamic and stochastic systems (Smith 1987). Early simulation models were coded in general-purpose procedural languages such as FORTRAN. This approach resulted in spending a large percentage of time developing, debugging, and verifying the model. The new languages alleviated most of these problems by significantly eliminating the need for programming. Most modern languages have an internal simulation clock, entity tracking, some form of statistical output, and animation capability. Typically, modern simulation languages have the ability to handle a broad range of problems and are flexible enough to handle complex decision rules.

1.2. Simulation Packages

Simulation packages have been in development for the last 15 years. They are a product of simulation companies' and simulation users' desire for the benefits of simulation and a belief that the simulation process should be accomplished more easily and results produced more quickly. All packages use some programming language as the base language, but some packages use a specific simulation language (e.g., Arena uses SIMAN, and AweSim uses Visual SLAM). All packages are designed to make the model-development phase much easier by limiting the number of constructs, providing a menu-driven format, presenting fill-in menus, and tying simulation and animation together. Today, all simulation packages are designed for specific industries such as manufacturing, communication, supply chain management, and business process improvement.

1.3. Complementary Products

Complementary products include all simulation support software, which are usually developed independently and later incorporated into a package as a front or end use. Some examples consist of curve-fitting programs, input data generators, scenario builders, programs for file management and data handling, graphics and animation builders, and optimization programs. The vendor of the package or the users of the package develop some upon a perceived need, such as specific material-handling extensions or applications. Some of the popular complementary simulation products are (Banks 1998a; Institute of Industrial Engineers 1999):

- ExpertFit (Averill M. Law & Associates): Automatically determines which of 39 probability distributions best represents a data set. It puts the selected distribution into the proper format for 30 simulation products. ExpertFit provides goodness-of-fit tests, more than 30 graphical plots, a distribution viewer, and batch mode operation.
- StatFit (Geer Mountain Software): Fits input data to one of 21 theoretical continuous and discrete distributions and provides relative comparisons between distribution types.
- Proof Animation (Wolverine Software Corporation): Provides animation for discrete event simulation software.

- OptQuest (OptTek Systems): An optimization tool that guides the process of selection of system inputs and then executes the model by running alternative scenarios for each set of inputs in order to enhance system performance (Laguna 1997; Glover and Kelly 1999). OptQuest is available with ARENA, Micro Saint, and Taylor ED.
- SimRunner: Given the input factor levels and system performance measures, SimRunner is employed in two stages. The first stage involves determining the important input factors, and the second stage involves determining the best factor levels to improve the system performance. It is included in ProModel's optimization module (Heflin and Harrel 1998).

2. SELECTION STRATEGY

Even with the tremendous increase in the number of simulation products and simulation companies, the selection of simulation tools need not be difficult if an organized approach is developed. This includes two key elements: the development of a strategy for the use of the products and the knowledge of the characteristics required in the products (Schwab 1987). The strategy has an impact on the features of the tool that is chosen. The features of the tool should be tailored for the people and uses of that tool. Strategy development for the use of the product is essential for the product's benefits to be fully exploited. A good strategy addresses how the tool will be used, who will use the tool, how the users will be trained, and how the users will be supported. In developing such a strategy, it would be useful to establish a team consisting of those who will use, maintain, and provide support for the product.

2.1. Establish the Uses of the Tool

First, the type of application for the tool should be specified. Types of application include:

- One product for one application
- One product for multiple applications
- Multiple products for multiple applications

Typically, individuals charged with investigating simulation tools start by selecting one product for one application. This allows for a fair evaluation of the specific simulation product in an application for which it was designed. This approach also minimizes start-up and training costs.

The one-product-for-multiple-applications approach is particularly helpful in integrating operations separated by distance or function. Selecting one product allows for a common communication tool that can be understood by everyone trained in the tool. In selecting only one tool for all applications, some sacrifice may be made in terms of modeling efficiency. Each product has features that make specific model development easy. As an example, CACI offers a product (COMNET) to design communication systems, and AutoSimulations (AutoMod) makes automatic guided vehicle systems design easy. Selecting only one product allows for standardization, model sharing, and communication but risks some inefficiency in model development.

The multiple-products-for-multiple-projects approach has the opposite advantages and disadvantages of the previous approach. It always allows for the selection of the most appropriate tool for the application. This approach is essential for consulting firms offering simulation services to diverse companies. A broad range of products and abilities is necessary to meet the diverse needs of clients. The disadvantage to this approach is a dilution of simulation personnel. If people must be trained on a number of products, then specialization in any one tool is sacrificed. As more people are cross-trained on various products, training costs increase. However, more flexibility and diversification is achieved.

Second, it is important to have an understanding of the specific use to which the tool will be put. The tool may be used for evaluation purposes such as the analysis of different layout alternatives for a manufacturing system. It may then be used for detailed design to define and fully describe the operational characteristics of the system. The tool may also be used in narrow application such as the design of communication systems or automatic guided vehicle systems. The product itself, however, may be used in broad applications from preliminary design to detailed design.

2.2. Select the Users of the Tool

Users of the simulation tool play an important part in its success. Matching the wrong people to the right tool is just as dangerous as matching the wrong tool to the project. Users of simulation may have a broad educational background and work experience. A flexible approach to the education and experience of the simulation user must be considered in the selection of a product. This also influences training requirements, which are discussed later. If personnel options are limited, then the training program for these individuals must be flexible enough to compensate for their varied backgrounds. On the other hand, if no time or money is available for training, then individuals must be chosen with enough previous training and experience to begin work on simulation projects immediately. Personnel selected for simulation projects should meet several criteria:

- Familiarity with the general process
- Analytical and computer skills
- Knowledge of probability, statistics, and design of experiments

Simulation is not just computer programming. A simulation analyst must be proficient in all of the above areas. Simulation tools are statistical tools; therefore, an understanding of statistics is essential to make effective use of any simulation product.

2.3. Design the Training Program

Even if the software supplier is to be the primary source of training, it is important to establish a comprehensive training plan as an integral part of an overall simulation strategy. The training plan should include the following major areas:

- *Fundamentals:* The fundamental issues related to the use of simulation tools might include training in the simulation process, data collection and preparation, model building, experimentation, and input and output data analysis. All of these areas can be taught by linking them to a specific simulation tool.
- *Tool training:* Training in the use of the tool is critical. Ideally, the training is accomplished over a number of days and includes an exposure to all the features of the tools. Instructors need to provide a wide range of examples and have extensive simulation experience in order to answer the questions of trainees effectively. Trainees should attend simulation classes with specific work-related projects and should begin working on them immediately after training. This helps cement the concepts of the class firmly in the minds of the trainees. Professional training and frequent use represent the only way to become proficient with any simulation tool.
- *Application training:* Many users of simulation products find it difficult to begin to use the product for a specific task even after they have received training in how to manipulate it. It may be necessary to show users how to use the product as it specifically applies to their work. An example of this would be a warehouse engineer who takes simulation training where most examples in the class are from manufacturing. The engineer may find it difficult to return to the job and immediately make application to the warehouse. This difficulty may be overcome by allowing the new simulation person to work with and be quided by an experienced simulation person. Another alternative might be to retain a consulting firm to develop the first model and allow the new simulation person to follow and assist in the entire model-development process. All new users should be taught the use of gradual complexity in the model-building process. They need to start with a simple representation of the process and get the model running. Then they can gradually add more and more complexity until the process has been modeled completely.
- *Programming:* Within the framework of modern simulation tools, what the users need to program today is more like programming an Excel spreadsheet than programming per se. However, some simulation tools provide use of an underlying language such as FORTRAN, Pascal, or C in addition to the syntax for the simulation language. This feature enables the user to program complex decision rules and other logic not capable of being handled by the simulation language itself.
- *Hardware and operating system:* New users may be unfamiliar with the hardware on which the simulation software runs, as well as the operating system utilized on the computer. A short course covering these topics enhances productivity.

2.4. Assess Support Needs

Care should be taken to determine in what manner the users will be supported. Obviously, if there is only one user, then the only reasonable approach is to depend on customer support from the software supplier. However, a site with multiple users presents some opportunities for self-support. As an example, one person at the site could be designated as the simulation support for everyone and could spend a certain number of hours each week learning more details about the simulation tool. This may result in faster problem resolution. Responsiveness to support needs, means of providing support, and geographical proximity of the support services need to be considered in selecting vendors

3. EVALUATING SIMULATION TOOL CHARACTERISTICS

Number of studies exist related to simulation software evaluation criteria and methodologies (Tumay and Harrington 2000; Banks 1998a; Nikoukaran et al. 1998; Hlupic 1997; Hlupic and Paul 1995; Pritsker 1995; Davis and Wiliams 1994; Law and Kelton 1991). The characteristics listed below should not be considered as all-inclusive but should be used as an aid in assembling a list for the software evaluation process. As an example, if the strategy calls for purchasing one package for only one computer, then portability is not a consideration. For a given set of criteria, there may be additional items that should appear on the list. The characteristics of any simulation tool can be divided into six major areas: language, operational, input, output, hardware, and vendor.

3.1. Language Characteristics

Language characteristics include the features used in writing code and providing useful external support tools. Some simulation tools allow the user to write complex rules. All the features in this section deal with writing simulation instructions.

- *Portability:* If the computing environment is made up of different hardware platforms, then the software should be usable on a wide variety of machines, from mainframes to PCs, and operating systems such as OS/2, Windows, and UNIX. Also important are factors such as special graphics card or driver requirements, ability to run on a network, special compiler needs, and memory (RAM) requirements.
- *Readability:* If the system generates or uses some type of computer code, it should be understandable and easily read by people as well as computers. Commands should be ‘‘English-like’’ statements to assist the writer and reviewer in debugging the model. The language should permit the use of comment statements within the code.
- *Documentation:* The documentation, consisting of installation instructions, test models, and user manual, must provide the user with all the information required to run and debug the models. It must also provide a thorough understanding of all aspects of the software and a large number of example problems and their detailed solutions. An extensive and well-organized index to all documents is a necessity.
- *Data requirements:* The software must allow for both deterministic and stochastic input. Sampling from system-supplied distribution functions should include theoretical statistical distributions such as exponential, normal, triangular, uniform, Poisson, beta, gama, erlang, and log-normal. In addition to these, user-defined distribution functions should be permitted.
- *Capability to support different worldviews:* Depending on the characteristics studied, a discrete or continuous change model may be suitable. In discrete change models, the variables of interest may change in a discrete fashion at any point in time or at certain points in time. In continuous change models, the variable of interest changes continuously at any point in time or at certain points in time (Pritsker and O’Reilly 1999). A project may need both discrete and continuous modeling. If so, they must be supported by the same software, or two different pieces of software must be purchased. Continuous modeling can be used with chemical-related processes such as tank filling and emptying and in discrete processes that can be modeled as continuous processes, such as high-speed canning.
- *Algorithms and modeling features:* A library of available algorithms can potentially save a significant amount of time in the model-building process. For example, automatic guided vehicles and automatic storage and retrieval systems modules are extremely useful. Additional features such as cost modeling, scheduling and schedule generation, breakdown generation, and maintenance planning also enhance modeling productivity.
- *Capability to have user-written interface:* The ability to interface with user-written routines permits the development of industry specific routines or decision logic, which can be filed in a library and reused. This allows flexibility and avoids duplication.
- *Compatibility with other software packages:* Integration is becoming more important with software packages. The ability to access other software packages allows the simulation person to work more efficiently and effectively. Interface to spreadsheet products is helpful in the collection and organization of data. An interface with computer-aided design software can be helpful to reduce the layout development time of animation. If a database package has already been chosen, the ability to integrate the simulation software with that package is extremely useful. An ability to interface as well with statistical analysis software adds additional powers and enhances user productivity.
- *Ability to communicate with other applications on a dynamic basis:* The concept of distributed interactive simulation that includes the ability to exchange data with other applications, including

other simulations, is becoming widely popular. For example, Micro Saint and Arena are both using Microsoft COM features that allow their tools to work with other applications that may even reside on different computers.

- *Ease of error recovery:* The software must be able to handle user-generated errors, especially in the interactive mode, and continue to function after receiving such an error.

3.2. Operational Characteristics

Operational characteristics involve the features in which the software interfaces with the user, except for code writing, which was considered above. These characteristics can make a significant contribution to the productivity of the modeler.

- *Applicability:* Ideally, a single piece of software should allow the novice to use the tool with a minimum amount of training and an absence of frustration. It should also allow the expert to use the tool to perform difficult, complicated simulations of a wide spectrum of systems.
- *Availability and use of debugging features:* The software must contain error diagnostics and debugging features, which help the modeler rapidly check out the model code. An interactive debugger with the ability to set stopping points and step through the model is useful in debugging complicated simulations. An important feature in model debugging is the ability to generate a detailed standard or customized trace to determine whether the model operates properly.
- *Ease of use and intuitiveness:* The software should have features to help automate the model-building process. Desirable features might include menu-driven operation, fill-in menus, and prompting for omitted values. The casual user requires online, context-sensitive help in the model-building process. The software should be accessible in an interactive mode.

  The software needs intuitive icons, menu arrangements, and keystroke sequences. Software that requires the user to enter obscure keystrokes to bring up the help screen prevents the user from making intuitive guesses about forgotten operations. Significant amounts of time are then wasted looking through a user's manual. Since most software programs use the F1 key to bring up the help screen, an intuitive system would have some form of help available by pressing the F1 key.
- *Ease of parameter variation:* The software should allow automatic parameter variation over a specified range of values and hence allow the running of multiple replications without requiring user interface. This feature can also aid in performing sensitivity analyses.
- *Ease of model revisions:* The software must allow changes to be made to the existing model quickly and easily. The best solution is rarely chosen first. The ability to change the model, the routing of the entities, and the parameters, and change them quickly, is important.
- *Interactive:* The software should allow changes in the model during the simulation run. Such changes are vital for doing "what-if" analysis with the model, although they invalidate the statistics collected during the run.
- *Availability of checkpoints:* The software should be able to print out data about the model, in tabular or graphical form, at specified intervals before the end of the simulation. This allows a look at intermediate results at critical times such as shift change.
- *Optimization:* The software should also have an optimization feature to guide the user in selecting system inputs and executing the model to enhance system performance.

3.3. Input Characteristics

The simulation model may require input from different sources. The following features may save significant time and effort in preparing the necessary input:

- Ability to read from external files, databases
- Ability to interact with popular spreadsheet, computer-aided design, and process-planning software
- Ability to trigger input functions throughout the simulation process
- Features to define, analyze, and display input in terms of statistical distributions
- Automated definition of model entities, their attributes, and rules for their interaction
- Ability to interact dynamically with other software applications

3.4. Output Characteristics

The results of simulation form the foundation for many conclusions regarding a process under consideration. Therefore, the availability and presentation of output results is important.

- *Reports:* The reports generated by the software must be clear and easy to interpret. The report format should provide graphical output such as simple bar charts and histograms as well as the standard statistical listing. Because all applications of simulation have some differences, the software must enable the user to generate special reports and control the amount, format, and type of output. A list of the standard statistics that may be included in an output report may consist of:
    1. Activity statistics: Average, minimum and maximum duration, standard deviation, entity count, current status
    2. Waiting statistics: Average, minimum and maximum length and wait time, standard deviation, queue current status
    3. Resource statistics: Average and maximum utilization, standard deviation, current status
- *Graphics:* The graphics should display a wide variety of graphs and charts (bar, pie, histogram) in a user-designated format. The ability to relate data across different scenarios helps during the analysis phase.
- *Animation:* Animation allows the user to represent the physical process graphically and describe and visualize the system being modeled. The animation portion should have a standard library of symbols and icons to assist the user in developing a picture of the process. Furthermore, an icon builder helps design shapes that are not included in the standard library. Animation can be helpful in the verification and validation of the model, as a selling tool to management or others not closely involved with the project as an operator-training tool, and as a technique to solicit ideas from experienced workers. Although experienced workers might never have used a computer product and might even feel intimidated if asked to do so, they could be drawn quickly into a discussion about the viability of a process change if they were shown an animated picture.
- *Output analysis assistance:* Automation of output analysis is certainly a characteristic of the new generation of simulation systems. Features that compare or relate data from replications or probable scenarios provide significant support to the user in reaching conclusions. Assistance in steady state analysis, confidence interval estimation, automatic stopping rules, sample size determination, and so on is very desirable.

3.5. Hardware Characteristics

Most organizations have hardware requirements that must be followed when selecting software. Therefore, the purchase of any software must be influenced by the hardware requirements of the organization and the flexibility and limitations of the simulation software. The compatibility of the software with the hardware already in place, file sharing capability between machines and facilities, and the capability to support the output devices already in place need to be considered.

3.6. Vendor Characteristics

The websites of the vendors are important sources of information. Many useful information about the companies and their products can be obtained from these sites. A critical examination of the company that develops the product requires significant attention. The process of acquiring a simulation tool, however, should not be viewed as a simple exchange of product for money. The company developing the product plays an important part in the successful use of the simulation product. Therefore, attention must be paid to certain features of the supplier. These features include:

- *Presence of ongoing software-development efforts:* The company should have an ongoing program of software development for both existing products and new applications. Otherwise, the user risks product obsolescence.
- *Training:* The company supplying the software must provide high-quality training programs that include both onsite and offsite classes. For large installations, training for systems personnel should also be available.
- *User support:* The company must have qualified technical personnel available during working hours to answer software questions. A functional user group with regular meetings is a good source of contacts and information. The user group can provide an avenue to influence the software-development priorities of the simulation company.
- *Company performance:* It is important that the company be established in the modeling and simulation business with an extensively field-tested software package that is in use by a significant number of corporate and academic clients. Nothing can be more frustrating than to have a supplier of simulation software go out of business and leave no one to call for support.
- *Cost/value relationship:* The price must be consistent with the job expected of the software, the features offered, and the training and support supplied with the software. In addition to

licensing, costs related to maintenance and upgrades, training, support, consultancy, and run times also need to be considered.

- *Business history:* Information such as the number of employees, the business focus (products vs. consultancy), experience, the customer base, the financial status, and references provide insight about the success and continuity of the vendor.

3.7. Evaluation Methodology

The abundance of simulation tools and the need to evaluate their numerous characteristics have turned tool selection into a major issue. Based on the May 1998 *IIE Solutions* "Last Word" reader survey on simulation (Garnett 1999), the capabilities of a simulation tool have been ranked in order of importance by 41 participants as follows:

1. Supply chain modeling
2. Ability to import CAD files
3. Customizable programming language
4. Cost modeling and analysis
5. Input distribution fitting and design of experiments

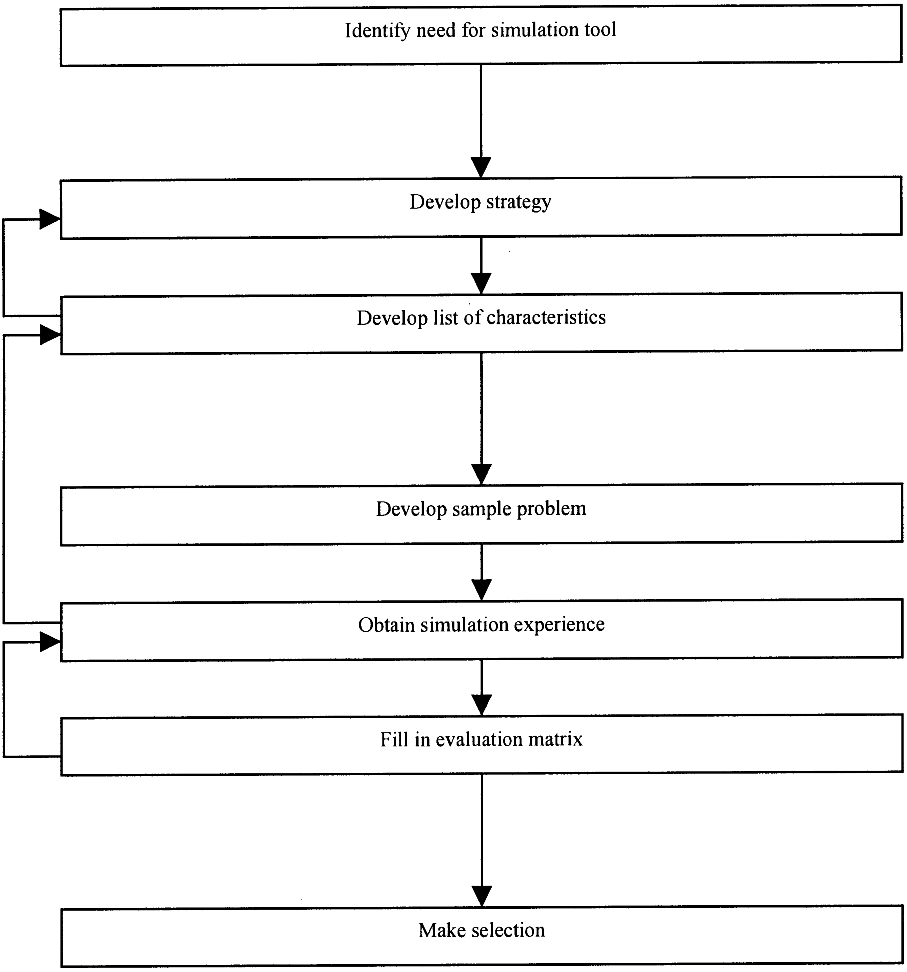

Identify need for simulation tool
Develop strategy
Develop list of characteristics
Develop sample problem
Obtain simulation experience
Fill in evaluation matrix
Make selection

Figure 1 Simulation Software-Selection Flowchart.

6. Process plan modeling
7. Output analysis systems
8. Graphical and programmable model-building capabilities

Based on the same survey, attributes of a simulation tool have been ranked in order of importance as follows:

1. Quality of graphics and animation
2. Availability of training and consulting
3. Price

TABLE 1 Software Evaluation Matrix

		Software 1		Software 2	
Importance (1)	Desired Features	Rank (2)	Score (1 × 2)	Rank (3)	Score (1 × 3)
	Language				
	Portability				
	Readability				
	Documentation				
	Data requirements				
	World views				
	Library of algorithms				
	User code interface				
	Interface with:				
	Databases				
	Spreadsheets				
	CAD software				
	Etc.				
	Compatibility				
	Operation				
	Applicability				
	Debugging features				
	Ease of use				
	Intuitive				
	Parameter revision				
	Model revision				
	Interactive				
	Check points				
	Reusability				
	Output				
	Reports				
	Graphics				
	Animation				
	Exportability				
	Hardware				
	Compatibility				
	File sharing				
	Output devices				
	Vendor				
	Ongoing development				
	Training				
	User support				
	Customer base				
	Cost/value relationship				
Total Score			Score 1		Score 2

4. Speed of execution
5. Flexibility
6. Ease of use

References cited in Section 3 also propose methodologies for evaluating simulation tools. The proposed methodology to select the appropriate software is outlined in Figure 1. Once the need for simulation software is identified, a strategy needs to be developed (Section 2).

Given the company objectives and constraints, a list of desired characteristics needs to be formulated. These characteristics are listed in Table 1. A simple yet realistic sample problem must be constructed to evaluate the software. Simulation experience must be gained either by using the software over a period of time or by participating in simulation activities with trained individuals. The software can be compared using an evaluation matrix like the one presented in Table 1. The decision maker must decide on the desired characteristics to be included in the study and the importance of these characteristics. During the experimentation with the sample problem, ranks can be assigned to the characteristics of the software subjected to evaluation. Once all the software is evaluated, the total score for each software can be calculated. The selection can be made comparing the total scores and the costs of each alternative. Since the assignment of importance factors and ranks is subjective, including more people in the evaluation process will strengthen the decision.

4. SIMULATION TOOLS

This section is prepared using the user's manuals/guides, demo software, articles, books, and promotional material obtained from simulation software vendors. The websites of the vendors were also very useful.

4.1. Languages

Once the simulation model is formulated, it needs to be translated into a computer language for execution. In the early days of simulation, general-purpose computer languages were the only alternatives. Regardless of the language used, simulation modelers have been burdened with the task of writing procedures and functions for creating and deleting entities (transactions), creating random numbers and variates, updating simulation time and system status, and recording and analyzing data for output (Schwab and Nisanci 1992). This inefficiency in the model translation stage has led to the development of powerful special purpose simulation languages. These languages are capable of supporting discrete, continuous, and combined modeling world views. Furthermore, discrete change simulation languages offer event scheduling, activity scanning, or process orientations in modeling systems. Uses of software engineering has led to reduced software development costs and improved flexibility. Significant advances have been made in the field of model translation.

The following section reviews some of the most popular simulation languages and their features.

AweSim is a descendant of SLAM II (O'Reilly and Lilegdon 1999a; Pritsker and O'Reilly 1998, 1999; Symix Systems/Pritsker Division 1999) that supports a wide range of tasks necessary to execute a simulation project. It is a general-purpose simulation system that is distributed by Symix Systems, Inc., formerly Pritsker Corporation. It takes advantage of the latest Windows technology and includes the Visual SLAM simulation language. AweSim provides the user with three frameworks (network, discrete, and continuous) that can be used independently or combined depending on the problem. This feature of AweSim enables virtually modeling any system or process. Programming is not needed for network models, yet flexibility is provided through allowing user coded inserts in C or Visual Basic. Network models can be combined with continuous and discrete-event models that can be developed using the object-oriented technology of Visual Basic, C, or Visual C++. AweSim's architecture has the openness and interconnectivity to store, retrieve, browse, and communicate with externally written software applications such as databases, spreadsheets, and word-processing programs.

A simulation project with AweSim consists of scenarios that represent alternative system configurations. Scenarios consist of component parts that are created by software programs, referred to as builders, that are provided by AweSim. Through an executive window, project, scenarios, components (network, subnetwork, control and user data, etc.), simulate, and report options are defined. Component builders are accessed through the Components menu. The project maintainer, which examines changes made to the current scenario, determines a model translation or compilation requirement. Based on the prompt from the project maintainer, the user initiates a translation task. AweSim allows multiple tasks to be performed in parallel while a simulation is executed as well as switching between tasks. Some of the features of AweSim are:

- *Building models:* A network is a basic component that graphically displays the flow of entities through the system being modeled. A network consists of nodes (processing locations) that are connected by activities that describe routing and operation time requirements. Nodes are used for functions such as removing or entering entities, seizing or freeing resources, changing variable values, etc. Entities may also be described by assigning attribute values such as weight, arrival time, etc. A network is built interactively in AweSim by selecting symbols from a graphical palette and dragging them to the desired location on a screen. Required informations about the symbols are specified by filling out a form that is supported by online error checking. Context-sensitive help and search capabilities are also provided.

- *Subnetworks:* These are reusable network objects that are provided for hierarchical models and are invoked from a calling network. This feature enables creating some form of a model depository that can be used by different modelers in different projects. Hence, it enables saving modeling time and sharing modeling expertise. Subnetwork builder works like the network builder but contains slightly different nodes.
- *Model output:* Output from various scenarios can be compared both graphically (in the form of bar charts, histograms, plots) and textually through a report browser. Multiple windows of graphical output, and side-by-side comparison of textual output are also provided. Information in the AweSim database can be exported to other Windows packages, such as Excel and Word.
- *Animation:* Multiple scenarios can be developed for a single scenario using "point and click." The animator manipulates graphical items one wants to move and the background on which they appear. Storing the symbols in standard Windows bitmap format allows them to be exchanged between programs using the Windows clipboard.
- *Interactive execution environment* (*IEE*)*:* Enables the modeler to examine, modify, save, or load the current system status. This feature facilitates debugging and verifying the model.
- *Scenario selector:* Allows screening a set of scenarios or selecting the best scenario.
- *Integration with other software:* AweSim is built on a relational database that can be accessed with tools such as Dbase, Access, FoxPro, and Excel. AweSim provides the capability to exchange data between its input and output files and these tools. Graphical elements produced by CAD, drawing, or paint programs can also be loaded into AweSim.

GPSS (General Purpose Simulation System) is a process-oriented special purpose simulation language (Crain 1998; Schriber and Brunner 1998; Banks et al. 1995, 1996; Pritsker 1995; Schriber 1991). Geoffrey Gordon developed GPSS about 1960. It is a language to model discrete systems. GPSS consists of a well-defined vocabulary (set of blocks) and a grammar. Each block carries information related to its location, operation, and operands. Operation descriptions define the tasks the blocks perform. Operands specify how these tasks need to be performed. A block may not have a location name and operands. Each block is represented by a differently shaped figure. A GPSS model consists of a selection of blocks connected with lines that describe the operation of the system model. GPSS defines dynamic and static model entities. Dynamic entities represent the units of traffic and are referred to as transactions. Static entities represent system resources. As the simulation runs, transactions enter the model and start moving over as many blocks as possible until they encounter a user-defined delay, are removed from the system, or are denied entry by the next block. The procedure summarized above describes the flow of entities in a GPSS model and the interactions between different types of model entities. Ease of learning the language and reduction of the modeling time have made GPSS one of the popular discrete-event simulation languages. Various versions of GPSS exist, such as GPSS/H (Wolverine Software Corporation) and GPSS/PC and GPSS World for Windows (both by Minuteman Software). GPSS World includes drag-and-drop modeling, a high-performance model translator, point-and-shoot debugging, an embedded programming language, built-in probability distributions, programmable experiments, interactive operation of all commands, and interactive graphical text views of running simulations.

SIMAN is the simulation language within Arena. First introduced in 1983 by Systems Modeling Corporation (Pegden et al. 1995; Kelton et al. 1998; Schriber and Brunner 1998), SIMAN provides the use with three distinct but fully integrated modeling frameworks. Each framework is divided into two sections: a system model framework and an experimental framework. The system model permits the user to describe the process, while the experimental frame permits the user to supply the parameters and characteristics describing the system operation. SIMAN provides modeling blocks for describing the system model and experimental elements for use in the experimental frame, including several blocks and elements that provide SIMAN with materials-handling modeling capabilities as part of the basic package. In the discrete-event framework, SIMAN provides a complement of user-definable and user-callable functions and subroutines to be used alone or in conjunction with the block framework. These subprograms allow the user to manipulate entity attributes and files easily and provide for complex decision rules. In the continuous framework, system state equations can be integrated over time while simultaneously integrating with the blocks and discrete-event frameworks.

CACI Products Company provides SIMSCRIPT II.5. Developed by Kiviat, Villanueva, and Markowitz, it is an event-oriented discrete-event simulation language (Kiviat et al. 1969; Russell 1983). SIMSCRIPT II.5 consists of features for developing time-driven simulation models. It includes both event- and process-oriented simulation constructs, multiple random-number generators, range of random distribution functions, automatic statistics collection, an integrated continuous simulation capability, and built-in graphics support for model input and output. Its world view facilitates the mapping of real-world modeling requirements directly onto its language features. Its general-purpose programming constructs, such as arithmetic operators, text manipulation, arrays, and subroutines, and the other features are accessed using English-like programming syntax that supports the self-documenting nature of the code. It provides extensive error checking through entity referencing, array bounds, and memory use. It contains a rich library of simulation objects that is designed to fulfill common simulation requirements, activity scheduling, and constrained resource modeling. Compilation and linking of all parts of a program are done automatically. Its graphic editor helps in creating and editing graphic images and charts and user interface objects such as menus and dialog boxes. Images and graphics are importable, and graphics library elements are fully portable. The debugger feature provides detection of programming and simulation errors, and a detailed trace reporting. SIMSCRIPT II.5 also provides integrated graphics and animation with scaling, rotating, and positioning capabilities. Using its graphics editor, presentation graphics such as pie charts, level meters, and bar graphs can be produced. The supported platforms consist of Windows 95/NT, SPARC, HP9000/700, SGI, DECAlpha, and IBM AIX.

4.2. Packages

Early offerings were no-programming products. Many were not versatile enough to handle complex decision-making rules. The biggest growth of simulation packages has been in the manufacturing area with products for PCs. Figure 2 lists the major components and functions of simulation packages. Some of the packages and their features are described below.

Arena (Sadowski et al. 1998; Kelton et al. 1998; Snowdon et al. 1998) is a graphical modeling and animation system. It is a product of Systems Modeling Corporation. Arena consists of a complementary family of products (Arena Product Suite) to meet the various needs for simulation in an enterprise through a common software interface and compatible features. These products share a common software foundation. The Business Edition (Arena BE) targets business and other systems. Arena Standard Edition (SE) targets detailed modeling of discrete and continuous systems and provides more flexibility. Arena Professional Edition enhances Arena SE with the capability to build custom simulation objects. Call$im targets call-center analysis, and HiSpeed$im targets high-speed production line modeling. Arena provides a hierarchical structure consisting of alternative and interchangeable templates of graphical simulation modeling and analysis modules. The user combines these modules in building a simulation model. Modules are grouped into a panel to compose a template. Different modeling constructs and capabilities are obtained by switching templates and bringing modules from different panels and templates. For example, modules from the SIMAN template can be pulled and SIMAN constructs can be mixed with the higher-level modules from another template. Complex decision rules can be coded using Visual Basic, FORTRAN, or C/C++. Each level in the hierarchical structure is managed with a single graphical user interface. Arena's application solution templates (collection of modules for an area, e.g., Call$im) provide applications for modeling specific areas. Arena is compatible with Microsoft Windows and Microsoft Office. It also enables integration with other technologies and applications such as databases, drawing/modeling (AutoCAD) products, and spread sheets such as Visio, Oracle, and Microsoft Office. Arena provides user interface features such as drag-and-drop, context-sensitive click menus, and customizable tool bar. It also provides a project bar for accessing modeling constructs and navigating through model hierarchy. Using Arena (Business Edition), process dynamics is represented in a hierarchical flowchart with system information stored in data spreadsheets. It also has a built-in activity-based costing and system performance data that enables analyzing business, manufacturing, service, and other systems. Arena also provides input and output analyzer for entering, analyzing and reviewing data. OptQuest for Arena is also provided for simulation optimization.

AutoMod version 8.7 (AutoSimulations 1999; Banks 1998a; Phillips 1998a, b) is an industrial-oriented simulation system provided by AutoSimulations Inc. It uses CAD-like graphics to define the physical components of a system. A procedural language is used to define the logical elements. This language is based on action statements that combine the power of a structured language with the ease of English-like syntax. If needed, the user may define C functions. The communications module allows communications between two or more models and third-party applications such as spreadsheets or Visual Basic programs. All graphics are represented in 3D space, and complex motion (kinematics and velocity) of equipment can be simulated. The kinematics module allows simulating the robots and other equipment containing moving parts and integrating them into a simulation model. AutoMod's CAD features are used to define layout and material-handling and storage systems by maintaining distance, space, and size characteristics in 3D. All graphics created can be viewed with unlimited control with respect to translation, rotation, scale, perspective, and so on. The animation runs in real time with the simulation. The model execution is interactive, which enables the user to

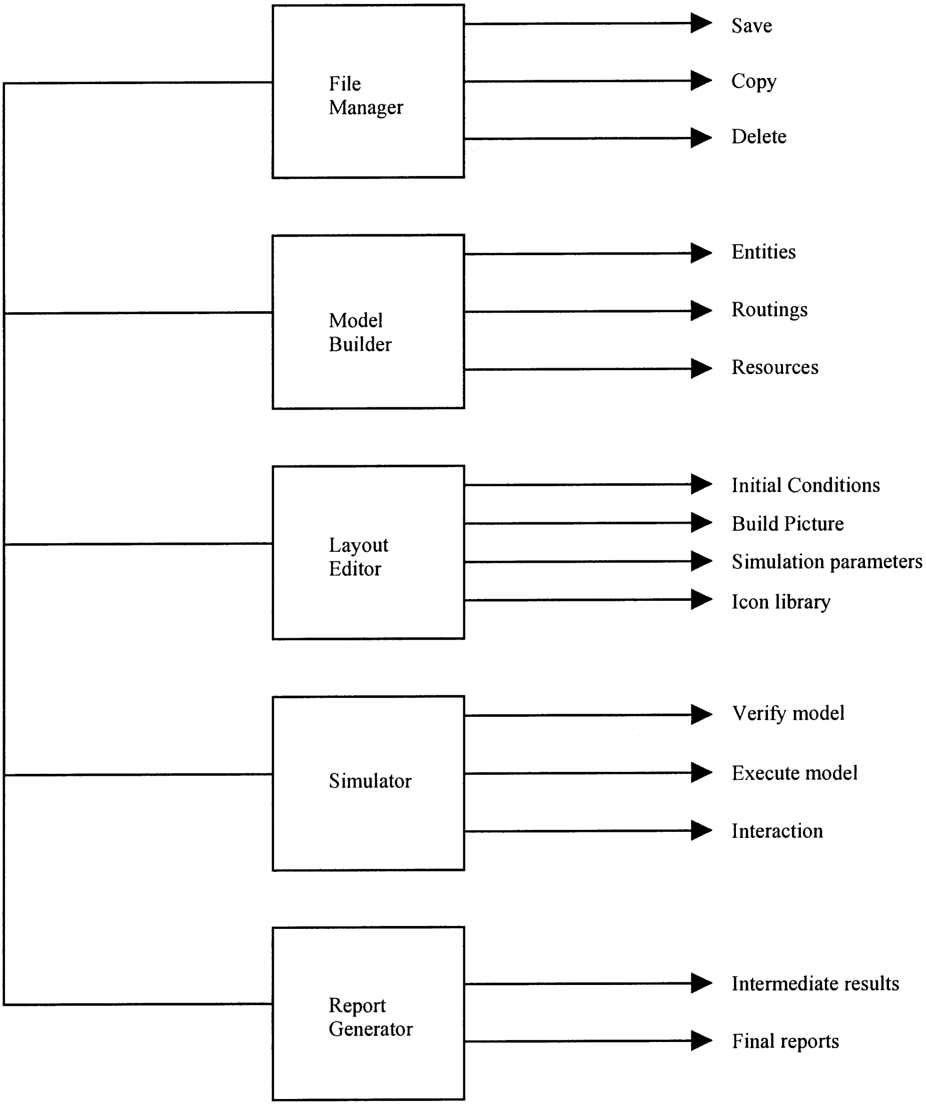

File Manager
Save
Copy
Delete
Model Builder
Entities
Routings
Resources
Layout Editor
Initial Conditions
Build Picture
Simulation parameters
Icon library
Simulator
Verify model
Execute model
Interaction
Report Generator
Intermediate results
Final reports

Figure 2 Major Components and Functions of Simulation Packages.

stop and resume the simulation, cancel animation, and select objects for detailed statistics. AutoMod has a material-handling orientation. Entities moving around the system are called loads, and loads are worked on by processes. The software features a well-developed array of predefined material-handling systems, including conveyors, automatic guided vehicle systems (AGVs), automated storage and retrieval systems (ASRS), bridge cranes, and power and free conveyors.

AutoMod also has support tools: AutoStat, Simulator, and AutoView. AutoStat provides enhanced analysis of output such as calculating confidence intervals, design of experiments capability, optimization, and steady-state analysis. The user defines the desired analysis through a point-and-click Windows interface. AutoView, a postprocess package, allows the creation of a directed walk-through of the output. It allows panning, zooming, and moving back and forth in time and space. The animation view can be attached to a moving model entity to get an inside view of the system. Simulator, which is shared with AutoSched AP (ASAP), provides a template for capacity planning. ASAP is a simulation-based, finite capacity planning and scheduling tool. It is used to increase throughput, reduce in-process inventory, and increase resource utilization. ASAP takes into consideration constraints such as shift schedules, work setup rules, batching, preventive maintenance, machine efficiency, and operator skill classes. Its built-in task-selection rules enable the modeling of selection of machines and personnel to enhance performance. Its output consists of reports, Gantt charts, and utilization graphs. In addition to scheduling, ASAP is also used as a planing tool for capacity analysis.

FACTOR/AIM (AIM) (Pritsker 1995; Symix Systems/Pritsker Division 1997; O'Reilly and Lilegdon 1999b) is a special-purpose simulation system for use in manufacturing decision support in areas such as engineering design, scheduling, and planning. A product of Symix Systems, Pritsker Division, it is used to build, simulate, and animate models and analyze graphically and statistically the simulation results obtained from various scenarios. AIM uses the language of manufacturing, which eliminates the need to learn a language-based syntax and abstract a system to fit syntax. AIM components, which have both operational and graphical characteristics, consist of machines, operators, materials, parts (loads), batches (group of loads), process plans or routings, conveyors, and so on. It also has nongraphical components such as work calendar, orders, shifts, and breakdowns. These components, coupled with a comprehensive set of predefined manufacturing rules, provide a powerful framework for modeling and simulation analysis of manufacturing systems. The modeler has the option of coding customized rules in C language if the built-in rules do not represent the desired logic.

The current version of AIM (8.1) supports user code with the latest version of Visual C++. AIM uses the Windows platform, where the model and the data are stored in a Microsoft Access database. This feature enables importing and exporting data and preparing customized reports. Access application wizards can be used in developing decision support applications with dialog boxes and menus. AIM consists of five modules: Executive, which triggers other modules, controls data collection, and supports database and scenario management; Simulator, which builds and runs simulation with animation; Batch Simulator, which supports experimentation through replications and multiple runs; Reporter, which provides textual reports; and Gantts, which provides detailed schedule information. A model is created through interactive point-and-click operations to select and locate graphical components that can be further refined using dialog boxes. Next, parts and their routings are defined. A model can also be prepared using automatic loading if the existing information is in electronic form. Animation of an AIM model is created automatically. An interactive run process provides support for verification and validation through automatic animation, reports and graphs summarizing system status, detailed information on each model component, and the ability to view all scheduled events. Multiple replicate and scenario runs can be made through the Batch Simulator module.

AIM provides an extensive set of standard output reports and graphs in addition to customized reports and graphs that can be created using Access report wizards. Furthermore, output can be moved to Excel Spreadsheet for additional processing. Schedule reports can be created and communicated using the GANTT chart module. AIM also has cost-modeling capability. Product costs in four groups, operating, overhead, inventory, and lateness, can be defined using fill-in-the-blank forms. For example, operating costs resulting from part production and flow, such as machining, fixture, operator, and consumables, can be tracked. Using the cost-modeling feature of AIM, cost-simulation and output reports can be generated. Cost reports such as average operating expense per part, value of inventory, and total expense incurred can be prepared. Cost reporting can be customized to generate cash flow reports and pro forma income statements. The user also has the option of selecting cost reports for the entire system, by part or resource type.

Mast Simulation Environment is a product of CMS Research Inc. (1999). It is an integrated software tool for addressing manufacturing system design and evaluation problems. It provides a single environment to aid decisions on design, acquisition, and operation of manufacturing systems. MAST offers a pull-down menu toolbar to access commonly used features, can launch multiple results, and has a cost modeling option. It utilizes CAD graphics for layout and animation and allows importing CAD layout as background. It has features such as exporting and importing data to and from spreadsheets and databases. It provides station worktable and animation libraries with zoom, scroll, scale, and speed adjustment options. Launching multiple results at one time is also an option. Mast is available for Windows 95/98/NT.

Micro Saint version 3.0 (Laughery 1999; Plott et al. 1999; Ernst et al. 1999; and Barnes and Laughery 1998) is a PC-based discrete-event simulation software program available from Micro Analysis & Design. With Micro Saint, any process can be modeled as a network of tasks that can be represented in a flow diagram. Sets of decision rules and algorithms consisting of algebraic and logical expressions specific to a process can be developed using Micro Saint's own C-like language. The user defines the terms that are used. A graphical user interface is used for model development and testing. The icons and background for the Action View animation and the flowcharting symbols are customizable. Drawings from other graphics programs can be imported through the clipboard. Micro Saint symbolically animates the network diagram as it executes the model, using the routing choices and random number seed to generate task times that are supplied by the user. The user also selects statistics charts or graphical representations to analyze the data collected during the simulation run. Queue, resource, and task statistics are collected automatically as well as the snapshots of user-defined variables at user-defined triggers. Micro Saint also has a resource wizard that can be used to create new resources quickly and allocate them to tasks in the process. The syntax checker searches the model for automatic detection and display of errors. OptQuest, included with Micro Saint, automatically searches for and finds optimal or near-optimal solutions to simulation problems and fine-tunes simulation parameters. Micro Saint has built-in features using Microsoft COM services that allow it to communicate dynamically (i.e., as the model runs) with other applications such as Visual Basic, C++ applications, and even other simulations. Micro Saint is available for Windows 95, 98, and NT, and a Unix/Linux version variant is available under the title of Integrated Performance Modeling environment.

MODSIM III is a product of CACI Products Company (Tumay and Wood 1999; Goble and Wood 1998; and Lewellen and Tumay 1998). MODSIM III is an object-oriented discrete-event simulation language that provides a complete development environment (built on top of Microsoft's Visual Studio environment) for developing models that are visual, interactive, and hierarchical. Its features include run-time libraries, graphical user interface and results presentation tools, database access, and hooks to HLA. Its components, such as Simulation Layer, Graphics Editor, Compilation Manager, and Interactive Debugger, support the development of advanced simulation models. Seamless interoperability with other tools is achieved through its ability to generate C++, which compiles using the platform's native compiler. Objects are defined in two separate blocks of code. The definition block defines the object type through its variables (everything the object knows) and methods (everything the object can do) which is used for interface specification. The implementation block defines the logic of what the objects do (their behaviors) and their affect on their state variables. MODSIM III sequences the execution of the methods of object instances to ensure correct sequences for the events. Its features to manage the scheduling, interaction, and synchronization of behaviors provide increased readability and consistency. Its features for modeling concurrent and interacting behaviors qualify it as a development tool.

MODSIM III provides a library of tested objects that can be customized. For example, objects may compete and queue for resources based on a priority while accumulating statistics on resource utilization, waiting time, and so on. Inheritance, which is a benefit of object-oriented software construction, provides object libraries that can be customized. New objects evolve through inheritance with additional capabilities resulting in enhanced software-development productivity. Inheritance also enriches the graphical properties of objects and facilitates the creation of interactive and graphically managed models. The graphical editor supports creation and configuration of icons, menus, dialog boxes, and presentation charts. Through interactive graphical editing, scenarios are defined on screen and animation and plots can be viewed while the simulation is running. The graphics environment allows easy access to animation, dynamic presentation graphics, and user interface toolkits using SIMDRAW graphics editor. Its DatabaseMod module enables access to ODBC-compliant databases and retrieving and committing data to databases. HLA is supported through run-time interface (RTI), which can be called from languages such as C++, Java, and Ada. RTI supports both local intranets and the global Internet. MODSIM III with SimGraphics is available on Windows 95/98/NT and Unix platforms. This enables load-balancing a simulation model through a network or displaying the user interface and animated graphical output on different and remote computers.

ProModel is a product of ProModel Corporation (Heflin and Harrell 1998; ProModel Corporation 1999). It is a menu-driven, mouse-activated discrete-event simulation and animation tool. The model development is graphical and object oriented. Modeling is done by selecting from a set of modeling elements such as resources, downtimes, and shifts and modifying the appropriate elements. All input is provided graphically, with information being grouped by object type and presented in a format similar to a spreadsheet. User coding in C, Pascal, or Basic can be linked to the models and called during the simulation. Its top-level menu allows the user to select model editor, run simulation, view output, or quit. The model editor permits full model definition, routing, scheduling, resource description, and transportation. All standard statistical distributions are available, and StatFit is provided to aid the user in selecting the proper distribution. The model editor also allows full control over the graphical presentation. A standard set of icons is provided and presented for selection by number. An icon editor is also provided to create special symbols. ProModel provides an automated model builder that prompts the user for all information required for modeling the process. It is organized logically and guides the user from creation of the model name to part and routing definition and finally to layout definition for the animation.

A pop-up menu is provided to prompt the user in defining a statement or expression. Its graphics editor allows scaling, rotating, copying, and so on, and drawings from other packages can be imported. Animation development is integrated with the model definition. Run simulation permits viewing of the animation while the simulation is running. The user retains full control over the animation. The simulation may be paused at any time to allow intermediate statistics to be viewed. ProModel is interactive during the run time, enabling model parameters to be changed and thus facilitating the "what-if" analysis. Basic statistics such as utilization and throughput are available, as well as cost statistics. It provides features to select reports and provides tabular and graphical reports on all system performance measures. Output reports from different simulation runs can be compared on the same graph. ProModel also provides model-merging capabilities. It can exchange data with external files and application programs such as Excel. SimRunner optimization capability is also provided.

QUEST is a product of Deneb Robotics, Inc. (Donald 1998). It is a discrete-event simulation package used to simulate discrete flow processes. QUEST uses 3D geometry and combines a graphical user interface with material-flow logic grouped in modules consisting of labor, conveyors, AGVs, kinematics, power and free conveyors, and ASRSs. For unique flow logic problems, QUEST's Simulation Control Language (SCL) can be used. It also provides a value-added costing module for activity-based costing analysis. Results can be viewed with graphical and numerical analysis capabilities. It consists of features to create a 3D virtual factory environment. Its other features consist of the ability to import and export 2D or 3D CAD geometry and data (e.g., spreadsheets, MRP systems), a graphical user interface and a programming language, and the ability to provide integrated solutions through interaction with other Deneb software. Quest also provides a virtual collaborative engineering (VCE) environment that enables modelers in remote locations to interact with the same model. QUEST integrates virtual reality (VR) devices such as head-mounted displays, stereo glasses, and cyber gloves to immerse the modelers within the factory simulation.

SDI Industry is a product of Simulation Dynamics (Siprelle et al. 1998). It is a discrete-event simulation package designed specifically for the process-oriented industries, such as chemicals, foods, consumer goods, and pharmaceuticals. SDI industry uses discrete rate technology to simulate processes in these industries. It adds an embedded database and specially designed functions to Extend for simulating high-speed and high-volume processes. Extend is the underlying simulation engine, developed by Imagine That Inc. SDI industry is a graphical interactive program that provides a hierarchical framework of model templates, flow blocks to model material flows, control blocks to manage flow, and an integrated database to store management information. A model is developed through four hierarchical levels, consisting of equipment (process), operations, systems, and plant. Each level is provided with a template and with a process management block to control the process at that level and interface with the other levels. A relational database is used to store and manage data centrally. Data requirements for the control blocks at each level are provided by standardized tables such as "materials and brands," and "shift schedule." A database viewer is provided to view and edit table data in Extend. Simulation Dynamics provides two more products: SDI Industry Pro provides scheduling and control tools, and SDI Supply Chain is used to model the dynamics of a supply chain from source to user.

SimEngine is a product of Industrial Modeling Corporation (1999). It is a 3D simulation tool that incorporates object-oriented programming. SimEngine builds on Delphi and C++ object-oriented languages, which feature inheritance, polymorphism, and encapsulation. Every SimEngine entity is an object with exposed properties, methods and events. Models are built by manipulating the objects in visual environment and in code. SimEngine also works with C++ Builder, which incorporates an ANSI standard version of C++. Delphi and C++ Builder share a development environment that is similar to Visual Basic. The environment features visual design tools, an object browser, code window, and debugging tools. SimEngine builds on these development tools by adding simulation objects such as orders, routings, resources, and vehicles. The basic model building is done via visually manipulating these objects in 3D simulation explorer. Objects are created and moved using drag-and-drop techniques, and their properties are set in the object inspector. Code writing becomes necessary only when implementing complex routing logic. Database components enable reading information directly from company databases. It has native links available to SQL databases. Its ActiveX feature enables posting of the models to a website.

SIMUL8 is a product of Visual Thinking International Inc. (1999). It is a discrete-event simulation package to simulate a business or a manufacturing process. SIMUL8 provides objects such as work centers, storage bins, conveyors, entry/exit points, resources, variables (including spreadsheets), labels (attributes), and work items. Objects are linked automatically. Models are built on the screen via point-and-click operations and by drawing them with the mouse. The drawings are detailed by using intuitively organized options. After dragging and dropping the parts of the process onto the screen, clicking on "run" shows the full animation of the products flowing through the system. All animation is automatic with full zoom and pan options and distance on the screen relates to distance and time in the model. By clicking on resources, entities, and so on, their parameters (speeds, setups, constraints, etc.) can be modified and status updates through counts and reports can be obtained. These capabilities are extended using Excel or VBA. SIMUL8 can also import Visio flowcharts to begin the simulation process. SIMUL8 also conforms to the SDX (simulation definition exchange) file standard, which allows interchange of structure and detailed data between simulation and other packages (e.g., CAD) that contain information related to simulations.

Taylor ED (Enterprise Dynamics) is the product of F&H Simulations (Hullinger 1999). It replaces the Taylor II (Nordgren 1998) software, which is being phased out. Taylor ED is an object-oriented software system for modeling, simulating, visualizing, and monitoring dynamic (discrete) flow process activities and systems. It uses an atom concept. An atom is defined as a four-dimensional object that has a location and speed (*x*, *y*, *z*) and a dynamic behavior (time). Atoms can inherit behavior, contain other atoms, and be created, destroyed, and moved onto one another. Atoms also communicate with each other. Within Taylor ED, everything, such as resources, products, a model, and a record, is referred to as an atom. Through its open architecture, it allows users to access standard libraries of atoms (logistics suite) to build models or create their own atoms. 4DScript language (atom editor) is used to create the atoms and to control atom's behavior, the user interface, and model visualization. A model is built by selecting, dragging, and dropping atoms from the library onto the screen. This is followed by connecting the atoms and defining routings of entities/atoms. Routing is defined by linking the output channels of an atom to the input channels of another one. Complex routing rules can be assigned using a pull-down menu. Finally, each atom is assigned a logic through defining atom's parameters, that is, its behavior and functionality. Clicking on the atom opens editing windows. Taylor ED includes experiment and optimizer (OptQuest) modules through which experimental and run conditions and performance measures can be defined. Results consisting of animation, predefined, and user-defined reports, and graphs can be viewed in 2D or 3D concurrently with simulation. Results can also be exported to external software programs. Taylor ED also monitors real-time systems by linking external programs to read and write information. Run time and real time can be synchronized, enabling the model to monitor the system in real time.

Call Center MAESTRO (Model Builders) is a decision support tool that is composed of a discrete-event simulation engine and is used to effectively staff and configure a call center. COMNET III (CACI) is a network simulation tool to predict LAN, WAN, and enterprise network performance. Its add-on modules consist of application profiler to predict the impact of a new client application, satellite and mobile module, and circuit-switched traffic module. SIMPROCESS (CACI) is a hierarchical and integrated simulation tool to enhance productivity for business process modeling and analysis. Packages such as ASSEMBLY, ERGO, and TELEGRIP are also offered by Deneb. These products are directed towards more specific applications and can be integrated with QUEST. Visual Simulation Environment (VSE) is a product of ORCA Computer, Inc. (Balci 1996). VSE is a discrete-event, general purpose, object-oriented, picture- and component-based visual simulation model development and execution product that can be used in solving complex problems such as air traffic control and communications networks.

4.3. Applications

Over 170 simulation products listed in the Directory of Simulation Software (Rodrigues 1994) published by the Society for Computer Simulation cover a wide range of application areas, such as aerospace, AI/Expert systems, business, communications, education, manufacturing, military, and networks. The annual Buyer's Guide of simulation software (Institute of Industrial Engineers 1999) lists 27 companies providing 40 simulation software products. In this list, the application areas are classified as general purpose, manufacturing, construction, service, health care, material handling, warehousing and distribution, and enterprise and business planning. Banks (1998a) classifies and gives examples of 27 simulation software products as general-purpose, manufacturing-oriented, and business process reengineering products. Simulation has also been used to study areas such as urban systems, economic systems, biological systems, environmental and ecological studies, computer and communication systems, project planning and control, and financial planning (Pritsker and O'Reilly 1999). Within each of these areas, a multitude of topics has been subjected to simulation analysis. For example, under transportation systems, topics include railroad system performance, truck scheduling and routing, air traffic control, terminal and depot operations, and issues related to supply chain management. The accelerated rate of change in simulation technology and the increasing complexity of problems associated with modern systems will undoubtedly widen the areas and the topics necessitating simulation studies (Nisanci 1997, 1998, 1999).

5. PITFALLS INHERENT IN THE USE OF SIMULATION TOOLS

Although there are many dangers inherent in the use of simulation products, only the major ones are addressed here (see Musselman 1998; Norman and Banks 1998; Rohrer and Banks 1998; Banks and Gibson 1996, 1998; Harrel and Tumay 1994; Pritsker Corporation 1993). Careful use of simulation will always be rewarded, while a disregard for the dangers may have disastrous consequences.

5.1. General Pitfalls

The general pitfalls consist of:

5.1.1. Having the Tool and Not Using It

Strange as this may seem, this is a real danger. Sooner or later, most project engineers will feel some pressure to reduce the time or expenditures required to complete a project. Simulation analysis is an easy target for elimination because the impact of its elimination is not noticed immediately. Cutting manpower or reducing the project scope has immediate and noticeable impact on the project. The project can move ahead, detailed design can be done, and equipment can be purchased and installed, all without any felt loss of simulation analysis. Usually the first time the lack of simulation analysis is experienced is during start-up when design flaws and oversights typically caught by simulation are revealed. When an inadequate design is revealed, the solution is usually expensive in redesign time and equipment cost. These costs, however significant, may still be dwarfed by the loss in revenue due to the failure to deliver the goods or services on time.

Complex systems often defy the intuition and the traditional simplified analysis (including rules of thumb) that have been faithfully passed down through the engineering generations. Complex systems with complex interactions demand detailed analysis using simulation tools. The project engineer assumes a large and unnecessary risk in not using available simulation tools when the project calls for them.

5.1.2. Failure to Establish and Adhere to Project Objectives

A clear and specific set of objectives must be established at the beginning of every simulation project. These objectives should address the questions to be answered by the project engineer using the simulation tool. Issues to consider include:

1. Configuration, quantity, utilization of nonconsumable resources
2. Quantity of consumable resources required
3. Total time to provide goods or services to the customer
4. Critical processes and bottlenecks
5. Process capability (throughput)

Once management personnel have agreed upon the objectives, they need to direct the project to its conclusion. Every new idea or suggested change to the project needs to be tested against the approved objectives. Value-adding changes should be incorporated into the objectives and reapproved. Suggested changes that add nothing to the project or do not support the objectives should be rejected. Periodically during the course of the project, the objectives should be reviewed to make certain that the project would answer all questions raised at the beginning. Failure to monitor the project in relation to the objectives may allow project personnel to pursue information and solutions that do not contribute to the objectives. This results in delays of the project's completion.

5.1.3. No Review or Inadequate Review

One of the greatest mistakes made by project engineers who have committed time and money to perform a simulation is neglecting to take the additional time to properly review the model. A review should take place after the completion of data collection and before the beginning of model building. A review of the anticipated modeling assumptions should be done in addition to a discussion of any problems with data collection. A process flowchart, providing a thorough description of the process being modeled and the method of modeling, should be developed and presented to all involved. As the project develops, additional interim progress reviews should be held as often as needed.

The final review should take place after completion of all the experimentation but before the submission of the project report. The final review process should include at least the four following key elements:

1. Thorough understanding of the assumptions. Particularly important would be a review of the elements of the process which were not modeled and why they were excluded from the analysis.
2. Review of all parameter values. Care should be taken to ensure that data used in the analysis is correct and up to date.
3. Line-by-line review of all code, including model input values. This is essential, especially if others do the modeling, e.g., simulation department or consultant.
4. Review of all conclusions and the supporting data. Careful review of the data that lead to the conclusions ensures that they are properly supported. The results and conclusions should also make reasonable sense.

At this point, simulation plays a key role in the development of the design parameters. Therefore, a thorough review is necessary to make certain that the model and the values used as a basis for the design are indeed correct. The engineer who depends on others to do simulation takes an enormous risk if the final product is not thoroughly reviewed. To assume that correct work is done by an internal simulation expert or a consultant is risky. Conceivably, one piece of information improperly translated, one parameter off by a factor of two, or one part of the process not modeled as it will operate can invalidate the results of the simulation. A thorough review of the model and the results is critical to the overall success of the project.

5.1.4. Not Having the Proper View of the Tool

The proper view of the simulation tool leads to proper expectations regarding what the tool can do and what results should be obtained.

5.1.5. Not Viewing Simulation as a Statistical Tool

Simulation is a statistical tool and must be used as such if reliable results are to be obtained. Its basic task is to collect data about how a system changes with time.

5.1.6. Thinking of Simulation as an Optimization Tool

Because simulation is a data-collection tool only, it should not be thought of as an optimization tool. Parameters are put into the tool and data are generated about the system (e.g., throughput, resource utilizations, congestion) and the components in their relationship as it was entered. Enhancement of any design only takes place as the engineer reviews the output from a simulation run and adjustments are made to the model and the model parameters. Improvement is therefore an external, manual process. However, incorporation of automated scenario-generation capabilities and optimization packages into simulation tools significantly enhances this process.

5.2. Pitfalls Related to Methodology

5.2.1. Failure to Establish Appropriate Model Boundaries and Parameters

Model boundaries become important in modeling because equipment and items not modeled are considered to be unlimited. Unlimited resources are not a problem if the cost of the resource is relatively low. However, resources having a significant impact on the process (e.g., fixtures that cost $50,000 and computer hardware) should all be modeled. Modeling can also be used to describe restrictions on the flow through the model. Not describing a restriction would mean that there is no restriction and therefore that flow is unlimited.

5.2.2. Improper Data Collection

In the simulation process, data collection may be the biggest problem faced by the project team. Critical data may not be available. If data are available, they may not be in a useful form.

5.2.3. Not Verifying Data

It is important to know the source of the data and how good they are. Downtime data carefully gathered over several years of maintenance are certainly more significant than a wild guess by a maintenance foreman.

5.2.4. Not Collecting the Proper Data

Because data collection is often a difficult task, data gatherers are often tempted to shortcut the process, especially if they do not understand the statistical implications. It is possible to believe that an average time for a transaction might be just as good (and of course easier to obtain) than the raw data reduced to a distribution. The use of average time leads to a deterministic solution and can invalidate the simulation study.

5.2.5. Avoidance of the Proper Level of Analysis

Because simulation is a statistical tool that collects statistical data about a system or process, statistical techniques must be used to examine and analyze the data. As an example, since each run of the simulation produces one sample point and since this process involves sampling for a distribution, it is impossible to determine the relationship of a single point to the true mean. Multiple runs are required to generate enough points to determine a calculated mean, standard deviation, and confidence intervals. Then and only then should the design proceed. Knowledge and use of statistical techniques with simulation is essential to produce reliable results.

5.2.6. Validation Problems

A successful simulation study must produce a solution credible enough to be used by the decision makers (Balci 1998). Validation may take more time than building the model.

6. CURRENT STATUS AND TRENDS

There have been continuous efforts to use computers in all stages of simulation besides model execution. These efforts have been significantly enhanced by the development in various technologies such as computer technology, animation, expert systems and artificial intelligence, optimization, and experimentation. Considering the stages of simulation, significant advances have been made in designing the simulation tools that have the following features:

- Assisting the modeler in problem and system definition. Intelligent user interfaces guide the user in producing a quicker and more complete definition of the problem. Complex manufacturing systems are defined quickly with the selection and placement of icons and computer-guided interrogation.
- Model building features in the form of interactive flow charting (block or network diagrams) with a one-to-one correspondence to the simulation program. These features also include a knowledge base consisting of a generic model that can be tailored to a specific model or submodels that can be combined to form the required model using object-oriented programming. Model building features tailored to specific industries allow access to simulation by nonsimulation personnel.
- Evaluation of data requirements based on problem and system definition and objectives. Automatic data-acquisition capability to prepare the data directly from an external company database.
- Model translation feature to provide automatic computer code generation using either program generators or data-driven simulators and thus eliminating the need to teach users a simulation language.
- Model verification features consisting of expert diagnosis programs for debugging models. Output and system animation features to evaluate the logical consistency of a model.
- Features to assist strategic and tactical planning. Automatic detection of steady-state and rundown periods, testing for autocorrelation, and checking for sample size sufficiency.
- Experimentation and automatic search features to find the best combination of resource configurations and levels and operating strategies.
- Automatic documentation of experimental conditions in graphical or tabular form in addition to standard simulation output.
- Highly developed graphical user interfaces that integrate the entire modeling process.

Successful new-generation simulation systems will be expected not only to automate the stages of simulation but to integrate them as well. The following developments are expected to impact future simulation software:

- The significant increase in research, development and application of object-oriented simulation. This technology is becoming the choice for modeling large, complex, and distributed systems. Modeling and simulation using an object-oriented language will increase due to its advantages in compilation, execution, and extensibility (Roberts and Dessouky 1998; Keller and Tchernev 1997).
- With high level architecture (HLA), an individual simulation or set of simulations that was developed for one purpose can be applied to another application. The HLA will provide a structure that will support reuse capabilities available in different simulations that will reduce the cost and time required. Through HLA, simulations can exchange data dynamically and interact with each other as they run. HLA will facilitate interoperability, software interface, and data exchanges. Simulation software will incorporate these developments, as is the case with MODSIM III and Micro Saint (Yu et al. 1998).
- Artificial intelligence and simulation have mutually beneficial connections. Model-based expert systems, simulations built out of neural networks, and smart interfaces to simulation tools are commonplace. It is expected that AI and simulation will join forces to model and simulate intelligent behavior (Wildberger 1999).
- Advances in parallel and distributed simulation are increasingly being applied to complex systems that require significant run times (George et al. 1999; Low et al. 1999; Fujimoto 1998).
- Web-based simulation to exploit Web technology is expected to grow. The number of languages such as Simjava (a discrete-event simulation library for Java) is expected to increase (Triadi and Barta 1999; Narayanan et al. 1997).
- Online simulation efforts for real-time management and control of complex systems will also grow (Gonzales and Davis 1997).
- More specialized products for specific environments will be developed, e.g., TRANSIM by Los Alamos Research Laboratories to simulate transportation and urban traffic.

- Links between virtual reality and simulation will grow stronger.
- Developments in Internet applications, enterprise resource planning applications, and embedded simulations will also impact the development of simulation tools (Banks 1998b).

REFERENCES

Autosimulations Inc. (1999), AutoMod Version 9.0 Student Manual, Bountiful, UT.

Balci, O. (1996), *Visual Simulation Environment User's Guide,* Orca Computer, Inc., Blacksburg, VA.

Balci, O. (1998), "Verification, Validation, and Testing," in *Handbook of Simulation,* J. Banks, Ed., John Wiley & Sons, New York, pp. 335–393.

Banks, J. (1998a), "Software for Simulation," in *Handbook of Simulation,* J. Banks, Ed., John Wiley & Sons, New York, pp. 813–835.

Banks, J. (1998b), "The Future of Simulation Software: A Panel Discussion," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 1681–1687.

Banks, J., and Gibson, R. R. (1996), "Getting Started in Simulation Modeling," *IIE Solutions,* Vol. 28, No. 11, pp. 34–39.

Banks, J., and Gibson, R. R. (1998), "Simulation Evolution," *IIE Solutions,* Vol. 30, No. 11, pp. 26–29.

Banks, J., Carson, J. S., and Sy, J. N. (1995), *Getting Started with GPSS/H,* 2nd Ed., Wolverine Software Corporation, Annandale, VA.

Banks, J., Carson, J. S., and Nelson, B. L. (1996), *Discrete Event Simulation,* 2nd Ed., Prentice-Hall, Inc., Englewood Cliffs, NJ.

Barnes, C. D., and Laughery, K. R. (1998), "Advanced Uses for Micro Saint Simulation Software," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 271–274.

CMS Research Inc. (1999), *MAST Simulation Environment Demonstration Manual,* Oshkosh, WI.

Crain, R. C. (1998), "Simulation with GPSS/H," in *Handbook of Simulation,* J. Banks, Ed., John Wiley & Sons, New York, pp. 235–240.

Davis, L., and Williams, G. (1994), "Selecting a Manufacturing Simulation System," *Integrated Manufacturing Systems,* Vol. 5, No. 1, pp. 23–32.

Donald, D. L. (1998), "A Tutorial on Ergonomic and Process Modeling Using QUEST and IGRIP," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 297–302.

Ernst, A. T., Krishnamoorthy, M., Nott, H., and Sier, D. (1999), "Micro Saint," *OR/MS Today,* April, pp. 58–62.

Fujimoto, R. M. (1998), "Parallel and Distributed Simulation," in *Handbook of Simulation,* J. Banks, Ed., John Wiley & Sons, New York, pp. 429–464.

Garnett, J. (1999), "The Last Word on Simulation," *IIE Solutions,* Vol. 31, No. 1, pp. 45–47.

George, A. D., Fogarty, R. B., Markwell J. S., and Miars M. D. (1999), "An Integrated Simulation Environment for Parallel and Distributed System Prototyping," *Simulation,* Vol. 72, No. 5, pp. 283–294.

Glover, F., and Kelly, J. P. (1999), "Combining Simulation and Optimization for Improved Business Decisions," Systems Modeling Corporation, Sewickley, PA.

Goble, J., and Wood, B. (1998), "MODSIM III: A Tutorial with Advances in Database and HLA Support," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 199–204.

Gonzales, F. G., and Davis, W. J. (1997), "A Simulation Based Controller for Distributed Discrete Event Systems with Application to Flexible Manufacturing," in *Proceedings of the Winter Simulation Conference,* (San Diego), S. Andradottir, K. J. Healy, D. H. Withers and B. L. Nelson, Eds., pp. 845–852.

Harrell, C. R., and Tumay, K. (1994), *Simulation Made Easy,* IIE Press, Norcross, GA.

Heflin, D. L., and Harrell, C. R. (1998), "Simulation Modeling and Optimization Using ProModel," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 191–197.

Hlupic, V. (1997), "Simulation Software Selection Using SimSelect," *Simulation,* Vol. 69, No. 4, pp. 231–239.

Hlupic, V., and Paul, R. J. (1995), "Manufacturing Simulators and Possible Ways to Improve Them," *International Journal of Manufacturing Systems Design,* Vol. 2, No. 1, pp. 1–10.

Hullinger, D. G. (1999), "Taylor Enterprise Dynamics," Internal Report, F&H Simulations Inc., Orem, UT.

Industrial Modeling Corporation (1999), *SimEngine Trial Edition Guide,* Seattle.

Institute of Industrial Engineers (1999), "Buyer's Guide: Simulation Software," *Solutions,* Vol. 31, No. 5, pp. 52–59.

Keller, P., and Tchernev, N. (1997), "Object Oriented Methodology for FMS Modeling and Simulation," *International Journal of Computer Integrated Manufacturing,* Vol. 10, No. 6, pp. 405–434.

Kelton, W. D., Sadowski, R. P., and Sadowski, D. A. (1998), *Simulation with Arena,* WCB/McGraw-Hill, New York.

Kiviat, P. J., Markowitz, H., and Villanueva, R. (1969), *SIMSCRIPT II Programming Language,* Prentice-Hall, Englewood Cliffs, NJ.

Krepchin, I. P. (1988), "We Simulate All Major Projects," *Modern Materials Handling,* August, pp. 83–86.

Laguna, M. (1997), "Optimization of Complex Systems with OptQuest," Graduate School of Business Management, University of Colorado, Boulder.

Laughery, R. (1999), "Using Discrete-Event Simulation to Model Human Performance in Complex Systems," in *Proceedings of the Winter Simulation Conference,* (Phoenix), P. A. Farrington and H. B. Nembhard, Eds.

Law, A. M., and Kelton, W. D. (1991), *Simulation Modeling and Analysis,* 2nd Ed., McGraw-Hill, New York.

Lewellen, M., and Tumay, K. (1998), "Network Simulation of a Major Railroad," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 1135–1138.

Low, Y., Lim, C., Cai, W., Huang, S., Hsu, W., Jain, S., and Turner, S. J. (1999), "Survey of Languages and Runtime Libraries for Parallel Discrete-Event Simulation," *Simulation,* Vol. 72, No. 3, pp. 170–186.

Musselman, K. J. (1998), "Guidelines for Success," in *Handbook of Simulation,* J. Banks, Ed., John Wiley & Sons, New York, pp. 721–743.

Nance, R. E. (1996), "A History of Discrete Event Programming Languages," in *History of Programming Languages,* T. J. Bergin, and R. Gibson, Eds., ACM Press, New York, and Addison-Wesley, Reading, MA, pp. 369–427.

Narayanan, S., Schneider, N. L., Patel, C., Carrico, T. M., and DiPasquale, J. (1997), "An Object Based Architecture for Developing Interactive Simulations Using Java," *Simulation,* Vol. 69, No. 3, pp. 153–171.

Nikoukaran, J., Hlupic, V., and Paul, R. J. (1998), "Criteria for Simulation Software Evaluation," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 399–406.

Nisanci, A. (1997), "Modeling of FMS Pallet/Fixture Contention Rules," in *Proceedings of Summer Computer Simulation Conference,* (Arlington, VA), M. S. Obaidat and J. Illgen, Eds., Society for Computer Simulation International, San Diego, pp. 323–326.

Nisanci, A. (1998), "Comparative Analysis of AGV and Multi-Layer Conveyor Systems Using Simulation," in *Proceedings of Summer Computer Simulation Conference,* (Reno, NV), M. S. Obaidat, F. Davoli, and D. DeMarinis, Eds., Society for Computer Simulation International, San Diego, pp. 15–20.

Nisanci, A. (1999), "Multi-Project and Multi-Resource Lot Sizing and Scheduling," in *Proceedings of Summer Computer Simulation Conference,* (Chicago), M. S. Obaidat, A. Nisanci, and B. Sadoun, Eds., Society for Computer Simulation International, San Diego, pp. 131–136.

Nordgren, W. B. (1998), "Taylor II Manufacturing Simulation Software," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 263–267.

Norman, V., and Banks, J. (1998), "Managing the Simulation Project," in *Handbook of Simulation,* J. Banks, Ed., John Wiley & Sons, New York, pp. 745–764.

O'Reilly, J. J., and Lilegdon W. R. (1999a), "Introduction to AweSim," in *Proceedings of the Winter Simulation Conference,* (Phoenix), P. A. Farrington and H. B. Nembhard, Eds.

O'Reilly, J. J., and Lilegdon, W. R. (1999b), "Introduction to FACTOR/AIM," in *Proceedings of the Winter Simulation Conference,* (Phoenix), P. A. Farrington and H. B. Nembhard, Eds.

Pegden, C. D., Shannon, R. E., and Sadowski, R. P. (1995), *Introduction to Simulation Using SIMAN,* 2nd Ed., McGraw-Hill, New York.

Phillips, T. (1998a), "AutoMod by Autosimulations," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 213–218.

Phillips, T. (1998b), "AutoSched AP by Autosimulations," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 219–222.

Plott, B., Wojciechowski, J. Q., and Kilduff, P. A. (1999), "Command and Control Human Performance Modeling," *CSERIAC GATEWAY,* Vol. 10, No. 1, pp. 10–11.

Pritsker, A. A. B. (1995), *Introduction to Simulation and SLAM II,* 4th Ed., John Wiley & Sons, New York.

Pritsker, A. A. B., and O'Reilly, J. J. (1998), "Introduction to AWESIM," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 249–256.

Pritsker, A. A. B., and O'Reilly, J. J. (1999), *Simulation with Visual SLAM and AweSim,* John Wiley & Sons, New York.

Pritsker Corporation (1993), *Simulation: A Decision Support Tool,* Pritsker Corporation, West Lafayette, IN.

ProModel Corporation (1999), *User's Guide,* Orem, UT.

Roberts, C. A., and Dessouky, Y. M. (1998), "An Overview of Object Oriented Simulation," *Simulation,* Vol. 70, No. 6, pp. 359–368.

Rodrigues, J. M. (1994), *Directory of Simulation Software,* Society for Computer Simulation, San Diego.

Rohrer, M., and Banks, J. (1998), "Required Skills of a Simulation Analyst," *IIE Solutions,* Vol. 30, No. 5, pp. 20–23.

Russell, E. C. (1983), *Building Simulation Models with SIMSCRIPT II.5,* CACI Inc., Los Angeles.

Sadowski, D., Bapat, V., and Drake, G. (1998), "The Arena Product Family: Enterprise Modeling Solutions," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 205–212.

Schriber, T. J. (1991), *An Introduction to Simulation Using GPSS/H,* John Wiley & Sons, New York.

Schriber, T. J. and Brunner, D. T. (1998), "How Discrete Event Simulation Works," in *Handbook of Simulation,* J. Banks, Ed., John Wiley & Sons, New York., pp. 765–811.

Schwab, R. E. (1987), "Development of Simulation at Caterpillar Inc.," Internal Report, Peoria, IL.

Schwab, R. E., and Nisanci, H. I. (1992), "Simulation Languages," in *Handbook of Industrial Engineering,* 2nd Ed., G. Salvendy, Ed., John Wiley & Sons, New York.

Siprelle, A. J., Phelps, R. A., and Barnes, M. M. (1998), "SDI Industry: An Extend Based Tool for Continuous and High Speed Manufacturing," in *Proceedings of the Winter Simulation Conference,* (Washington, DC), D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 349–358.

Smith, J. M. (1987), *Mathematical Modeling and Digital Simulation for Engineers and Scientists,* John Wiley & Sons, New York.

Snowdon, J. L., El-Taji, S., Montevecchi, M., MacNair, E., Callery, C. A., and Miller, S. (1998), "Avoiding the Blues for Airline Travelers," in *Proceedings of the Winter Simulation Conference,* (Washington, DC) D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds., pp. 1105–1112.

Symix Systems/Pritsker Division (1999), *AweSim Demonstration Software Guide, Version 3.0,* West Lafayette, IN.

Symix Systems/Pritsker Division (1997), *FACTOR AIM: Working Model Primer, Version 8.0,* West Lafayette, IN.

Triadi, M. N., and Barta, T. M. (1999), "Experience with a Web-Based Discrete-Event Simulator," in *Proceedings of Summer Computer Simulation Conference,* (Chicago), M. S. Obaidat, A. Nisanci, and B. Sadoun, Eds., *Society for Computer Simulation International,* San Diego, pp. 93–96.

Tumay, K., and Harrington, H. (2000), Simulation Modeling Methods, McGraw-Hill, New York, NY.

Tumay, K., and Wood, B. (1999), "MODSIM III and Its Applications," in *Proceedings of the Winter Simulation Conference,* (Phoenix), P. A. Farrington and H. B. Nembhard, Eds. Visual Thinking International Inc. (1999), *SIMUL8 Manual and Simulation Guide,* Reston, VA.

Visual Thinking International Inc. (1999), *SIMUL8 Manual and Simulation Guide,* Restin, VA.

Wildberger, A. M. (1999), "AI and Simulation," *Simulation,* Vol. 72, No. 2, pp. 115–116.

Yu, L. C., Steinman, J. S., and Blank, G. E. (1998), "Adapting Your Simulation for HLA," *Simulation,* Vol. 71, No. 6, pp. 410–420.

CHAPTER 95

Statistical Analysis of Simulation Results

Barry L. Nelson
Northwestern University

1. INTRODUCTION 2469
2. OVERVIEW 2470
3. EXAMPLES 2470
  3.1. Static Simulation: Acceptance Sampling 2471
  3.2. Terminating Simulation: License Plate Removal 2471
  3.3. Steady-State Simulation: Inventory Model 2471
4. RANDOMNESS IN SIMULATIONS 2472
5. SIMULATION STATISTICS 2473
6. POINT ESTIMATORS 2475
  6.1. Mean Estimation 2475
  6.2. Probability Estimation 2476
  6.3. Quantile Estimation 2476
7. INITIAL-CONDITION BIAS 2477
  7.1. Remedial Measures 2478
    7.1.1. Data Deletion 2478
    7.1.2. Intelligent Initialization 2478
  7.2. Detection of Initial-Condition Bias 2479
    7.2.1. Mean Plot 2479
    7.2.2. Cusum Plot 2480
    7.2.3. Initial-Condition-Bias Test 2482
  7.3. A Comment on Probabilities and Quantiles 2483
8. MEASURES OF ERROR 2483
  8.1. Standard Error 2483
  8.2. Confidence Intervals 2485
    8.2.1. Confidence Intervals for Means 2485
    8.2.2. A Confidence Interval for Quantiles 2485
  8.3. A Note on Experiment Planning 2487
9. OPTIMIZATION AND SENSITIVITY 2487
  9.1. Multiple Comparisons 2488
  9.2. Metamodels 2490
10. VARIANCE REDUCTION 2492
REFERENCES 2494

1. INTRODUCTION

The design and control of many industrial and service systems requires the industrial engineer to account for uncertainty: uncertainty about the demand for products, the reliability of a machine, the arrival of customers, or the availability of components, for example. The language of probability is the primary tool for modeling uncertain or *stochastic* systems. Mathematical analysis, numerical analysis and computer simulation are techniques for analyzing stochastic models. Chapter 83 of the Handbook covers mathematical analysis of stochastic models. Chapters 93 and 94 describe the process of modeling a stochastic system and translating that model into a simulation program, respectively. This chapter provides guidelines for the design and analysis of computer simulation experiments. For a more extensive textbook treatment of these topics see Banks et al. (2000).

Compared to the other techniques, simulation is a brute force approach: using pseudorandom numbers to reproduce the uncertainty in the model, sample model behavior is generated and analyzed,

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

much in the same way that sample behavior of the physical system might be collected and analyzed. Simulation is perhaps the most general analysis technique, but the price of generality is that simulations provide only *estimates* of model performance parameters. The best that can be guaranteed is that these estimates converge to the true performance parameters as the simulation becomes infinitely long.

The emphasis in this chapter is on problems that are common in, or unique to, computer simulation relative to statistical experiments as a whole. The reader should refer to Chapters 83 to 87 of the Handbook for basic statistical methods. No specific computer hardware or software is assumed other than a language for programming the simulation experiment and possibly a statistical-analysis package capable of standard procedures such as least-squares regression.

Where possible, algorithms and procedures are provided rather than referenced. In that sense the chapter is self-contained. When references are provided, textbooks are often cited rather than original research papers, since they are more accessible.

Due to space constraints, only one or two methods are included for any problem, which does not imply that they are necessarily the best methods for any particular problem. The methods chosen are theoretically sound, empirically tested, but often conservative. The chapter is up to date, but it is not a guide to current research or emerging techniques.

One omission is the use of animation for simulation output analysis. Animation can be extremely helpful for developing a simulation model, verifying and validating the model, communicating results, and studying unusual model behavior in detail. However, it is not a substitute for a thorough numerical and graphical analysis that encompasses a much longer, and more representative, simulation run than anyone can reasonably view as animation.

The chapter is organized around three examples that are described below. The examples are simple enough that they can be presented in detail (except for the computer code).

2. OVERVIEW

This section is an annotated index to the chapter so that, if desired, the user can find the information of critical interest directly. However, all users should read Section 3, which introduces the three examples that are used as illustrations throughout the chapter, and most users should read Section 4 and Section 5, which describe the source of randomness and the standard statistics available in many simulation languages. A brief description of the other sections in this chapter is given below:

- Section 6 describes estimators for means, probabilities, and quantiles.
- Section 7 describes methods for detecting and reducing the effect of the initial conditions in a simulation experiment that estimates long-run performance.
- Section 8 describes ways to evaluate the goodness of estimates, including standard-error and confidence-interval procedures.
- Section 9 describes methods for evaluating alternative systems.
- Section 10 describes a way to improve the precision of simulation estimators.

Many of the sections contain programming-language-independent algorithms for design and analysis procedures. To understand the algorithms, it is useful to know the following:

The symbol ← indicates assignment; for example, $a \leftarrow 5$ assigns the value 5 to the variable a.

The scope of *for* and *if* are delimited by *endfor* and *endif,* respectively.

The notation $\lfloor x \rfloor$ means the greatest integer less than or equal to x.

The subscripts on variables that are vectors or arrays are indicated by square brackets; e.g., $y[i,j]$ for y_{ij}.

Addition, subtraction, multiplication and division are indicated by +, −, * and /, respectively.

3. EXAMPLES

Three examples that will be used throughout the chapter to illustrate simulation design and analysis techniques are introduced in this section. Appropriate design and analysis depends on characteristics of the model and on the questions being asked, so these three examples have been chosen to illustrate important classes of problems. Although they are simpler than many practical models—in fact, they are so simple that simulation is not actually needed—the techniques are no more difficult to apply in complex models.

3.1. Static Simulation: Acceptance Sampling

Acceptance sampling is a widely used technique for economically assessing the quality of a "lot" of items. A single-sampling attributes plan of the form (n, c) specifies that n items will be sampled and the lot will be accepted if no more than $c < n$ defective items are discovered. The values of n and c are chosen based on the operating characteristic (OC) curve they imply. The OC curve is the probability of accepting a lot as a function of the lot quality, where quality is usually stated in terms of the probability of a defective item (see Chapter 69 of the Handbook).

OC curves for standard acceptance-sampling plans are derived under the assumption that the quality of items can be modeled as independent and identically distributed (i.i.d.) Bernoulli random variables. Although this model is often plausible, the quality of items produced by some processes exhibit statistical dependence. The goal of this simulation experiment is to estimate the OC curve for sampling plan (10, 1) when item quality is dependent.

Let $X_1, X_2, \ldots, X_n$ represent a sample of n items, where $X_i = 1$ if the ith item is defective and $X_i = 0$ if the ith item is acceptable. The probability that an item is defective is p, and the quality of any item may be dependent on the other items. Specifically, the items are assumed to have a joint Pólya distribution with pairwise correlation $\rho = 0.08$. Standard tables of sampling plans are not appropriate for this situation.

Let $Y = \sum_{i=1}^{n} X_i$ be the number of defective items in the sample. The performance parameter of interest is $\theta(p) = \Pr\{Y \leq c|p\}$, the probability that there are c or fewer defective items (i.e., that the lot is accepted) as a function of p.

This simulation experiment is called a *static simulation* because there is no explicit modeling of the passage of time ($X_1, X_2, \ldots, X_n$ need not even be arranged by order of selection). Although static simulations are conceptually the easiest to design and analyze, they nevertheless present important design and analysis problems. In addition, this example illustrates estimating a probability, $\theta(p)$.

3.2. Terminating Simulation: License Plate Renewal

A small city will allow auto owners to renew their license plates by mail, with each owner's renewal taking place during the month of his or her birth. The mail-in renewal applications will be processed by a clerk. It is anticipated that the rate of receipt of applications will increase steadily throughout the month, but that the load during all months is about the same. Mail-in renewals that are postmarked after the 27th day of the month are returned to the applicant without being processed. The city is interested in determining the performance level the clerk must attain, in terms of renewals processed per day, to prevent excessive delays or carryover from month to month.

The arrival of mail-in renewals is modeled as a Poisson process with time-dependent arrival rate $\lambda(t) = t$ renewals per day, where time $0 \leqslant t \leq 27$ is measured in days. In other words, on the first day of the month applications arrive at a rate of 1 per day, but by the end of the month they are arriving at a rate of 27 per day. The time to process an application is modeled as an exponentially distributed random variable with mean $1/\mu$ days, so that μ is the processing rate in applications per day.

Let Y_i be the delay in processing the ith application received during the month, and let N be the number of applications received during the month. A performance parameter of interest is

$$\theta(\mu) = E_\mu \left[\frac{1}{N} \sum_{i=1}^{N} Y_i \right]$$

the expected average delay for mail-in renewals as a function of the processing rate, μ.

This simulation experiment is called a *terminating simulation* because the parameter of interest is defined with respect to a finite time horizon, the time to process one month of renewal applications in this case. The example illustrates estimating transient or time-dependent performance, and also examining the effect of a continuous design variable, the processing rate μ.

3.3. Steady-State Simulation: Inventory Model

An (s,S) inventory system involves the periodic review of the level of inventory of some discrete unit. If the inventory position (units in inventory plus units on order minus units backordered) at a review is found to be below s units, then enough additional units are ordered to bring the inventory position up to S units. When the inventory position at a review is found to be above s units, no additional units are ordered. One possible goal is to select the values of s and S that minimize inventory cost. The third example is the (s, S) inventory model in Koenig and Law (1985) and Law and Kelton (2000).

TABLE 1 Inventory Policies

ℓ	s	S
1	20	40
2	20	80
3	40	60
4	40	100
5	60	100

Let $\{I_t;\ t = 1, 2, \ldots\}$ represent the inventory position just after a review at period t, and let $\{X_t;\ t = 1, 2, \ldots\}$ represent the demand for units of inventory in period t. The inventory position I_t changes in the following way:

$$I_{t+1} = \begin{cases} S, & \text{if } I_t - X_t < s \\ I_t - X_t, & \text{if } I_t - X_t \geq s \end{cases}$$

For convenience, the initial inventory position is taken to be $I_1 = S$; that is, the inventory position is initially at its maximum. The demand $\{X_t;\ t = 1, 2, \ldots\}$ is modeled as a sequence of i.i.d. Poisson random variables with mean 25 units.

In each period there are costs associated with the inventory position. If $I_t - X_t < s$, then in period $t + 1$ a cost of $32 + 3[S - (I_t - X_t)]$ is incurred, which is a fixed cost plus a per-unit cost of bringing the inventory position up to S. In period $t + 1$, if $I_{t+1} \leq X_{t+1}$, then a holding cost of $(I_{t+1} - X_{t+1})$ dollars is incurred; otherwise a shortage cost of $5(X_{t+1} - I_{t+1})$ dollars is incurred.

Let $Y_t^{(\ell)}$ be the cost incurred in period t under inventory policy ℓ; the inventory policies under consideration are given in Table 1. The performance parameters of interest are the long-run expected cost per period of each inventory policy, with a smaller expected cost being preferred. The assumption behind such a long-run analysis is that $Y_t^{(\ell)}$ converges in distribution to $Y^{(\ell)}$ as $t \rightarrow \infty$, which means that the cost per period converges to a limiting random variable whose distribution does not depend on time (period) t. The parameter of interest is $\theta^{\ell} = \mathrm{E}\,[Y^{(\ell)}]$, the long-run expected cost per period for policy ℓ.

This simulation experiment is called a *steady-state simulation* because the parameter of interest is associated with a limiting random variable. This example illustrates estimating a limiting quantity from a simulation experiment that is (necessarily) finite, and also comparing alternative systems (inventory policies) with the goal of selecting the best system.

4. RANDOMNESS IN SIMULATIONS

Modern simulation languages contain *pseudorandom number generators* that produce sequences of numbers—typically numbers in the interval (0,1)—that are difficult to distinguish from independent and identically distributed (i.i.d.) random numbers. In other words, an analyst subjecting a sequence of pseudorandom numbers to a battery of statistical tests would be unlikely to recognize that they were produced by a deterministic algorithm rather than by a random process. However, if the same simulation program is executed on the same computer at any time, identical results are obtained. This is a useful property because debugging programs would be difficult if results changed randomly.

This method of incorporating randomness into computer simulations has a profound impact on the design and analysis of simulation experiments. Most importantly, it means that different simulation runs will be dependent if they employ the same pseudorandom numbers. This can be good, yielding sharper comparisons between alternative systems, or bad, invalidating the assumptions behind statistical procedures that assume independent observations.

An important feature of most simulation languages is that they permit control of the pseudorandom numbers through random number *streams* or *seeds,* which permit the user to access different, and thus apparently independent, portions of the pseudorandom number sequence. Such control allows the user to induce dependence where desired or obtain independence where necessary.

To be more precise, suppose that a simulation language includes one pseudorandom number generator that produces an ordered sequence of pseudorandom numbers $u_1, u_2, \ldots, u_g$. The length of this sequence is necessarily finite; for many pseudorandom number generators $g < 2^{31} \approx 2$ billion, although generators that produce much longer sequences are becoming common.

Let the seeds or streams be denoted by $s_1, s_2, \ldots, s_h$. The seeds or streams are nothing more than reference points within the sequence $u_1, u_2, \ldots, u_g$. For instance, s_1 might correspond to starting the sequence at u_{2137}. Thus, the number of seeds $h \leq g$, and is typically much less than g with the reference points spaced widely apart. Note that some simulation languages allow the user to specify the *offset* between seeds, rather than having to specify the seeds themselves. In any event, since the entire sequence $u_1, u_2, \ldots, u_g$ appears to be a sequence of i.i.d. random numbers, the subsequences between seeds also appear to be independent of each other. This property implies an important design and analysis principle:

Design and Analysis Principle 1. The assignment of random number streams or seeds is part of the design of a simulation experiment. Assigning the same streams or seeds to different simulations induces dependence, while assigning different seeds or streams to different simulations induces independence between simulation results.

The simulation-language user seldom works directly with the pseudorandom numbers $u_1, u_2, \ldots, u_g$, but rather specifies the probability distributions of the simulation *inputs*. A convention in this chapter is that input random variables are denoted generically by X. Examples of simulation inputs are:

- The item quality random variables $X_1, X_2, \ldots, X_n$ in the acceptance-sampling model, which have a joint Pólya distribution
- The arrival and processing times of renewal applications in the license-renewal model, which have time-dependent Poisson and exponential distributions, respectively
- The demand for inventory in period t, X_t, in the inventory model, which has a Poisson distribution

Observations or realizations of the inputs are obtained by transforming the pseudorandom numbers. One or more pseudorandom numbers may be required to produce each input, depending on what transformation is used. The inputs, X, are functions of the pseudorandom numbers, u, so they are completely determined by the seed or stream, s, say $X = X(s)$.

The purpose of performing a simulation experiment is to observe model performance. The observed model performance, called the *output,* is derived from realizations of the inputs and the (often complex) logic of the model. A convention in this chapter is that output random variables are denoted generically by Y. Since the outputs are functions of the inputs, they are also functions of the seeds or streams, say $Y = Y[X(s)]$. Examples of simulation outputs are:

- The number of defective items discovered in a sample, Y, in the acceptance-sampling model
- The delay experienced by the ith renewal application, Y_i, in the license-renewal model
- The cost in period t for inventory policy ℓ, $Y_t^{(\ell)}$, in the inventory model

In most simulation experiments a large number of outputs are generated. They are summarized by a *statistic,* which is often a sample average of the outputs, denoted $\bar{Y}$. The value of the statistic is used to estimate system performance, so statistics are also called *estimators*. The statistics are functions of the outputs, so they are also completely determined by the random number seeds or streams.

Perhaps it seems strange that statistical methods are employed to analyze a completely deterministic process, which a simulation is after the seeds are specified (particularly since many users accept the default seeds or streams). However, if the pseudorandom numbers cannot easily be distinguished from random numbers, then treating functions of these numbers—the inputs, outputs and statistics—as random variables will not be misleading.

The role of the pseudorandom number streams or seeds is important. This method of representing randomness, and the corresponding control it permits, is the primary difference between simulation experiments and classical statistical experiments.

5. SIMULATION STATISTICS

Many simulation languages compute statistics and generate reports automatically. To understand these statistics it is necessary to distinguish between within- and across-replication statistics (some simulation languages refer to a replication as a "run"). To make the difference concrete, consider the license-renewal simulation. One replication or run of this simulation represents one month and generates outputs $Y_1, Y_2, \ldots, Y_N$, which are processing delays for the N applications received during the month. The sample average of these delays is

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

which is a *within-replication statistic,* meaning a summary of the outputs within one replication or run.

If k months are simulated, then each month yields a sample average delay, say $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$. The sample average of these sample averages is

$$\bar{\bar{Y}} = \frac{1}{k}\sum_{j=1}^{k} \bar{Y}_i$$

which is an *across-replication statistic,* meaning a summary of the statistics across several replications.

Nearly all simulation languages compute within-replication statistics automatically. Many compute across-replication statistics. The standard statistics provided are sample averages, variances (or standard deviations), maximum and minimum observed values, and number of observations.

Some care must be exercised when interpreting the within-replication statistics because the outputs within a replication *are typically neither independent nor identically distributed,* and "i.i.d. data" is a common assumption behind many statistical procedures.

For example, in the license-renewal simulation, the within-replication sample variance of the processing delays, $Y_1, Y_2, \ldots, Y_N$, would be calculated by many simulation languages as

$$S^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N - 1}$$

Familiarity with classical statistics might lead one to use S^2 as an estimator of $\mathrm{Var}[Y_i]$, the (assumed) common variance of the processing delays, and S^2/N as an estimator of $\mathrm{Var}[\bar{Y}]$, the variance of the average delay.

However, the observed delays within a replication are not identically distributed, since applications received later in the month tend to be delayed longer than applications received earlier in the month. Thus, there is no common variance of the delays within a replication.

Also, the observed delays within a replication are not independent, since applications that experience long (short) delays tend to be followed by applications that experience long (short) delays. The validity of S^2/N as an estimator of $\mathrm{Var}[\bar{Y}]$ rests on the assumption of i.i.d. data, and the estimator can be significantly biased when this assumption is violated. Unfortunately, the estimator is often biased low, which means that $\bar{Y}$ appears to be more precise than it actually is and confidence intervals based on S^2/N are inappropriately narrow.

Finally, the number of applications received during the month, N, is a random variable, so S^2 is a ratio of the random variables $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ and N. The properties of ratio estimators are not straightforward or easy to summarize, and they are typically different from the corresponding estimator when the denominator is not a random variable.

In summary, the within-replication sample variance S^2 is not a useful measure in this example, but there is no way for the simulation language to know that and notify the user.

On the other hand, most simulation languages ensure that statistics across replications are i.i.d. because by default they initialize each replication in the same way, implying identically distributed outputs, but they use different random numbers (by continuing in the same pseudorandom number sequence), implying independent outputs. In the license-renewal simulation the replication averages $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$ are i.i.d., so applying classical statistical analysis is appropriate.

Design and Analysis Principle 2. Outputs within a replication may be neither independent nor identically distributed, and the number of observations may be random. Statistical analysis should therefore typically be based on multiple replications, with each replication supplying one observation of the performance measure of interest.

Rather than using automatically generated statistics, the author favors saving the raw data from a simulation experiment and performing the statistical analysis a posteriori using a general-purpose analysis package that may be separate from, or integrated into, the simulation language. There are three reasons for this preference. First, statistical-analysis packages can perform more procedures than most simulation languages; regression analysis, for instance. Second, using a package on the raw data facilitates exploratory analysis that may suggest what additional analysis is appropriate. For example, the normality of the output data can be empirically checked to see if an inference based on the normal distribution is reasonable. Finally, the raw data may be reanalyzed if new questions arise for which the necessary statistics were not computed during the simulation runs. For example, in a service system, the mean customer delay may be estimated from the individual customer delays. If it is later determined that the probability of a customer being delayed beyond some threshold value is of interest, then this can be estimated without rerunning the simulation if the raw data are available.

The primary arguments against saving all of the data are the storage requirement, which can be excessive, the decrease in simulation execution speed due to electronically writing all of the output data, and the difficulty of transforming the simulation data into a form suitable for an analysis package. Language designers have addressed these problems by developing integrated database and analysis software and exporting data in standard formats that can be imported by spreadsheet or statistical-analysis software.

6. POINT ESTIMATORS

The decisions resulting from simulation studies are often based on point estimates of system performance parameters. The term *point estimate* means a single number that serves as a best candidate for the unknown performance parameter. Although the point estimates are important, they should always be accompanied by a measure of the error; error estimation is covered in Section 8.

This section describes estimating θ, a parameter of a probability distribution, F. Although F is usually not known, it is assumed that simulation outputs Y can be generated from F. The three cases considered are:

1. When θ is the *mean* of Y; i.e., $\theta = \mathrm{E}[Y] = \int_{-\infty}^{\infty} y\,dF(y)$
2. When θ is a *probability* associated with Y; e.g., $\theta = \Pr\{a < Y \leq b\} = F(b) - F(a)$ for given values $a < b$
3. When θ is a *quantile* of Y; i.e., $q = \Pr\{Y \leq \theta\} = \int_{-\infty}^{\theta} dF(y)$ for a given probability q

These cases arise in static and terminating simulation experiments. In steady-state simulation, outputs from F cannot be observed directly (because the distribution of interest is a limiting distribution), which introduces bias into the estimators. Estimation of parameters of a steady-state distribution is treated in Section 7.

6.1. Mean Estimation

Perhaps the most commonly reported performance parameter is the mean (expected value) of a system performance measure. The term *expected value,* which defines a parameter, will be used rather than the synonym *mean,* because *mean* may be confused with the *sample mean,* which is an estimator of a parameter.

In the license-renewal simulation, the parameter

$$\theta(\mu) = \mathrm{E}_{\mu}\left[\frac{1}{N}\sum_{i=1}^{N} Y_i\right] \tag{1}$$

is an expected value: the expected average delay for all applications received during one month when the processing rate is μ.

To estimate $\theta(\mu)$ the simulation is designed to generate k i.i.d. replications of one month of renewal processing. Each replication yields a within-replication average

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Let $\overline{Y}_j$ be the within-replication average from the jth replication. Since $\theta = \mathrm{E}[\overline{Y}_j]$ for each j by definition, an unbiased estimator of θ is the across-replication average

$$\overline{\overline{Y}} = \frac{1}{k}\sum_{j=1}^{k} \overline{Y}_j$$

Within- and across-replication averages for $k = 5$ replications when $\mu = 20$ applications per day are given in the second column of Table 2.

The standard error of $\overline{\overline{Y}}$ as an estimator of $\theta(\mu)$ is $\sigma/\sqrt{k}$, where σ^2 is the common variance of the replication averages $\overline{Y}_j$. Estimating measures of error is covered in Section 8. The important concept is that the error decreases as the number of replications increases.

Design and Analysis Principle 3. The sample average of i.i.d. outputs is an unbiased estimator of their common expectation. The standard error of the estimate decreases at rate $1/\sqrt{k}$, where k is the number of replications.

Given i.i.d. outputs with a common expectation, mean estimation is straightforward, as illustrated above. However, as discussed in Section 5 some care must be exercised when defining the expected value of interest in terminating simulations because performance parameters may be time dependent.

TABLE 2 Simulation Results for Five Replications of the License-Renewal Simulation with $\mu = 20$ (units are days)

Replication j	$\bar{Y}_j$	Y_{Nj}
1	0.34	1.25
2	0.72	2.20
3	0.32	1.04
4	0.46	1.68
5	0.42	1.08
$\bar{\bar{Y}}$	0.45	1.45

For example, the expected delay of an individual renewal application is not well defined because applications have different expected delays, depending on their order of arrival during the month. The parameter $\theta(\mu)$ in (1) circumvents this problem by averaging all the delays during a month and defining $\theta(\mu)$ to be the expectation of this average. However, $\theta(\mu)$ masks the time-dependent behavior of the delays, which may be important.

A time-dependent parameter that is well defined is $\theta_N(\mu) = \mathrm{E}[Y_N]$, the expected delay of the last application received during the month. To estimate $\theta_N(\mu)$, which is also an expected value, the delay for the last application, Y_N, is the within-replication statistic, and the across-replication average of these delays is an estimator of $\nu_N(\mu)$. The third column of Table 2 gives the delay for the last application received in each of the five replications and the across-replication average of these delays.

6.2. Probability Estimation

Probabilities can be represented as expected values, so the discussion in Subsection 6.1 is applicable to estimating probabilities as well. However, probability estimation is required so frequently that it is worth treating as a separate topic.

In the acceptance-sampling simulation, the parameter of interest is $\theta(p) = \Pr\{Y \leq c|p\}$, where Y is the number of defective items in a sample of n items and the quality of the items has a joint Pólya distribution with marginal probability p of a defective. As a function of p, $\theta(p)$ is the OC curve for sampling plan (n, c).

The key to estimating probabilities is the following principle:

Design and Analysis Principle 4. A probability θ can be represented as the expected value of an indicator (0, 1) random variable, where the value 1 corresponds to occurrence of the event of interest. The variance of an indicator random variable is $\sigma^2 = \theta(1 - \theta)$.

For the acceptance-sampling simulation, define an indicator random variable I as follows:

$$I = \begin{cases} 1, & \text{if } Y \leq c \\ 0, & \text{otherwise} \end{cases}$$

Then $\theta(p) = \mathrm{E}[I]$, and all of the principles of mean estimation can be applied to the indicator random variable, I.

Figure 1 shows an estimate of the OC curve for sampling plan $(n, c) = (10, 1)$ that was obtained by estimating $\theta(p)$ at points $p = 0.01, 0.05, 0.1, 0.15$ and $0.2, 0.3, \ldots, 0.9$. Each point estimate is the average of 10,000 replications of I. Since $\theta(p) = \mathrm{E}[I]$, the point estimators are unbiased, and the standard error of the estimators is $\sqrt{\theta(p)[1 - \theta(p)]/10000}$.

6.3. Quantile Estimation

Quantiles are points on the distribution of a random variable. The most familiar quantile is the *median,* which is the 0.5 quantile. The 0.25 and 0.75 quantiles are also called the *quartiles*. Extreme quantiles, such as the 0.9, 0.95, and 0.99 quantiles, are one way to characterize the tail of a distribution.

Subsection 6.1 described estimating the expected value of $\bar{Y}$, the average delay of renewal applications received during a month. The 0.9 quantile of $\bar{Y}$, denoted $\theta_{0.9}$, is the value such that $\Pr\{\bar{Y} \leq \theta_{0.9}\} = 0.9$. In other words, the average delay in 9 out of 10 months is less than or equal to $\mu_{0.9}$.

The procedure below, called *algorithm quantile estimation,* estimates the q quantile of a random variable given k i.i.d. replications:

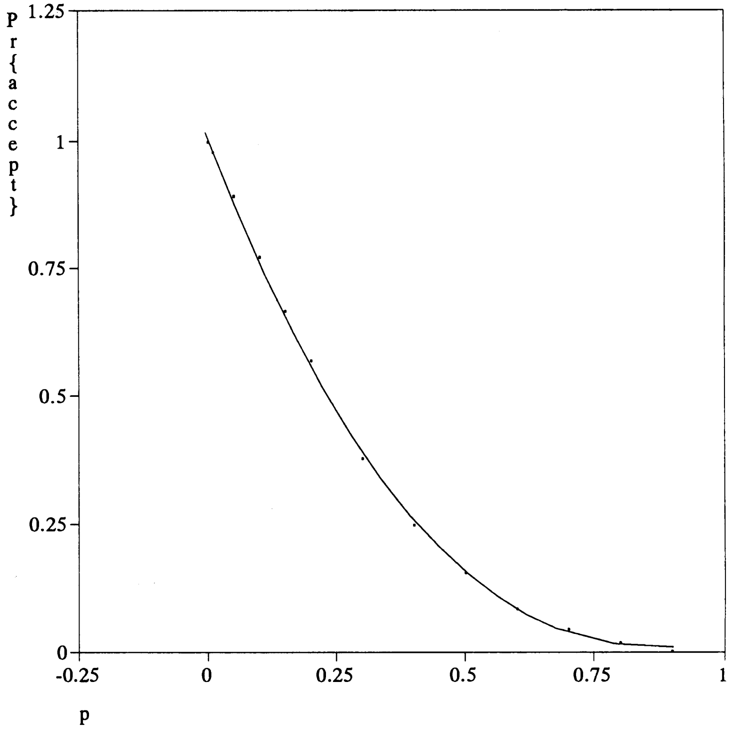


Figure 1 Plot of Estimated OC Curve for Sampling Plans (10, 1) Derived from $k = 10{,}000$ Replications.

```
1. Initialization: Number of replications, $k$; quantile probability, $0 < q < 1$.
2. Data: $y[j]$, output from the $j$th replication.
3. Compute order statistics: sort $y[j]$ from smallest to largest so that $y[1] \leq y[2] \leq \ldots \leq y[k]$
   (comment: these sorted values are called the order statistics of the sample).
4. $i \leftarrow \lfloor (k + 1)*q \rfloor$.
   if $i < 1$ output $y[1]$
   if $i \geq k$ output $y[k]$
   otherwise $f \leftarrow (k + 1)*q - i$; output $(1 - f)*y[i] + f*y[i + 1]$
```

Quantile estimators are biased in general, but the bias (as well as the variance) of the estimator decreases as the number of replications increases. The procedure *algorithm quantile estimation* interpolates between order statistics to reduce bias.

For the license-renewal simulation the data are $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$, the average delays from k replications. For $k = 200$ replications and $\mu = 20$ applications per day, an estimate of $\theta_{0.9}$ is $(0.1)*y[180] + (0.9)*y[181] = 0.88$ days.

7. INITIAL-CONDITION BIAS

Performing a steady-state simulation experiment implies that the analyst is interested in long-run performance of the model that is independent of the initial conditions of the simulation replication or run. In the inventory simulation the parameter of interest is $\theta^{(\ell)}$, the long-run expected cost per period of inventory policy ℓ, which does not depend on the inventory position in period 1.

The length of a simulation run is necessarily finite, so residual effects of the initial conditions will often be present in the outputs. These residual effects manifest themselves as *bias* in the statistics, meaning that the expectation of the statistic is not equal to the value of the performance parameter of interest. In some simulations, the effect of the initial conditions can be so overwhelming that the results are meaningless unless appropriate remedial measures are taken. This section covers mean estimation for steady-state simulation. For more general estimation problems, see Law and Kelton (2000), Schruben (1981), and the comments at the end of this section.

Consider (s, S) inventory policy 1—which is (20, 40)—and let $Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{200}^{(1)}$ be the observed inventory costs in the first 200 periods of a replication of this policy. An estimator of the long-run expected cost per period, $\theta^{(1)}$, is the sample average

$$\overline{Y}^{(1)} = \frac{1}{200} \sum_{t=1}^{200} Y_t^{(1)}$$

The effect of the initial condition $I_1 = S$ is to cause the expected value of $\overline{Y}^{(1)}$ to be less than $\theta^{(1)}$ because little or no ordering and setup cost is incurred until the inventory has been depleted somewhat. This means that, on average, the estimate will indicate that policy 1 is less costly than it actually is.

The appropriate remedial measure depends on whether it is feasible to perform some preliminary or pilot analysis of the model prior to making the final runs. When a large number of similar systems are to be investigated (e.g., different inventory policies for the same inventory system), then preliminary analysis of one system may be worthwhile since the results can be extrapolated to the other systems. However, if only one system is to be investigated, or if each run is very expensive, then it may be desirable to use the same experiment to remedy the initial-condition bias and estimate the parameters of interest. This section describes both approaches and two remedial measures, data deletion and intelligent initialization.

7.1. Remedial Measures

For most estimation and inference problems, increasing the number of replications is beneficial. The initial-condition bias problem is an exception.

Design and Analysis Principle 5. Increasing the number of independent replications does not decrease the bias of the sample average. If the number of replications is increased at the expense of decreasing the length of each replication, then increasing the number of replications may even increase the bias.

This principle suggests that in the presence of initial-condition bias and a tight budget, the number of replications should be small and the length of each replication should be long. When bias is severe, making only one very long replication may be advantageous. Experiment designs that call for only one replication require more sophisticated strategies for computing a measure of error, as discussed in Section 8.

Even with only a few long replications, remedial measures may be needed. Two of the many remedial measures, which can be used together, are described below.

7.1.1. Data Deletion

If there is a standard remedial measure, it is to delete or discard some of the outputs from the beginning of each replication where the effects of the initial conditions are most pronounced. Returning to the inventory simulation, if $Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{200}^{(1)}$ are the observed inventory costs in the first 200 periods, then another estimator of $\theta^{(1)}$ is the truncated sample average

$$\overline{Y}^{(1)}(d) = \frac{1}{200 - d} \sum_{t=d+1}^{200} Y_t^{(1)}$$

where d, the number of outputs deleted, could be 0, 1, . . . , 199. Typically, the bias of the truncated sample average decreases as d increases, which is good, but its variance increases with d, which is bad. Methods for choosing d are discussed in a later subsection.

Many simulation languages make it convenient to delete all outputs up to a fixed point in simulated time or up to an event time rather than deleting a fixed number of outputs, d. This makes d a random variable but does not significantly change the properties of the estimators, provided the runs are not too short.

7.1.2. Intelligent Initialization

The initial conditions for a simulation experiment are the starting values assigned to the variables in the model and the events scheduled at the beginning of each run or replication. Since the parameters of interest in a steady-state simulation do not depend on the initial conditions, initial conditions are often chosen for programming convenience. In the inventory simulation the initial inventory position is $I_1 = S$, for example. Some care in setting the initial conditions can greatly reduce the bias they cause.

Any model that has a steady-state distribution has a "correct" initialization procedure so that there is no initial-condition bias. This result is of more theoretical interest than practical importance because knowing the correct procedure typically implies already knowing the values of the parameters of interest. However, setting initial conditions close to steady-state mean or mode conditions, while not correct, can be very effective in reducing bias. There are at least two ways to approximate steady-state conditions:

1. *Approximate models:* Steady-state distributions and parameters are known for many stochastic processes; e.g., queueing, inventory, Markov chains. These results can be used to approximate the simulation model. For example, a service system can be approximated by a Markovian queue to determine the expected number of customers in the system. This value can be used to set the initial number of customers in the system for the simulation, rather than using the (convenient) initial condition of an empty system. Chapter 81 of the *Handbook* is a good source of approximations. Even cruder approximations, such as replacing a random quantity by its expectation, can also be used.
2. *Pilot runs:* The debugging runs or preliminary pilot runs of the simulation can provide information for setting initial conditions more intelligently. For instance, printing out the average inventory position during debugging runs of the inventory simulation showed that taking $I_1 = 36$ is closer to steady-state conditions than $I_1 = 40$ for inventory policy 1.

The problem of setting optimal initial conditions has been studied, but there is no single recommendation. Significant improvement is possible even if the initial conditions are not optimal, especially when the convenient initial conditions are far from optimal. The point is to use all available knowledge about the model, including knowledge gained from experimentation, to refine the initial conditions.

7.2. Detection of Initial-Condition Bias

The methods described in this subsection can be used to detect the presence of initial-condition bias and to determine the amount of data to delete.

7.2.1. Mean Plot

Let $\theta_t^{(1)} = \mathrm{E}[Y_t^{(1)}]$ be the expected cost of inventory policy 1 *in period t.* The assumption behind steady-state simulation is that $\theta_t^{(1)} \to \theta^{(1)}$ as t increases. The mean plot estimates $\theta_t^{(1)}$ as a function of t so that the rate of convergence can be determined. This method is appropriate when some preliminary study of the bias is possible.

An estimate of $\theta_t^{(1)}$ can be obtained by making a number of replications of the experiment, averaging across replications, and plotting the resulting values. To be specific, suppose that 30 replications of 200 periods each were generated for inventory policy 1. Let $Y_{tj}^{(1)}$ be the cost in period t for replication j, $t = 1, 2, \ldots, 200$ and $j = 1, 2, \ldots, 30$. Then the across-replication averages

$$\overline{Y}_t^{(1)} = \frac{1}{30}\sum_{j=1}^{30} Y_{tj}^{(1)}$$

are estimates of $\theta_t^{(1)}$, since $\overline{Y}_t^{(1)}$ is the average cost in period t across 30 replications. A plot of these averages as a function of t is given in Figure 2. Since the data are still quite variable, a smoothing curve was fitted to make the mean function more apparent (a smoothing spline was used in Figure 2; Welch 1983 recommends a moving average). Another way to smooth the plot is to increase the number of replications. The procedure below, called *algorithm mean-plot,* calculates the across-replication averages needed for the plot:

```
1. Initialization: number of replications, $k$; length of each replication, $m$; array $a[t]$ of length $m$;
   array $s[t] \leftarrow 0$ for $t \leftarrow 1, 2, \ldots, m$
   (comment: $k$ and $m$ are problem dependent and require experimentation)
2. Data: $y[t, j]$, the $t$th observation from replication $j$
3. Calculate averages:
   (a) for $t \leftarrow 1$ to $m$:
           for $j \leftarrow 1$ to $k$: $s[t] \leftarrow s[t] + y[t, j]$, endfor
       endfor
   (b) for $t \leftarrow 1$ to $m$: $a[t] \leftarrow s[t]/k$, endfor
4. Output: plot $a[t]$ vs. $t$, smoothing the plot if necessary
```

The mean plot in Figure 2 shows that bias is clearly present—although it dissipates rapidly in this example—and the early cost values have a negative bias (biased low). A deletion point d can

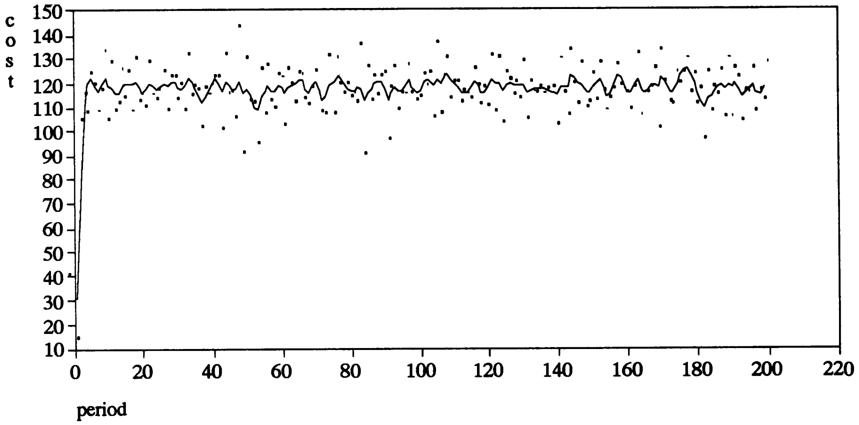


Figure 2 Mean Plot for Inventory Example.

be obtained by looking for a point where the curve seems to become nearly horizontal, perhaps at $d = 20$ periods in this example. The selection of d using the mean plot is subjective.

7.2.2. Cusum Plot

Schruben derived a plot that can be formed from a single, long replication and that is particularly sensitive to the presence of initial-condition bias (Barton and Schruben 1989). In terms of the inventory simulation, define $S_0 = 0$, and let

$$S_j = \sum_{t=1}^{j} (\bar{Y}^{(1)} - Y_t^{(1)})$$

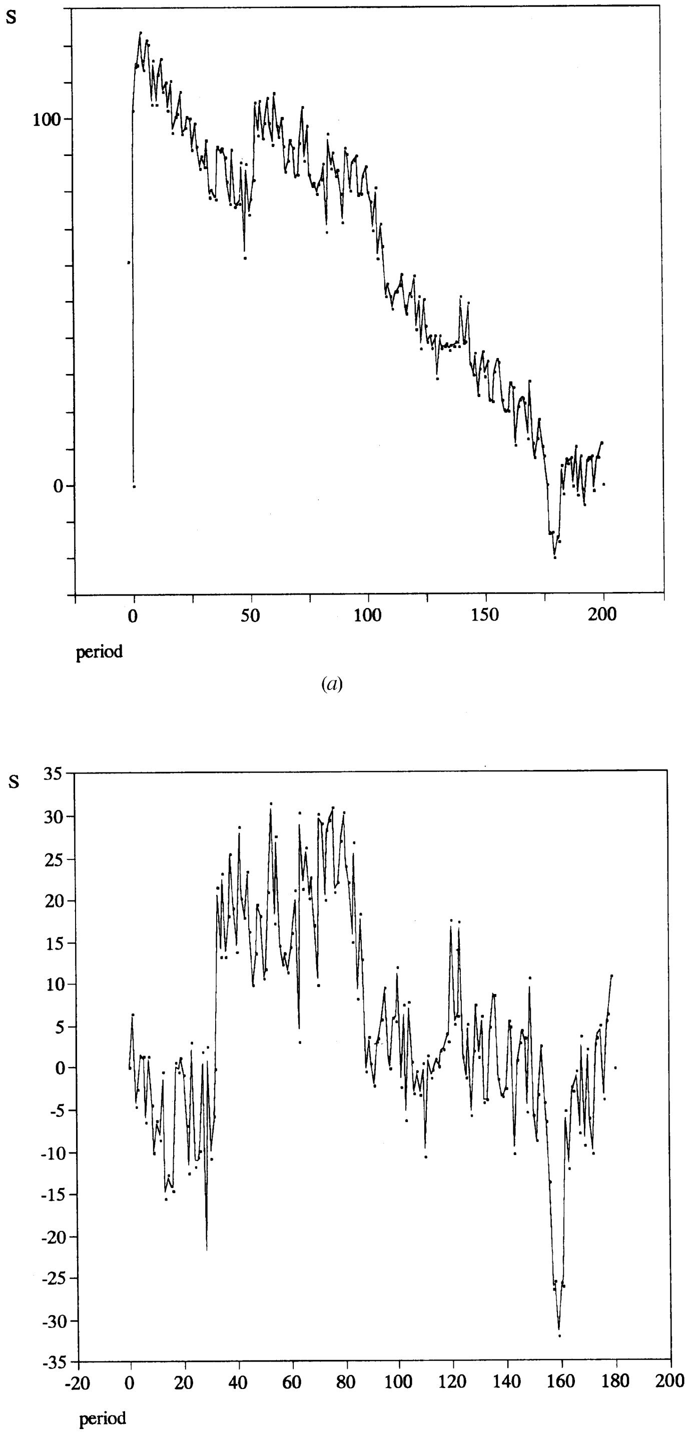
for $j = 1, 2, \ldots, m$, where m is the number of periods simulated. The *cusum plot* is a graph of S_j vs. j.

Notice that $S_0 = S_m = 0$. If there is no initial-condition bias, then $\mathrm{E}[S_j] = 0$ for all values of j in between, so the plot of S_j will tend to cross zero several times. However, when bias is present the values of S_j will all tend to be on one side of zero, depending on whether the bias is positive or negative.

Figure 3 shows such a plot for $m = 200$ periods of the inventory simulation before and after 20 periods of output were deleted (this plot was formed after averaging across several replications to make the characteristic behavior even more pronounced). In the first case the plot tends to be above zero, indicating negative bias, while after deletion no bias appears to be present.

To smooth the plot, batching can be used. This is done in *algorithm cusum:*

```
1. Initialization: length of the replication, $m$; batch size $b$; number of batches $k \leftarrow \lfloor m/b \rfloor$; $a \leftarrow 0$; array $s[j] \leftarrow 0$ for $j \leftarrow 0, 1, \ldots, k$
   (comment: $m$ is problem dependent but should be large; Schruben recommends $b \leftarrow 5$)
2. Data: $y[t]$, the $t$th observation from a single replication
3. Batch the outputs:
   (a) for $j \leftarrow 1$ to $k$:
           $y[j] \leftarrow y[(j - 1)*b + 1]$
           for $t \leftarrow 2$ to $b$: $y[j] \leftarrow y[j] + y[(j - 1)*b + t]$, endfor
       endfor
   (b) for $j \leftarrow 1$ to $k$: $y[j] \leftarrow y[j]/b$, endfor
   (comment: batch averages are now stored in $y[1], y[2], \ldots, y[k]$; if the original data must be saved then a separate array should be used)
4. Compute overall average:
   for $j \leftarrow 1$ to $k$: $a \leftarrow a + y[j]$, endfor
   $a \leftarrow a/k$
```



(*a*)

(*b*)

Figure 3 Cusum Plot for Inventory Model before and after Deleting Outputs.

```
5. Generate cusum plot: for $j \leftarrow 1$ to $k$: $s[j] \leftarrow s[j-1] + a - y[j]$, endfor
6. Output: plot $s[j]$ vs. $j$
```

The procedure *algorithm cusum* assumes that all of the outputs $y[t]$ are available before generating the plot. If an additional batch average, $y[k+1]$, is generated, then the plot values can be updated using the following steps:

```
$s[j] \leftarrow s[j] + j*(y[k+1] - a)/(k+1)$ for $j \leftarrow 1, 2, \ldots, k$
$s[k+1] \leftarrow 0$
$a \leftarrow (k * a + y[k+1])/(k+1)$
```

In the next subsection a test for initial-condition bias based on the characteristic behavior of the cusum plot is given.

7.2.3. Initial-Condition-Bias Test

Schruben (1982) used the characteristic behavior of the cusum plot to derive a statistical test for the presence of initial-condition bias. The test would typically be performed on an output process after remedial measures, such as data deletion, have been applied. The null hypothesis of the test is that there is no negative initial-condition bias (biased low) in the mean of the output process.

Loosely speaking, the test works as follows. The output of a single replication is divided in half. If the run is long enough, then any bias is most prevalent in the first half. The cusum values from each half are compared in terms of the location and magnitude of their maximum deviation from zero. If the behavior of the first half is significantly different from the second half, then the hypothesis of no initial-condition bias is rejected.

There are several versions of this test that have power against specific alternatives (Schruben et al. 1983; Goldsman et al. 1989). The version given in *algorithm bias test* is conservative in the sense that it tries to ensure that all of the asymptotic assumptions behind the test will be valid:

```
1. Initialization: length of the replication, $m$; batch size $b$; number of batches $k \leftarrow \lfloor m/b \rfloor$; arrays
   $a[j] \leftarrow 0$, $s[j] \leftarrow 0$, $smax[j] \leftarrow 0$ and $\ell[j] \leftarrow 0$ for $j = 1, 2$
   (comment: $m$ is problem dependent but should be large; Schruben recommends $b \leftarrow 5$)
2. Data: $y[t]$, the $t$th observation from a single replication
   (comment: test assumes negative bias; if positive bias, modify step 6 as indicated; both tests
   can be performed if the direction of bias is uncertain)
3. Batch the outputs:
   (a) for $j \leftarrow 1$ to $k$:
           $y[j] \leftarrow y[(j-1)*b + 1]$
           for $t \leftarrow 2$ to $b$: $y[j] \leftarrow y[j] + y[(j-1)*b + t]$, endfor
       endfor
   (b) for $j \leftarrow 1$ to $k$: $y[j] \leftarrow y[j]/b$, endfor
   (comment: batch means are now stored in $y[1], y[2], \ldots, y[k]$; if the original data must be
   saved, then a separate array should be used)
4. $n \leftarrow \lfloor k/2 \rfloor$
   (comment: $n$ is half the batched process; the algorithm ignores the last batch if $k$ is odd)
5. Calculate sample mean of each half:
   for $i \leftarrow 1$ to $n$:
       $a[1] \leftarrow a[1] + y[i]$
       $a[2] \leftarrow a[2] + y[n+i]$
   endfor
   for $i \leftarrow 1$ to 2: $a[i] \leftarrow a[i]/n$, endfor
6. Locate maximum from each half:
   (comment: if positive bias is suspected, replace all >'s with <'s in this step)
   for $i \leftarrow 1$ to $n$:
       $s[1] \leftarrow s[1] + a[1] - y[i]$
       if $s[1] > smax[1]$
           $\ell[1] \leftarrow i$ and $smax[1] \leftarrow s[1]$
       endif
       $s[2] \leftarrow s[2] + a[2] - y[n+i]$
       if $s[2] > smax[2]$
           $\ell[2] \leftarrow i$ and $smax[2] \leftarrow s[2]$
       endif
   endfor
```

```
7. Calculate test statistic:
   (comment: if $\ell[1] = 0$ test for bias of the opposite sign; if $smax[2] = 0$ then $m$ is too small)
   $f \leftarrow \ell[2] * (n - \ell[2]) * smax[1] * smax[1] / ((\ell[1] * (n - \ell[1]) * smax[2] * smax[2])$
8. If $f > 9.28$, then reject the hypothesis of no initial-condition bias
   (comment: 9.28 is the critical value for a 5% significance level; use 5.39 or 29.5 for 10% or 1% significance levels, respectively; to obtain other significance levels use the $F$ distribution with (3,3) DOF).
```

Applying the test to the output process in Figure 3 before data deletion yields an f value of 88.66, which is significant even at the 1% level. However, after deleting the first 20 outputs from the process the f value is only 2.93, which is not significant.

7.3. A Comment on Probabilities and Quantiles

Long-run probabilities and quantiles of an output process can be estimated from a single replication of a steady-state simulation using the estimators given in Section 6. However, the convergence of probabilities and quantiles to their limiting values is often much slower than convergence of the mean. Thus, approximate convergence of the mean cannot be used as a substitute for establishing the convergence of probabilities and quantiles if they are the parameters of primary interest.

8. MEASURES OF ERROR

Design and Analysis Principle 6. A point estimator should be accompanied by a measure of its potential error.

A simple example illustrates why this principle is so important: Measures of error can be interpreted as the number of meaningful digits in a point estimate. If the estimated expected cost of an inventory policy is 1.1 million, it probably matters whether or not the second digit is meaningful. Without a measure of error, however, the analyst cannot even be sure if the *first* digit has meaning! A true expected cost of 2.7 million may mean the company is out of business.

This section describes measures of error for estimates of a performance parameter, θ, that can be represented as a mean (expected value), probability or quantile; point estimators for such parameters are covered in Section 6.

Two measures of error are the *standard error* and *confidence interval*. Whether or not it is difficult to derive a measure of error is related to whether or not the experiment design calls for i.i.d. replications: static and terminating experiments always specify replications, but steady-state simulations may not, depending on the severity of the initial-condition bias (see Section 7).

8.1. Standard Error

The *standard error* of a point estimator is an average error, where the average is with respect to repeated experiments of the same type. More precisely, it is the standard deviation of the point estimator. This subsection begins with a basic result for estimating the standard error of a mean or probability estimate and then illustrates it using the examples.

Let $Y_1, Y_2, \ldots, Y_k$ be i.i.d. output random variables with mean (expected value) θ, and let $\bar{Y}$ be their average as defined in Section 6. Then the standard error of $\bar{Y}$ as an estimator of θ is $\sigma/\sqrt{k}$, where σ^2 is the common $\text{Var}[Y_i]$. An estimator of $\sigma/\sqrt{k}$ is

$$\frac{S}{\sqrt{k}} = \sqrt{\frac{\sum_{i=1}^{k} (Y_i - \bar{Y})^2}{k(k-1)}} = \sqrt{\frac{\sum_{i=1}^{k} Y_i^2 - (\sum_{i=1}^{k} Y_i)^2/k}{k(k-1)}} \tag{2}$$

where $S^2 = (k-1)^{-1} \sum_{i=1}^{k} (Y_i - \bar{Y})^2$ is the sample variance, an estimator of σ^2.

The estimator $S/\sqrt{k}$ is also valid when θ is a probability and the outputs are indicator variables I_i (as defined in Section 6), in which case (2) reduces to

$$\frac{S}{\sqrt{k}} = \sqrt{\frac{\bar{I}(1-\bar{I})}{k-1}}$$

For example, in the acceptance-sampling simulation I_i could be an indicator of whether or not the ith sample of size 10 was accepted (had 1 or fewer defectives), so that $\bar{I}$ estimates $\theta(p)$, the probability of accepting the sample. Table 3 gives the estimated standard errors for estimates of $\theta(p)$, where each estimate is based on $k = 10{,}000$ i.i.d. replications as described in Section 6.

TABLE 3 Estimated Points on OC Curve and Associated Standard Error of the Estimate

p	$\bar{I}$	$S/\sqrt{10000}$
0.00	1.00	0.000
0.01	0.98	0.001
0.05	0.89	0.003
0.10	0.77	0.004
0.15	0.67	0.005
0.20	0.57	0.005
0.30	0.38	0.005
0.40	0.25	0.004
0.50	0.16	0.004
0.60	0.09	0.003
0.70	0.05	0.002
0.80	0.02	0.001
0.90	0.01	0.001

Design and Analysis Principle 7. The first nonzero digit in the standard error indicates an uncertain digit in the corresponding decimal place of the point estimator.

This principle is conservative; it does not imply that the first uncertain digit in the point estimator has no meaning, but only that it may not be correct. Table 3 displays two decimal places for each $\bar{I}$ because the first nonzero digit of the standard error is in the third decimal place; however, the third decimal place of $\bar{I}$ was used for rounding.

Formula (2) is applicable when the point estimator is an average across k i.i.d. replications. As discussed in Section 5, it is seldom applicable to outputs within a single replication because they may be neither independent nor identically distributed. This causes no problem in static or terminating simulations because the natural experiment design is to generate independent replications.

Replications can also be generated in steady-state simulations, although remedial measures (such as deletion, see Section 7) must be applied to each replication. When initial-condition bias is not severe, replications are recommended and formula (2) can be applied to the across-replication statistics.

When only a single replication is generated from a steady-state simulation, an estimator of the standard error can be derived using the method of *batch means*. To illustrate the discussion, let $Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_m^{(1)}$ be the output from a single replication of the inventory simulation for policy 1 *after* the first 20 periods have been deleted; thus, $Y_t^{(1)}$ is the observed cost in period $t + 20$. The sample average, $\bar{Y}^{(1)}$, is used to estimate $\theta^{(1)}$, the long-run expected cost per period for inventory policy 1. An estimate of the standard error of $\bar{Y}^{(1)}$ is required. For convenience, drop the superscript (1) and let the output and statistic be simply Y_t and $\bar{Y}$, respectively.

Let $k \leq m$ be the *number of batches*, $b = \lfloor m/k \rfloor$ the *batch size*, and define the jth batch mean (average) to be

$$\bar{Y}_j(k) = \frac{1}{b} \sum_{t=1}^{b} Y_{(j-1)b+t}$$

for $j = 1, 2, \ldots, k$. The principle behind batch means is that $\bar{Y}_1(k), \bar{Y}_2(k), \ldots, \bar{Y}_k(k)$ are more nearly independent than the original outputs, $Y_1, Y_2, \ldots, Y_m$, for most simulation output processes. If the remedial measures for initial-condition bias have been effective, then the batch means are also nearly identically distributed. Thus, the batch means are (approximately) i.i.d. statistics, even though they are within-replication statistics, and formula (2) applies. The batch means are written as functions of k because selecting the number of batches is a decision that the analyst must make.

The procedure below, called *algorithm batch_means,* computes an estimate of the standard error of a within-replication average $\bar{Y}$. It can be used to compute the standard error for the across-replication average of k replications (formula (2)) by omitting step 3(a) and setting $m = k$:

1. *Initialization:* length of the replication, m; number of batches k; batch size $b \leftarrow \lfloor m/k \rfloor$; *sum* $\leftarrow 0$; and *sumsq* $\leftarrow 0$.
(comment: m is problem dependent but should be large; choosing k is discussed below)
2. *Data:* $y[t]$, the tth observation from a single replication *after* applying remedial measures.

```
3. Batch the outputs:
   (a) for $j \leftarrow 1$ to $k$:
         $y[j] \leftarrow y[(j-1)*b+1]$
         for $t \leftarrow 2$ to $b$: $y[j] \leftarrow y[j] + y[(j-1)*b+t]$, endfor
       endfor
   (b) Compute the standard error:
       for $j \leftarrow 1$ to $k$:
         $y[j] \leftarrow y[j]/b$
         $sum \leftarrow sum + y[j]$
         $sumsq \leftarrow sumsq + y[j]*y[j]$
       endfor
       (comment: batch means are now stored in $y[1], y[2], \ldots, y[k]$; if the original data must be
   saved, then a separate array should be used)
4. Output: $\mathrm{sqrt}\{(sumsq - sum*sum/k)/(k(k-1))\}$.
```

To illustrate the effect of batching, 2020 periods of the inventory simulation were generated and the first 20 periods of data were deleted, leaving $m = 2000$ outputs. The sample mean of the outputs was $\bar{Y} = 118.48$ dollars per period. The estimated sample correlation between successive periods, $\mathrm{Corr}[Y_t, Y_{t+1}]$, was -0.51, clearly showing dependence. After batching the cost data into $k = 20$ batches of size $b = 100$, the estimated correlation between successive batch means, $\mathrm{Corr}[\bar{Y}_j(20), \bar{Y}_{j+1}(20)]$, was -0.09, which appears less correlated. The standard error of $\bar{Y}$ calculated by *algorithm batch means* was 0.29, an average error of 29 cents.

The most difficult design decision when using the method of batch means is choosing k, the number of batches. Several studies have concluded that no matter how long the replication is, it should be divided into a relatively small number of batches, say $10 \le k \le 30$. Diagnostic tests, such as computing the correlation between successive observations, can be employed to verify that batching has been effective. Formal batching algorithms are also available (Bratley et al. 1987; Fishman 1978; Fishman and Yarberry 1997; Law and Kelton 2000).

8.2. Confidence Intervals

Confidence intervals are constructed in hopes that they contain the unknown performance parameter of interest, θ. The "confidence" associated with a confidence interval is, strictly speaking, a statement about the *procedure* used to construct the interval, not the interval itself. The confidence interval either contains θ or it does not. The width of the confidence interval is a measure of error.

There are many confidence-interval procedures based on a variety of different assumptions, too many to describe here (see Chapter 86 of the Handbook or any introductory statistics text). This subsection contains some general comments regarding the use of confidence-interval procedures for estimating means, and a specific confidence-interval procedure for quantile estimation.

8.2.1. Confidence Intervals for Means

The validity of a confidence-interval procedure frequently depends on a number of conditions. One such condition is that the output data are i.i.d. As discussed in Subsection 8.1, i.i.d. outputs can be obtained by generating replications, or, in steady-state simulation, by using the method of batch means.

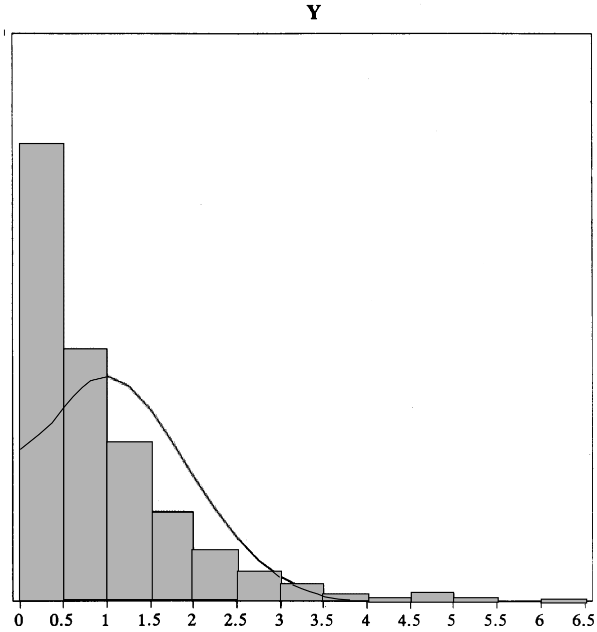
A second common condition is that the simulation outputs have a normal distribution. The assumption of normality can be severely violated in simulations. Diagnostic checks using histograms, quantile plots, or hypothesis tests are recommended to verify the normality of the output data.

The normality of the output data can be improved by using the method of batch means, even if the data are already i.i.d. This is because averages tend to be normally distributed, as shown in the central limit theorem. For example, Figure 4 shows two histograms for the same 500 i.i.d. simulation outputs. The first histogram shows raw data, while the second displays the same data after batching them into $k = 100$ batch means (batch size $b = 5$). In both cases a normal curve is superimposed over the histogram. Clearly the sample distribution of the batch means is more nearly normal.

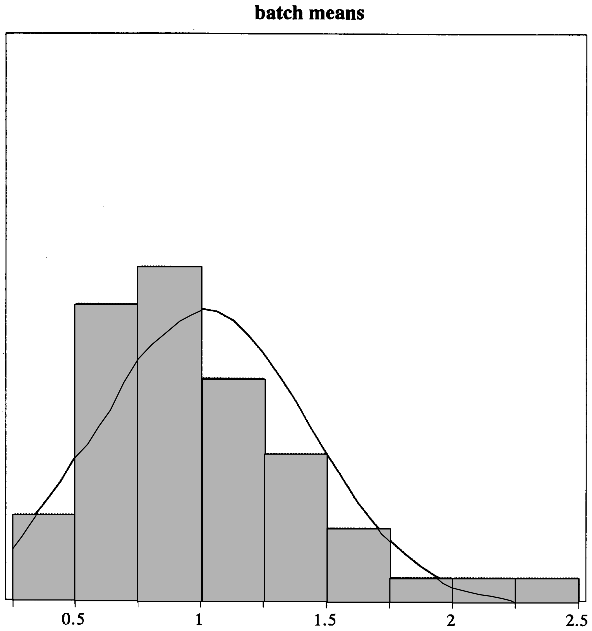
Although batching reduces the number of degrees of freedom for the confidence-interval procedure—since there are fewer batch means than total outputs—the penalty is slight provided that the number of batch means is not too small. Schmeiser (1982) showed that, for the standard t confidence-interval procedure, batching to $10 \le k \le 30$ batch means does not degrade the performance of the procedure substantially *even if the original data are precisely i.i.d. normal.* On the other hand, when the data are not normal, batching can improve the performance of the procedure in terms of delivering the requested level of confidence.

8.2.2. A Confidence Interval for Quantiles

A point estimator for the q quantile, θ_q, of a random variable Y was given in Section 6. The point estimator and the confidence-interval procedure given here assume that k i.i.d. replications of Y are



(*a*)



(*b*)

Figure 4 Simulation Outputs before and after Batching.

TABLE 4 Standard Normal Quantiles

$1 - \alpha$	$z_{1-\alpha/2}$
0.90	1.6449
0.95	1.9600
0.99	2.5758

available, denoted generically by $Y_1, Y_2, \ldots, Y_k$. Let the sorted values of Y, called the order statistics, be denoted $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(k)}$.

Although a standard error estimate is difficult to derive, deriving a confidence interval for θ_q is relatively easy. In addition, this confidence-interval procedure is nonparametric, meaning that the distribution of Y is not a factor.

A $(1 - \alpha)100\%$ confidence interval for θ_q is $(Y_{(\ell)}, Y_{(u)})$, where $1 \leq \ell < u \leq k$ are integers such that the binomial probability

$$\sum_{j=\ell}^{u-1} \binom{k}{j} q^j (1 - q)^{k-j} \approx 1 - \alpha$$

The confidence interval is nonparametric since the constants ℓ and u depend only on k and q. Determining ℓ and u is difficult when the number of replications, k, is large (and k should be large for quantile estimation), but large k allows a normal approximation to the binomial distribution. The approximation leads to setting

$$\ell = \lfloor kq - z_{1-\alpha/2}\sqrt{kq(1 - q)} + 1/2 \rfloor \text{ and} \tag{3}$$

$$u = \lfloor kq + z_{1-\alpha/2}\sqrt{kq(1 - q)} + 1/2 \rfloor + 1 \tag{4}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. The standard normal quantiles for 90%, 95%, and 99% confidence intervals are given in Table 4.

Section 6 illustrated estimating the 0.9 quantile, $\theta_{0.9}$, of $\bar{Y}$, the average delay for renewal applications received during a month in the license-renewal simulation. Based on 200 replications, the point estimate was 0.88 days. Substituting $k = 200$ and $q = 0.9$ into (3) and (4) gives $\ell = 172$ and $u = 189$ for a 95% confidence interval. The corresponding interval is $(Y_{(172)}, Y_{(189)}) = (0.82, 0.98)$.

8.3. A Note on Experiment Planning

The discussion above assumes that the analyst is interested in computing a measure of error *after* completing the simulation replications or runs. Experiment planning includes determining the number of replications required to achieve a prespecified level of error.

Consider the standard error. If the process variance σ^2 is known, then it is easy to determine the number of replications (or batch means) required to achieve a standard error less than or equal to, say, δ, by solving

$$\frac{\sigma}{\sqrt{k}} \leq \delta \tag{5}$$

for k. Typically, σ^2 is not known, but it may be estimated during debugging or preliminary pilot runs and the estimate can be substituted into (5).

Law and Kelton (2000) contains a discussion of methods for determining k sequentially (while the simulation is in progress) to guarantee a prespecified level of error.

9. OPTIMIZATION AND SENSITIVITY

Industrial engineers frequently use simulation experiments to compare the performance of alternative systems and, ideally, to optimize system performance. When a system is modeled as a stochastic process, the objective is often to optimize expected performance, where "expected" means the mathematical expectation of a random variable. This section describes methods for optimization via simulation, using the problem of selecting the inventory policy that minimizes long-run expected cost per period as an illustration.

Even if optimization is not desired or feasible, the analyst may be interested in the sensitivity of system performance to certain controllable factors. This section also discusses sensitivity analysis, using the sensitivity of the expected average delay $\theta(\mu)$ to the processing rate μ in the license-renewal simulation as an illustration.

The scope of this section is limited because methods for optimization and sensitivity analysis are still evolving. See Andradóttir (1998).

9.1. Multiple Comparisons

A common optimization problem is to choose the system with maximum or minimum expected performance from among a small number of systems, say 2 to 10. Assuming that estimates of expected performance are available from simulation experiments, the question of which system to select is easily answered: all else being equal, select the system with the sample best performance. A more relevant question is whether the observed ranking of the systems is due to estimation error or actual differences in performance. Thus, optimization is an extension of the problem of estimating and controlling error, as discussed in Section 8.

The case of comparing two systems is covered in standard statistics texts. For three or more systems, two classes of statistical procedures are widely used: ranking-and-selection procedures and multiple-comparison procedures.

Ranking-and-selection procedures treat the optimization problem as a decision problem, typically either deciding which system is the best (indifference-zone ranking) or on a subset of systems that contains the best system (subset selection). The decisions are guaranteed to be correct with a prespecified probability. Achieving this goal often requires two-stage sampling, which means restarting simulation experiments after initial runs of all systems. A summary of the many ranking and selection procedures is given by Bechhofer et al. (1995), Gupta and Panchapakesan (1979), and Law and Kelton (2000). Extensions of these procedures have been made specifically to stochastic simulation (Koenig and Law 1985; Clark and Yang 1986; Goldsman 1985; Goldsman and Nelson 1998; Iglehart 1977; Sullivan and Wilson 1989).

Multiple-comparison procedures, on the other hand, treat the optimization problem as an inference problem on the performance parameters of interest. An important property of multiple-comparison procedures is that inference about the relative performance of all systems is provided. In addition, multiple-comparison procedures can be implemented in a single-stage of sampling. This subsection describes a multiple-comparison procedure that is useful for optimization.

In the inventory simulation smaller $\theta^{(\ell)}$ is preferred, where $\theta^{(\ell)}$ denotes the long-run expected cost per period for policy ℓ, $\ell = 1, 2, \ldots, 5$. For policy ℓ, the parameter $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)}$ is policy ℓ cost minus the minimum of the other policies' costs. The parameters $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)}$, for $\ell = 1, \ldots, 5$, are the appropriate parameters to estimate for optimization. If $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)} < 0$, then policy ℓ is the best because all other policies have larger expected cost. On the other hand, if $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)} > 0$, then policy ℓ is not the best, since there is another policy with smaller expected cost. Even when $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)} > 0$, if $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)} < \Delta$, where Δ is a positive number, then policy ℓ is within Δ of the best. Simultaneous statistical inference on $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)}$, for $\ell = 1, \ldots, 5$, is termed *multiple comparisons with the best* (MCB). In problems where larger expected performance is preferred, the parameter of interest is $\theta^{(\ell)} - \max_{i\neq\ell}\theta^{(i)}$, so that $\theta^{(\ell)} - \max_{i\neq\ell}\theta^{(i)} > 0$ indicates the best system.

For the general case of r systems, Hsu (1984) derived simultaneous $(1 - \alpha)100\%$ confidence intervals for $\theta^{(\ell)} - \max_{i\neq\ell}\theta^{(i)}$ (equivalently $\theta^{(\ell)} - \min_{i\neq\ell}\theta^{(i)}$) for $\ell = 1, 2, \ldots, r$. The form of the intervals is

$$\left[(\overline{Y}^{(\ell)} - \max_{i\neq\ell}\overline{Y}^{(i)} - H)^-, (\overline{Y}^{(\ell)} - \max_{i\neq\ell}\overline{Y}^{(i)} + H)^+\right]$$

where $\overline{Y}^{(\ell)}$ is the sample average of the outputs from system ℓ, $x^- = \min\{0, x\}$, $x^+ = \max\{0, x\}$ and H is a quantity that depends on r, k and the sample standard error.

The assumptions behind these confidence intervals, and their interpretation in simulation experiments, are listed below. To aid the explanation let $Y_j^{(\ell)}$ be generic for the jth output from system ℓ.

- For each system ℓ, the output data consist of k i.i.d. outputs $Y_1^{(\ell)}, Y_2^{(\ell)}, \ldots, Y_k^{(\ell)}$ with common expected value $\theta^{(\ell)}$. This will be true if the outputs are from across replications, or approximately true if they are batch means from within a single replication (see Section 8).
- The outputs from different systems $\ell = 1, 2, \ldots, r$ are independent of each other. This will be true if different random number streams or seeds are specified for the simulation of each system (see Section 4).
- All the outputs $Y_j^{(\ell)}$ $j = 1, 2, \ldots, k$ and $\ell = 1, 2, \ldots, r$ are normally distributed. This may be approximately true if $Y_j^{(\ell)}$ is an average of many outputs, but it should be verified through graphical or statistical analysis.
- $\mathrm{Var}[Y_j^{(\ell)}]$ is the same for all j and ℓ. This may be true if the different systems are similar, but it should be verified through graphical or statistical analysis. If the variances across systems are widely different then variance-stabilizing transformations can be applied.

The procedure below, called *algorithm mcb*, calculates MCB intervals for the maximization case; to obtain results for the minimization case replace all the >'s with <'s in step 6 of the algorithm as indicated.

```
1. Initialization: number of independent observations, $k$; number of systems, $r$; arrays $sum[\ell] \leftarrow 0$ and $sumsq[\ell] \leftarrow 0$ for $\ell = 1, 2, \ldots, r$; $s \leftarrow 0$; array $ybar[\ell]$ of length $r$; critical value $d \leftarrow d^{\alpha}_{r(k-1),r}$ (see Table 5)
2. Data: $y[\ell, j]$, the $j$th independent observation from system $\ell$
3. Compute sums and sums of squares:
   for $\ell \leftarrow 1$ to $r$:
      for $j \leftarrow 1$ to $k$:
         $sum[\ell] \leftarrow sum[\ell] + y[\ell, j]$
         $sumsq[\ell] \leftarrow sumsq[\ell] + y[\ell, j]*y[\ell, j]$
      endfor
   endfor
4. Compute sample averages and pooled standard error estimate:
   for $\ell \leftarrow 1$ to $r$:
      $ybar[\ell] \leftarrow sum[\ell]/k$
      $s \leftarrow s + sumsq[\ell] - sum[\ell]*sum[\ell]/k$
   endfor
   $s \leftarrow \mathrm{sqrt}\{s/(r(k-1))\}$
5. Confidence interval halfwidth: $h \leftarrow d * s * \mathrm{sqrt}(2/k)$
6. Find first- and second-best sample performance:
   (comment: for minimization problems replace all >'s by <'s in this step)
   $findex \leftarrow 1$; $first \leftarrow ybar[1]$
   $sindex \leftarrow 2$; $second \leftarrow ybar[2]$
   if $ybar[2] > ybar[1]$
      $findex \leftarrow 2$; $first \leftarrow ybar[2]$
      $sindex \leftarrow 1$; $second \leftarrow ybar[1]$
```

TABLE 5 Critical Values $d^{0.05}_{r(k-1),r}$ for MCB Intervals at 95% Confidence[a]

	Number of Systems, r							
k	3	4	5	6	7	8	9	10
2	2.938	2.885	2.849	2.827	2.815	2.809	2.807	2.807
3	2.337	2.417	2.466	2.502	2.532	2.558	2.581	2.601
4	2.180	2.287	2.356	2.407	2.448	2.482	2.511	2.537
5	2.108	2.227	2.305	2.362	2.408	2.445	2.478	2.506
6	2.067	2.193	2.274	2.335	2.384	2.424	2.458	2.488
7	2.041	2.170	2.255	2.318	2.368	2.410	2.445	2.476
8	2.022	2.154	2.241	2.306	2.357	2.400	2.436	2.467
9	2.008	2.142	2.230	2.296	2.349	2.392	2.429	2.461
10	1.998	2.133	2.222	2.289	2.342	2.386	2.424	2.456
11	1.989	2.126	2.216	2.284	2.337	2.382	2.419	2.452
12	1.982	2.120	2.211	2.280	2.333	2.378	2.416	2.449
13	1.977	2.115	2.207	2.275	2.330	2.375	2.413	2.446
14	1.972	2.111	2.203	2.272	2.327	2.372	2.411	2.444
15	1.968	2.107	2.200	2.269	2.324	2.370	2.408	2.442
16	1.964	2.104	2.197	2.267	2.322	2.368	2.407	2.440
17	1.961	2.101	2.195	2.265	2.320	2.366	2.405	2.439
18	1.959	2.099	2.193	2.263	2.319	2.365	2.404	2.438
19	1.956	2.097	2.191	2.261	2.317	2.363	2.402	2.437
20	1.954	2.095	2.189	2.260	2.316	2.362	2.401	2.435
30	1.941	2.084	2.180	2.251	2.308	2.355	2.394	2.429
40	1.935	2.078	2.175	2.246	2.304	2.351	2.391	2.426
50	1.931	2.075	2.172	2.244	2.301	2.349	2.389	2.424
60	1.928	2.073	2.170	2.242	2.300	2.347	2.388	2.423
120	1.922	2.067	2.165	2.238	2.296	2.344	2.384	2.420
∞	1.916	2.062	2.161	2.234	2.292	2.341	2.381	2.417

[a] These values were calculated using a program provided by Jason C. Hsu

```
    endif
    for $\ell \leftarrow 3$ to $r$:
        if $ybar[\ell] > first$
            $sindex \leftarrow findex$; $second \leftarrow first$
            $findex \leftarrow \ell$; $first \leftarrow ybar[\ell]$
        else
            if $ybar[\ell] > second$
                $sindex \leftarrow \ell$; $second \leftarrow ybar[\ell]$
            endif
        endif
    endfor
7. Output: For system $\ell$, the lower endpoint, point estimate, and upper endpoint for $\theta^{(\ell)} - \max_{i \neq \ell} \theta^{(i)}$ (or $\theta^{(\ell)} - \min_{i \neq \ell} \theta^{(i)}$)
    for $\ell \leftarrow 1$ to $r$:
        $point \leftarrow ybar[\ell] - first$
    if $\ell = findex$ then $point \leftarrow ybar[\ell] - second$
        $lower \leftarrow \min\{point - h, 0\}$
        $upper \leftarrow \max\{point + h, 0\}$
        output $\ell$, $lower$, $point$, $upper$
    endfor
```

The critical values $d^{\alpha}_{r(k-1),r}$ for 95% confidence intervals ($\alpha = 0.05$) are given in Table 5. Critical values for other confidence levels can be found in Table 4 of Appendix 3 in Hochberg and Tamhane (1987).

MCB intervals are constrained intervals, meaning that they either contain 0 or one of their endpoints is 0. In a maximization problem, if the confidence interval for $\theta^{(\ell)} - \max_{i \neq \ell} \theta^{(i)}$ contains 0 it means that system ℓ is not significantly different from the best system, and may be the best. If the upper endpoint of the interval is 0, then system ℓ is not the best system. However, if the lower endpoint is 0, then system ℓ is the best system; at most one system will have lower endpoint 0. These statements are made with confidence level $1 - \alpha$.

The minimization case is illustrated by the inventory example. For each inventory policy a single replication of 2020 periods was generated and the first 20 periods of data were deleted to reduce initial-condition bias (see Section 7). A different random number stream was used to generate demands for each policy so that the results across policies are independent. To obtain approximately i.i.d. outputs from within each replication, the 2000 outputs were batched into $k = 20$ batch means, and the batch means were used as the basic data for each inventory policy (see Section 8). Thus, the critical value from Table 5 is $d^{0.05}_{5(19),5} = 2.189$, which comes from the $k = 20$ row and $r = 5$ column. The 95% simultaneous confidence intervals for $\theta^{(\ell)} - \min_{i \neq \ell} \theta^{(i)}$ computed by *algorithm mcb* are given in Table 6.

Since the lower endpoints for policies 3, 4, and 5 are 0, they are significantly more expensive than the least-cost policy; in other words, the difference between the cost of any one of these policies and the minimum cost of the other policies is greater than 0. Policy 2 is the sample best policy, but it is not statistically significantly different from policy 1 since the intervals for both policies 1 and 2 contain 0. However, if policy 2 is not the best, the upper endpoint of its interval indicates that it is no more than 1.836 per period more expensive than the best. Similarly, policy 1 is no more than 2.526 from the best. If these differences are practically insignificant, then policy 2 should be selected. If more precision is required, then additional replications or batches will sharpen the comparison.

9.2. Metamodels

In some simulation studies, the analyst is interested in determining the effects of controllable system factors, perhaps to optimize system performance with respect to the controllable factors or to examine

TABLE 6 MCB Intervals for $\theta^{(\ell)} - \min_{i \neq \ell} \theta^{(i)}$ in the Inventory Simulation

ℓ	Policy	Lower Endpoint	Point Estimate	Upper Endpoint
1	(20,40)	−1.836	0.345	2.526
2	(20,80)	−2.526	−0.345	1.836
3	(40,60)	0	15.409	17.590
4	(40,100)	0	18.574	20.755
5	(60,100)	0	34.064	36.245

their relative impact. For example, in the license-renewal simulation, the application processing rate, μ, may be controllable through the equipment and staff available to the clerk, and the analyst may be interested in its effect on the average processing delay for applications.

The sample performance of a simulated system can be viewed as a complicated function of the controllable factors, which include the random number seeds or streams assigned to the experiment. A *metamodel* is a simpler function that approximates the relationship between system performance and the controllable factors. The primary benefit of a metamodel is that it can be studied using straightforward mathematical analysis.

One approach that is often used to derive a metamodel is to propose a function that has some unknown parameters and then estimate the parameters via least-squares regression. For example, the following metamodel might be postulated to describe the relationship between the average monthly delay for renewal applications, $\overline{Y}$, and the processing rate, μ:

$$\overline{Y} = \beta_0 + \beta_1\mu + \beta_2\mu^2 + \epsilon(s) \tag{6}$$

where β_0, β_1, and β_2 are unknown parameters and $\epsilon(s)$ is a mean 0 random variable that represents the variability in the system. Notice that $\epsilon(s)$ is a function of s, the random number stream or seed assigned to the simulation experiment. Metamodel (6) implies that $\theta(\mu)$, the expected average delay when the processing rate is μ, can be approximated by the function

$$\theta(\mu) = \mathrm{E}_\mu[\overline{Y}] \approx \beta_0 + \beta_1\mu + \beta_2\mu^2$$

Metamodels, such as (6), become useful when appropriate values are assigned to the unknown parameters. Experimental design—choosing the levels of the controllable factors at which simulation experiments will be performed—is needed to ensure good parameter estimates; see Chapter 85 of the Handbook for guidance. Once the experiments have been performed, least-squares regression is a standard procedure for estimating the parameters; see Chapter 87 of the Handbook.

The methods and procedures discussed in Chapters 85 and 87 are as relevant to simulation experiments as they are to statistical experiments in general. There is one experiment design and analysis issue that is unique to simulation, however:

Design and Analysis Principle 8. Statistical inference on a metamodel is typically based on the assumption that outputs from different design points (factor settings), and the replications at a single design point, are statistically independent. Thus, different random number streams or seeds should be assigned to each design point unless the analyst plans to account directly for the dependence induced by using common streams.

Although there are potential advantages from using common streams or seeds across design points, methods for statistical analysis under such designs are still evolving (Hussey, et al. 1987a, b; Kleijnen 1988; Nozari, et al. 1987; Schruben and Margolin 1978). See also Section 10.

Table 7 shows the experiment design used to fit (6). Ten independent replications were made at each design point, and the parameters were estimated via least-squares regression. The resulting metamodel for the expected average delay is

$$\theta(\mu) \approx 12.01 - 0.97\mu + 0.02\mu^2 \tag{7}$$

Because each design point was simulated independently, standard tests for significance of the parameters and lack of fit could be used to validate the fitted model.

The derivative of (7) with respect to μ gives the approximation

TABLE 7 Experiment Design for Fitting License-Renewal Processing Metamodel

Design Point	Processing Rate, μ	Random Number Stream
1	15	3
2	20	1
3	25	4

$$\frac{d\theta(\mu)}{d\mu} \approx -0.97 + 0.04\mu$$

The derivative of a metamodel is one way to examine the sensitivity of the expected average delay to the processing rate. For example, at $\mu = 20$ the derivative is -0.17, implying that an increase of one application per day in the processing rate should result in a decrease of 0.17 days in the expected average delay.

The extent to which a metamodel is valid is necessarily limited. Extrapolating a metamodel beyond the factor settings used to fit the model should always be done with caution. For instance, the metamodel (7) predicts that the expected average delay will begin to increase if the processing rate exceeds 25 applications per day!

Many methods, in addition to metamodel estimation, have been proposed for examining the sensitivity of system performance to controllable factors. For recent surveys see Andradóttir (1998), Glynn (1989), and Jacobson and Schruben (1989) and the references therein.

10. VARIANCE REDUCTION

The goal of variance reduction is to decrease the error of a point estimate, which should lead to a smaller estimated standard error or a narrower confidence interval (see Section 8 for a discussion of measures of error). This section describes the variance reduction technique known as *common random numbers* (CRN), which is useful for reducing the error in comparing the expected performance of two or more systems. The presentation is based on Nelson (1987).

For simplicity, suppose that there are two competing systems, and let $\theta^{(1)}$ and $\theta^{(2)}$ be the (unknown) expected performance of system 1 and system 2, respectively. If there are more than two systems, then the methods described below can be applied to each pair of systems. When comparing the performance of two systems, the single parameter $\theta^{(1)} - \theta^{(2)}$ often captures the essential information about their relative performance.

For example, consider comparing the expected cost per period of inventory policies 1 and 2 of the inventory example. Since a smaller expected cost is preferred, $\theta^{(1)} - \theta^{(2)} < 0$ implies that inventory policy 1 is superior, while $\theta^{(1)} - \theta^{(2)} > 0$ implies that inventory policy 2 is better. The absolute value of $\theta^{(1)} - \theta^{(2)}$ is the amount by which one of the policies is better than the other.

Generically, let $Y_1^{(\ell)}, Y_2^{(\ell)}, \ldots, Y_k^{(\ell)}$ be outputs from k i.i.d. replications (or possibly batch means, see Section 8) from system ℓ, for $\ell = 1, 2$, so that $\theta^{(\ell)} = \mathrm{E}[Y_j^{(\ell)}]$; that is, $Y_j^{(\ell)}$ is the sample performance of system ℓ on replication j. Let $D_j = Y_j^{(1)} - Y_j^{(2)}$ for $j = 1, 2, \ldots, k$, the difference between the sample performance of the two systems on the jth replication. A point estimator of $\theta^{(1)} - \theta^{(2)}$ is

$$\overline{D} = \frac{1}{k}\sum_{j=1}^{k} D_j = \overline{Y}^{(1)} - \overline{Y}^{(2)}$$

An estimator of the standard error of $\overline{D}$ is $S_D/\sqrt{k}$, and an approximate $(1 - \alpha)100\%$ confidence interval for $\theta^{(1)} - \theta^{(2)}$ is

$$\overline{D} \pm t_{1-\alpha/2,k-1} S_D/k$$

where

$$S_D^2 = \frac{1}{k-1}\sum_{j=1}^{k} (D_j - \overline{D})^2$$

is the sample variance of the differences and $t_{1-\alpha/2,k-1}$ is the $1 - \alpha/2$ quantile of the t distribution with $k - 1$ degrees of freedom (a table of t-distribution quantiles can be found in any statistics text; if no such text is available, the normal distribution quantiles $z_{1-\alpha/2}$ from Table 4 can be used provided $k \geq 30$). The condition that makes this confidence interval valid is that the D_j are i.i.d. normally distributed random variables. If the $Y_j^{(\ell)}$ are actually within-replication averages or batch means, then this condition may be approximately satisfied.

The size of the estimated standard error or width of the confidence interval determines whether or not a difference in system performance can be detected in the presence of the random variation in the simulation results. To be able to detect a difference, it is necessary that the potential error in the point estimate $\overline{D}$ be small with respect to the difference $\theta^{(1)} - \theta^{(2)}$.

The true standard error of $\overline{D}$ can be written as $\sigma_D/\sqrt{k}$, where $\sigma_D^2 = \mathrm{Var}[D_j]$, the common variance of the differences. Thus, the size of the estimated standard error and the width of the confidence interval depend on the underlying variance σ_D^2. The variance of the difference is

$$\sigma_D^2 = \mathrm{Var}\,[Y_j^{(1)}] + \mathrm{Var}\,[Y_j^{(2)}] - 2\mathrm{Cov}[Y_j^{(1)}, Y_j^{(2)}]$$
$$= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \qquad (8)$$

where $\sigma_\ell^2 = \mathrm{Var}\,[Y_j^{(\ell)}]$ and ρ is the correlation between pairs of observations from the two systems on the same replication, $Y_j^{(1)}$ and $Y_j^{(2)}$. Typically, the variances σ_1^2 and σ_2^2 are properties of the systems being simulated and are not easily altered. However, the correlation ρ may be controlled, and Eq. (8) shows that making $\rho > 0$ will result in a variance reduction. The following design and analysis principle describes how to accomplish this:

Design and Analysis Principle 9. Positive correlation between the sample performance of two systems can often be induced by assigning the same ("common") random number streams or seeds to the simulation of each system (see Section 4 for a description of random number streams and seeds). The magnitude of the correlation can be increased further by synchronizing the pseudorandom numbers, as described below.

The intuition behind the use of CRN is straightforward: a fairer comparison between competing systems is obtained if they are both subjected to the same experimental conditions. Assigning the same random number streams or seeds to each system means that the influence of randomness is similar for both systems and the observed differences in performance are due to differences in the systems, not in the pseudorandom numbers.

The goal of synchronization is to guarantee that each pseudorandom number is used *for the same purpose* in the simulation of each system. The effect of CRN can be greatly enhanced by synchronizing the pseudorandom numbers. Synchronization is achieved through the random number stream assignments to the input processes.

Simulations contain one or more input processes, which are frequently i.i.d. sequences of random variables (see Section 4 for the definition of simulation inputs). For example, in the inventory simulation there is one input process, the sequence of demands in each period. In the license-renewal simulation there are two input processes, the successive times between arrivals of applications and the successive processing times for applications.

One of the best ways to synchronize is to dedicate a distinct random number stream or seed to each input process and then to use the same stream assignment for all simulated systems. In the license-renewal simulation, assigning different random number streams to the two input processes guarantees that exactly the same sequence of application arrival times is experienced by all systems and that the processing times will be positively correlated for all processing rates, μ. In the inventory simulation, CRN guarantees that each inventory policy experiences exactly the same sequence of demands. Synchronization is the primary reason for having more than one random number stream in a simulation language.

Table 8 illustrates the effect of CRN on the inventory simulation when the goal is to estimate $\theta^{(1)} - \theta^{(2)}$, the expected difference between the cost per period of inventory policies 1 and 2. The experiment design is the same one described in Section 9, so the basic data for each inventory policy are $k = 20$ approximately i.i.d. normal batch means (for the purposes of this illustration they can be thought of as k i.i.d. replications). The table gives the point estimate, estimated standard error, and a 95% confidence interval for the expected difference.

The estimated correlation between outputs when using CRN was 0.58, which is positive as desired. Notice that the estimated standard error and the width of the confidence interval are reduced relative to the experiment with independent runs (different random number streams for each policy).

More critically, the point estimate derived without CRN is positive, indicating that policy 2 yields the smallest expected cost, while the point estimate derived with CRN is negative, indicating that policy 1 is superior. For this simple model it can be shown that policy 1 does have the smaller expected cost, so the use of CRN leads to the correct decision. However, in both experiments the 95% confidence interval for the difference contains 0, so it cannot be stated conclusively that policy 1 is best based on the data. Additional batches are needed to reduce the error even further.

TABLE 8 Estimated Difference in Expected Cost With and Without CRN

Design	D	Standard Error	95% Confidence Interval
Independent	0.3	0.57	(−0.85, 1.54)
CRN	−0.3	0.31	(−0.98, 0.35)

ACKNOWLEDGEMENTS

The author gratefully acknowledges the helpful comments and recommendations of Gordon Clark, Lynne Goldsman, Jason C. Hsu, W. David Kelton, A. Alan B. Pritsker, Charles Reilly, Bruce Schmeiser, Lee Schruben, Fataneh Taghaboni, Mingjian Yuan, and the 1990 IND ENG 854 class at Ohio State on the 2d edition. He also acknowledges the support of NSF Grant DMI-9622065.

REFERENCES

Andradóttir, S. (1998), "Simulation Optimization," in *Handbook of Simulation,* John Wiley & Sons, New York.

Banks, J., Carson, J. S., Nelson, B. L., and Nicol, D. M. (2000), *Discrete-Event System Simulation,* 3rd Ed., Prentice Hall, Upper Saddle River, NJ.

Barton, R. R., and Schruben, L. W. (1989), "Graphical Methods for the Design and Analysis of Simulation Experiments," in *Proceedings of the 1989 Winter Simulation Conference,* E. MacNair, K. Musselman, and P. Heidelberger, Eds., pp. 51–61.

Bechhofer, R. E., Santner, T. J., and Goldsman D. (1995), *Design and Analysis for Statistical Selection, Screening and Multiple Comparisons,* John Wiley & Sons, New York.

Bratley, P., Fox, B. L., and Schrage, L. E. (1987), *A Guide to Simulation,* 2nd Ed., Springer, New York.

Clark, G. M., and Yang, W. (1986), "A Bonferroni Selection Procedure when Using Common Random Numbers with Unknown Variances," in *Proceedings of the 1986 Winter Simulation Conference,* J. Wilson, J. Henriksen, and S. Roberts, Eds., pp. 313–315.

Fishman, G. S. (1978), *Principles of Discrete Event Simulation,* John Wiley & Sons, New York.

Fishman, G. S., and Yarberry, L. S. (1997), "An Implementation of the Batch Means Method," *INFORMS Journal on Computing* Vol. 9, pp. 296–310.

Glynn, P. W. (1989), "Optimization of Stochastic Systems via Simulation," in *Proceedings of the 1989 Winter Simulation Conference,* E. MacNair, K. Musselman, and P. Heidelberger, Eds., pp. 90–105.

Goldsman, D. (1985), "Ranking and Selection Procedures Using Standardized Time Series," in *Proceedings of the 1985 Winter Simulation Conference,* D. Gantz, G. Blais, and S. Solomon, Eds., pp. 120–124.

Goldsman, D., and Nelson, B. L. (1998), "Comparing Systems via Simulation," in *Handbook of Simulation,* John Wiley & Sons, New York.

Goldsman, D., Schruben, L. and Swain, J. J. (1989), "Tests for Transient Means in Simulated Time Series," *Naval Research Logistics,* Vol. 41, pp. 171–187.

Gupta, S. S. and Panchapakesan, S. (1979), *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations,* John Wiley & Sons, New York.

Hochberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures,* John Wiley & Sons, New York.

Hsu, J. C. (1984), "Constrained Two-Sided Simultaneous Confidence Intervals for Multiple Comparisons with the Best," *Annals of Statistics,* Vol. 12, pp. 1136–1144.

Hussey, J. R., Myers, R. H., and Houck, E. C. (1987a), "Correlated Simulation Experiments in First-Order Response Surface Design," *Operations Research,* Vol. 35, pp. 744–758.

Hussey, J. R., Myers, R. H., and Houck, E. C. (1987b), "Pseudorandom Number Assignment in Quadratic Response Surface Designs," *IIE Transactions,* Vol. 19, pp. 395–403.

Iglehart, D. L. (1977), "Simulating Stable Stochastic Systems, VII: Selecting the Best System," *TIMS Studies in the Management Sciences,* Vol. 7, pp. 37–49.

Jacobson, S. H., and Schruben, L. W. (1989), "Techniques for Simulation Response Optimization," *Operations Research Letters,* Vol. 8, pp. 1–9.

Kleijnen, J. P. C. (1988), "Analyzing Simulation Experiments with Common Random Numbers," *Management Science,* Vol. 34, pp. 65–74.

Koenig, L. W., and Law, A. M. (1985), "A Procedure for Selecting a Subset of Size *m* Containing the *l* Best of *k* Independent Populations, with Applications to Simulation," *Communications in Statistics—Simulation and Computation,* Vol. 14, pp. 719–734.

Law, A. M., and Kelton, W. D. (2000), *Simulation Modeling and Analysis,* 3rd Ed., McGraw-Hill, New York.

Nelson, B. L. (1987), "Variance Reduction for Simulation Practitioners," in *Proceedings of the 1987 Winter Simulation Conference,* A. Thesen, H. Grant, and W. Kelton, Eds., pp. 43–51.

Nozari, A., Arnold, S. F., and Pegden, C. D. (1987), ‘‘Statistical Analysis for Use with the Schruben and Margolin Correlation Induction Strategy,’’ *Operations Research,* Vol. 35, pp. 127–139.

Schmeiser, B. (1982), ‘‘Batch Size Effects in the Analysis of Simulation Output,’’ *Operations Research,* Vol. 30, pp. 556–568.

Schruben, L. (1981), ‘‘Control of Initialization Bias in Multivariate Simulation Response,’’ *Communications of the ACM,* Vol. 24, pp. 246–252.

Schruben, L. (1982), ‘‘Detecting Initialization Bias in Simulation Output,’’ *Operations Research,* Vol. 30, pp. 569–590.

Schruben, L. W., and Margolin, B. H. (1978), ‘‘Pseudorandom Number Assignment in Statistically Designed Simulation and Distribution Sampling Experiments,’’ *Journal of the American Statistical Association,* Vol. 73, pp. 504–525.

Schruben, L. W., Singh, H., and Tierney, L. (1983), ‘‘Optimal Tests for Initialization Bias in Simulation Output,’’ *Operations Research,* Vol. 31, pp. 1167–1178.

Sullivan, D. W., and Wilson, J. R. (1989), ‘‘Restricted Subset Selection Procedures for Simulation,’’ *Operations Research,* Vol. 37, pp. 52–71.

Welch, P. D. (1983), ‘‘The Statistical Analysis of Simulation Results,’’ in *The Computer Performance Modeling Handbook,* S. Lavenberg, Ed., Academic Press, New York, pp. 286–328.

CHAPTER 96

Virtual Reality for Industrial Engineering: Applications for immersive Virtual Environments

HANS-JÖRG BULLINGER
RALF BREINING
MARTIN BRAUN
Fraunhofer Institute of Industrial Engineering

1. **OVERVIEW** **2497**
   - 1.1. Virtual Reality in General 2497
   - 1.2. Field of Use for VR in Engineering and Science 2497
     - 1.2.1. Virtual Prototyping 2498
     - 1.2.2. 3D Representation of Complex Data 2498
     - 1.2.3. Simulation of Work Sequences and Activities 2498
     - 1.2.4. Substitution of Complex Physical Prototypes and Tools 2498
     - 1.2.5. Modeling and Control of Business and Production Processes 2499
     - 1.2.6. Cooperation in Virtual Work Environments 2499
   - 1.3. Virtual Reality in the Product Development Process 2499
2. **DEFINITIONS** **2499**
   - 2.1. Virtual Environment 2499
   - 2.2. Digital Mock-up 2501
   - 2.3. Virtual Prototypes 2501
   - 2.4. Rapid Product Development 2501
   - 2.5. Augmented Reality 2501
3. **VIRTUAL ENVIRONMENT HARDWARE AND SOFTWARE** **2501**
   - 3.1. Structure of VE Systems 2501
   - 3.2. Hardware 2501
     - 3.2.1. Display Systems 2502
     - 3.2.2. Position and Orientation Systems 2502
     - 3.2.3. Interaction and Manipulation Systems 2503
     - 3.2.4. Computation Systems 2503
     - 3.2.5. Networks 2503
   - 3.3. Software 2503
     - 3.3.1. Modeling 2503
     - 3.3.2. Simulation Control 2504
     - 3.3.3. Communication 2504
     - 3.3.4. Rendering 2504
4. **HUMAN–COMPUTER INTERACTION FOR VE** **2504**
   - 4.1. Visualization Basics 2504
     - 4.1.1. Resolution 2505
     - 4.1.2. Stereoscopic Visualization and Depth 2505
     - 4.1.3. Field of vies (FOV) 2505
     - 4.1.4. Brightness and Luminance 2505
     - 4.1.5. Contrast 2506
     - 4.1.6. Color Saturation 2506
     - 4.1.7. Masking 2506
     - 4.1.8. Performance 2506
     - 4.1.9. Refresh Rate 2506
   - 4.2. Visualization Systems 2506
     - 4.2.1. Fully Immersive VE 2507
     - 4.2.2. Projection VE 2507
     - 4.2.3. Augmented Reality 2507
     - 4.2.4. Monitor VR 2507
     - 4.2.5. Responsive Workbench 2507
   - 4.3. Methods for Human–Computer Interaction 2507
     - 4.3.1. Formal Language Interaction 2508

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

- 4.3.2. Natural Language Interaction 2508
- 4.3.3. Direct Manipulative Interaction 2508
- 4.3.4. Gesture Interaction 2508
- 4.3.5. Combined Interaction 2509

5. VR APPLICATIONS 2509

- 5.1. Virtual Prototyping 2509
  - 5.1.1. Immersive Design Review 2509
  - 5.1.2. Tool Evaluation 2511
  - 5.1.3. Immersive Postprocessing 2511
- 5.2. Process Simulation/Factory Planning 2512
- 5.3. Education and Training 2513
- 5.4. Scientific Visualization 2514

6. PROCESS INTEGRATION OF VR APPLICATIONS 2514

- 6.1. Analyzing Engineering Tasks for VE Applications 2514
- 6.2. Requirements from the Engineering Side 2516
- 6.3. Data Requirements 2517
  - 6.3.1. Partitioning of VR Data Representation 2518
  - 6.3.2. Material: Color, Textures 2518
  - 6.3.3. Performance Optimization 2518
  - 6.3.4. The Training Concept 2518

7. CONCLUSIONS 2518

REFERENCES 2519

ADDITIONAL READING 2520

1. OVERVIEW

This chapter focuses on virtual reality (VR) applications in the fields of engineering and industrial engineering. It gives definitions of the main keywords for the field of engineering and describes the hardware and software and the specific human–computer interaction aspects for virtual environments (VE). Some typical applications are specified that show the field-tested use of VR, and the basics for the integration of such applications in the development process are described.

1.1. Virtual Reality in General

By means of VR, an intuitive working environment can be created in which the multiple aspects of human intelligence are addressed. With sufficient perception and interaction qualities to give the user a feeling of intuitive handling of virtual objects, the following VR principles can be attained.

- Visualization, evaluation, and design of 3D objects
- Advancement of creative abilities by free design and communication of imagined objects
- Interactive demonstration of objects and events
- Mediation of (abstract) knowledge based on the user's experiences (e.g., exploring a molecular structure)
- Mediation of skills and training of behavior in dangerous situations and environments
- Knowledge mediation beyond human physical limitations (e.g., exploring how a gasoline engine functions by opening up the combustion chamber)
- Exploration of distant places or past epochs

1.2. Field of Use for VR in Engineering and Science

The main application area for VE is the simulation of objects and processes. This suggests that VE can be used as a presentational simulation technique. Aside from the presentation of perceptible sensory objects, a special use is in the expression of nonmaterial and invisible objects or in the saving of physical products. The presentation can rely on a static scene as well as on a dynamic process. With a pure presentational simulation, the degree of interaction is limited because the emphasis is on the presentation of knowledge. However, with control simulation by means of VE, the user takes a more intensive part in the events, while the qualities of presentation have been preserved. Apart from the use of control simulation for entertainment purposes, uses are in handling critical situations, controlling devices in inaccessible areas, and working with meager resources. The third application type is particularly characterized by its interaction forms, where the virtual objects are not just moved but also used for design simulation. These applications support decision making in the product-development process.

In addition to process simulation, VE systems can be used as mediators for other, generally real, objects. The purpose of presentation of any object is communication. Furthermore, VE can be used as a medium for human–machine interaction. Communication with an unreal object, such as another virtual representation, is possible. Finally, it is possible to integrate virtual information within a real environment.

A third dimension of possible VE use is telepresence applications, which include simulations that do not take place at the application location, as well as communication. These applications are founded on real spatial distances. Application areas are differentiated into two areas: telepresence, which is a spatially distanced presentational simulation, and teleoperation, which is a type of simulation for a manipulation or design purpose.

1.2.1. Virtual Prototyping

Among other things, the duration of the development process is determined by the number of the physical prototypes that are required for evaluation of the desired product properties. Virtual prototypes are defined as a computer-based simulation of a technical system or subsystem with a degree of functional behavior that is comparable to corresponding physical prototypes (Haug et al. 1993).

The visualization of virtual prototypes calls for the processing of the data with dependence on the visualization techniques. Apart from purely realistic visualization, complexity-reducing models and symbolic visualization are used as well. Mixed models of both methods are most widely spread. Metaphors for visualization of simulation results, for example, are used in FEM analysis (overlaying of paint leveling) or the representation of paint coat thickness in robot simulation (Brown 1996).

Research activities for the evaluation of virtual prototypes with techniques of VR are encountered in the investigation of flow response of flying objects (Bryson et al. 1997) and automobile chassis. The geometry model is overlaid with the results of flow simulation such as isosurfaces or as paint particles. The user can freely move cutting planes or virtual trails of smoke. Investigation of the capability for assembling components is also found in aircraft construction and vehicle construction, including the evaluation of the free construction room available, accessibility, and the ergonomic design of manual assembly workplaces.

Apart from the evaluation of the product properties, a further-reaching use of virtual prototypes also requires consideration of the production processes of the products.

Aspects of asynchronous and synchronous object management in a distributed environment and the data models for the manipulation of lead time have thus far been in the foreground in systems for designing objects in a virtual work environment. Here, the data models are oriented by existing CAD standards or they focus on application protocols for the product data model STEP, with these protocols still to be developed. These systems are also not fully immersive, but they are monitor based in connection with relevant 3D input tools (e.g., space mouse, trackball). Handling aspects of virtual prototypes in an immersive VE have been studied, for example, at the University of Wisconsin-Madison (Dani and Gadh 1996).

1.2.2. 3D Representation of Complex Data

The 3D representation of complex data is applied to the examination of comprehensive geometry data records, such as in architecture or in the scientific visualization field, the representation of mathematically computed data. In the visualization of building data in architecture, the architect would like to walk through the virtual building together with the customer and detect design errors. Research and development work is conducted at a large number of research institutions and enterprises. The visualization of mathematically computed data has the goal of rendering simulation results visible, audible, or tactile. Outstanding work on this subject has been done by NASA Ames Research (Bryson et al. 1997). For evaluation of the outside shape of flying objects, they simulate the flow response in a virtual wind tunnel.

1.2.3. Simulation of Work Sequences and Activities

A near-reality perception (immersion) in a virtual environment is used to plan and investigate work sequences and activities. Immersion is achieved by 3D representation of information in real time by integrating several human senses, with the user perceiving himself or herself as part of the scene. The interaction differs from the input devices used so far in that direct intervention in the 3D space is possible (Durlach 1997). The goal is to adapt the work environment to human responses and physical capabilities. Simulation of the work environment and human interaction with work surroundings is necessary to produce a near-reality perception. Ergonomic investigations can be conducted with subjects without a physical test setup.

1.2.4. Substitution of Complex Physical Prototypes and Tools

Physical prototypes are used for testing and evaluation of product properties in the early development stages of product development. VR permits testing and evaluation to be conducted in virtual prototypes. The objective is to avoid physical prototypes if possible. Research is being conducted in the automotive and aerospace industries (Leston et al. 1996). Apart from purely design-oriented tasks, efforts are also being made to conduct physical materials tests, such as crash tests, in virtual prototypes. Additional approaches deal with the production engineering view of prototypes and products up to the planning of production facilities and production plants. These also include approaches to the substitution of complex resources for the support of production, such as in the production of cable harnesses in the aerospace industry (Ellis et al. 1997).

1.2.5. Modeling and Control of Business and Production Processes

Apart from product- and production-oriented spheres of application, VR technologies are also used in product modeling and control. Research approaches can also be encountered in information technology, for the analysis and maintenance of communication networks, and in the modeling of production engineering processes.

1.2.6. Cooperation in Virtual Work Environments

In addition to spoken language, human communication is based on expression by the body (gestures) and the face (miming). Research in near-reality virtual work environments investigates human communication in immersive VR environments. This research is focused on computer-supported cooperative work (CSCW) (Pandzic et al. 1996), the mapping of human gestures and movement, and the realization of synthetic behavior of virtual human beings (Shawver 1997).

1.3. Virtual Reality in the Product-Development Process

The use of computer-based tools and methods in the development process is essential with respect to quality, time, and costs. The general business conditions are similar worldwide: product life cycles are declining drastically, and technological leadership of the highest possible quality is required in order to react to constantly increasing innovation dynamics. Engineers have to decrease the time for construction and evaluation of the prototypes. The mainly computer-based development of products using 3D CAD systems offers many advantages in optimization of time, costs, and quality, but it leads quickly to a restricted, reduced perception of the product during the development process. Therefore, digital models require an immersive VR-based working environment to improve the perception of the computer-based models.

Another factor is that new processes like rapid product development (RPD) focus on the short time between the final determination of the construction and the start of sale. Although the complexity of most products and of the product development process is growing, the first sketches must be more detailed and the basis for valid decisions should be guaranteed. Therefore, the use of VE has two main goals: to enhance the degree of freedom of predefinition and to achieve a higher level of elaboration in the early development phases for better decision support.

Therefore, the use of immersive projection technology (IPT) has two main goals: to enhance the degree of freedom of the construction constraints and to achieve a higher level of elaboration in the early development phases for better decision support. Figure 1 shows how these goals are benefitted by the use of VE.

Tools and methods that are used in VE and that permit information to be obtained about the product in the early development phases offer attractive optimization potentials (time, costs, quality) in engineering. Among other things, in the foreground here are product aspects and features that essentially permit qualitative information to be obtained (e.g., formal aspects) and can be primarily assessed by evaluation (e.g., motion spaces in connection with capability of assembly). Human interaction with the virtual prototypes and human perception and immersion in the VE is decisive for handling the tasks on hand (Figure 2). The verification of qualitative information has normally been possible so far only on the basis of real, physical prototypes.

2. DEFINITIONS

2.1. Virtual Environment

An application-orientated definition that might state the most accurate minimal demand describes VE systems as a combination of computer-based display and interaction techniques.

Ellis (1995) defines a virtual environment as a synthetic, interactive, illusory environment perceived when a user wears or inhabits an appropriate apparatus, providing a coordinated presentation of sensory information in imitation of a physical environment.

All VE applications are founded on the generation, perception, and manipulation of naturalistic or abstract virtual worlds without any physical equivalent. Objects existing within virtual worlds can possess various qualities and behaviors. Examples are graphics, sound, and force feedback. By multiple addressing of the human senses, the attempt is made to generate the greatest possible intuitiveness of virtual environments; VE can be experienced through visualization, marked out by 3D object representations and real-time-orientated interaction modes.

To distinguish VE from the multitude of computer-based visual simulation techniques, the following minimal requests must be fulfilled:

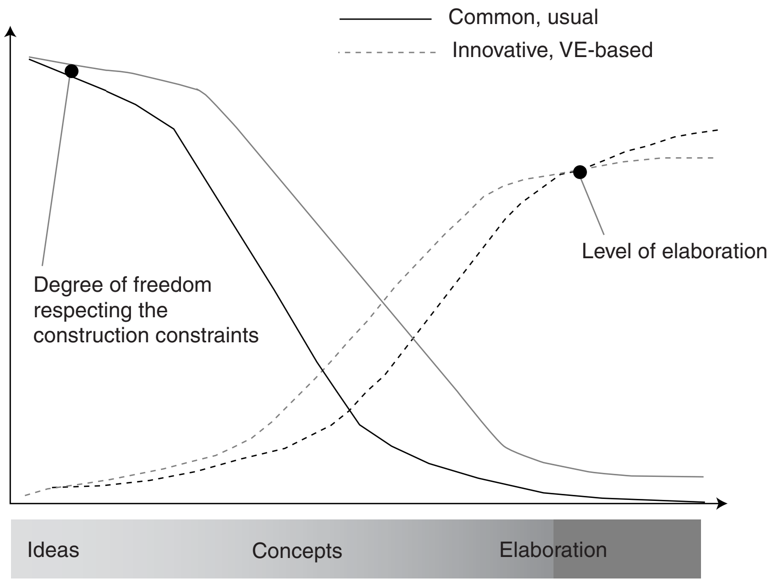


Figure 1 Benefits of use of VE in the Development Process.

- Graphical display dependent on the position and orientation of the user.
- 360° visualization complies with all three coordinate axes
- Realistic 3D behavior of the modeled objects
- The most possible intuitive interaction modes with objects, adapting to human experiences and behavior
- Object manipulation in all real or requested degrees of freedom
- Quasi-real-time-orientated object response

VE systems can be defined as computer-based information technologies for interactive, real-time orientated simulation and multisensory representation of objects, processes, and their results. With

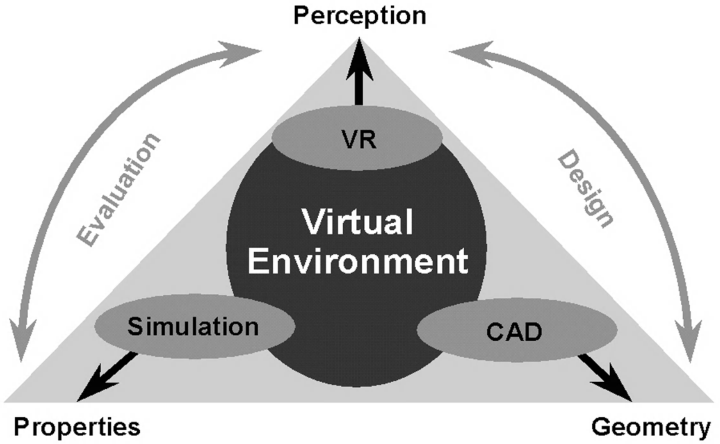

Perception
VR
Virtual Environment
Evaluation
Design
Simulation
CAD
Properties
Geometry

Figure 2 System for Evaluating and Designing Virtual Prototypes.

the surmounting of conventional input and output interaction forms, the user should be given the impression that he or she is situated in a virtual environment. The human's central cognitive abilities, such as the interpretation of visual data and the recognition of complex structures, are thus supported.

To describe computer-generated worlds, terms in addition to *virtual environment* have been established, such as *virtual reality, artificial reality, telepresence,* and *cyberspace*. Generally accepted is the use of *virtual reality* as an umbrella term for computer-generated, interactive 3D scenes, *cyberspace* as a place of communication connecting several users together, and *telepresence* for an experience that overcomes space and time. *Virtual reality* tends to suggest the perception of a meticulously detailed computer-generated duplicate of a natural environment. In contrast, *virtual environment* stresses the secluded nature of the application, whose abstract representations depend more on the user's ideas than on real models. *Virtual reality* suggests utopian visions, and hence for reasons of precise description, experts prefer *virtual environment*.

The degree of immersion indicates the extent to which a user is tied to a virtual environment. Immersion can be described as an opportunity for real-time-orientated perception and interaction. Immersion is characterized by the believability of the environment. Immersion has a technical dimension, a content-representative dimension, and an individual–psychic dimension. These dimensions each contribute to an immersive experience.

2.2. Digital Mock-up

Digital mock-up (DMU) is a digital product description for development, design, and manufacturing. The aim at present is to enhance the DMU in order to replace physical mock-ups with a digital one that will require only one verification prototype at the end of the design process. DMU is based on the management of a complete product model. DMU functions for the assembly, fitting, and operation of a virtual product, simulation of tolerances, ergonomics, and flexible parts will be available. In parallel, DMU checks the virtual product with regulation constraints. DMU is required to be an integral part of design systems, available from any location and from any activity point of view. Virtual reality techniques will be used for the stereoscopic visualization of the DMU.

2.3. Virtual Prototypes

Virtual prototypes are defined as a computer-based simulation of a technical system or subsystem with a degree of functional behavior that is comparable to corresponding physical prototypes. The visualization of virtual prototypes calls for the processing of the data in dependence on the visualization techniques. Apart from purely realistic visualization, complexity-reducing models and symbolic visualization are used as well. Mixed models of both methods are widespread.

2.4. Rapid Product Development

Rapid product development (RPD) shortens the feedback control cycles concerned with product data generation and the associated management processes. RPD exploits the potential of modern information and communication tools in order to support the necessary dynamic cooperation structures. Development times are systematically shortened by means of a holistic integration of man, organization, and technology. The learning processes can be systematically relocated to early product development phases. As a result of generative methods of production and virtual reality, physical and virtual prototypes can be made quickly available.

2.5. Augmented Reality

Augmented reality (AR) combines real worlds/objects with virtual worlds/objects. AR is a novel approach to the interaction between human and machine. It is possible, for example, to view information using a head-mounted display. The information is displayed context sensitive, which means that it depends on the observed objects, such as a part of an assembly. The engineer can now display job-related assembly data while viewing the real object.

3. VIRTUAL ENVIRONMENT HARDWARE AND SOFTWARE

3.1. Structure of VE Systems

In order to be able to design and use effective VE systems, it is necessary to understand the technical concepts related to virtual environments, to be aware of the limitations of the available technology, and know the design approaches that lead to the creation of successful virtual environments. Although there are many VE systems in use, areas open for research are not just the system and device design, but also the ground use concepts. The basic functioning of a VE system is shown in Figure 3.

3.2. Hardware

To create immersive experiences in VEs, VE systems integrate a combination of several hardware technologies. These technologies can be grouped under the following categories:

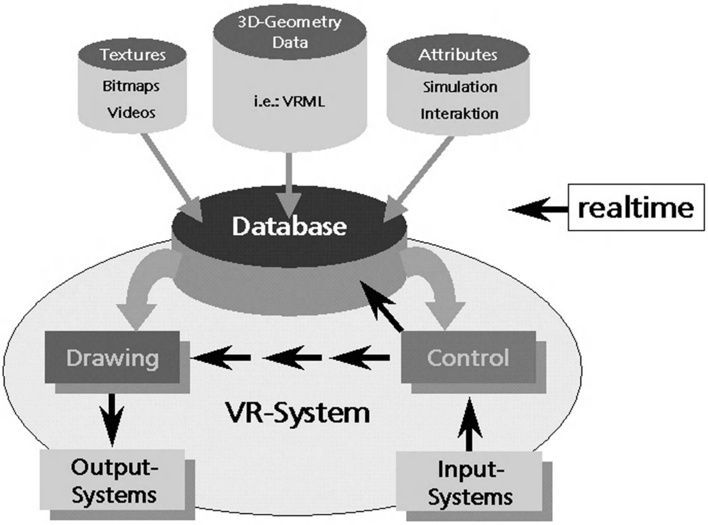


Figure 3 Basic Structure of a VR System.

- *Display systems:* present the virtual environment to the user.
- *Position and orientation systems:* track the user's position and orientation in the virtual environment. They are also used for interaction purposes.
- *Interaction and manipulation systems:* provide manipulation of the virtual environment.
- *Computation systems:* perform the computations required for the generation of a virtual environment.
- *Networks:* allow integration of several distributed user systems in one common environment.

3.2.1. Display Systems

The main elements of VE interfaces are displays that mediate visual, acoustic, and tactile sense stimuli. Of essential significance are visual and acoustic displays.

Visual displays are devices that present the virtual environments to the user. The degree of immersion given by a particular VE system depends greatly on the visual interface display. Several kinds of visual displays are currently available: monitors, head-mounted displays (HMDs), head-coupled displays (HCDs), and projection systems. All of these systems are capable of producing wide-angle stereoscopic views of the scene, although in some cases monoscopic vision is also used.

HMDs, which are the most broadly used visual displays in immersive VE systems, and HCDs place a pair of display screens directly in front of the user's eyes. In HMDs, the screens are mounted on a helmet that the user wears while staying in virtual environments. The HCD is like a pair of binoculars freely attached to a flexible swivel-arm construction, which can be easily handled by the hand through open space. Both HMDs and HCDs are coupled with a tracking system to determine the viewer's position in space. The virtual environment is displayed in stereo from the user's point of view, which serves a high degree of immersion; users are completely surrounded by the virtual environment. Other kinds of visual displays are projection-based systems. In such systems the user's position and actions are tracked and the corresponding virtual scene is projected onto large screens.

Acoustic displays can be used to provide feedback concerning the virtual environment. Sound plays an important role in localization and interaction. Due to the application of simple audio in many other areas outside of VE, it is known as a mature technology. Synthesizers creating, mixing, and reproducing sounds and systems for speech in- and output have been made available.

3.2.2. Position and Orientation Systems

Tracking is a critical component of an immersive environment. Tracking includes the measurement of the user's head position and orientation as well as the measurement of other body parts, such as the user's hand and fingers. The continuous measurement of the user's head position and orientation is of high importance because it is used to produce the correct environmental view from the user's point of view, which is critical to obtaining a high degree of immersion. Tracking of body parts and movements allows interaction with and control of the virtual environment to take place.

In principle, tracking is suitable for any technology. Thus, the three-dimensional position and orientation of an object can be determined as free of delay and reliable. Electromagnetic, kinematics, acoustic, optical, image work processing, and inertial procedures find use for tracking tasks. Currently available body-tracking systems fall into two classes (Bryson 1993):

- *Position tracking:* devices that detect the absolute position and usual orientation of a tracker in 3D space
- *Angle measurement:* devices that detect the angle of bend of some body part; usually found in data gloves to measure the angle of finger joints

3.2.3. Interaction and Manipulation Systems

Interaction and manipulation systems are devices that allow manual exploration of the virtual environment and the manipulation of virtual objects. These devices, functionally based on a tracking system, measure the position and forces of the user's hand and other body parts and can apply forces to the user to produce the sensation that the virtual objects are real (Cruz-Neira 1993).

Interaction and manipulation devices have been classified as:

- Pointing and selection devices
- Force and torque feedback devices
- Tactile devices
- Devices to produce stimulus, such as temperature

The best-known interaction device is the data glove. After having been calibrated to the user, the data glove allows the computer to determine the user's hand gesture, such as fist, open hand, or pointing. Often the hand's physical position in a 3D space is also determined. The glove's coordinates are overlaid into the virtual environment, where the physical glove position controls a 3D virtual hand in this environment. The virtual hand can be used to pick up virtual objects, push virtual buttons, and so on (Iovine 1994). Numerous interface devices also exist, such as the free-flying joystick and spaceball, which were developed into 3D devices from conventional interface devices, as well as systems for tactile and force feedback.

Another interface technology for effective control and interaction in VEs is speech control. Speech control in virtual environments will facilitate tasks and exploration. Not tied to tapping onto a keyboard or moving, a control device will free the user's hands and add to the feeling of immersion (Iovine 1994).

3.2.4. Computation Systems

Computation systems are the computer hardware used to control the overall operation of a virtual environment. Computer hardware has to handle several tasks: generation of graphics for the scene, computation of the state of the environment, and control of input and output devices. In a VE system, all these tasks have to be integrated and synchronized.

3.2.5. Networks

Networks are used to exchange data between different virtual environments. They lay the foundations for distributed VE systems. Distributed VE systems make it possible to connect people located in different places to let them take part in a joint communication and design process. With the appropriate tools, such systems can be powerful platforms for interdisciplinary work, independent of location. This is known as computer-supported cooperative work (CSCW).

3.3. Software

An internal computer functional hierarchy exists for processing information input and generating appropriate output. Essential functions of this hierarchy are represented by the modules of modeling, simulation control, and rendering. Built onto appropriate hardware structures, this functionality is realized by software technology.

3.3.1. Modeling

The virtual world of an application is constructed principally through modeling. As a rule, object data are converted from CAD data and completed through color and surface data. In addition to geometric description, modeling also includes function modeling, beyond the CAD applications, in which the total number of functions and the geometric free degrees of a respective object are determined.

The structure of geometric objects can be described through parameters (object data). The sum of the computer-stored object data produces the internal computer representation of real objects. The mathematical model of a real object is composed not only from data structures but also from algorithms that operate on these structures. It forms the base for all built-up modules, such as simulation control and rendering.

3.3.2. Simulation Control

During VE application, the arrangement of objects is constantly calculated by putting the contents of the databases in relation with the interaction instructions of the user or the behavioral parameters of autonomous objects. Simulation control is closely linked to the communication of the data processing. The basis for coordinated running is real-time management.

Simulation requires the determination of certain functions. Therefore, interaction parameters or collision identification have to be defined. Complex simulations require very high computing performance, which may lead to bottlenecks within real-time management.

3.3.3. Communication

Communication includes data transformation, that is, the coordination of data transfer between input/output devices and simulation control as well as data exchange between users and units involved in a multiuser system.

With the data transformation, the transformation software interprets the interaction instructions of the user and passes the input commands to the simulation control, where the real-time orientated computation of the virtual environment occurs. Additionally, the data representing the virtual environment are reconverted by the transformation software into appropriate interface signals and presented as simulation output. Thus, the data transformation primarily functions to transfer data between the user interfaces and the simulation control.

Within a multiuser application, the communication coordinates the data exchange between the active participants and units involved in a network. The real-time-oriented data exchange informs all participants about relevant interaction processes as well as the actualization of the databases.

3.3.4. Rendering

The term *rendering* or *drawing* includes different procedures of the transformation of parametric data models into discrete images and sounds. Therefore, rendering is part of the communication within data processing, but due to its complexity and importance in VE it is specially dealt with. The performance requirements of real-time rendering are high. In VE applications the performance bottleneck is mostly in rendering, both visual and acoustic.

The visual presentation normally occurs with a pair of images, which must be rendered with a minimum frame rate of 20 frames/sec in order to avoid picture disturbance.

Image generation is based on lighting models, which are divided into two types: local and global. Local models take into account the reflections from single surfaces, independent of the environmental lighting situation. Global models take into account the independence of lighting and reflections and therefore seem more realistic, but they are more difficult to generate.

The acoustic presentation within a virtual environment orients itself towards the geometry and surface formation of the room being modeled. Acoustic simulation is suitable for two methods. First is the image source method, which calculates virtual sound inclusive of the reflection behavior of the surfaces being modeled. From this, the sound energy coming towards the receiver can be defined. Second, with the particle-tracing method, the transmitted sound impulses are registered by detectors, where the impulse response can be measured.

4. HUMAN–COMPUTER INTERACTION FOR VE

4.1. Visualization Basics

The light transmitted from a source reflected by an object is sent to the eyes. By refraction and collection, the light is concentrated on the retina, where the photoreceptors are embedded. The nerves are then aroused, and the signal is escorted along the optical track and analyzed by the brain. The eye accommodates itself to the distance of the object being seen. By adapting to dark and light, the eyes are able to change their acuity within a wide range. Color vision results from the stimulation of the retina by light of various wavelengths in the range of 400–780 nm.

When a 3D object is seen, a separate 2D image is projected onto each retina. The two images are converted into impulse patterns in the brain, where they are united to create a 3D object. The convergence angle between the eyes' axis and the accommodation is used to determine the distance of an object from the eye.

Visual perception is the most important for feedback in a virtual environment. Optical information is presented by optical displays. Along with the application, display devices with various inserted performance profiles are used, making it possible to create anything from simple monochrome graphics and line models to a photorealistic rendering. The specifications of the visualization technologies, which can be varied in depending to the application, are discussed below.

4.1.1. Resolution

Optical resolution is the angular size of an object that can be individually resolved. Resolution is defined as the angular size of a picture element. Resolution increases as the angular size of the picture element decreases. Optical resolution is closely related to screen resolution, the number of pixels on a screen, which in turn determines picture element resolution. The effective optical resolution is determined not only by the picture element resolution of the screen devices but also by the optics through which the screen is viewed (Bryson 1993).

Eye resolution depends on many factors, including color, brightness, contrast, and length of exposure. On the axis, resolutions of around 5 arcmin are required to reach the region of peak sensitivity. Acuity increases rapidly as the object moves outside the central 2° region. It is principally sufficient only to provide the central field of vision with detailed pictures because the natural resolution of the eye in the periphery strongly decreases. At 10° off-axis eccentricity, acuity drops around 10 arcmin (Helman 1993).

However, the marginal areas are particularly sensitive to low light intensities and movement. To achieve natural resolution capabilities, screens with 6000 picture lines are suitable. For a convincing visual display with a field of view of 150°/60° at a distance of 250 mm, a resolution of, ideally, 9000 × 3600 pixels is recommended.

4.1.2. Stereoscopic Visualization and Depth

Stereoscopic visual presentation usually takes place using a pair of pictures with a slightly shifting convergence angle. The limit of stereo vision typically occurs at a binocular disparity of 12 arcsec. The methods of stereoscopic visual presentation can be divided into time parallels (both-eye simultaneous visual presentation) and time-multiplexed systems (non-perceptible alternating visual presentation). Time parallel stereoscopic systems produce the displacement of the optical convergence axis either by different monitors for each eye or by two half pictures differentiating in color and perspective, which are looked at through toned or polarized 3D glasses. Time-multiplexed systems insert optical or mechanical called shutter systems for the viewing of alternating pictures displayed on a monitor.

Three-dimensional visual perception through object-differentiated accommodation cannot occur with an HMD, because the monitor has a flat surface. When the eyes look at stereoscopic computer-generated imagery, their accommodation and convergence often do not match because they must focus on the screen or the image plane defined by the optics of an HMD with an angle for stereoscopic view dictated by the rendered images. Without proper calibration (or a monoscopic system), neither focus nor convergence may reflect the actual position of the virtual object relative to the viewer. Because of these inconsistencies, many users of stereoscopic systems have trouble fusing stereo images. To place the images at infinity, thereby making convergence and focus match closely for distance vision, collimated optics can be used. Computer graphics applications that do not need to depict the depth of scale acurately can artificially adjust the parallax to allow more comfortable viewing. For virtual environments requiring close-up manipulation of objects or accurate registration of virtual objects with the physical world, all variables affecting stereo viewing should be considered. Size, image distance, and overlap of the system must be chosen carefully to match the task and operation distance.

However, other solution approaches try to realize the adaptation of depth with changing of the focus. The eye's focus is measured by means of a laser beam reflected by the retina, and in consequence the depth of focus is followed by image visualization (Aukstakalnis and Blatner 1992).

4.1.3. Field of View (FOV)

Each eye has approximately a 150° FOV horizontally and 120° FOV vertically. The binocular overlap when focused at infinity is approximately 120°. With VE display devices, a wide field of view is very desirable for conveying a feeling of immersion. For HMDs a 120° horizontal and 60° vertical FOV is minimally recommended. The trade-offs for higher FOV are lower effective resolution and usually more distortion in the periphery of the picture. Distortion can be avoided by using complex optics. However, distortion is a big problem with see-through HMDs because the world provides a reference for "straightness". With the relatively small FOVs of HMDs, large overlaps of more than 50% have been found useful.

4.1.4. Brightness and Luminance

Including dark adaptation, the eye has a dynamic range of around seven orders of magnitude, far greater than any current display device. The eye is sensitive to ratios of intensities rather than to absolute differences. At high illuminations the eye can detect differences in luminance as small as 1%. Brightness is a basic precondition for the perception of objects.

Brightness is the available luminance for each picture element. A typical HMD based on a cathode ray tube (CRT) can display no more than about 400 perceptible luminance levels. Sufficient brightness is particularly a problem for liquid crystal displays (LCD).

In order to minimize disturbing light conditions and differences of contrasts between the physical environment and the virtual simulation, high brightness is especially important for see-through HMDs.

4.1.5. Contrast

Contrast is the dynamic range of the luminance that the display supports. Contrast is important to the perception of structure in an image. Low-contrast systems are difficult to interpret. Low-contrast and low-brightness displays do not serve a high degree of immersion. With use of screen displays, a 5:1 contrast ratio for scenery and 25:1 for light points is recommended.

With LCDs, the display brightness and contrast depend greatly on the viewing angle. The display viewing angle is characterized as the angle between the normal to the display surface and the line between the center of the display and the user's eye. For high image quality, the viewing angle should be as small as possible.

4.1.6. Color Saturation

Color is a significant feature of visual images. The quality of color can be characterized in terms of the decomposition of a color signal into three primary components. Usually these components will be red, green, and blue, the primary colors, measured by their luminance. Alternatively, color components could include the hue, luminance, and value components.

In display systems, color is usually attained by grouping pixels into sets of color components. For example, a red pixel, blue pixel, and green pixel are grouped into a triple that comprises one full-color picture element. While this method works when the optical pixel resolution is very high, in wide-field displays the individual component pixels are easily individually visible. Another method of attaining color, called time-multiplexed color, is to use a monochromatic display screen with three rotating colored filters, one for each primary color. According to the activated filter, the monochromatic image with the illumination of the correspondent colored filter is displayed.

The quality of the color signal depends on several aspects, such as quality of color component pixels. The dynamic range of the luminance value for each component will determine the full range of colors that can be achieved.

Apart from color displays, monochrome displays are in use for several purposes. In general, monochrome displays offer higher resolution for an equivalent price.

4.1.7. Masking

A display quality issue closely related to color is the problem of masking, or nonabutting pixels, which occurs when the pixels are separated by a blank space. Masking is a problem because wide-field optics magnify the display screen and so magnify the space between the pixel elements. To avoid the problem of visible pixel elements and masking, the image in wide-field displays is often intentionally degraded, which is usually done by a diffusion screen that blurs the pixel.

4.1.8. Performance

In order to avoid picture disturbance, the frame rate must at the least be 20 frames per second. The frequency at which modulation is no longer perceptible varies from 15 Hz up to around 70 Hz for high illumination levels. Bright displays with large FOV can require frame rates up to 85 Hz. In order to minimize movement dizziness, the perceptional latency must be below 0.1 seconds.

4.1.9. Refresh Rate

The refresh rate is the time required for the picture elements to change state (on or off). These times need not be the same. Typically, a pixel takes longer to go off than to go on. If the refresh time of a pixel is too long, ghosting effects will occur in rapidly changing images.

4.2. Visualization Systems

Depending on the grade of immersion and the extent of the interactions, graded concepts from fully immersive VEs down to extended reality can be realized. VE applications for entertainment purposes are intended principally to simulate secluded fantasy worlds. Although development at the beginning relied almost exclusively on fully immersive concepts, today, because of the shortcomings of those concepts, an increasing number of partly immersive VE system applications are being developed.

4.2.1. *Fully Immersive VE*

With a fully immersive VE, the user wears a head-mounted display (HMD) equipped with headphones and a visual stereo display directly in front of the eyes. The user is visually and acoustically sealed off from the physical environment and primarily perceives impressions the virtual environment.

HMD fully immersive display is made up of two color monitors that project the image directly in front of the eyes and a lens system that widens the image to the natural field of view. Because the convergence angles, the screen arrangement presents a certain appearance of depth.

4.2.2. *Projection VE*

With projection VE, stereo pictures are projected by means of a special projection system onto surrounding walls. Projection VE systems allow the user a higher degree of free movement because 3D glasses are the only device that has to be worn. Such a system make it possible to link up several users within a VE application. However, orientation and interaction within the virtual environment is more difficult because of the system's low grade of immersion.

For an example of projection VE, see Cruz-Neira et al. (1992), who describe the CAVE system (Audio Visual Experience Automatic Virtual Environment).

4.2.3. *Augmented Reality*

With augmented reality, semitransparent data glasses are used, which make it possible for computer-generated objects and information to be linked by superimposing them with the perception of the physical environment. Parts of the physical environment remain perceptible at the same time that virtual elements contribute to an enrichment of information.

4.2.4. *Monitor VR*

In contrast to immersive VE systems, monitor VR portrays a low-priced alternative for 3D visualization of virtual worlds. As with projection VE, the user wears 3D glasses, which give him or her a steroscopic view of the virtual world on a monitor. The user's head movements are measured by a tracking system. The tracking data are acquired by the computation of the orientation-dependent displayed vision. Monitor VR makes the viewing of virtual objects from different directions and distances possible. A decided advantage of monitor VR over fully immersive VE is the possibility of simultaneous interaction with the physical environment. In addition, in contrast to head-mounted displays, present monitors make substantially higher visual resolution available. Because monitor VR is limited to a visual frame, only a slight feeling of immersion is possible.

4.2.5. *Responsive Workbench*

Responsive workbenches enable direct interaction with virtual objects appearing in physical environments. At a responsive workplace, several users move around a table with a glass top. By means of stereoscopic visual projection from the under side of the table, 3D virtual objects are constructed that appear to the users as if they were resting on the workbench. The users perceive the virtual object and the workbench as well as the individual person and colleagues. Responsive workbench applications are used in medicine and architecture (Krüger and Fröhlich 1994).

4.3. Methods for Human–Computer Interaction

Foley and Silbert (1989) define a human–computer interface as the determination of all user inputs into a computer, the determination of all computer outputs to the user, and the determination of sequences of inputs/outputs made accessible to the user.

The fundamental design principle of VE interfaces is to support human mental processes through an extensive communication and interaction environment and therefore increase the range of human handling and decision making. Following from these principles is the reduction of the degree of enforced handling sequences and the use of a computer for establishment and expression of relations. According to Brooks (1988), from these principles the following differentiated design criteria for the development of user interfaces and computer-based tools comply with ergonomic system design can be derived:

- Three-dimensionality of the modeled objects
- Direct manipulative and intuitive interaction instead of formal interaction
- Interactivity rather than sequence professed routines
- Multisensory stimulation rather than purely visual perception

These requirements derive from the basic idea that human–computer interaction should be comparable to human communication by means of speech, gestures, or body movements. For human access to a computer, all the human sensory channels should be included in the interaction, if possible.

The principal technical challenge in the design of interactive VE systems is in the development of human–computer interfaces, which make it possible to convert internal data structures into sensory-perceptible representations that possess consistent and (for the user) understandable behavior. In the same way, development of devices that translate human movements into computer commands.

In the context of VEs, generic modes of human–computer interaction can be classified into formal language interaction, natural language interaction, direct manipulative interaction, and gesture interaction. These interaction modes are complemented by combined interaction.

4.3.1. Formal Language Interaction

Formal interaction languages are classified into programming languages, command languages, and formal query languages. With command languages, the meanings are predominantly laid down in the vocabulary. An accurate, distinguished volume of available commands and parameters exists. The meaning of a command arises from the sequence and the relative position within an expression. Interactions based on formal languages are technically effective but are unfamiliar to most users. Because formal languages must be learned before they can be used, they are suitable only for a limited number of users. Programming languages are established for system implementation and form the basis of all other interaction languages.

4.3.2. Natural Language Interaction

Natural language systems can use conventional methods of electronic data processing or interfere with knowledge-based systems. In the framework of human–computer interaction, natural language is an adequate means for expressing references to objects, actions, and abstract facts. Natural language offers possibilities of expressing things that can be expressed by other forms of interaction only incompletely or with a large expenditure. The use of natural language in user interfaces increases the number of possible users considerably, particularly unpracticed users. Spoken language, however, changes quickly and cannot be called up in format as written language can. When analyzed, natural language proves inefficient and inaccurate and is manipulable and applicable only with selected dialogs.

Natural speech interaction occurs by means of appropriate input/output speech devices. The technical precondition for speech input/output is voice recognition. Voice recognition systems are classified into speaker-dependent and speaker-independent systems. Speaker-dependent systems, which are capable of achieving a high command count, are trained by the individual who uses the system. The most common drawback of this approach is that the system responds accurately only to the individual who trained the system. A speaker-independent system is trained to respond to a word, regardless of who is speaking. Therefore, the system must respond to a large variety of speech patterns of the target word. The command word count is usually lower than the speaker-dependent word count (Iovine 1994).

Special hardware and software for speech output exists that enables one great amounts of acoustic information to be efficiently stored and reproduced. As yet, the quality of synthetic speech is insufficient because the generation of correctly accentuated and pronounced speech has not been satisfactory achieved.

4.3.3. Direct Manipulative Interaction

Direct manipulative interaction techniques, which are applied with graphical user interfaces, make use of familiar metaphors of daily life. The dialogue of direct manipulative interaction techniques is based on a permanent visual presentation of all relevant objects and function undertaking by single-stage reversible operations. The impacts of actions on the relevant objects are received on a direct visual feedback. Direct manipulative systems make easy learning, use, and extension of system functions possible. The syntax of direct manipulation shows a clear, standardized structure of objects, functions, and attributes. The steps of dialogue with direct manipulation are slight in their complexity and range, as with natural or command languages. Disadvantages of direct manipulative systems are high implementation expenditures and the impossibility of activating objects that are nonvisible on a graphical user interface.

4.3.4. Gesture Interaction

Gesture interaction can be defined as a command-based tool kit that allows the user to interact through nonverbal, nonsymbolic commands and instructions, using gestures, hand signals, and movements. A special kind of gesture interaction is manual interaction. Manual interaction forms are distinguished by grips and movements of virtual objects, corresponding to a physical environment. With the application of natural and intuitive gesture interaction modes, improvement of efficiency of human–computer interaction is aimed at.

The naturalism of gesture interaction forms results from the inclusion of human sensory characteristics and abilities as well as the integration of gestures, which have been culturally conditioned. The quality of a gesture, in particular manual interaction, is orientated towards maximal usability of adaptability and dexterity of the human hand. Consequent allocation of hand movements and accompanying actions contributes to the understanding of gesture interaction.

Gesture interaction makes the specification of certain commands and parameters with high expressive abilities possible—for example, pointing out directions, gripping objects, controlling complex kinematics movements, and parameterizing object qualities. Trivial gesture interaction is easy to learn and does not assume linguistic knowledge. A further advantage is the directness of gesture interaction, in that the hand serves as an immediate medium.

With gesture interaction, every command must be represented by a certain hand gesture clearly distinguishable from other gestures. Particularly for complex commands and control processes, not enough gestures are available, so that gesture interactions based on arbitrary gestures are inefficient. Discrete values and vague expressions are scarcely. Complex gesture interaction, such as for shaping and design applications, has proven somewhat inaccurate and applicable only with selected commands.

In addition, for gesture interaction within the control process, sensory feedback is an essential decision-making criterion.

4.3.5. Combined Interaction

Combined forms of interaction do not represent generic interaction modes, but rather an application-oriented combination of existing interaction modes, primarily gesture, direct manipulative, and speech input.

The isolated use of some interaction modes often leads to a one-sided application and performance profile. With the symbiosis of gesture-based, natural language and direct manipulative forms of interaction, the scope of human–computer interaction and the functionality of the information input can be increased, making it possible to maximize usability and efficiency.

The interaction modes applied within virtual environments in the historical context of human communication are not new. In fact, the tendency to limit communication to screen and keyboard, driven during the last decades by computer technology, will be reversed and communication will be shifted back to a human standard.

5. VR APPLICATIONS

5.1. Virtual Prototyping

Prototyping includes numerous technical, methodical, and organizational measures, from concept formulation to a finished draft. In the face of increasing task complexity and shorter innovation cycles, the various requirements, design areas, methods, and planning participants can however only be integrated in a common development, design, and communication process with considerable computer support. For this task, computer-based planning and design tools are needed that possess a high degree of stimulation at their human–machine interface. VE technologies are well suited as the methodical foundation for the prototyping.

5.1.1. Immersive Design Review

3D CAD systems generate continuous surface and volume descriptions often referred to as solids. VR rendering systems deal with polygons or, in some cases, voxels. The process of converting CAD data to VR data is called tessellation. Tessellators exist for all major CAD formats, including the neutral formats IGES, VDAFS, and STEP. Different problems occur with this conversion:

- Insufficient model quality (wrong-side normals, cracks through not considering topology, LODs too coarse)
- Unneeded model complexity (wrong SAG values, improper algorithms, tessellation of invisible details, LODs too fine)
- Poor image-rendering performance (poor culling structure, missing LODs, unneeded model complexity)
- Missing data structure (loss of logical structure, trade-off to culling structure)
- Typical CAD conversion problems with neutral formats such as IGES or VDAFS.

Data exchange properties other than geometrical ones (surface description, constraints, kinematics, etc.) is by no means standardized. Importing those properties into a VR system currently requires a proprietary interface. However, for consistent and efficient work, all properties considered need to be imported and altered definitions need to be sent back to the CAD system.

Performance problems with complex model data will likely be handled by intelligent software algorithms (occlusion culling, motion LODs) in the near future.

One example of CAD data visualization is the virtual design review, which is used to evaluate, compare, and optimize exterior car designs in very early stages of the development process. The benefits gained are clear: saving physical design models can dramatically reduce time and costs. In Figure 4, a picture of a nonexistent car can be seen. Key factors for the designers in this type of application are very realistic rendering, especially of surface properties (in this case SGI′s ClearCoat technology), and the ability to render a very high number of polygons.

Scaling accuracy and high resolution for good visual perception are another must. This example was created using the Fraunhofer IAO software Lightning (Blach et al. 1998). The application properties and demands here are:

- *Goal:* evaluation of design/surface/proportions
- *Data:* surface geometry (very complex), surface appearance (very realistic)
- *Visualization:* highest possible projection quality, 1:1 scale, medium field of view power wall or CAVE (especially for interior models)
- *Graphics power:* highest possible
- *Application level:* basic

Of course, it is not possible to replace all physical design prototypes with virtual ones, but even saving of one physical prototype represents a great success for the first generation of virtual design

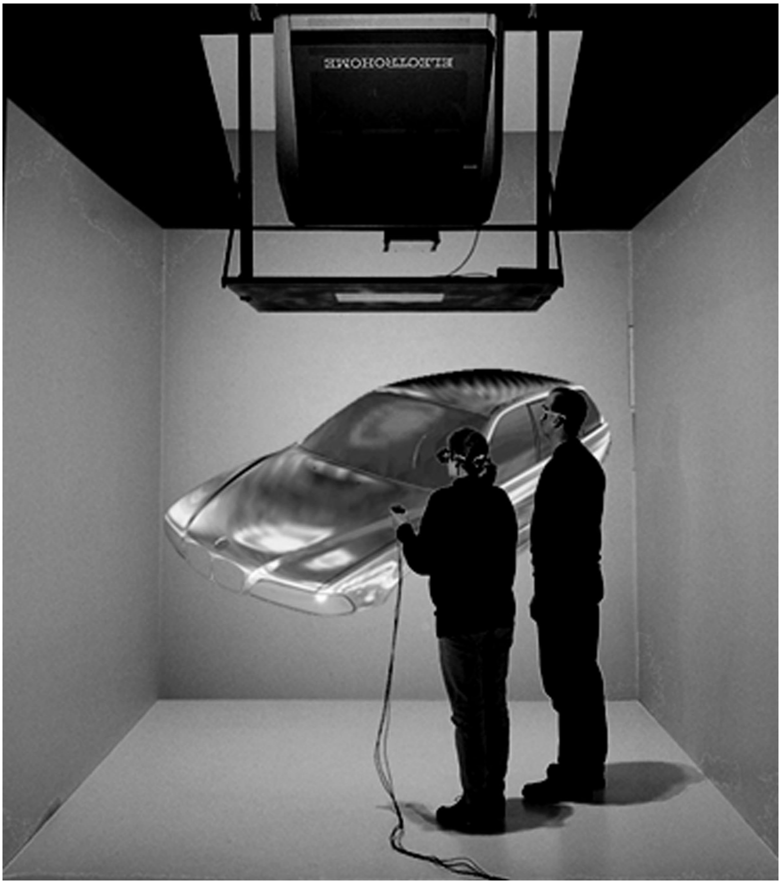


Figure 4 Example of Immersive Design Review.

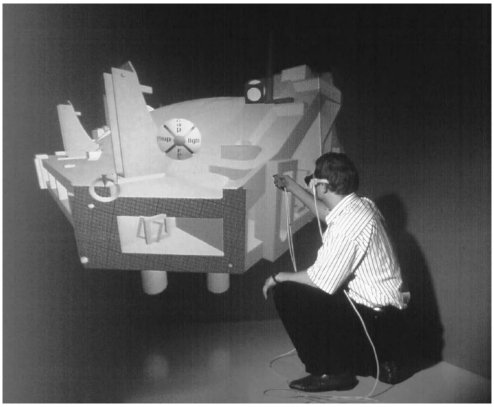


Figure 5 Tool Evaluation.

review tools. High-resolution power walls matching a 1:1 scale model seem to be the best output device here.

5.1.2. Tool Evaluation

The main task in this application example was to give specialized engineers a tool for improving the construction and evaluation process with sheet metal forms. The application was developed by Fraunhofer IAO for BMW:

- *Goal:* evaluation of CAD data
- *Data:* surface geometry (medium sized, CATIA solids), surface appearance (simple)
- *Visualization:* medium field of view, 1:1 scale, medium projection quality, single-wall rear projection or multiwall rear projection
- *Graphics power:* medium
- *Application level:* complex

The evaluation of tools can, in some cases, be greatly simplified by completely replacing the physical prototype with a virtual one.

Automatically creating a VR data set from the CAD solids reduces the preparation time for a tool review session to minutes. The required level of image-rendering speed and projection quality in this special application is medium, so the actual IPT system is moderate in price.

Special care has been given to creating an easy-to-learn, easy-to-use, and therefore simple interface for end users working with the application up to six hours a day. The main functionality of the application is as follows:

- Cutting of the data to explore collisions, etc.
- Annotation function to mark critical areas
- Snapshot function for automatic documentation (creating HTML pages for the intranet)
- Movie player for animating FEA data
- Virtual light source with dimmer function
- Measurement of sizes and angles

5.1.3. Immersive Postprocessing

The field of simulation (mostly FEM and CFD) deals with engineering and the natural sciences, which in many cases are to complex to be discussed exclusively on the basis of numerical values or texts. Multidimensional data sets need adequate visualization and presentation. An interdisciplinary discussion of simulation results is supported by a spatial, multidimensional representation of the data using VR-based techniques. An example of an industrial VR-based postprocessing application can be seen in Figure 6, where the thermal comfort in a car cabin is analyzed by examining the results of a stationary fluid flow simulation (using STAR CD).

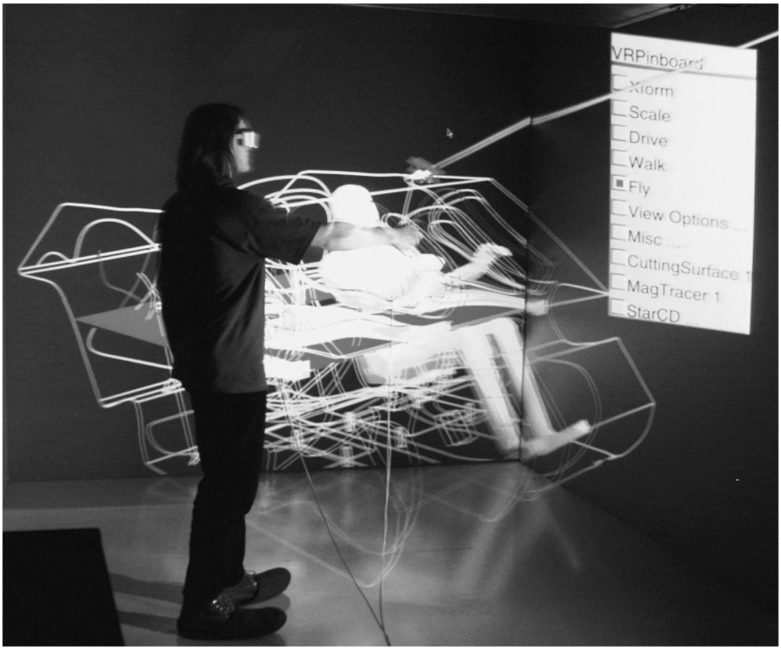


Figure 6 Fluid Dynamic Visualization in a VE. (*Source:* HLRS and Fraunhofer IPA)

- *Goal:* visualization of simulation results
- *Data:* surface geometry (simulation meshes), surface appearance (simple), visual appearance of simulation data (high resolution, medium quality)
- *Visualization:* high field of view, variable scale, medium projection quality CAVE
- *Graphics power:* high
- *Application level:* complex

The software used, COVISE (Rantzau and Lang 1998), is in principle based on a data flow paradigm found in visualization packages. In contrast to most pure data flow packages, COVISE uses the concept of data and function objects (modules) that can be arbitrarily distributed across machines. The data objects are transferred between hosts using special request broker modules (one for each host). A central controller module is responsible for proper synchronization and execution of a module network constructed beforehand in a visual application builder.

5.2. Process Simulation/Factory Planning

Near-reality perception (immersion) in virtual environment is used to plan and investigate work sequences and activities. Immersion is achieved through 3D representation of information in real time by integrating several human senses, with the user perceiving himself or herself as part of the scene. The interaction differs from input devices used so far in the possibility of direct intervention in the 3D space. The goal is to adapt the work environment to human responses and physique. The simulation of the work environment and the interaction of humans with their work surroundings is necessary to produce a near-reality perception. It is possible in the surroundings to conduct ergonomic investigations with subjects without a physical test setup.

Figure 7 shows a snapshot of a robot simulation for an automotive welding cell, DaimlerChrysler A-Class manufacturing plant in Rastatt, Germany. This application is based on the Fraunhofer IAO software Lightning attached to a kinematics simulation package from Fraunhofer IPA:

- *Goal:* visualization of the production process
- *Data:* surface geometry (medium), surface appearance (medium), logistic data kinematics
- *Visualization:* highest possible field of view, 1:1 scale, medium projection quality CAVE
- *Graphics power:* high

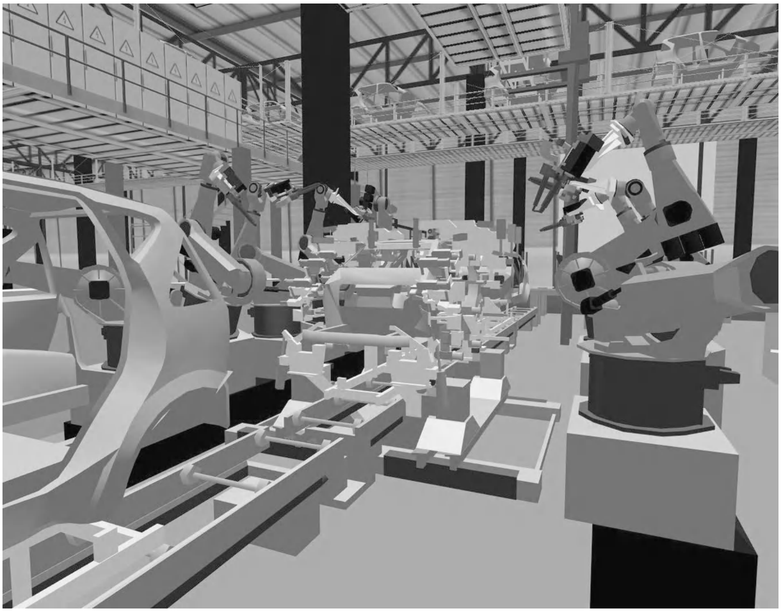


Figure 7 Robot Welding Cell of a Production System. (*Source:* HLRS and Fraunhofer IPA)

- *Application level:* advanced

The immersive visualization of production and workflow simulations greatly increases the clearness of the production cycles and helps to identify problem areas quickly. Therefore, a large field of view is essential. The ideal IPT system used here is a CAVE-like multiwall projection system. Currently, most of these applications are used for presentation purposes only. Offline teaching of robots in VR poses a number of problems, most involving accuracy and usability.

5.3. Education and Training

By means of virtual environments a learning context can be created where the multiple aspects of human intelligence are addressed. With sufficient perception and interaction qualities to give the learner a feeling of intuitive handling of virtual objects to the learner, the following VE applications are possible in the area of apprenticeship.

- Interactive demonstration of objects and events
- Mediation of (abstract) knowledge based on the user's experience (e.g., exploring a molecular structure)
- Mediation of skills and training of behavior in dangerous situations and environments
- Knowledge mediation beyond humans physical limitations (e.g., exploring how a gasoline engine functions by opening up the combustion chamber)
- Exploring distant places or past epochs
- Advancement of creative abilities by free design and impart of imagined objects

The use of VE systems in teaching, education, and training can be very effective because it agrees with the psychological and pedagogical knowledge, which is that the greatest success in learning can result from the practical application of educational programs and case studies. The success of learning can be increased by parallel addressing which means frequently changing the sensory channels, and by the inclusion of the learner's own actions. Concentration and learning success increase when thinking structures are fixed in the imagination or the ideas of a third person are made intelligible.

Particularly in the artistic field, VE should contribute to supporting students' creativity. In addition to this, when VE is is used in language education, the simulation of a suitable environment will make the learning of a foreign language more intensive and therefore more effective.

5.4. Scientific Visualization

The 3D representation of complex data is applied to the examination of comprehensive geometry data records and, especially in the scientific visualization field, the representation of mathematically computed data. The visualization of mathematically computed data has the goal of rendering simulation results visible, audible, or tactile.

6. PROCESS INTEGRATION OF VR APPLICATIONS

VE must be integrated in the development process with respect to the available 3D geometry database, the communication structure between departments, and the specific characteristics of the product.

The use of VE-based tools in a virtual immersive work environment with real-time capability speeds up finding and reviewing concepts in the early stages of the development process. Product aspects and features of prototypes can be examined in the virtual environment that so far have been verifiable only in real, physical prototypes (Figure 8). For this purpose, product aspects and features can be represented directly or indirectly by means of metaphors. In a passenger car, for example, the shape and the room required by individual units are directly presentable product aspect and features, in contrast to the dynamic load of the chassis (which is indirectly presentable by means of metaphors). Connecting these steps in the early development with quality, time, and cost management will create several advantages.

6.1. Analyzing Engineering Tasks for VE Applications

The field of engineering, apart from planning tasks, above all handles complex contents of engineering and the natural sciences, which will be discussed in the global engineering network on the basis of CSCW in the future. Unlike planning tasks, engineering and the natural sciences can in many cases not be discussed exclusively on the basis of numerical values or texts. Frequently they involve complex, multidimensional problem scopes. Therefore, a discussion of these subjects is supported by a spatial, multidimensional representation of the contents.

The major weak point of the computer-assisted simulation and design tools that have been available so far is that ideas of 3D geometry and/or functions are handled by 2D input/output media. This disadvantage is intensified by the growing complexity and multidimensionality of the problem at hand. A thoroughly interdisciplinary discussion of the scope of the problem is hampered by the existing tools, which are usually very specialized, whereas the multisensory and immersive systems of virtual reality (VR) support quick and interdisciplinary visualization of contents.

The use of conventional computer systems in such complex processes as the design of product shapes or the evaluation of assembly processes requires the user to have a great ability to think in abstract terms. To provide more intuitive access to these problem scopes, the problems can be handled in a VE.

Virtual prototypes (VP) are defined as a computer-based simulation of a technical system or subsystem with a degree of functional behavior which is comparable to corresponding physical pro-

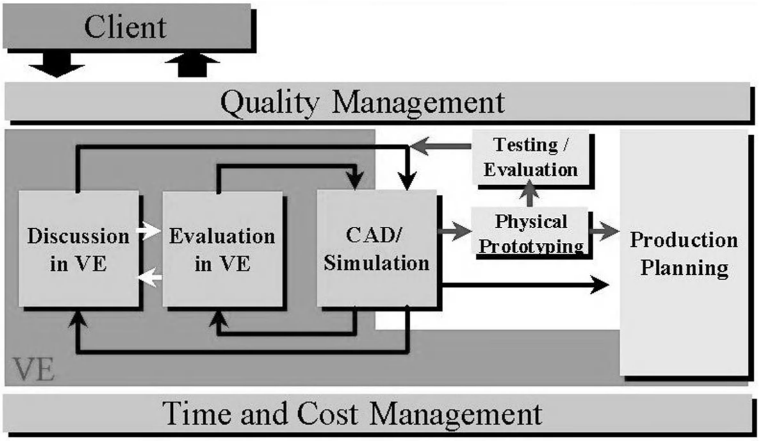


Figure 8 Use of VE in the Development Process.

totypes (Haug et al. 1993). Practically speaking, VP is a 3D graphical representation of an object that can be manipulated in its geometry, surface, behavior, and other degrees of freedom in real time. The source of the information can be parts from the traditional design process, such as hand-made models and drawings that are digitized in all their features, such as by 3D laser scanners with videocapture capabilities for catching surface information. Another option is the use of digital databases, which can store nearly everything from CAD files on materials and colors up to rules and standards for the generation of a new virtual prototype. Because all the data used inside the virtual prototype are digital, changes to the virtual prototype can be reflected quickly in the digital model without physical effort. This same model can be used in different ways to get feedback on various aspects of the design and its performance.

Are there any specific characteristics of the tasks that guarantee a maximum benefit when using immersive VR systems, or rather an IPT system? If we analyze engineering tasks focusing on the use of immersive projection environments, we should do so with regard to aspects of the following areas:

1. Representation of the product (model, prototype) and the user
2. Communication between users about the engineering task
3. The Engineering task itself

In addition, we have to check our requirements for costs, quality, and time. For costs, key words are investment, the costs for physical models, the assessment of earlier time-to-market, decision support through more realistic models, and checking of customer acceptance of reduction of risks. The quality can be enhanced through more variants (digital mock-ups), the integration of implicit know-how (as well as experience) by assessing the digital models, and the reduction of bugs early in the development process. This can also be a relevant factor in shortening the time of the whole development process, in addition to there being fewer real prototypes necessary and the digital manufacturing process needing less time to achieve the final quality of the series.

Most tasks can be done by transferring the tasks in immersive VR systems. Therefore, the user can choose different IPT systems based on the tables above. But each task has its own specific characteristics and has its specific benefits. Below are some typical tasks based on examples and experiences regarding benefits:

- Visual evaluation of the form
- Visual evaluation of the position of different parts
- Visual evaluation of proportions
- Assessment of surface quality
- Evaluation of color concepts
- Evaluation of colors
- Modification of geometry
- Modification of colors/textures/material appearance

TABLE 1 Ranking of IPT Task Requirements

		CAVE-Like	Power Wall	Benches	HMD	HCD
Scale	micro	+	0	+	+	+
	1:1	++	+	0	+	+
	macro	++	++	−	+	+
Simulation of human activities by the user		++	0	0	+	−
Immersion of the user in the VE		++	+	−−	+	+
Immersion of the user in the product		++	0	−−	+	+

−−: not possible
−: not suitable
0: possible
+: suitable
++ : very suitable

- Assessment of kinematics
- Auditory evaluation

Tasks that allow visual spatial perception for a subjective assessment of product characteristics are especially convenient for the use of immersive projection environments. Tasks with high complexity (especially in spatial aspects) that have to be discussed in an interdisciplinary team also have interesting advantages for using VR.

Because we use visual perception especially in immersive projection technology (IPT), we have to ask which product characteristics we can assess there. The product characteristics for which we can achieve reliable assessment differ between the real world and a virtual environment. Therefore, Table 2 ranks characteristics and their relative usefulness for direct representation and assessment in the different IPT systems.

The goal of digital mock-ups is the assessment of products early stage in the development process. For direct assessment, the digital mock-up should have nearly the same characteristics as the final product. But if we think only in such a direct way of assessment we get only a small part of the possibilities and advantages of digital mock-ups. As in other fields of science and technology, we have to learn to think of assessment in an indirect way. That means that it is not possible to measure or assess a specific parameter or characteristic of the product, but it is possible to measure or assess a number of other parameters or characteristics from which we can derive the specific characteristic. With the less useful ranked characteristics in Table 1 we should try this strategy of substitution or indirect assessment. Therefore, we have to look for adequate metaphors.

6.2. Requirements from the Engineering Side

For a quick development process, the definition of geometry must be, in principle, closely linked to the evaluation of particular product features that are influenced through geometry. For this reason, design and evaluation form iteration loops that reach a higher level of design of the prototype after each iteration step. The tools necessary for executing of the described tasks in the virtual environment are being researched, conceived, and developed within the scope of the subproject. In addition to the tools for design and evaluation, communication and cooperation in the virtual environment are primary requirements. Within this subproject, the fundamentals of the representation of the actor are researched. The representation model, which is based on the above-mentioned fundamentals, is integrated into the virtual working environment. The tools for design and evaluation are the basis for an interdisciplinary project by experts on the basis of CSCW in a virtual working environment. A user guidance system adapted to this environment has to be developed for system control.

The advantages of VE are:

- 1:1 scale
- Immersion

TABLE 2 Ranking of Product Characteristics with Respect to Representation in IPT

		CAVE-Like	Power Wall	Benches	HMD[a]	HCD
Geometry	Form	++	+	+	+	+
	Continuity of the surface	+	+	++	+	+
	Proportion	++	+	0	0	0
Material	Color	0	0	0	+	++
	Surface structure		+	0	+	++
	Haptic characteristics	−−	−−	−−	0[b]	0[b]
Kinematics		++	+	+	+	+
Acoustics		++	+	+	+	+
Functionality	User interface	+[c]	0[c]	0[c]	−	−−
	Physical effects	−	−	−	+[d]	0[d]

−−: not possible
−: not suitable
0: possible
+: suitable
++: very suitable

[a] Especially LCD-based HMD in comparison to CRT-based BOOM/PUSH.

[b] In combination with a force feedback device (i.e., PHANTOM).

[c] In an augmented reality environment.

[d] In combination with special systems for physical effects (i.e., temperature device).

- Interactivity
- Intuitiveness

3D visualization from the individual line of sight and the most varied interaction possibilities through gestures, speech input, and new input/output media are still to be developed. As a result of the free positioning and scaling of the user inside and outside the object to be planned, it will be possible for a large number of persons to work simultaneously on an object. The object will not have to be broken down into its elements.

Taking advantage of a new process chain the whole process of product development and product improvement can benefit, like other fields of work, from VE-based tools:

- Reduction of cost and time for building prototypes, making it possible to test and evaluate more variants,
- Faster testing and evaluation through virtual testbeds
- Scalable test persons
- Online changes of parameters
- Enhancement of quality and effectiveness of project drawing and design performance
- Improvement of the product quality, e.g., by integrating the customer earlier and getting better feedback from the customer
- Better ecological balance through saving resources during iteration cycles

6.3. Data Requirements

3D CAD data consist of continuous area and volume description and can therefore not be used directly as a VR database. The data have to be transformed in a polygonal database by a process called tessellation. For most of the current proprietary CAD formats there are one or more data filters, called tessellators, that can also be used for generic formats such as IGES, VDA-FS, and STEP. But the best way in general is as close as possible to the generated data from CAD or other modeling tools.

Aspects of the VR database include:

- *Geometry:* continuous (CAD, STEP, IGES, etc.), discrete (polygons, voxels, STL, VRML, etc.)
- *Surfaces:* textures, reflection properties, acoustic properties
- *Kinematics*
- *Material properties*

The most important guideline for tessellation is variation of the chordal height. This value indicates the maximum permissible variation of polygon data from CAD data. However, this guideline does not fix the tessellation. There are many tessellation algorithms that can guarantee such a maximum variation. From a VR point of view, the quality assessment of the data produced concentrates on the following criteria:

- The number and kind of the produced polygons are directly linked with the frame rates achieved by visualization.
- The maintenance of the topology of CAD objects guarantees the avoidance of cracks and other inconsistencies in the VR database.
- Only the indication of correct normals in every point of polygon representation enables the correct assessment of the tessellated surface.

Most tessellators are applications within the CAD system and not independent data filters on the demand line. The reasons are as follows:

- Some relevant information is stored in proprietary CAD data only as a reference in library data of the CAD system. The tessellator must have access to these data.
- The same CAD systems on different basic programs may produce different data with the same geometric information (e.g., codepage problem with CATIA)

Tessellators of generic data formats such as IGES and VDA-FS can work on a stand-alone basis because all necessary information is contained in these data.

6.3.1. Partitioning of VR-Data Preparation

The preparation of the database (especially the 3D Geometry-inclusive attributes for surfaces, kinematics, etc.) should be distributed to the departments where the data are generated. This means that every user of the IPT system is responsible for its own VR data. The know-how for generating VR data must first be taught.

Usually there are several strategies for generating VR data, and a compromise should be made regarding the visual quality (degree of detail of the objects) and the performance of the application. This decision could best be optimized by the users who are responsible for the results.

Apart from pure geometrical information (points, polygons, normals) detailed surface information such as color, material (in the understanding of VR), and texture are needed in order to visualize CAD data close to reality. Many CAD systems already offer the possibility of defining such properties in a CAD record. However, in construction this is often ignored. The evaluation of such information, if available, is not carried through by every tesselator.

6.3.2. Material: Color, Textures

To deposit geometric data with color and material qualities is not a difficult problem and can be automated if the parameters are known. Because with every CAD construction the material in use is determined, it would be relatively simple to construct a valid databank that can link VR material qualities with material or equipment. The information needed could then be extracted from this databank, in principle. The challenge is integration in a PDM system. Therefore, the data structure of the VR objects should be designed to be flexible and modular.

The use of textures offers the possibility of reproducing surface qualities such as roughness in a VR system. How far this process can be automated must be examined. Probably it does not make sense to replace CAD geometry as low LOD stage by textures.

6.3.3. Performance Optimization

There are several aspects of performance optimization. In the construction process the tessellation of the CAD data is one example. The number of polygons produced determines the speed of real-time visualization. The finer the tessellation, the worse the graphic performance.

One problem in controlling the complexity is the goal in tessellating. In general, it is not possible to set the number of polygons that have to be produced (= complexity of VR databases) in advance but a quality criterion, namely the variation of the chordal height, is set in advance. This is ideal from the user's point of view. The estimation of a resulting polygon number from only the variation of the chordal height is extremely unreliable. In order to use the complexity as a line function, the tessellator has to be iterated until the desired complexity is reached.

However, the tessellation algorithm used has an even stronger influence on the complexity than the variation of the chordal height. If one follows the argument of the developers of OpenGL Optimizer (SGI), huge differences in complexity of the variation of the chordal height are possible. It is therefore not unrealistic to minimize the database to 20% by determining the CAD topology and total analysis of CAD surfaces. The challenge here is also integration in a PDM system where different models for high-end IPTs, desktop visualization, video-based animation, printing quality, and so on can be managed.

6.3.4. The Training Concept

It is well known that CAD requires a great deal of training, especially 3D CAD. Although the functionality of the current IPT-based tools is less complex, their handling must be taught.

Table 3 gives a brief overview of the migration from existing applications. Because a migration strategy depends very strongly on the available products, we show here only a very abstract version of migration.

7. CONCLUSIONS

The development of products mainly will base on virtual prototypes in future. Many of the virtual prototype characteristics can only be evaluated in a VE, which can provide developer or customer with an adequate perception. Changes to the virtual prototype can be quickly reflected in the digital model without physical effort—a great advantage for the engineering process. This same model can be used in different ways to get feedback on various aspects of the design and its performance. VE-based tools are being used as a vehicle for clear communication and understanding, such as in seeing how assemblies will fit together without a real physical prototype.

Immersive projection technology (IPT) is ready for productive applications. Despite the complexity of such systems, it is possible to provide them as a productive tool. Analysis and understanding of the special requirements is necessary to deliver the most appropriate system. An intuitive user

TABLE 3 Migration of Data from Existing Model

Type of Application	Type of Data	Migration
3D CAD/modeler	Free-form surfaces, parametric geometry, trim curves, surface/material properties, assembly information	Tessellation (static dynamic), data reduction (LOD), surface properties −> textures, materials, reflection mapping
Prototypes	3D scanning data sets, surface reconstruction −> CAD data sets	Data repair, data reduction, color information −> texture
Simulation (FEM/GEM, etc.)	Geometry, additional simulation data, temporal changes	Geometrical representation of simulation data, intelligent structures
Functionality	Interaction human–model, operating/behavior	Description language, feedback, intuitive user interaction

interface is very important for user productivity and acceptance. The design of 3D user interfaces is clearly progressing, at least special-purpose user interfaces created for very specific tasks. This, together with careful VR workplace design, is dramatically increasing user productivity and acceptance.

While the return on investment can be calculated immediately in some cases, in other cases will only see some excited engineers playing with the new technology.

Migration and integration of IPT and VE in standard engineering processes is a part of the second principal stream. This includes plug-ins for established VR software as well as the support of data transfer and translation of data from legacy applications. This work will depend on available standards. On the other hand, there will be a trend toward the integration of VR functionality into existing CAD/CAE packages.

All innovative media of the future will contain three major components: interactivity, co-operation, and new possibilities in receiving experience by including several senses at once. The further development of virtual reality (and all technologies included under it, such as IPT systems) must be carried out on the basis of an user-centered approach—that is, the problems and limits of a person working in an immersive VE must be considered. This is how the necessary user acceptance can be achieved.

REFERENCES

Aukstakalnis, S., and Blatner, D. (1992), *Silicon Mirage: The Art and Science of Virtual Reality,* Peachpit Press, Berkeley, CA.

Blach, R., Landauer, J., Rösch, A., and Simon, A. (1998), “A Highly Flexible Virtual Reality System,” in *Future Virtual Environments,* Elsevier, Amsterdam.

Brooks, F. P. (1998), “Grasping Reality through Illusion—Interactive Graphics Serving Science,” in *Proceedings of CHI ’88* (Washington, DC).

Brown, R. G. (1996), “An Overview of Virtual Manufacturing Technology,” in *Proceedings of Advisory Group for Aerospace Research and Development* (May 6–10), AGARD, Sesimbra, Portugal, pp. 1–11.

Bryson, S. (1993), Implementing Virtual Reality,” in *Course Notes 43, ACM SIGGRAPH ’93* (Chicago), ACM, New York.

Bryson, S., Johan, S., and Schlecht, L. (1997), “An Extensible Interactive Visualization Framework for the Virtual Windtunnel,” in *Proceedings of Virtual Reality International Symposium ’97* (Albuquerque, March 1–5), IEEE Computer Society Press, Washington, DC, pp. 106–113.

Cruz-Neira, C. (1993), “Virtual Reality Overview,” in *Course Notes 23, ACM SIGGRAPH ’93* (Anaheim, CA, January 1–11), ACM, New York.

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R., and Hart, J. C. (1992), “The CAVE, Audio Visual Experience Automatic Virtual Environment,” *Communications of the ACM,* Vol 35, No. 6, pp. 64–72.

Dani, T. H., and Gadh, R. (1996), “COVIRDS: A Framework for Conceptual Shape Design in a Virtual Environment,” in *Proceedings of NSF Grantees 1996 Symposium* (also at smartcad.me.wisc.edu/~tushar/nsf/nsf.html).

Durlach, N. (1997), "VR Technology and Basic Research on Humans," *in Proceedings of Virtual Reality International Symposium '97* (Albuquerque, March 1–5), IEEE Computer Society Press, Washington, DC, pp. 2–3.

Ellis, S. (1995), "Human Engineering in Virtual Environments," in *Proceedings of Virtual Reality World '95* (Stuttgart, February 21–23), IDG, Munich, pp. 295–301.

Ellis, S. R., Breant, F., Menges, B., Jacoby, R., and Adelstein, B. D. (1997), "Factors Influencing Operator Interaction with Virtual Objects Viewed via Head-Mounted See-Through Displays," in *Proceedings of Virtual Reality International Symposium '97* (Albuquerque, March 1–5), IEEE Computer Society Press, Washington, DC, pp. 138–145.

Foley, J., and Silbert, J. (1989), "User-Computer Interface Design," *in Lecture Notes, No. 1, CHI '89 Conference* (Austin, TX).

Haug, E. J., Kuhl, J. G., and Tsai, F. F. (1993), "Virtual Prototyping for Mechanical System Concurrent Engineering," in *Concurrent Engineering: Tools and Technologies for Mechanical System Design,* E. J. Haug, Ed., Springer, Berlin.

Iovine, J. (1994), *Step into Virtual Reality,* McGraw-Hill, New York.

Krüger, W., and Fröhlich, B. (1994), "Responsive Workbench," in *Virtual Reality '94: Anwendungen und Trends* (Stuttgart, February 9–10), H.-J. Warnecke and H.-J. Bullinger, Eds., Springer, Berlin, pp. 73–80.

Leston, J., Ring, K., and Kyral, E. (1996), *Virtual Reality: Business Applications, Markets and Opportunities,* OVUM Reports, Ovum Ltd., London.

Pandzic, I., Sunday, C., Tolga, K., Thalmann, N. M., and Thalmann, D. (1996), "Towards Communication in Networked Collaborative Virtual Environment," in *Proceedings of the Conference of the FIVE Working Group,* PERCRO, Scuola Superiore S. Anna, Pisa, pp. 37–47.

Rantzau, D., and Lang, U. (1998), "A Scalable Virtual Environment for Large Scale Scientific Analysis," *Future Generation Computer Systems,* Vol 14, pp. 215–222.

Shawver, D. M. (1997), "Virtual Actors and Avatars in a Flexible User-Determined-Scenario Environment," in *Proceedings of Virtual Reality International Symposium '97* (Albuquerque, March 1–5), IEEE Computer Society Press, Washington, DC, pp. 170–179.

ADDITIONAL READING

Bauer, W., Breining, R., and Rössler, A., "Co-operative, Virtual Planning and Design," in *Proceedings of Virtual Reality World '95* (Stuttgart, February 21–23), IDG, Munich, 1995, pp. 213–223.

Bergamasco, M., "The GLAD-In-ART Project," in *Virtual Reality '93,* IPA-/IAO-Forum (Stuttgart, February 4–5), Springer, Berlin, 1993, pp. 251–258.

Bricken, W., "VEOS: Preliminary Functional Architecture," *in ACM SIGGRAPH '91 Course Notes, 18th International Conference on Computer Graphics and Interactive Techniques* (Las Vegas), 1991, pp. 46–53.

Brückmann, R., and Gottlieb, W., "Spatial Perception of Vehicle Interior," in *Proceedings of Virtual Reality World '95* (Stuttgart, February 21–23), IDG, Munich, 1995, pp. 459–461.

Burdea, G., "Research on Portable Force Feedback Masters for Virtual Reality," in *Proceedings of Virtual Reality World '95* (Stuttgart, February 21–23), IDG, Munich, 1995, pp. 317–324.

Cruz-Neira, C., Ed., "Applied Virtual Reality," in *Course Notes 14, SIGGRAPH '98.*

Dai, F., Ed., *Virtual Reality for Industrial Applications,* Springer, Berlin, 1997.

Earnshaw, R., Gigante, M., and Jones, H., Eds., *Virtual Reality Systems,* Academic Press, London, 1993.

Iwata, H., "A Six-Degree-of-Freedom Pen-Based Force Display," in Proceedings of the Fifth International Conference on Human-Computer Interaction (Orlando, FL, 1993), pp. 651–656.

Kalawasky, R.*, The Science of Virtual Reality and Virtual Environments,* Addison-Wesley, Reading, MA, 1993.

King, D., "Heads up," *Computer Graphics World,* Vol. 16, No. 11, 1993, pp. 41–46.

Odegard, O., "Social Interaction in Televirtuality," *in Proceedings of Virtual Reality World '96* (Stuttgart, February 13–15), Computerwoche, Munich, 1996, pp. 1–7.

Pausch, R., and Bryson, S., "Interface Methods in Virtual Reality," *in Course Notes 2, ACM SIGGRAPH '94* (Anaheim, CA, May 1–17), ACM, New York, 1994.

Satava, R. M., Morgan, K., Sieburg, H. B., Mattheus, R., and Christensen, J., Eds., *Interactive Techology and the New Paradigm for Healthcare,* IOS Press, Amsterdam, 1995.

Sutherland, I. E., "The Ultimate Display," in *Proceedings of IFIP Congress,* 1965.

Wexelblat, A., Ed., *Virtual Reality: Applications and Explorations,* Academic Press, Boston, 1993.

V.D

Optimization

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

CHAPTER 97

Linear Optimization

A. “Ravi” Ravindran
Pennsylvania State University

Roy Marsten
Cutting Edge Optimization

1. **INTRODUCTION** **2524**
2. **FORMULATION OF LINEAR MODELS** **2524**
    - 2.1. Basic Steps in Formulation 2524
    - 2.2. Product-Mix Problem 2524
        - 2.2.1. Example 1 (Product-Mix Problem) 2524
3. **BASIC ASSUMPTIONS OF LINEAR MODELS** **2525**
    - 3.1. Proportionality 2525
    - 3.2. Additivity 2525
4. **HANDLING NONLINEARITIES BY LINEAR PROGRAMMING** **2526**
    - 4.1. Piecewise Linear Functions 2526
    - 4.2. Max-Min Problems 2527
    - 4.3. Handling Absolute Value Functions 2527
5. **SIMPLEX ALGORITHM** **2527**
    - 5.1. Basic Principles 2528
        - 5.1.1. Example 2 2528
    - 5.2. General Steps 2528
    - 5.3. Computational Efficiency of the Simplex Method 2529
6. **INTERIOR POINT METHOD** **2530**
    - 6.1. Newton’s Method 2530
    - 6.2. Lagrange Multiplier Method 2531
    - 6.3. Fiacco and McCormick Algorithm 2531
    - 6.4. Application to Linear Programming 2532
    - 6.5. Computational Efficiency of the Interior Point Method 2534
7. **COMPUTER SOLUTION OF LINEAR PROGRAMS** **2534**
    - 7.1. Evolution of Commercial Packages for LP 2534
    - 7.2. PC Software for LP 2535
    - 7.3. High-End LP Software 2535
    - 7.4. LP Modeling Languages 2535
    - 7.5. LP Software on the Internet 2536
8. **SENSITIVITY ANALYSIS IN LINEAR PROGRAMMING** **2536**
    - 8.1. Reasons for Sensitivity Analysis 2536
    - 8.2. Practical Uses 2536
    - 8.3. Simultaneous Variations in Parameters 2537
        - 8.3.1. 100% Rule for Objective Function Coefficients 2537
        - 8.3.2. 100% Rule for RHS Constants 2538
9. **APPLICATIONS OF LINEAR PROGRAMMING** **2538**

REFERENCES **2538**

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

Linear programming (LP) defines a particular class of optimization problems that meet the following two conditions:

1. The objective function to be optimized can be described by a linear function of the decision variables.
2. The operating rules or constraints governing the process (e.g., limited resources) can be expressed as a set of linear equations or inequalities.

LP techniques are widely used to solve a number of military, economic, industrial, and social problems.

1. A large variety of problems in diverse fields can be represented (within reasonable accuracy) as LP models.
2. Efficient techniques for solving LP problems are available.
3. Data variation (sensitivity analysis) can be handled with ease through LP models.

2. FORMULATION OF LINEAR MODELS

2.1. Basic Steps in Formulation

The three basic steps in constructing an LP model are:

Step 1: Identify the unknown variables to be determined (design or decision variables) and represent them in terms of algebraic symbols.

Step 2: Identify all the restrictions or constraints in the problem and express them as linear equations or inequalities, which are linear functions of the unknown variables.

Step 3: Identify the objective or criterion and represent it as a linear function, which is to be maximized or minimized.

The following example illustrates these basic steps.

2.2. Product-Mix Problem (Ravindran et al. 1987)

2.2.1. Example 1 (Product Mix Problem)

A company manufactures three products that require three resources—labor, material, and administration. The company's production engineering department has furnished the following data:

	Products		
	1	2	3
Labor (hr / unit)	1	1	1
Material (lb / unit)	10	4	5
Administration (hr / unit)	2	4	6
Profit ($ / unit)	10	6	4

The supply of raw material is restricted to 600 lb / day. The daily availability of manpower is 100 hr. There are 300 hr of administration. Formulate a linear programming model to determine the daily production levels of the various products in order to maximize the total profit.

The formulation is as follows:

Step 1. Identify the decision variables. The unknown activities to be determined are the daily rates of production on the three products. Represented by algebraic symbols, they are

x_1 = daily production of product 1
x_2 = daily production of product 2
x_3 = daily production of product 3

Step 2. Identify the constraints. In this problem the constraints are the limited availability of the three resources—labor, material, and administration. Product 1 requires 1 hr of labor for each unit, and its production quantity is x_1. Hence the requirement of labor for product 1 alone will be x_1 hr (assuming a linear relationship). Similarly, products 2 and 3 will require x_2 and x_3 hr, respectively. Thus, the total requirement of labor will be $x_1 + x_2 + x_3$ which should not exceed the available 100 hr. So the labor constraint becomes

$$x_1 + x_2 + x_3 \leq 100$$

The raw material requirements will be $10x_1$ lb for product 1, $4x_2$ lb for product 2, and $5x_3$ lb for product 3. Thus, the raw material constraint is given by

$$10x_1 + 4x_2 + 5x_3 \leq 600$$

Similarly, the constraint for administration becomes

$$2x_1 + 2x_2 + 6x_3 \leq 300$$

In addition, we restrict the variables x_1, x_2, and x_3 to having only nonnegative values. This is called the nonnegativity constraint, which the variables must satisfy. Most practical LP problems will have this nonnegative restriction on the decision variables.

Step 3. Identify the objective. The objective is to maximize the total profit from sales. Assuming that a perfect market exists for the products such that all that is produced can be sold, the total profit from sales becomes

$$Z = 10x_1 + 6x_2 + 4x_3$$

Thus, the LP model for our product mix problem is to find numbers x_1, x_2, x_3 that will maximize

$$Z = 10x_1 + 6x_2 + 4x_3$$

subject to the constraints

$$\begin{aligned} x_1 + x_2 + x_3 &\leq 100 \\ 10x_1 + 4x_2 + 5x_3 &\leq 600 \\ 2x_1 + 2x_2 + 6x_3 &\leq 300 \\ x_1 \geq 0 \quad x_2 \geq 0 \quad x_3 &\geq 0 \end{aligned}$$

3. BASIC ASSUMPTIONS OF LINEAR MODELS

The LP approach to modeling a system under study is to decompose the system into its elementary functions, or "activities." In Example 1 there were three activities—manufacture of one unit of product 1, manufacture of one unit of product 2, and manufacture of one unit of product 3. The decision variables merely define the levels at which these activities are to be carried out. Of course, the aim of the LP model is to determine the optimal activity levels. To change the activity level, the input and output flows into each activity have to be changed. These flows are called "items." In Example 1 the input flows were labor, material, and administration, and the output was profit in dollars.

There are two basic assumptions in the formulation of all LP models: proportionality and additivity.

3.1. Proportionality

This guarantees that the flow of items into and out of an activity is directly proportional to its activity level. If one unit of product 1 requires 1 hr of labor, 10 lb of material, and 2 hr of administration, then to make x units of product 1 will require x hr of labor, $10x$ lb of material, and $2x$ hr of administration for any x. Similarly, the unit profit from selling product 1 is always \$10, irrespective of how many units are sold.

3.2. Additivity

This assumption implies that the total usage of an item is equal to the sum of the item usages of each individual activity at its specified level. In Example 1 the total material consumption was equal to the sum of the material consumed by the individual products.

For a more detailed discussion on the input–output approach to modeling and LP assumptions. the reader is referred to Dantzig 1963. The proportionality and additivity assumptions imply that all the constraints of the LP problem can be expressed as linear equations or inequalities and that the objective is a linear function of the decision variables. It is common to find some practical problems violating one or more LP assumptions. In such cases, by using clever formulations or good approximations, one could still use LP. Some of these are discussed in the next section.

4. HANDLING NONLINEARITIES BY LINEAR PROGRAMMING (Ravindran, et al., 1987; and Murty 1995)

Nonlinearities can arise in a number of ways in optimization problems in either the objective function or the constraints. Some of the nonlinearities can be handled by LP methods, whereas the rest have to be solved by specialized nonlinear programming methods.

4.1. Piecewise Linear Functions

A piecewise linear function arises when the per-unit contribution (cost) depends on the level of sales (production). For example, consider a product whose profit contribution is \$10/unit for the first 40 units, \$8/unit for the next 60 units, and \$5/unit for the rest. The nonlinearity of the profit function is apparent if a graph is plotted between total profit and quantity sold. This is illustrated by Figure 1, which is called a piecewise linear function since it is linear in the region (0, 40), (40, 100), and (100,∞). Partitioning the quantity sold into three activities means that the profit function could be expressed as a linear function as follows:

x_1 = quantity sold at \$10/unit profit
x_2 = quantity sold at \$8/unit profit
x_3 = quantity sold at \$5/unit profit

The amount of product sold is $x_1 + x_2 + x_3$, and the objective function is to maximize $Z = 10x_1 + 8x_2 + 5x_3$. Since there is a limit on how many units can be sold for a certain profit, we need the following constraints:

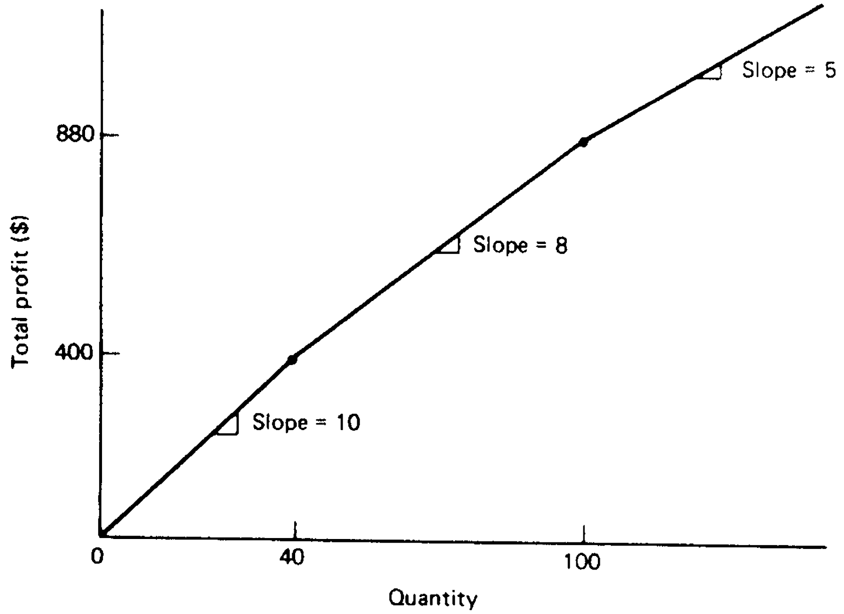


Figure 1 Example of a Piecewise Linear Function.

$$0 \leq x_1 \leq 40 \quad 0 \leq x_2 \leq 60 \quad 0 \leq x_3$$

When the objective function is maximized, it is easy to see that x_2 will not become positive until x_1 reaches its limit of 40. Similarly, x_3 cannot be positive until $x_1 = 40$ and $x_2 = 60$.

One could extend this idea to handle smooth nonlinear profit functions by approximating them into piecewise linear functions. It is important to realize that LP methods can be used successfully to handle piecewise linear functions as long as the following hold:

1. In a maximization problem the piecewise linear function must have decreasing slope or be a concave function (i.e., the per-unit contribution must be decreasing or at least nonincreasing).
2. In a minimization problem the function must have increasing slope or be a convex function (i.e., the per unit cost must be increasing or at least nondecreasing).

If these properties do not hold, then one has to use the more complicated integer programming formulations.

4.2. Max-Min Problems

In some optimization problems one may encounter a nonlinear objective that is to maximize the minimum of several variables or functions. Consider an assembly made up of three different parts. Let x_1, x_2 and x_3 be the decision variables denoting the number of parts 1, 2, and 3 produced, respectively. If management wishes to maximize the number of assemblies, then the objective function becomes

$$\text{Maximize [minimum of } (x_1, x_2, x_3)]$$

Even though this is a nonlinear function, it can be linearized as follows: Let y denote the number of assemblies made. Then the linear objective function is

$$\text{Maximize } y \tag{1}$$

Since y is the minimum of x_1, x_2, x_3, we get the three additional constraints

$$y \leq x_1 \tag{2}$$

$$y \leq x_2 \tag{3}$$

$$y \leq x_3 \tag{4}$$

The inequalities (2), (3), and (4) in conjunction with (1) are equivalent to maximizing the minimum of x_1, x_2 and x_3.

4.3. Handling Absolute Value Functions

Absolute value sign in the constraint can be handled by replacing it by two constraints. For example, a nonlinear constraint of the type

$$|x_1 - x_2| \leq 30 \tag{5}$$

is equivalent to the following two linear constraints:

$$x_1 - x_2 \leq 30$$

$$-x_1 + x_2 \leq 30$$

Nonlinear constraints of the type given in inequality (5) occur frequently in machine balancing. If x_1 represents the daily utilization of machine 1 in minutes and x_2 is the utilization for machine 2, then inequality (5) is equivalent to the machine balancing constraint that no machine run more than 30 min/day longer than the other machine.

5. SIMPLEX ALGORITHM

The simplex algorithm as developed by G. B. Dantzig in 1947 is an iterative procedure for solving LP problems. The theory of the simplex algorithm and its many computational refinements are fully presented in several outstanding textbooks (Dantzig 1963; Murty 1995; Ravindran et al. 1987; Gass 1975) The intent of this section is to describe briefly the basic principles of the simplex method.

5.1. Basic Principles

5.1.1. Example 2

Consider the following LP problem (Ravindran et al. 1987):

$$\begin{aligned} \text{Minimize } Z &= 40x_1 + 36x_2 \\ \text{Subject to } \quad & x_1 \leq 8 \\ & x_2 \leq 10 \\ & 5x_1 + 3x_2 \geq 45 \\ & x_1 \geq 0 \qquad x_2 \geq 0 \end{aligned}$$

In this problem we are interested in determining the values of the variables x_1 and x_2 that will satisfy all the restrictions and give the least value for the objective function. As a first step in solving this problem, we want to identify all possible values of x_1 and x_2 that are nonnegative and that satisfy the constraints. For example, a solution $x_1 = 8$ $x_2 = 10$ is positive and satisfies all the constraints. Such a solution is called a feasible solution. The set of all feasible solutions is called the feasible region. Solution of a linear program is nothing but finding the best feasible solution in the feasible region. The best feasible solution is called an optimal solution to the linear programming problem. In our example an optimal solution is a feasible solution that minimizes the objective function $40x_1 + 36x_2$. The value of the objective function corresponding to an optimal solution is called the optimal value of the linear program.

To represent the feasible region in a graph, every constraint may be plotted, and all values of x_1, x_2 that will satisfy these constraints can be identified. The nonnegativity constraints imply that all feasible values of the two variables will lie in the first quadrant. The constraint $5x_1 + 3x_2 \geq 45$ requires that any feasible solution (x_1, x_2) to the problem should be above the straight line $5x_1 + 3x_2 = 45$ (see Figure 2). Similarly, the constraints $x_1 \leq 8$ and $x_2 \leq 10$ are plotted. The feasible region is given by the shaded region *ABC* as shown in Figure 2. Obviously there are an infinite number of feasible points in this region. Our objective is to identify the feasible point with the lowest value of *Z*. The feasible points *A, B,* and C are called the comer points of the feasible region.

Observe that the objective function given by $Z = 40x_1 + 36x_2$ represents a straight line if the value of Z is fixed a priori. Changing the value of Z essentially translates the entire line to another straight line parallel to itself. To determine an optimal solution, the objective function line is drawn for a convenient value of *Z* such that it passes through one or more points in the feasible region. Initially *Z* is chosen as 600. By moving this line closer to the origin, the value of *Z* is further decreased (Figure 2). The only limitation on this decrease is that the straight line $40x_1 + 36x_2 = Z$ contain at least one point in the feasible region *ABC*. This clearly occurs at the comer point *A* given by $x_1 = 8$, $x_2 = 5/3$. This is the best feasible point giving the lowest value of *Z* as 380. Hence $x_1 = 8$, $x_2 = 5/3$ is an *optimal solution* and $Z = 380$ is the *optimal value* for the linear program.

In our example one of the comer points of the feasible region (namely, *A*) was an optimal solution. As a matter of fact, the following property is true for any LP problem: if there exists an optimal solution to an LP problem, then at least one of the corner points of the feasible region will always qualify to be an optimal solution.

This is the fundamental property on which the simplex method for solving LP problems is based. Even though the feasible region of an LP problem contains an infinite number of points, an optimal solution can be determined by merely examining the finite number of comer points in the feasible region. In LP terminology the comer point feasible solutions are known as basic feasible solutions.

Hence the simplex method for solving general LP problems is simply an orderly procedure for generating and examining different basic feasible solutions. In problems involving just two variables, one can easily draw the feasible region in a graph and identify the comer points that are basic feasible solutions. In practice, the LP problems involve hundreds of constraints and several thousand variables, and we need an algebraic procedure for generating the basic feasible solutions. The simplex method uses the classical Gauss-Jordan elimination scheme for generating the basic feasible solution. The Gauss–Jordan elimination can be represented by a sequence of vector–matrix operations and hence easily implemented on a digital computer.

5.2. General Steps

The general steps of the matrix-based simplex method are as follows:

1. Start with an initial basic feasible solution.

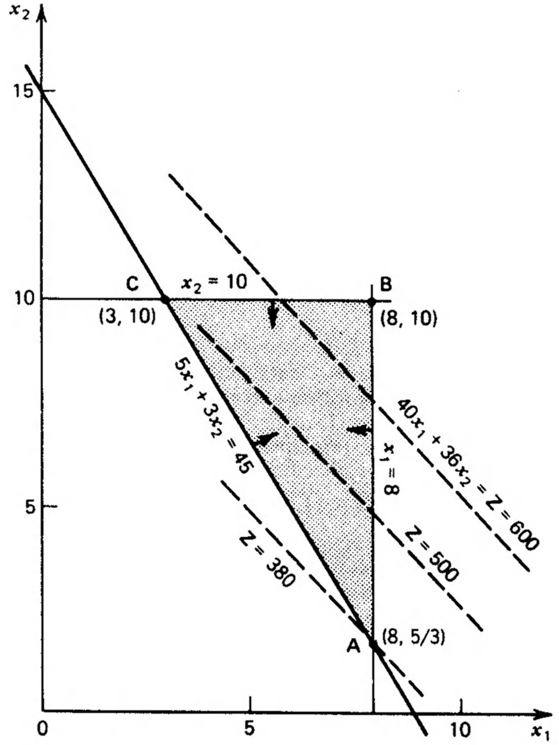


Figure 2 Graphical Solution of Example 2.

2. Improve the initial solution, if possible, by finding another basic feasible solution with a better objective function value. At this step the simplex method implicitly eliminates from consideration all those basic feasible solutions whose objective function values are worse than the present one. This makes the procedure more efficient than the naive approach, which would examine all the basic feasible solutions.
3. Continue to find better basic feasible solutions, improving the objective function values. When a particular basic feasible solution cannot be improved further, it becomes an optimal solution and the simplex method terminates.

Finding an initial basic feasible solution can be accomplished by temporarily loosening the definition of "feasible" in such a way that the origin becomes feasible. In Example 2, the simplex algorithm would start at the origin, make its first step to the temporarily feasible point *D*, and make its second step to the truly feasible and optimal point *A*. (The fact that the first truly feasible point is optimal is coincidental.)

5.3. Computational Efficiency of the Simplex Method

The computational efficiency of the simplex method depends on (1) the number of iterations (basic feasible solutions) to go through before reaching the optimal solution and (2) the total computer time to solve the problem. Much effort has been spent in studying the computational efficiency with regard to the number of constraints and the decision variables in the problem.

Empirical experience with thousands of practical problems shows that the number of iterations of a standard linear program with m constraints and n variables varies between m and $3m$, the average being $2m$. A practical upper bound for the number of iterations is $2(m + n)$. (Occasionally some problems have violated this bound.)

If every decision variable had a nonzero coefficient in every constraint, then the computational time would increase approximately in relation to the cube of the number of constraints, m^3. In practical large-scale models, however, typically fewer than 1% of the matrix coefficients are nonzero.

The use of sparse matrix techniques makes computation times unpredictable but far better than the m^3 would suggest.

It is to be noted that the computational efficiency of the simplex method is more sensitive to the number of constraints than to the number of variables. Hence the general recommendation is to keep the number of constraints as small as possible by avoiding unnecessary or redundant constraints in the formulation of the LP problem.

6. INTERIOR POINT METHOD

In 1984, a new and very different way of solving linear programs was introduced (Karmarkar 1984). Announced with much fanfare by Karmarkar's employer, AT&T Bell Laboratories, the new method was claimed to be 50 times faster than the simplex method. By 1990, Karmarkar's seminal work had spawned hundreds of research papers and a large class of what are now called interior point methods. It has become clear that while the initial claims were somewhat overenthusiastic, interior point methods dominate the simplex method for very large problems and for certain special classes of problems that have always been particularly difficult for the simplex method. Two such classes are highly degenerate problems (in which many different algebraic basic feasible solutions correspond to the same geometric comer point) and multiperiod problems that involve decisions in successive time periods that are linked together by inventory transfers. For both of these classes, the number of iterations taken by the simplex method can far exceed the $2m$ rule of thumb.

The definition of the class of very large problems for which interior point methods dominate is changing almost daily as computer implementations of the interior point method become more and more efficient. However, since 1984 there have been dramatic improvements in the computer implementations of the simplex method as well, largely spurred by the competition between the two methods. As a result, there is not currently much reason to prefer either method for LP models with a few hundred constraints. Beyond a few thousand constraints, however, the interior point method leaves the simplex method further and further behind as problem size grows.

Karmarkar's derivation was quite original, but it has since become clear that the whole family of interior point methods is equivalent to some classical ideas from the field of nonlinear programming (Gill et al. 1986). We shall present here a brief introduction to the current framework for these methods. A more detailed tutorial can be found in Marsten et al. 1990.

The classical building blocks that we need are Newton's method (Newton 1687) for unconstrained optimization, Lagrange's method (Lagrange 1788) for optimization with equality constraints, and Fiacco and McCormick's barrier method (Fiacco and McCormick 1968) for optimization with inequality constraints. Let us review these. A good general reference is Bazarra and Shetty 1979.

6.1. Newton's Method

One of the foundations of numerical mathematics is Newton's method for finding a zero of a function of a single variable: $f(x) = 0$. Given an initial estimate x^0, we compute a sequence of trial solutions.

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}$$

for $k = 0, 1, 2, \ldots$, stopping when $|f(x^k)| < \varepsilon$ where ε is some stopping tolerance, for example, $\varepsilon = 10^{-8}$.

Suppose we have n equations in n variables: $f(x) = 0$, where

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} \text{ and } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The Jacobian at the point x, $J(x)$, is defined as the matrix with (i, j) component

$$\frac{\partial f_i}{\partial x_j}(x)$$

and Newton's method looks like

$$x^{k+1} = x^k - [J(x^k)]^{-1} f(x^k)$$

Or, if we let

$$dx = x^{k+1} - x^k$$

denote the displacement vector and move the Jacobian matrix to the other side

$$J(x^k)dx = -f(x^k) \tag{6}$$

Newton's method can be applied to the unconstrained minimization problem as well: to minimize $g(x)$, take $f(x) = g'(x)$ and use Newton's method to search for a zero of $f(x)$. Each step of the method can be interpreted as constructing a quadratic approximation of $g(x)$ and stepping directly to the minimum of this approximation.

If g is a real valued function of n variables, a local minimum of g will satisfy the following system of n equations in n variables:

$$\frac{\partial g}{\partial x_1}(x) = 0$$

$$\frac{\partial g}{\partial x_n}(x) = 0$$

In this case the Newton iteration (6) looks like

$$\nabla^2 g(x^k)dx = -\nabla g(x^k)$$

where ∇g is the gradient of g and $\nabla^2 g$ is the Hessian of g, that is, the matrix with (i, j) component

$$\frac{\partial^2 g}{\partial x_i \partial x_j}(x)$$

If g is convex, then any local minimizer is also a global minimizer. If x^* is a local minimizer of $g(x)$, that is, $\nabla g(x^*) = 0$, and if $\nabla^2 g(x^*)$ has full rank, then Newton's method will converge to x^* if started sufficiently close to x^*.

6.2. Lagrange Multiplier Method

Lagrange discovered how to transform a constrained optimization problem, with equality constraints, into an unconstrained problem. To solve the problem

$$\begin{aligned}&\text{Minimize } f(x)\\&\text{Subject to } g_i(x) = 0 \quad \text{for } i = 1, \ldots, m\end{aligned}$$

form a Lagrange function

$$L(x, y) = f(x) - \sum_{i=1}^{m} y_i g_i(x)$$

and then minimize the unconstrained function $L(x, y)$ by solving the system of $(n + m)$ equations in $(n + m)$ variables:

$$\begin{aligned}\frac{\partial L}{\partial x_j} &= \frac{\partial f}{\partial x_j}(x) - \sum_{i=1}^{m} y_i \frac{\partial g_i}{\partial x_j}(x) = 0 \quad \text{for } j = 1, \ldots, n\\ \frac{\partial L}{\partial y_i} &= -g_i(x) = 0 \quad \text{for } i = 1, \ldots, m\end{aligned}$$

These equations can be solved by Newton's method.

6.3. Fiacco and McCormick Algorithm

General inequality constraints can be converted to equations by adding nonnegative slack (or surplus) variables. So the only essential inequalities are the nonnegativity conditions: $x \geq 0$. The idea of the barrier function approach is to start from a point in the strict interior of the inequalities ($x_j^0 > 0$ for all j) and construct a barrier that prevents any variable from reaching the boundary ($x_j^0 = 0$).

For example, adding $-\log(x_j)$ to the objective function will cause the objective to increase without bound x_j as approaches 0. Of course, if the constrained optimum is on the boundary (i.e., some $x_j^* = 0$ that is always true for linear programming), then the barrier will prevent us from reaching it. The solution is to use a barrier parameter that balances the contribution of the true objective function against that of the barrier function.

A minimization problem with nonnegativity conditions can be converted into a sequence of unconstrained minimization problems in the following way. The problem

$$\text{Minimize } f(x)$$
$$\text{Subject to } x \geq 0$$

is replaced by the family of unconstrained problems

$$\text{Minimize } B(x|\mu) = f(x) - \mu \sum_{j=1}^{n} \log(x_j)$$

which is parameterized on the positive barrier parameter μ. Let $x(\mu)$ be the minimizer of Fiacco and McCormick show that $x(\mu) \to x^*$ the constrained minimizer, as $\mu \to 0$. The set of minimizers is called the central trajectory.

As a simple example, consider the problem

$$\begin{aligned} &\text{Minimize } (x_1 + 1)^2 + (x_2 + 1)^2 \\ &\text{Subject to } x_1 \geq 0 \quad x_2 \geq 0 \end{aligned}$$

The unconstrained minimum would be at $(-1, -1)$, but the constrained minimum is at the origin $(x_1^*, x_2^*) = (0,0)$. For any $\mu > 0$:

$$B(x|\mu) = (x_1 + 1)^2 + (x_2 + 1)^2 - \mu \log(x_1) - \mu \log(x_2)$$

$$\frac{\partial B}{\partial x_1} = 2(x_1 + 1) - \frac{\mu}{x_1} = 0$$

$$\frac{\partial B}{\partial x_2} = 2(x_2 + 1) - \frac{\mu}{x_2} = 0$$

$$x_1(\mu) = x_2(\mu) = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + 2\mu}$$

which approaches (0,0) as $\mu \to 0$.

In general, we cannot get a closed-form solution for the central trajectory, but can use the following algorithm.

1. Choose $\mu_0 > 0$, set $k = 0$.
2. Find $x^k(\mu_k)$, the minimizer of $B(x|\mu_k)$.
3. If $\mu_k < \varepsilon$, stop. Otherwise, choose $\mu_{k+1} < \mu_k$.
4. Set $k = k + 1$ and go to step 2.

Then as $x^k(\mu_k) \to x^*$ as $\mu \to 0$.

In step 2 we can find $x(\mu)$ by using Newton's method to solve the system of n equations in n variables:

$$\frac{\partial B}{\partial x_j}(x|\mu) = \frac{\partial f(x)}{\partial x_j} - \frac{\mu}{x_j} = 0 \quad \text{for } j = 1, \ldots, n$$

In practice, we do not have to compute $x(\mu)$ very accurately before reducing μ.

6.4. Application to Linear Programming

The classical ideas reviewed can be applied to LP in the following way. If nonnegative slack and/or surplus variables have been used to convert inequalities into equations, then the general LP problem can be written as

$$\begin{aligned} \text{Minimize} \quad & c^T x \\ \text{Subject to} \quad & Ax = b \\ & x \geq 0 \end{aligned}$$

where A is $m \times n$. Let A_j denote column j of A, for $j = 1, \ldots, n$. We can replace the nonnegativity conditions with a barrier function:

$$\begin{aligned} &\text{Minimize } c^T x - \mu \sum_{j=1}^{n} \log(x_j) \\ &\text{Subject to } Ax = b \end{aligned}$$

This is now a problem with equality constraints, so we can use Lagrange's method. The Lagrangian function is

$$L(x, y) = c^T x - \mu \log(x_j) - y^T (Ax - b)$$

and the partial derivatives are

$$\begin{aligned} \frac{\partial L}{\partial x_j} &= c_j - \frac{\mu}{x_j} - A_j^T y = 0 \quad j = 1, \ldots, n \\ \frac{\partial L}{\partial y} &= -(Ax - b) = 0 \end{aligned}$$

The first set of equations gives

$$c_j - A_j^T y = \frac{\mu}{x_j} > 0$$

since $\mu > 0$ and $x_j > 0$. If we define

$$z_j = c_j - A_j^T y \quad \text{for } j = 1, \ldots, n$$

and require $z \geq 0$, then the system of equations that need to be solved is

$$(*) \begin{cases} Ax = b \\ A^T y + z = c \\ x_j z_j = \mu \quad \text{for } j = 1, \ldots, n \end{cases}$$

So we have $(m + 2n)$ equations in $(m + 2n)$ variables. Notice that the last n equations are nonlinear.

Starting from any point x^0, y^0, z^0 that satisfies $x^0 > 0$ and $z^0 > 0$, the idea is now to use Newton's method to solve the *nonlinear* system (*) for a given $\mu > 0$, and then reduce μ. In fact μ could be reduced after some fixed number of Newton steps (perhaps only one!). Various methods will differ in how many Newton steps are taken before reducing μ and in how much μ is reduced.

Forming the Jacobian of (*), we see that making a single Newton step requires the solution of the *linear* system of equations:

$$(**) \begin{bmatrix} A & 0 & 0 \\ 0 & A^T & I \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \begin{bmatrix} b - Ax^0 \\ c - A^T y^0 - z^0 \\ \mu e - xZe \end{bmatrix}$$

where

$$\begin{aligned} X &= \text{diag}(x_1^0, \ldots, x_n^0) \\ Z &= \text{diag}(z_1^0, \ldots, z_n^0) \\ e &= (1, \ldots, 1)^T \end{aligned}$$

The iteration count for the interior method is the total number of Newton steps, that is, the number of times the *linear* system (**) is solved.

The interior point method sketched here, which is simply an application of the classical nonlinear barrier method to LP, was actually tried in the late 1960s. At that time it was not even close to being competitive with the simplex method. The reason was that the efficient numerical linear algebra for solving (**) had not yet been developed. This was done during the 1970s and early 1980s for the purpose of solving large-scale finite-element problems. One way to solve (**) involves the Cholesky factorization of the symmetric positive-definite matrix: $AZ^{-1}XA^T$. Factoring large, sparse, symmetric, positive-definite matrices is the central computational core of finite-element methods. See George and Liu (1981) and Duff et al. (1987).

6.5. Computational Efficiency of the Interior Point Method

The truly remarkable thing about the interior point method is that the number of iterations (Newton steps) is almost independent of problem size. For all models solved to date, the number of iterations has been less than 100, and is usually between 20 and 40. (Note, however, that one Newton step involves much more computation than one simplex step.) There are theoretical and empirical reasons to believe that the number of iterations increases with the log of the number of variables, $\log(n)$. Indeed, Marsten et al. (1990) report a family of problems with from 35,000 to 2,000,000 variables for which a regression of iterations vs. $\log(n)$ gave an $R^2 = 0.979$.

Given this surprising behavior, the computation time depends on how efficiently the individual steps can be executed. This is a very active area of research and one that is exploring new parallel and vector computer architectures. (As with the simplex method, if the A matrix were completely dense, the computation time would increase with the cube of the number of constraints, but sparse matrix techniques make this rather meaningless.)

To give some comparison between the interior point method and the simplex method, consider the following set of multiperiod oil refinery planning models. The number of periods is the number of days in the planning horizon. This is a class of problems, as mentioned earlier, that becomes particularly hard for the simplex method as the number of periods increases. In Table 1, OB1 is an implementation of the interior point method and XMP is an implementation of the simplex method. They were both written entirely in FORTRAN by the same programmer and run on the same computer, a DECstation 3100.

One surprising consequence of the independence between iteration count and problem size is the fact that it takes about 20 Newton steps to solve even very small problems. For Example 2, the interior method takes 18 steps to obtain eight digits of accuracy, while the simplex method takes only 2 steps! The interior point method has to converge to the vertex, while the simplex method, being essentially combinatorial, hops exactly (within machine precision) onto the vertex. The sequence of trial solutions generated by the interior method, rounded to three decimal places, are given in Table 2. Steps 14 to 18 are required for eight digits of accuracy and are omitted. Recall that the solution is at (8, 5/3).

7. COMPUTER SOLUTION OF LINEAR PROGRAMS

7.1. Evolution of Commercial Packages for LP

The discovery of the simplex method and the birth of the digital computer occurred at about the same time, in the late 1940s. Because of the prodigious amount of computation required to solve all but the smallest problems, the simplex method would be of no practical use without the digital computer. Solving large LP models has always been one of the most challenging computational tasks for each new generation of computers. This remains true today, as the meaning of "large" expands with the capabilities of each new computer. It is interesting that the interior point method appeared at about the same time as the widespread availability of vector and parallel computers, since it seems much better suited to these new architectures than the simplex method.

TABLE 1 OB1 vs. XMP on Multiperiod Refinery Models

	Number of Time Periods in the Model					
	1	2	3	4	8	16
Equation	442	883	1,255	1,659	3,275	6,507
Variables	1,032	1,945	2,644	3,506	6,874	13,610
Nonzeros	7,419	14,280	19,494	25,727	50,659	100,253
XMP[a]	40	123	418	683	3,207	16,331
OB1[a]	62	139	217	306	656	1,441

[a] These are CPU seconds on a DECstation-3100.

TABLE 2 Trial Solutions for Example 2 Using the Interior Point Method

Iteration	x_1	x_2
1	2.000	2.000
2	3.997	3.327
3	1.258	2.657
4	5.211	7.629
5	5.375	7.611
6	5.590	7.598
7	5.723	7.476
8	5.794	6.797
9	7.503	2.562
10	7.981	1.734
11	7.988	1.692
12	7.998	1.672
13	7.999	1.667

7.2. PC Software for LP

Up until the mid-1980s, almost all industrial LP models were solved on mainframe computers, with software provided by the computer manufacturers or by a very small number of specialty LP software companies. In the late 1980s, the situation changed drastically. LP models with a few hundred constraints can now be solved on personal computers (PCs using Intel Pentium chips). This has expanded the use of LP dramatically. There are even LP solvers that can be invoked directly from inside spreadsheet packages. For example, Microsoft Excel and Microsoft Office for Windows contain a general purpose optimizer for solving small linear, integer, and nonlinear programming problems. The LP optimizer is based on the simplex algorithm. There are now at least a hundred companies marketing LP software for personal computers. They can be found in any issue of *OR/MS Today,* a publication of the Institute for Operations Research and Management Science. For a 1999 survey of LP software, see Fourer (1999). A web version of the survey can be reached via Fourer's home page at http://iems.nwu.edu/~4er/.

7.3. High-End LP Software

As we entered the 1990s, another dramatic change became apparent. The workstation class of machines offered mainframe speed at a fraction of the mainframe cost. The result is that virtually all currently used LP models can now be solved on workstations.

We now have the capability to solve truly enormous LP models using interior point methods on supercomputers, but at the moment users are just beginning to realize this and to think of larger models. For example, an oil company that has been solving single-period production planning models for each refinery can start to piece together a multiperiod, multirefinery model that incorporates distribution as well as production planning. Another sure source of very large LP models is stochastic programming, which attempts to deal with uncertainty in the context of optimization.

In 1989, three new high-end LP software systems were introduced that dominated all earlier systems and redefined the state of the art. The first is OSL from IBM. OSL incorporates both the simplex method and the interior point method. The simplex part is a very substantial improvement over IBM's earlier LP software. The second is CPlex, a simplex code written by Robert Bixby of Rice University and marketed by CPlex Optimization. The third is OB1, an interior point code written by Roy Marsten and David Shanno and marketed by XMP Software.

7.4. LP Modeling Languages

During the 1980s, there was also great progress in the development of computer tools for building LP models. The old mainframe LP systems had matrix generator languages that provided some assistance in constructing LP coefficient matrices out of collections of data tables. The focus was on the matrix, however, rather than on the algebraic form of the model. The matrix is a very large, very concrete object that is of little real interest to the human model builder.

There are now modeling languages that allow the user to express a model in a very compact algebraic form, with whole classes of constraints and variables defined over index sets. Models with thousands of constraints and variables can be defined in a couple of pages, in a syntax that is very close to standard algebraic notation. The algebraic form of the model is kept separate from the actual data for any particular instance of the model. The computer takes over the responsibility of transforming the abstract form of the model and the specific data into a specific coefficient matrix. This has greatly simplified the building, and even more the changing, of LP models. Several modeling languages are available for PCs. The two high-end products are GAMS (General Algebraic Modeling System) and AMPL (A Mathematical Programming Language). GAMS is marketed by GAMS Development. For a reference on GAMS, see Brooke et al. (1988). AMPL is marketed by AT&T. For a general introduction to modeling languages, see Fourer (1983), and for an excellent discussion of AMPL see Fourer et al. (1990).

7.5. LP Software on the Internet

A complete list of optimization software available for LP problems is available at the following NEOS website: http://www.mcs.anl.gov/home/otc. This site not only provides access to the software guide but also to the other optimization related sites that are continually updated. The NEOS guide on optimization software is based on More and Wright (1993), an excellent resource for those interested in a broad review of the various optimization methods and their computer codes. The book is divided into two parts. Part I has an overview of algorithms for different optimization problems, categorized as unconstrained optimization, nonlinear least squares, nonlinear equations, linear programming, quadratic programming, bound-constrained optimization, network optimization, and integer programming. Part II includes product descriptions of 75 software packages that implement the algorithms described in Part I. Much of the software described in this book is in the public domain and can be obtained through the Internet.

8. SENSITIVITY ANALYSIS IN LINEAR PROGRAMMING

8.1. Reasons for Sensitivity Analysis

In all linear programming models, the coefficients of the objective function and the constraints are supplied as input data or as parameters to the model. The optimal solution obtained by the simplex method is based on the values of these coefficients. In practice, the values of these coefficients are seldom known with absolute certainty because many of the coefficients are functions of some uncontrollable parameters. For instance, the future demands, the cost of raw materials, or the cost of energy resources cannot be predicted with complete accuracy before the problem is solved. Hence the solution of a practical problem is not complete with the mere determination of the optimal solution.

Each variation in the values of the data coefficients changes the linear programming problem, which may in turn affect the optimal solution found earlier. To develop an overall strategy to meet the various contingencies, one has to study how the optimal solution will change as a result of changes in the input (data) coefficients. This is sensitivity analysis or postoptimality analysis. Other reasons for performing a sensitivity analysis are as follows:

1. There may be some data coefficients or parameters of the linear program that are controllable, for example, availability of capital, raw material, or machine capacities. Sensitivity analysis enables one to study the effects of changing these parameters on the optimal solution. If it turns out that the optimal value (profit/cost) changes (in our favor) by a considerable amount for a small change in the given parameters, then it may be worthwhile to implement some of these changes. For example, if increasing the availability of labor by allowing overtime contributes to a greater increase in the maximum return as compared to the increased cost of overtime labor, then one might want to allow overtime production.
2. In many cases the values of the data coefficients are obtained by statistical estimation procedures on past figures, as in the case of sales forecasts, price estimates, and cost data. These estimates, in general, may not be very accurate. If we can identify which of the parameters affect the objective value most, then we can obtain better estimates of these parameters. This will increase the reliability of our model and the solution.

8.2. Practical Uses

Example 1, discussed at the beginning of this chapter, can be used to help illustrate the practical uses of sensitivity analysis. A computer output of the solution of this problem is given in Table 3. Note from the optimal solution that the optimal product mix is to produce products 1 and 2 only at levels 33.33 and 66.67 units, respectively.

The shadow prices give the net impact on the maximum profit if additional units of certain resources can be obtained. Labor has the maximum impact, providing a $3.33 increase in profit per every hour of increase in labor. Of course, the shadow prices on the resources apply as long as their variations stay within the prescribed ranges on right-hand side (RHS) constants given in Table 3. In other words, a $3.33/hr increase in profit is achievable as long as the labor hours are not increased beyond 150 hr. Suppose it is possible to increase the labor hours by 25% by scheduling overtime that incurs an additional labor cost of $50. To see whether it is profitable to schedule overtime, we

TABLE 3 Computer Solution of Example 1

Optimal solution	$x_1 = 33.33, \quad x_2 = 66.67, \quad x_3 = 0$
Optimal value	Maximum profit = \$733.33
Shadow prices	For row 1 = \$3.33, for row 2 = \$0.67, for row 3 = 0
Opportunity costs	For $x_1 = 0$, for $x_2 = 0$, for $x_3 = 2.67$

Ranges on Objective Function Coefficients

Variable	Lower Limit	Present Value	Upper Limit
x_1	6	10	15
x_2	4	6	10
x_3	$-\infty$	4	6.67

Ranges on RHS Constants

Row	Lower Limit	Present Value	Upper Limit
1	60	100	150
2	400	600	1000
3	200	300	∞

first determine the net increase in maximum profit due to 25 hr of overtime as (25) (3.33) = \$83.25. Since it is more than the total cost of overtime, it is economical to schedule overtime. It is important to note that, when any of the RHS constants is changed, the optimal solution will change. However, the optimal product mix will be unaffected as long as the RHS constant varies within the specified range. In other words, we will still be making products 1 and 2 only, but their quantities may change.

The ranges on the objective function coefficients given in Table 3 exhibit the sensitivity of the optimal solution with respect to changes in the unit profits of the three products. It shows that the optimal solution will not be affected as long as the unit profit of product 1 stays between \$6 and \$15. Of course, the maximum profit will be affected by the change. For example, if the unit profit on product 1 increases from \$10 to \$12, the optimal solution will be the same, but the maximum profit will increase to \$733.33 + (12 − 10) (33.33) = 799.99.

Note that product 3 is not economical to include in the optimal product mix. Hence a further decrease in its profit contribution will not have any impact on the optimal solution or maximum profit. Also, the unit profit on product 3 must increase to \$6.67 (present value + opportunity cost) before it becomes economical to produce.

8.3. Simultaneous Variations in Parameters (Bradley et al., 1977)

The sensitivity analysis output on profit and RHS ranges is obtained by varying only one of the parameters and holding all other parameters fixed at their current values. However, it is possible to use the sensitivity analysis output when several parameters are changed simultaneously. This is done with the help of the 100% rule.

8.3.1. 100% Rule for Objective Function Coefficients

The 100% rule for the objective function coefficients is given by

$$\sum_j \frac{\delta c_j}{\Delta c_j} \leq 1 \tag{7}$$

where δc_j is the actual increase (decrease) in the objective function coefficient of variable x_j and Δc_j is the maximum increase (decrease) allowed by sensitivity analysis. As long as inequality (7) is satisfied, the optimal solution to the LP problem will not change. For example, suppose the unit profit on product 1 decreases by \$1, but increases by \$1 for both products 2 and 3. This simultaneous variation satisfies the 100% rule, since $\delta c_1 = -1$, $\Delta c_1 = -4$, $\delta c_2 = 1$, $\Delta c_2 = 4$, $\delta c_3 = 1$, $\Delta c_3 = 2.67$, and

$$\frac{-1}{-4} + \frac{1}{4} + \frac{1}{2.67} = 0.875 < 1$$

Hence the optimal solution will not change, but the maximum profit will change by $(-1)(3.33) + 1(66.67) + 1(0) = 33.34$.

8.3.2. 100% Rule for RHS Constants

The 100% rule for the RHS constants is given by

$$\sum_{j} \frac{\delta b_i}{\Delta b_i} \leq 1 \tag{8}$$

where δb_1 is the actual increase (decrease) in the RHS constant of the ith constraint and Δb_i is the maximum increase (decrease) allowed by sensitivity analysis. If inequality (8) is satisfied, then the optimal product mix remains the same and the shadow prices apply, but the optimal solution and maximum profit will change. Of course, the net change in the maximum profit ran be obtained using the shadow prices.

Warning: The failure of the 100% rule does not automatically imply that the LP solution will be affected.

9. APPLICATIONS OF LINEAR PROGRAMMING

Linear programming models are widely used to solve a number of military, economic, industrial, and social problems. The oil companies have been and still are one of the foremost users of very large LP models in petroleum refining, distribution, and transportation. The number of LP applications has grown so much in the last 20 years that it would be impossible to survey all the different applications. Instead, the reader is referred to two excellent textbooks, Gass (1970) and Salkin and Saha (1975) which are devoted solely to LP applications in such diverse areas as defense, industry, commercial-retail, agriculture, education, and the environment. Many of the applications also contain a discussion of the experiences in using the LP models in practice.

An excellent bibliography on LP applications is available in Gass (1975). It contains a list of references arranged by area (e.g., agriculture, industry, military, production, transportation). In the area of industrial application, the references have been further categorized by industry (e.g., chemical, coal, airline, iron and steel, paper, petroleum, railroad). For additional bibliographies on LP applications, readers may refer to a survey by Gray and Cullinan-James (1976). For more recent applications of LP in practice, readers should check the recent issues of *Interfaces, AIIE Transactions, Decision Sciences, European Journal of Operational Research, Management Science, Operations Research, Operational Research* (United Kingdom), *Naval Research Logistics Quarterly,* and *Op-Search* (India).

REFERENCES

Bazaraa, M. S., and Shetty, C. M. (1979), *Nonlinear Programming: Theory and Algorithms,* John Wiley & Sons, New York.

Bradley, S. P., Hax, A. C., and Magnanti, T. L. (1977), *Applied Mathematical Programming,* Addison-Wesley, Reading, MA, pp. 220–229.

Brooke, A., Kendrick, D., and Meeraus, A. (1988), *GAMS: A User's Guide,* Scientific Press, Redwood City, CA.

Dantzig, G. B. (1963), *Linear Programming and Extensions,* Princeton University Press, Princeton, NJ.

Duff, I. S., Erisman, A. M., and Reid, J. K. (1987), *Direct Methods for Sparse Matrices,* Oxford University Press, New York.

Fiacco, A. V., and McCormick, G. P. (1968), *Nonlinear Programming: Sequential Unconstrained Minimization Techniques,* John Wiley & Sons, New York.

Fourer, R. (1999), "Linear Programming Software Survey," *OR/MS Today,* Vol. 26, No. 4, pp. 64–71.

Fourer, R. (1983), "Modeling Languages versus Matrix Generators for Linear Programming," *ACM Transactions on Mathematical Software,* Vol. 9, pp. 143–183.

Fourer, R., Gay, D. M., and Kernighan, B. W. (1990), "A Modeling Language for Mathematical Programming," *Management Science,* Vol. 36, No. 5, pp. 519–554.

Gass, S. (1970), *An Illustrated Guide to Linear Programming,* McGraw-Hill, New York.

Gass, S. (1975), *Linear Programming,* 4th Ed., McGraw-Hill, New York.

George, A., and Liu, J. W.-H. (1981), *Computer Solution of Large Sparse Positive Definite Systems,* Prentice Hall, Englewood Cliffs, NJ.

Gill, P. E., Murray, W., Saunders, M. A., Tomlin, J. A., and Wright, M. A. (1986), "On Projected Newton Barrier Methods for Linear Programming and an Equivalence to Karmarkar's Projective Method," *Mathematical Programming,* Vol. 36, No. 2, pp. 183–209.

Gray, P., and Cullinan-James, C. (1976), "Applied Optimization—A Survey," *Interfaces,* Vol. 6, No. 3, pp. 24–41.

Karmarkar, N. (1984), "A New Polynomial-Time Algorithm for Linear Programming," *Combinatorica,* Vol. 4, No. 4, pp. 373–395.

Lagrange, J. L. (1788), *Mecanique Analytique,* Paris.

Marsten, R., Subramanian, R., Saltzman, M., Lustig, I., and Shan-No, D. (1990), "Interior Point Methods for Linear Programming: Just Call Newton, Lagrange, and Fiacco & McCormick!" *Interfaces,* Vol. 20, No. 4, pp. 105–116.

More, J. J., and Wright, S. J. (1993), *Optimization Software Guide,* SIAM, Philadelphia.

Murty, K. G. (1995), *Operations Research: Deterministic Optimization Models,* Prentice Hall, Englewood Cliffs, NJ.

Newton, I. (1687), *Principia,* London.

Ravindran, A., Phillips, D. T., and Solberg, J. J. (1987), *Operations Research: Principles and Practice,* 2nd Ed., John Wiley & Sons, New York.

Salkin, H. M., and Saha, J. (1975), *Studies in Linear Programming,* Amsterdam, and Elsevier, North-Holland, New York.

CHAPTER 98
Nonlinear Optimization

Tom M. Cavalier
Pennsylvania State University

1. INTRODUCTION 2540
2. CONVEXITY 2543
  2.1. Convexity and the Hessian Matrix 2546
3. UNCONSTRAINED OPTIMIZATION 2546
  3.1. Classical Unconstrained Results 2546
  3.2. Line Search Techniques 2547
    3.2.1. Golden Section Method 2547
  3.3. Multidimensional Search Techniques 2549
    3.3.1. Multidimensional Search Techniques without Using Derivatives 2549
    3.3.2. Multidimensional Search Techniques Using Derivatives 2550
  3.4. Conjugate Gradient Methods 2552
4. CONSTRAINED OPTIMIZATION 2553
  4.1. Lagrange Multipliers 2553
  4.2. Karush–Kuhn–Tucker Conditions 2554
    4.2.1. KKT Necessary Conditions 2554
  4.3. Quadratic Programming 2555
  4.4. Separable Programming 2556
  4.5. Geometric Programming 2558
  4.6. Methods of Feasible Directions 2559
  4.7. Sequential Unconstrained Minimization Techniques 2560
    4.7.1. Penalty Function Methods 2560
    4.7.2. Barrier Function Methods 2561
    4.7.3. Augmented Lagrangian Methods 2561
  4.8. Successive Linear Programming 2562
  4.9. Successive Quadratic Programming 2562
  4.10. Nonsmooth Optimization 2562
5. ONLINE SOURCES OF INFORMATION ON OPTIMIZATION 2563
6. NONLINEAR PROGRAMMING CODES 2563
  6.1. Optimization Software 2563
REFERENCES 2565
ADDITIONAL READING 2567

1. INTRODUCTION

Nonlinear programming and nonlinear optimization are generally considered synonymous terms, where in this context the term *programming* refers to the process of determining an optimum program or solution. Optimization can be thought of as a very broad extension of the simple calculus problem of finding the extrema of a given function. This process of finding the "best" solution to a mathematical model of some real-world system has a wealth of practical applications (see, e.g., Bracken and McCormick 1968.)

The basic elements of a mathematical program are *decision variables,* an *objective function,* and *constraints* or restrictions. To help fix ideas, consider the simple problem of finding the dimensions of a rectangle that has maximum area and whose perimeter is at most 4. This problem can be posed as the following mathematical program:

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

Example 1

$$\text{maximize } f(\mathbf{x}) = x_1x_2 \tag{1}$$

$$\text{subject to } g_1(\mathbf{x}) = 2x_1 + 2x_2 - 4 \leq 0 \tag{2}$$

$$g_2(\mathbf{x}) = -x_1 \leq 0 \tag{3}$$

$$g_3(\mathbf{x}) = -x_2 \leq 0 \tag{4}$$

where $\mathbf{x} = (x_1, x_2)^t \in E_2$, Euclidean 2-space. (The notation $\mathbf{v}^t$ represents the *transpose* of the vector **v**.) The decision variables in this case are x_1 and x_2, which represent the length and width of the rectangle, respectively. The objective function is $f(\mathbf{x}) = x_1x_2$, which represents the area of the rectangle. $g_1(\mathbf{x}) \leq 0$ is the perimeter constraint, whereas $g_1(\mathbf{x}) \leq 0$ and $g_1(\mathbf{x}) \leq 0$ represent nonnegativity restrictions on the variables. These nonnegativity restrictions, as well as simple bounds on the variables, are handled implicitly by some mathematical programming techniques (e.g., the simplex algorithm for linear programming). Here it is assumed that they are included as explicit constraints.

In general, a mathematical program can be represented in the generic form:

$$\text{(P) optimize } f(\mathbf{x}) \tag{5}$$

$$\text{subject to } g_i(\mathbf{x}) \leq 0 \text{ for } i = 1, \ldots, m \tag{6}$$

$$h_j(\mathbf{x}) = 0 \text{ for } j = 1, \ldots, p \tag{7}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t \in E_n$, Euclidean n-space, and f, g_i, h_j: $E_n \rightarrow E_1$. Generally, throughout much of this chapter, "optimize" will be taken to mean "minimize," since maximization problems can be addressed using the simple transformation, maximize $f(\mathbf{x}) = -\text{minimize } (-f(\mathbf{x}))$.

If all of the functions f, g_i, h_j are linear functions of **x**, then (P) is called a *linear program.* Otherwise (P) is a *nonlinear program.* Note that Example 1 is a nonlinear program since the objective function (1) is a nonlinear function. Actually, as will be seen later, this problem can be classified as a *quadratic program* since the objective function is a quadratic function and the constraints are all linear functions.

The region in E_n defined by constraints (6) and (7) is referred to as the *feasible region*; the feasible region for Example 1 is illustrated in Figure 1. The *objective function contours* (or level curves) are also identified in Figure 1, and it is clear that the point, $\mathbf{x}^* = (x_1^*, x_2^*)^t = (1, 1)^t$, where the objective contour is tangent to the boundary of the feasible region, is the global optimal solution. The constraint $g_1(\mathbf{x}) \leq 0$ is said to be *binding* (tight or active) at the optimal solution since $g_1(\mathbf{x}^*) = 0$, whereas $g_2(\mathbf{x}) \leq 0$ and $g_3(\mathbf{x}) \leq 0$ are *nonbinding* (loose or inactive) at the optimal solution.

Unlike linear programming, in which an optimum, if one exists, can be found at an extreme point of the feasible region, solutions of nonlinear programming problems can occur at any feasible point. Whereas Figure 1 shows an example where the optimum lies on the boundary of the feasible region, Figure 2 illustrates a case where the optimum is an interior point of the feasible region. In the latter

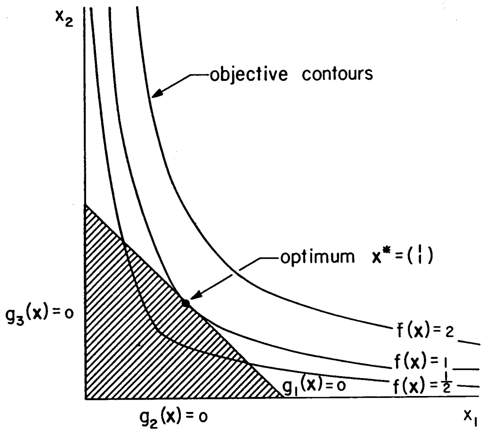


Figure 1 Graphical Solution of Example 1.

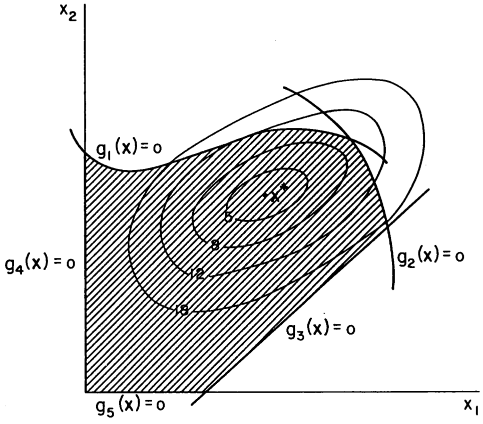


Figure 2 Example of a Nonlinear Programming Solution—No Active Constraints.

case, the constraints have no influence on the optimal solution, that is, the unconstrained optimum is also the constrained optimum.

This is just one of the difficulties encountered in nonlinear optimization. In contrast to linear programming, no single method is suitable for all types of nonlinear programs. And generally, it is only practical to obtain a *local minimum* that is not guaranteed to be a *global minimum*. (See Figure 3.)

Generally speaking, optimization methods can be divided into direct and indirect methods. Classical optimization methods are frequently referred to as *indirect methods* since these methods rely on analytic techniques for determining the optimal solution. This indirect approach does not involve a direct search for a particular solution but rather specifies a set of general conditions that must be satisfied by all solutions to the problem. As such, these methods do not lend themselves to computer implementations; however, they form the foundation for many of the computer-oriented direct methods. *Direct methods* of optimization seek to find a particular solution to a specified problem in a direct, iterative manner. The iterative nature of these procedures allows for effective computer implementations. Since real-world problems may involve many variables and constraints, the major challenge of nonlinear programming has been the development of efficient computational and algorithmic techniques. Although many algorithms have been developed, only a relative few have enjoyed continued success in real-world applications.

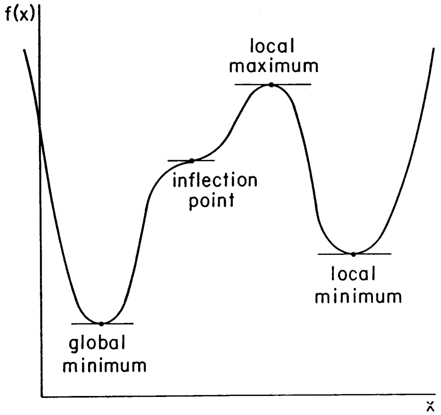


Figure 3 Local and Global Extrema.

This chapter outlines the fundamentals of nonlinear optimization. A review of some basic concepts and convexity is followed by a section addressing unconstrained optimization. This section presents classical optimization results as well as several direct search algorithms for solving univariate and multivariate problems. Next, constrained optimization problems are addressed. A review of Lagrange multipliers and the Karush–Kuhn–Tucker conditions precedes the presentation of several algorithms for solving different classes of nonlinear programming problems. There are a multitude of direct search algorithms in the literature, and the author has selected several representative methods to review. This is by no means an exhaustive review. It is simply an attempt to expose the reader to a cross-section of algorithmic methods. Finally, a listing of some nonlinear programming codes and some online resources is provided.

2. CONVEXITY

As indicated previously, nonlinear programming algorithms can generally only find a local optimal solution. Under these conditions, it becomes very important to know when a local minimum is actually guaranteed to be a global minimum over a given feasible region. One set of conditions that ensures such an outcome is that the feasible region is a convex set and the objective function is a convex function. Such a problem is called a *convex program.*

A *convex set* satisfies the following property: if $\mathbf{x}_1, \mathbf{x}_2 \in S$, then $\bar{\mathbf{x}} = \lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in S$ for all $\lambda \in [0, 1]$. $\bar{\mathbf{x}}$ is referred to as a *convex combination* of $\mathbf{x}_1$ and $\mathbf{x}_2$. Geometrically, the above property simply means that if $\mathbf{x}_1, \mathbf{x}_2 \in S$, then the line segment joining $\mathbf{x}_1$ and $\mathbf{x}_2$ is also contained in S. An *extreme point* of a convex set S is any point in S that cannot be written as a strict convex combination (i.e. $\lambda \in (0, 1)$) of two distinct points in S. Figure 4 illustrates a convex and a nonconvex set. A set is *closed* if it contains all of its boundary points.

A function $f : S \rightarrow E_1$ defined on a convex set $S \subset E_n$ is a *convex function* if $f(\bar{\mathbf{x}}) \leq \lambda f(\mathbf{x}_1) + (1 - \lambda)f(\mathbf{x}_2)$ for all $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $\lambda \in [0, 1]$. This property is illustrated geometrically in Figure 5(*a*) and means that the line segment joining the points $(\mathbf{x}_1, f(\mathbf{x}_1))$ and $(\mathbf{x}_2, f(\mathbf{x}_2))$ lies above the graph of $f(\mathbf{x})$. Also, f is a *strictly convex function* if $f(\bar{\mathbf{x}}) < \lambda f(\mathbf{x}_1) + (1 - \lambda)f(\mathbf{x}_2)$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{S}$, $\mathbf{x}_1 \neq \mathbf{x}_2$, and $\lambda \in [0, 1]$. Similarly, $g : S \rightarrow E_1$ is (*strictly*) *concave* if $-g$ is (strictly) convex.

As a special case, if the feasible region of a linear program with nonnegative variables is nonempty, the feasible region is always a convex polyhedral set with a finite number of extreme points. If an optimal solution exists, an optimal extreme point among this finite set can be found quite efficiently using either the simplex algorithm, which was developed by Dantzig in 1947, or, more recently, by Karmarkar's interior point algorithm (Karmarkar 1984). An analogous approach is not possible with general nonlinear programs, since the feasible region cannot readily be reduced to a finite set of candidate solutions and a global optimal solution cannot necessarily be found by relying only on local information, as in the simplex method. The effects of convexity on the outcome of different mathematical programming problems can be summarized as follows.

1. *Minimizing an unconstrained convex objective function:* In this case, any local minimum will also be a global minimum. Furthermore, if the objective function is differentiable, any stationary point (i.e., a point at which all the first-order derivatives vanish) will be a global minimum.
2. *Maximizing an unconstrained concave objective function:* Any stationary point will be a global maximum, and any local maximum will also be a global maximum.
3. *Minimizing a convex objective function over a closed convex feasible region:* Again, any local minimum will be a global minimum. As mentioned earlier, this class of problems is labeled convex programming problems, and a global minimum, if one exists, may occur at either an interior point or a boundary point of the feasible region.

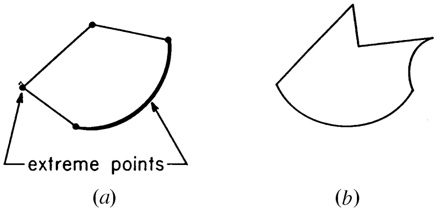
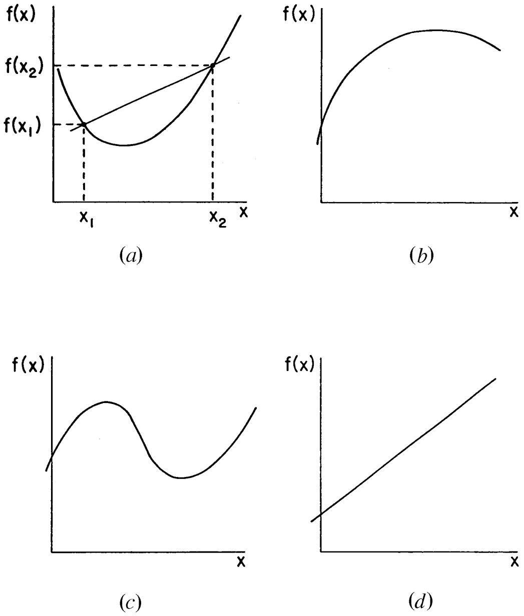


Figure 4 Examples of (*a*) Convex Set and (*b*) Nonconvex Set.



(*a*) (*b*) (*c*) (*d*)

Figure 5 Examples of (*a*) Convex function, (*b*) Concave function, (*c*) Neither convex nor concave, (*d*) Both convex and concave.

4. *Maximizing a convex objective function over a closed bounded convex feasible region:* In this case, a maximum, if one exists, will be found at an extreme point of the feasible region. However, a local maximum need not be a global maximum.
5. *Maximizing a concave objective function over a closed convex feasible region:* This case is similar to case 3. That is, a local maximum is a global maximum.
6. *Minimizing a concave objective function over a closed bounded convex feasible region:* This case is analogous to case 4 and does not guarantee that a local minimum is a global minimum. However, a global minimum, if one exists, will occur at an extreme point of the feasible region.
7. *Maximizing or minimizing an objective function over a nonconvex set:* This is a difficult case that may produce a local minimum or maximum that is not also a global solution. Figure 6 illustrates that this can occur even in the simple case when the objective function is linear. Observe that extreme point **B** is a local minimum since all feasible points near **B** have an objective value that is greater than that at **B**.

In general, the task of proving that an arbitrary subset of E_n is convex can be quite difficult. However, the feasible region of problem (P) will be a convex set if each function $g_i(\mathbf{x})$ is a convex function and each function $h_j(\mathbf{x})$ is a linear function. Actually, these conditions on the functions $g_i(\mathbf{x})$, $h_{j(}\mathbf{x})$ can be relaxed somewhat using the concepts of generalized convexity. A nice summary of generalized convexity is provided by Avriel (1976).

Verifying the convexity of an arbitrary function $f : E_n \rightarrow E_1$ can also be a difficult task. In the case when f is twice differentiable, however, the concept of quadratic form is useful for investigating the convexity or concavity of f.

A *quadratic form* $q(\mathbf{x})$ is defined by

$$q(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x} \tag{8}$$

where $\mathbf{x} \in E_n$ and $\mathbf{A}$ is an $n \times n$ real symmetric matrix. The matrix $\mathbf{A}$ is said to be:

1. *Positive definite* if and only if $\mathbf{x}^t\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$
2. *Negative definite* if $\mathbf{x}^t\mathbf{A}\mathbf{x} < 0$ and only if for all $\mathbf{x} \neq \mathbf{0}$

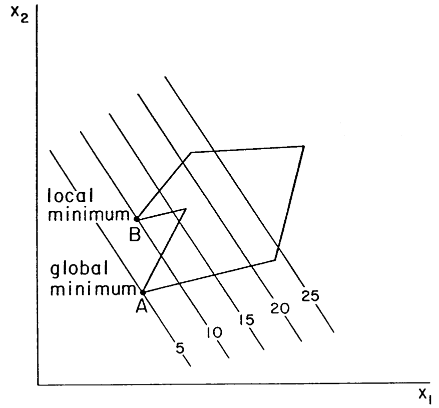


Figure 6 Minimizing a Linear Objective over a Nonconvex Set.

3. *Positive semidefinite* if and only if $\mathbf{x}^t\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$
4. *Negative semidefinite* if and only if $\mathbf{x}^t\mathbf{A}\mathbf{x} \leq 0$ for all $\mathbf{x}$
5. *Indefinite* if $\mathbf{x}^t\mathbf{A}\mathbf{x} > 0$ for some $\mathbf{x}$ and $\mathbf{x}^t\mathbf{A}\mathbf{x} < 0$ for some $\mathbf{x}$

Example 2. Consider the 2 × 2 symmetric matrix $\mathbf{A} = \begin{pmatrix} 1 & -3 \\ -3 & 11 \end{pmatrix}$. $\mathbf{A}$ is positive definite since

$$q = \mathbf{x}^t\mathbf{A}\mathbf{x} = (x_1, x_2)\begin{pmatrix} 1 & -3 \\ -3 & 11 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 - 6x_1x_2 + 11x_2^2 = (x_1 - 3x_2)^2 + 2x_2^2 > 0 \text{ for all } \mathbf{x} \neq \mathbf{0}$$

Positive and negative definiteness can also be checked using the leading principal minor test, although semidefiniteness cannot be verified in this manner. Using standard matrix notation, let

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{9}$$

Define

$$A_1 = a_{11}, A_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

with $A_3, \ldots, A_n$ defined similarly. Note that $A_n = |\mathbf{A}|$, that is, A_n is the determinant of $\mathbf{A}$. A_i is called the ith *leading principal minor* of $\mathbf{A}$. The results of the *leading principal minor test* can be summarized as follows:

1. If $A_i > 0$ for all $i = 1, \ldots, n$, then $\mathbf{A}$ is positive definite.
2. If A_i, $i = 1, \ldots, n$, alternate in sign with $A_1 < 0$, then $\mathbf{A}$ is negative definite.
3. If 1 and 2 are not satisfied and $A_i \neq 0$ for all $i = 1, \ldots, n$, then $\mathbf{A}$ is indefinite.
4. If $A_i = 0$ for some i, then the test fails.

In Example 2, $A_1 = 1 > 0$ and $A_2 = 2 > 0$, therefore $\mathbf{A}$ is positive definite. A more comprehensive polynomial-time algorithm for checking both definiteness and semidefiniteness is presented in Bazaraa et al. (1994, pp. 96–97). The algorithm uses Gauss–Jordan reduction and works toward systematically reducing the matrix to upper triangular form until a conclusion is reached. Definiteness and semidefiniteness can also be determined by examining the eigenvalues of the matrix.

2.1. Convexity and the Hessian Matrix

Let $f : E_n \rightarrow E_1$ be a twice differentiable function and define the *Hessian matrix,* $\mathbf{H}(\mathbf{x})$, to be the matrix of second partial derivatives, $f_{ij}(\mathbf{x}) = \partial^2 f(\mathbf{x})/\partial x_i \partial x_j$.

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} f_{11}(\mathbf{x}) & f_{12}(\mathbf{x}) & \dots & f_{1n}(\mathbf{x}) \\ f_{21}(\mathbf{x}) & f_{22}(\mathbf{x}) & \dots & f_{2n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1}(\mathbf{x}) & f_{n2}(\mathbf{x}) & \dots & f_{nn}(\mathbf{x}) \end{pmatrix} \tag{10}$$

Let $\bar{\mathbf{x}} \in E_n$. Then by examining the nature of the quadratic form of $\mathbf{H}(\mathbf{x})$, it is possible to determine local characteristics of the function f.

1. If $\mathbf{H}(\bar{\mathbf{x}})$ is positive definite, then f is locally convex at $\bar{\mathbf{x}}$.
2. If $\mathbf{H}(\bar{\mathbf{x}})$ is negative definite, then f is locally concave at $\bar{\mathbf{x}}$.
3. If $\mathbf{H}(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$, then f is a convex function.
4. If $\mathbf{H}(\mathbf{x})$ is negative semidefinite for all $\mathbf{x}$, then f is a concave function.

If f is a univariate function, that is, $f : E_1 \rightarrow E_1$, then the Hessian matrix simply reduces to the scalar, $f''(x)$ and the above conditions simplify to the following:

1. If $f''(\bar{x}) > 0$, then f is locally convex at $\bar{x}$.
2. If $f''(\bar{x}) < 0$, then f is locally concave at $\bar{x}$.
3. If $f''(x) \geq 0$ for all x, then f is a convex function.
4. If $f''(x) \leq 0$ for all x, then f is a concave function.

These conditions are used in the following section to develop optimality conditions in unconstrained optimization.

3. UNCONSTRAINED OPTIMIZATION

An unconstrained mathematical programming problem is a problem of the form: minimize $f(\mathbf{x})$ where $f : E_n \rightarrow E_1$. This is the type of problem that is typically addressed in a first course in calculus, and this class of problems has many applications, including maximum-likelihood or least-squares estimation problems in statistics. The classical fundamentals and algorithmic techniques used to solve these problems are often used as a basis for constructing efficient procedures for solving more general problems. This section presents some fundamental classical results, as well as algorithmic strategies for finding the solution of unconstrained optimization problems.

3.1. Classical Unconstrained Results

Classical methods of optimization are based on differential calculus, and it is generally assumed that the function to be optimized is continuous and differentiable (smooth). For a function of one variable, $f : E_1 \rightarrow E_1$, a *necessary condition* for a local extremum (either a local maximum or local minimum) to occur at a point $x^* \in E_1$ is that the first derivative vanishes at x^*, that is,

$$f'(x^*) = 0 \tag{11}$$

However, a local extremum does not necessarily occur at every point that satisfies (11); that is, (11) is not a *sufficient condition* for optimality. In practice, necessary conditions are used to identify *stationary points,* which are candidate extrema, whereas sufficient conditions are used to classify the stationary points as local maxima, local minima, or saddle points (inflection points in E_2). Once all local extrema are found, the global extrema can be found by selecting the absolute maximum or minimum. The necessary and sufficient conditions for determining and classifying the stationary points of a function of one variable are summarized in Table 1. These conditions are easily derived using a Taylor series expansion.

As indicated in the previous section, the sufficiency conditions in Table 1 are essentially examining the local properties of the function f. For example, $f''(\bar{x}) > 0$ indicates that the function is convex at the point $\bar{x}$, and thus, if $f'(\bar{x}) = 0$, a local minimum occurs at $\bar{x}$. Similarly, if $f''(\bar{x}) < 0$, then f is concave at $\bar{x}$.

To illustrate the use of the results presented in Table 1 consider the following simple problem.

Example 3. Find the extrema of the function $f(x) = x^3\,(3x - 8) + 20$. First, the stationary points are identified by finding the roots of $f'(x) = 12x^2\,(x - 2) = 0$. Clearly $x = 0, 2$ are the only stationary points. Next, these candidate points are classified by examining higher-order derivatives. In particular,

TABLE 1 Necessary and Sufficient Conditions for Local Extrema of a Univariate Function

	Necessary Condition	Sufficient Condition
Local minimum at $\bar{x}$	$f'(\bar{x}) = 0$	$f^{(i)}(\bar{x}) = 0,\ i = 1, \ldots, m - 1,$ $f^{(m)}(\bar{x}) > 0$ and m is even
Local maximum at $\bar{x}$	$f'(\bar{x}) = 0$	$f^{(i)}(\bar{x}) = 0,\ i = 1, \ldots, m - 1$ $f^{(m)}(\bar{x}) < 0$ and m is even
Inflection point at $\bar{x}$	$f'(\bar{x}) = 0$	$f^{(i)}(\bar{x}) = 0,\ i = 1, \ldots, m - 1$ $f^{(m)}(\bar{x}) \neq 0$ and m is odd

for this example, it is necessary to determine $f''(x) = 36x^2 - 48x$ and $f'''(x) = 72x - 48$. A local minimum occurs at $x = 2$ since $f''(2) = 48 > 0$ and $m = 2$ is even. Also, $f''(0) = 0$ and $f'''(0) = -48$. Therefore, an inflection point occurs at $x = 0$ since $f'''(0) < 0$ and $m = 3$ is odd. In addition, upon graphing, it can be seen that the local minimum at $x = 2$ is also a global minimum, even though the function f is not convex.

For a function of n variables, $f : E_n \rightarrow E_1$, a necessary condition for an extremum to occur at a point $\mathbf{x}^* \in E_n$ is that the gradient vector, $\nabla f(\mathbf{x})$, the vector of first partial derivatives, vanishes at $\mathbf{x}^*$. Thus, in this case, the task of finding the stationary points is more difficult, in that it is necessary, in general, to solve a system of n simultaneous nonlinear equations in n unknowns,

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = 0 \quad \text{for } i = 1, \ldots, n \tag{12}$$

Sufficiency conditions are determined by examining the Hessian matrix $\mathbf{H}(\mathbf{x})$, the matrix of second partial derivatives. Based on the discussion in Section 2, Table 2 gives the necessary and sufficiency conditions for identifying and classifying stationary points in the multivariate case. In the event that the sufficiency conditions are not satisfied, higher-order derivative information must be used.

Example 4. Find the extrema of the function $f(\mathbf{x}) = x_1^3 - 6x_1x_2 + 24x_1 + x_2^2$. First, obtain first and second order derivative information.

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 3x_1^2 - 6x_2 + 24 \\ -6x_1 + 2x_2 \end{pmatrix} \tag{13}$$

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 6x_1 & -6 \\ -6 & 2 \end{pmatrix} \tag{14}$$

Setting $\nabla f(\mathbf{x}) = 0$, the stationary points are $\mathbf{x}_1 = (2, 6)^t$ and $\mathbf{x}_2 = (4, 12)^t$. From Table 2, $\mathbf{x}_1$ is a saddle point since $H_1(\mathbf{x}_1) = 12$ and $H_2(\mathbf{x}_1) = -12$, that is, $\mathbf{H}(\mathbf{x}_1)$ is indefinite. Similarly, $H_1(\mathbf{x}_2) = 24$ and $H_2(\mathbf{x}_2) = 12$ implies that $\mathbf{H}(\mathbf{x}_2)$ is positive definite and thus, a local minimum occurs at $\mathbf{x}_2$.

3.2. Line Search Techniques

There are several direct search techniques for minimizing a function of one variable. The methods generally start from an initial estimate and sequentially move toward the minimum. Univariate or line search techniques play a major role in solving subproblems in more complex direct search algorithms.

3.2.1. Golden Section Method

An example of a line search technique that does not use derivatives is the golden section method, which seeks to find the minimum of a function $f(x)$ on an interval $[a, d]$. The interval $[a, d]$ is called

TABLE 2 Necessary and Sufficient Conditions for Local Extrema of a Multivariate Function

	Necessary Condition	Sufficient Condition
Local minimum at $\bar{\mathbf{x}}$	$\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$	$\mathbf{H}(\bar{\mathbf{x}})$ is positive definite
Local maximum at $\bar{\mathbf{x}}$	$\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$	$\mathbf{H}(\bar{\mathbf{x}})$ is negative definite
Saddle point at $\bar{\mathbf{x}}$	$\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$	$\mathbf{H}(\bar{\mathbf{x}})$ is indefinite

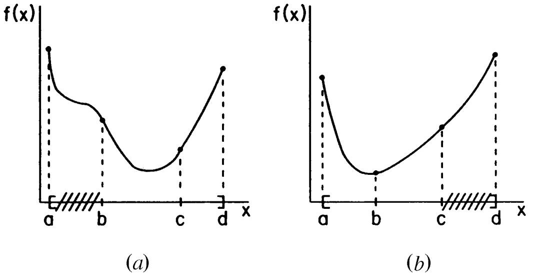


Figure 7 Narrowing the Interval of Uncertainty.

the *interval of uncertainty,* and the basic strategy of the method is to successively decrease the length of the interval, assuming that the function is unimodal, that is, the function has only one minimum on the interval $[a, d]$.

Let $b, c \in [a, d]$ such that $a < b < c < d$ and compute $f(b)$ and $f(c)$. Since f is unimodal on $[a, d]$, it is possible to narrow the interval of uncertainty by comparing $f(b)$ and $f(c)$. That is, if $f(b) > f(c)$, then the true minimum cannot lie on the interval $[a, b]$ [Figure 7(*a*)], and if $f(b) < f(c)$, the minimum cannot lie on the interval $[c, d]$ [Figure 7(*b*)]. Thus, in either case, the length of the interval of uncertainty is reduced and the entire process can be repeated until some desired accuracy is achieved.

Let $a_k < b_k < c_k < d_k$, where $[a_k, d_k]$ is the interval of uncertainty at iteration k and assume that $[a_1, d_1] = [a, d]$. Since functional evaluations are the most expensive step in the process, the golden section method reduces the amount of overall work by intelligently choosing symmetric points b_k and c_k so that they can be reused on successive iterations, as illustrated in Figure 8. This is achieved by using the relationships $b_k = \lambda a_k + (1 - \lambda)d_k$ and $c_k = (1 - \lambda)a_k + \lambda d_k$, where $\lambda = 0.618$. Observe that b_k and c_k are simply expressed as convex combinations of a_k and d_k. A summary of the *golden section algorithm* follows:

1. Choose a tolerance $\varepsilon > 0$ for the length of the final interval of uncertainty. Choose an initial interval of uncertainty $[a_1, d_1]$ and set the iteration counter $k = 1$. Compute $b_1 = \lambda a_1 + (1 - \lambda)d_1$ and $c_1 = (1 - \lambda)a_1 + \lambda d_1$.
2. If $d_k - a_k < \varepsilon$, stop; the interval $[a_k, d_k]$ contains the minimum. Otherwise, continue with step 3.
3. If $f(b_k) > f(c_k)$, go to step 4. Otherwise go to step 5.
4. Let $a_{k+1} = b_k$, $b_{k+1} = c_k$, $d_{k+1} = d_k$ [see Figure 8(*a*)]. Compute $c_{k+1} = (1 - \lambda)a_{k+1} + \lambda d_{k+1}$. Replace k by $k + 1$ and go to step 2.
5. Let $a_{k+1} = a_k$, $c_{k+1} = b_k$, $d_{k+1} = c_k$. [see Figure 8(*b*)]. Compute $b_{k+1} = \lambda a_{k+1} + (1 - \lambda)d_{k+1}$. Replace k by $k + 1$ and go to step 2.

Other line search methods that involve only function evaluations, that is, no derivative calculations, are the dichotomous search, the Fibonacci search (Kiefer 1957), and the quadratic fit line search. The Fibonacci search is the most efficient derivative-free line search technique in the sense that it requires the fewest function evaluations to attain a prescribed degree of accuracy. The quadratic fit method

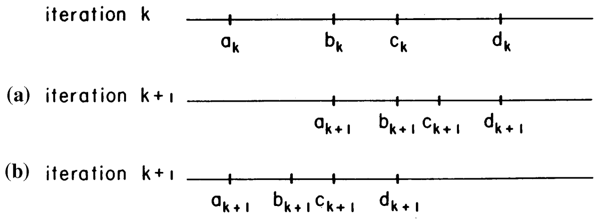


Figure 8 Successive Iterations of the Golden Section Method.

fits a quadratic function to three points on the function and then finds the unique minimizing point of the resulting quadratic function. A new interval of uncertainty is determined and the process is repeated until convergence is achieved (see, e.g., Bazaraa et al. 1994, pp. 279–281). Examples of line search techniques that utilize derivative information are the bisection search method and Newton's method. The more general multivariate version of Newton's algorithm will be discussed in the next section. In practice, a line search is typically used to find the step length during an iterative step of a more general algorithm. As such, it may be impractical to try to find the exact minimum point and it may be appropriate to apply an inexact method such as Armijos' rule (Armijo 1966) or simply to terminate the line search procedure before it has converged.

3.3. Multidimensional Search Techniques

This section is a natural extension of the previous section and addresses the problem of minimizing a function of several variables, that is, minimize $f(\mathbf{x})$ where $f : E_n \rightarrow E_1$. The general process behind multidimensional search techniques may be expressed as follows: Given a current point $\mathbf{x}_k$, determine a direction $\mathbf{d}_k$ and a step size α_k to yield a new point,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{15}$$

The step size α_k is determined by either solving the line search problem in the variable α,

$$\text{Minimize } f(\mathbf{x}_k + \alpha \mathbf{d}_k) \tag{16}$$

where typically $\alpha \in E_1$, $\alpha \in [0, \infty)$, or $\alpha \in [a, b]$, or by taking prescribed discrete steps along the search directions. The strategy for choosing the search directions and the step sizes determines the different methods.

3.3.1. Multidimensional Search Techniques without Using Derivatives

The *cyclic coordinate search* method successively uses search directions that are parallel to the coordinate axes along with line search problems to determine the step sizes. This method is conceptually simple, but the sequence of iterates generated by this method tend to zigzag if the optimal solution lies in a valley.

To help overcome this problem, the method of Hooke and Jeeves (1961) involves both exploratory moves and pattern moves (acceleration moves) with discrete steps along the search directions. The discrete steps eliminate the need for a line search. In the exploratory move phase, starting at a point $\mathbf{x}_k$, a modified cyclic coordinate search is performed in which, if possible, the objective function is reduced once along each of the coordinate directions using a prescribed discrete step length. This leads to a new point $\mathbf{x}_{k+1}$ and establishes a direction of improvement. A pattern move is then performed in the direction $\mathbf{x}_{k+1} - \mathbf{x}_k$, leading to an intermediate point $\mathbf{y}$. Now, starting from $\mathbf{y}$, another exploratory move yields the point $\mathbf{x}_{k+2}$. If $f(\mathbf{x}_{k+2} < f(\mathbf{x}_{k+1})$, then an improvement has been found and the process is continued with a pattern move in the direction $\mathbf{x}_{k+2} - \mathbf{x}_{k+1}$. Otherwise, $\mathbf{x}_{k+1}$ is considered the new initial point and the process is begun anew. In the instance when no improving exploratory step can be made, the discrete step size is reduced and the process is repeated. When the step size becomes smaller than some prescribed tolerance, the search terminates. A graphical interpretation of the algorithm is presented in Figure 9.

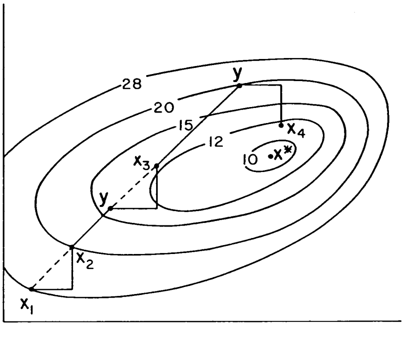


Figure 9 The Method of Hooke and Jeeves.

Other methods of multidimensional search without using derivatives include Rosenbrock's method (1960) and the simplex method of Spendley et al. (1962), which was later modified by Nelder and Meade (1974). Although it has the same name, this simplex method is not the same algorithm as that used for linear programming; it is a polytope algorithm that requires only functional evaluations and requires no smoothness assumptions.

3.3.2. Multidimensional Search Techniques Using Derivatives

In this section, optimization techniques are discussed that use derivative information in determining the search directions. As in the preceding discussion, a basic step in the algorithmic process consists of choosing a direction $\mathbf{d}_k$ and a step length α_k to arrive at a new point,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{17}$$

Recall from calculus that the gradient, $\nabla f(\mathbf{x})$, of a differentiable function $f : E_n \rightarrow E_1$ provides local information concerning the rate of change of the function. In fact, given a point $\bar{\mathbf{x}}$, the gradient of f evaluated at $\bar{\mathbf{x}}$, $\nabla f(\bar{\mathbf{x}})$, is the direction of steepest ascent at $\bar{\mathbf{x}}$ and $-\nabla f(\bar{\mathbf{x}})$ is the direction of steepest descent at $\bar{\mathbf{x}}$. This fundamental result leads to the following algorithm.

3.3.2.1. The Method of Steepest Descent This is one of the most basic procedures for minimizing an unconstrained differentiable function, and the process is defined by simply specifying $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$ in Eq. (17) to get

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \tag{18}$$

where α_k is determined by solving the line search problem

$$\underset{\alpha \geq 0}{\text{Minimize}}\ f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)) \tag{19}$$

The path generated by the algorithm for an example problem is illustrated in Figure 10. Note that successive directions are orthogonal and at each point $\mathbf{x}_k$, $\nabla f(\mathbf{x}_k)$ is normal to the objective contour at $\mathbf{x}_k$. A typical stopping rule is when the Euclidean norm of $\nabla f(\mathbf{x}_k)$, $(\|\nabla f(\mathbf{x}_k)\|)$, is less than some prescribed small positive number. The performance of the method of steepest descent is quite good during early iterations, but the convergence tends to slow excessively during later iterations and exhibits zigzagging tendencies near stationary points. Thus, it is not well used in practice.

3.3.2.2. Newton's Method The classical Newton's method is a technique that instead of specifying a step length at each iteration uses the inverse of the Hessian matrix, $\mathbf{H}(\mathbf{x})^{-1}$, to deflect the direction of steepest descent. The method assumes that $f(\mathbf{x})$ may be approximated locally by a second order Taylor approximation and is derived quite easily by determining the minimum point of this quadratic approximation. Assuming that $\mathbf{H}(\mathbf{x}_k)$ is nonsingular, then the algorithmic process is defined by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k) \tag{20}$$

The classical method does not require a line search at each iteration, but it does require second order derivative information and a matrix inversion, or equivalently, the solution of a system of linear

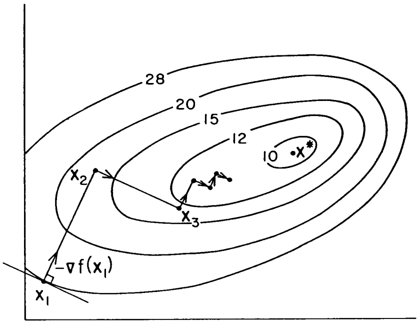


Figure 10 The Steepest Descent Method.

equations. In general, the performance of Newton's method is quite erratic and the sequence of points generated may not converge. Thus, it is not well suited for general use. However, Newton's method performs quite well if the initial point is sufficiently close to the optimal solution. This is because near an optimal solution, the local contours of the approximating quadratic function are usually a good approximation of those of the function.

Another interesting feature is that the method is not a true descent procedure, since the objective function is not guaranteed to decrease at each iteration. However, the direction, $-\mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$, is guaranteed to be a descent direction if $\mathbf{H}(\mathbf{x}_k)$ is positive definite. Thus, adding a line search to control the step size means the objective will improve at each iteration provided that $\mathbf{H}(\mathbf{x}_k)$ is positive definite. In the case when $\mathbf{H}(\mathbf{x}_k)$ is indefinite, some other corrective action must be taken. One alternative is to give the Newton search direction a bias towards the steepest descent direction as suggested by Levenberg (1944) and Marquardt (1963). Another method, suggested by Fiacco and McCormick (1990, pp. 167–169), involves computing a negative curvature descent direction when $\mathbf{H}(\mathbf{x}_k)$ is indefinite. For further discussion of these methods, see, for example, McCormick (1983).

Example 5. Use classical Newton's method to minimize $f(\mathbf{x}) = (x_1^2 - x_2)^2 + (1 - x_1)^2$. Clearly, since $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, a unique minimum occurs at $\mathbf{x}^* = (1, 1)^t$. First, compute the gradient vector, $\nabla f(\mathbf{x})$, and the Hessian matrix, $\mathbf{H}(\mathbf{x})$.

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 4x_1^3 - 4x_1x_2 + 2x_1 - 2 \\ -2x_1^2 + 2x_2 \end{pmatrix} \tag{21}$$

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 12x_1^2 - 4x_2 + 2 & -4x_1 \\ -4x_1 & 2 \end{pmatrix} \tag{22}$$

Choosing $\mathbf{x}_1 = (-1, 1)^t$, then from (20),

$$\mathbf{x}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} - \begin{pmatrix} 10 & 4 \\ 4 & 2 \end{pmatrix}^{-1} \begin{pmatrix} -4 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \end{pmatrix} \tag{23}$$

$$\mathbf{x}_3 = \begin{pmatrix} 1 \\ -3 \end{pmatrix} - \begin{pmatrix} 26 & -4 \\ -4 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 16 \\ -8 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{24}$$

Observe that $f(\mathbf{x}_1) = 4$, $f(\mathbf{x}_2) = 16$, and $f(\mathbf{x}_3) = 0$. That is, even though the minimum was found in two iterations, the objective function did not improve from iteration 1 to iteration 2.

3.3.2.3. Quasi-Newton Methods In some sense, quasi-Newton methods are an attempt to combine the best features of the steepest descent method with those of Newton's method. Recall that the steepest descent method performs well during early iterations and always decreases the value of the function, whereas Newton's method performs well near the optimum but requires second order derivative information. Quasi-Newton methods are designed to start like the steepest descent method and finish like Newton's method while using only first order derivative information. The basic idea was originally proposed by Davidon (1959) and subsequently developed by Fletcher and Powell (1963). An additional feature of quasi-Newton methods is that the minimum of a convex quadratic function $f : E_n \rightarrow E_1$ can be found in at most n iterations if exact line searches are used. The basic step of the algorithm is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{D}_k \nabla f(\mathbf{x}_k) \tag{25}$$

where $\mathbf{D}_k$ is a deflection matrix requiring only first order derivative information and α_k is determined using a line search. The method of Davidon–Fletcher–Powell can be summarized as follows:

1. Choose a small positive number $\varepsilon > 0$ to test convergence and choose an initial point $\mathbf{x}_1$. Let the initial deflection matrix $\mathbf{D}_1 = \mathbf{I}$ and set the iteration counter $k = 1$.
2. Compute $\nabla f(\mathbf{x}_k)$. If $\|\nabla f(\mathbf{x}_k)\| < \varepsilon$, stop with $\mathbf{x}_k$ as the optimal solution. Otherwise, set $\mathbf{d}_k = -\mathbf{D}_k \nabla f(\mathbf{x}_k)$ and continue with step 3.
3. Let $\alpha = \alpha_k$ be a solution to the univariate problem, $\underset{\alpha \geq 0}{\text{minimize}}\, f(\mathbf{x}_k + \alpha \mathbf{d}_k)$.
4. Compute $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$.
5. Find the updated deflection matrix $\mathbf{D}_{k+1}$ using

$$\mathbf{D}_{k+1} = \mathbf{D}_k + \frac{\mathbf{u}_k \mathbf{u}_k^t}{\mathbf{u}_k^t \mathbf{v}_k} - \frac{\mathbf{D}_k \mathbf{v}_k \mathbf{v}_k^t \mathbf{D}_k}{\mathbf{v}_k^t \mathbf{D}_k \mathbf{v}_k} \tag{26}$$

where $\mathbf{u}_k = \alpha_k \mathbf{d}_k$ and $\mathbf{v}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$. Replace k by $k + 1$ and go to step 2.

Since $\mathbf{D}_1 = \mathbf{I}$, the initial iteration is a steepest descent step. Also, if f is a convex quadratic function, it can be shown that at termination $\mathbf{D}_k$ is the inverse of the Hessian matrix.

There are several other methods for updating $\mathbf{D}_k$ in step 5. Another important method is based on the BFGS formula, which was developed by Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970) and is given by

$$\mathbf{D}_{k+1} = \mathbf{D}_k + \left(1 + \frac{\mathbf{v}_k^t \mathbf{D}_k \mathbf{v}_k}{\mathbf{u}_k^t \mathbf{v}_k}\right) \frac{\mathbf{u}_k \mathbf{u}_k^t}{\mathbf{u}_k^t \mathbf{v}_k} - \frac{\mathbf{u}_k \mathbf{v}_k^t \mathbf{D}_k + \mathbf{D}_k \mathbf{v}_k \mathbf{u}_k^t}{\mathbf{u}_k^t \mathbf{v}_k} \tag{27}$$

The BFGS formula is generally preferred to (26) since computational results have shown that it requires considerably less effort, especially when inexact line searches are used. Quasi-Newton methods, also referred to as variable metric methods, are much more widely used than either the steepest descent or Newton's method. For additional details and computational comparisons, see Fletcher (1987, pp. 44–74).

Example 6. Find the minimum of $f(\mathbf{x}) = x_1^2/2 - x_1 x_2 + x_2^2 - x_1 - x_2$ using the method of Davidon–Fletcher–Powell. Choose an initial point $\mathbf{x}_1 = (0, 0)^t$, an initial deflection matrix $\mathbf{D}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and determine the gradient vector,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} x_1 - x_2 - 1 \\ -x_1 + 2x_2 - 1 \end{pmatrix} \tag{28}$$

Next, compute $\nabla f(\mathbf{x}_1) = (-1, -1)^t$, and as in step 3, solve the line search problem:

$$\underset{\alpha \geq 0}{\text{minimize}}\ f\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} - \alpha \begin{pmatrix} -1 \\ -1 \end{pmatrix}\right) = \alpha^2/2 - 2\alpha \tag{29}$$

to find $\alpha_1 = 2$. Therefore, from (25),

$$\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \tag{30}$$

Now compute the updated deflection matrix as in (26).

$$\mathbf{D}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{\begin{pmatrix} 2 \\ 2 \end{pmatrix} (2 \quad 2)}{(2 \quad 2) \begin{pmatrix} 2 \\ 2 \end{pmatrix}} - \frac{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix} (0 \quad 2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}{(0 \quad 2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix}} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \tag{31}$$

Then,

$$\mathbf{x}_3 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} - 1 \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \tag{32}$$

Since $\nabla f(\mathbf{x}_3) = 0$, the algorithm terminates with $\mathbf{x}^* = (3, 2)^t$. Note that $\mathbf{D}_2$ is precisely the inverse of the Hessian matrix, $\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$, of the convex quadratic objective function $f(\mathbf{x})$.

3.4. Conjugate Gradient Methods

The nonzero vectors $\mathbf{d}_1, \ldots, \mathbf{d}_k$ are said to be *conjugate* with respect to the positive definite matrix $\mathbf{H}$ if they are linearly independent and $\mathbf{d}_i^t \mathbf{H} \mathbf{d}_j = 0$ for $i \neq j$. A method that generates such directions when applied to a quadratic function with Hessian matrix $\mathbf{H}$ is called a *conjugate direction method.* These methods will locate the minimum of a quadratic function in a finite number of iterations, and they can also be applied iteratively to optimize nonquadratic functions.

The algorithm proposed by Fletcher and Reeves (1964), is probably the best-known example of a conjugate gradient algorithm. Starting at some initial point $\mathbf{x}_1$, the algorithm begins like the steepest descent method by searching in the direction $\mathbf{d}_1 = -\nabla f(\mathbf{x}_1)$ to determine $\mathbf{x}_2$. Subsequent directions are computed using the expression

$$\mathbf{d}_{k+1} = -\nabla f(\mathbf{x}_{k+1}) + \alpha_k \mathbf{d}_k \tag{33}$$

where

$$\alpha_k = \frac{\nabla f(\mathbf{x}_{k+1})^t \nabla f(\mathbf{x}_{k+1})}{\nabla f(\mathbf{x}_k)^t \nabla f(\mathbf{x}_k)} \tag{34}$$

Note from (33) that the method actually deflects the current direction of steepest descent by adding on a positive multiple of the direction vector used in the previous step. In fact, it is easily shown that $\mathbf{d}_{k+1}$ is essentially a convex combination of $-\nabla f(\mathbf{x}_{k+1})$ and $\mathbf{d}_k$.

Quasi-Newton methods are generally preferable to conjugate gradient methods because they have been shown to be more efficient. However, conjugate gradient methods have the advantage of requiring no matrix operations and having minimal storage requirements for computer implementation.

4. CONSTRAINED OPTIMIZATION

In this section, methods for addressing constrained optimization problems are discussed. First the classical method of Lagrange multipliers is presented, followed by the Karush–Kuhn–Tucker conditions. Both of these techniques are indirect methods, and general optimality conditions are presented. Then several algorithmic strategies are presented for addressing specific classes of nonlinear programming problems.

4.1. Lagrange Multipliers

The method of Lagrange multipliers is a classical indirect method for solving nonlinear programming problems in which the objective is to optimize a differentiable function subject to a set of equality constraints comprised of differentiable functions. The basic technique involves transforming the constrained nonlinear problem into an unconstrained nonlinear problem by forming what is called the Lagrangian function. Consider the mathematical programming problem with equality constraints,

$$\text{(NLPE) Minimize } f(\mathbf{x}) \tag{35}$$

$$\text{Subject to } h_j(\mathbf{x}) = 0 \text{ for } j = 1, \ldots, p \tag{36}$$

The *Lagrangian function,*

$$L(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \sum_{j=1}^{p} v_j h_j(\mathbf{x}) \tag{37}$$

is formed by introducing a vector of Lagrange multipliers, $\mathbf{v} = (v_1, \ldots, v_p)^t$. v_j is referred to as the Lagrange multiplier, dual multiplier, or dual variable for constraint $h_j(\mathbf{x}) = 0$. Since each feasible point $\mathbf{x}$ must satisfy $h_j(\mathbf{x}) = 0$ for all $j = 1, \ldots, p$, the addition of the term $\Sigma_{j=1}^{p} v_j h_j(\mathbf{x})$ to the objective function $f(\mathbf{x})$ does not change the value of the function. As in section 3, a necessary condition for a stationary point is that the partial derivatives with respect to x_i for $i = 1, \ldots, n$ and v_j for $j = 1, \ldots, p$, equal zero. That is,

$$\frac{\partial L(\mathbf{x}, \mathbf{v})}{\partial x_i} = \frac{\partial f(\mathbf{x})}{\partial x_i} + \sum_{j=1}^{p} v_j \frac{\partial h_j(\mathbf{x})}{\partial x_i} = 0 \quad \text{for } i = 1, \ldots, n \tag{38}$$

$$\frac{\partial L(\mathbf{x}, \mathbf{v})}{\partial v_j} = h_j(\mathbf{x}) = 0 \quad \text{for } j = 1, \ldots, p \tag{39}$$

A stationary point for a general Lagrangian function may or may not be a local extremum. If, as described in Section 2, suitable convexity conditions hold, then the method of Lagrange multipliers will yield a global minimum.

Example 7

$$\text{Minimize } f(\mathbf{x}) = x_1^2 + 2x_2^2 + 4x_1x_2 - 12x_1 - 15x_2 \tag{40}$$

$$\text{Subject to } h_1(\mathbf{x}) = 2x_1 - x_2 - 5 = 0 \tag{41}$$

First, form the Lagrangian function,

$$L(\mathbf{x}, \mathbf{v}) = x_1^2 + 2x_2^2 + 4x_1x_2 - 12x_1 - 15x_2 + v_1(2x_1 - x_2 - 5) \tag{42}$$

Then, using (38) and (39),

$$\frac{\partial L(\mathbf{x}, \mathbf{v})}{\partial x_1} = 2x_1 + 4x_2 - 12 + 2v_1 = 0 \tag{43}$$

$$\frac{\partial L(\mathbf{x}, \mathbf{v})}{\partial x_2} = 4x_1 + 4x_2 - 15 - v_1 = 0 \tag{44}$$

$$\frac{\partial L(\mathbf{x}, \mathbf{v})}{\partial v_1} = 2x_1 - x_2 - 5 = 0 \tag{45}$$

Solving the system of linear equations, (43) through (45), yields $x_1 = 3$, $x_2 = 1$, and $v_1 = 1$. Thus, the optimal solution is $\mathbf{x}^* = (3, 1)^t$.

4.2. Karush–Kuhn–Tucker Conditions

Consider the general nonlinear programming problem,

$$\text{(NLP) Minimize } f(\mathbf{x}) \tag{46}$$

$$\text{Subject to } g_i(\mathbf{x}) \leq 0 \text{ for } i = 1, \ldots, m \tag{47}$$

$$h_j(\mathbf{x}) = 0 \text{ for } j = 1, \ldots, p \tag{48}$$

where $f, g_i, h_j : E_n \rightarrow E_1$ are continuously differentiable functions. The *Karush–Kuhn–Tucker (*KKT*) conditions* are essentially an extension of the method of Lagrange multipliers to problems involving inequality constraints. The results were derived independently by Karush (1939), and Kuhn and Tucker (1951). Although these necessary conditions are not typically used to derive an optimal solution, they do provide insight into solution behavior and form the basis for many nonlinear programming algorithms. They can also be used to verify that a given solution is a candidate for an optimal solution.

4.2.1. KKT Necessary Conditions

If $\mathbf{x}^*$ is a local minimum of problem (NLP) and some constraint qualification holds (e.g. the vectors $\nabla g_i(\mathbf{x}^*)$ for all i such that $g_i(\mathbf{x}^*) = 0$, and $\nabla h_j(\mathbf{x}^*)$ for $j = 1, \ldots, p$, are linearly independent), then there exists vectors $\mathbf{u} = (u_1, \ldots, u_m)^t$ and $\mathbf{v} = (v_1, \ldots, v_p)^t$ such that

$$g_i(\mathbf{x}^*) \leq 0 \text{ for } i = 1, \ldots, m \tag{49}$$

$$h_j(\mathbf{x}^*) = 0 \text{ for } j = 1, \ldots, p \tag{50}$$

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_k} + \sum_{i=1}^{m} u_i \frac{\partial g_i(\mathbf{x}^*)}{\partial x_k} + \sum_{j=1}^{p} v_j \frac{\partial h_j(\mathbf{x}^*)}{\partial x_k} = \quad \text{for } k = 1, \ldots, n \tag{51}$$

$$u_i \geq 0 \text{ for } i = 1, \ldots, m \tag{52}$$

$$u_i g_i(\mathbf{x}^*) = 0 \text{ for } i = 1, \ldots, m \tag{53}$$

Conditions (49) and (50) are called *primal feasibility,* conditions (51) and (52) are called *dual feasibility,* and condition (53) is called *complementary slackness*. To interpret these results geometrically, it is helpful to rewrite (51) in a more compact notation using gradients,

$$—\nabla f(\mathbf{x}^*) = \sum_{i=1}^{m} u_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} v_j \nabla h_j(\mathbf{x}^*) \tag{54}$$

Given a point $\mathbf{x}^*$, condition (53) asserts that the dual variable u_i will be zero if $g_i(\mathbf{x}^*) < 0$ (i.e., the constraint $g_i(\mathbf{x}) \leq 0$ is nonbinding at $\mathbf{x}^*$). Thus, from (53) and (54), the KKT conditions are specifying that, at an optimal solution, $-\nabla f(\mathbf{x}^*)$ can be written as a linear combination of the gradients of the binding constraints (i.e., those constraints satisfied as equalities). In the case involving only constraints of the form $g_i(\mathbf{x}) \leq 0$, $-\nabla f(\mathbf{x}^*)$ must lie in the cone spanned by the gradients of the binding constraints since $u_i \geq 0$ for $i = 1, \ldots, m$. This geometric interpretation is illustrated in Figure 11 using the following example.

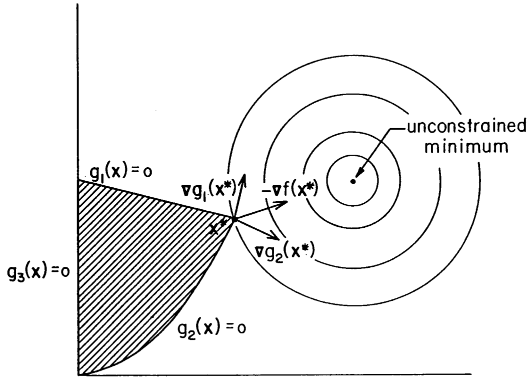


Figure 11 Graphical Solution of Example 8 Illustrating the KKT Conditions.

Example 8

$$\text{Minimize } f(\mathbf{x}) = (x_1 - 14)^2 + (x_2 - 12)^2 \tag{55}$$

$$\text{Subject to } g_1(\mathbf{x}) = x_1 + 4x_2 - 40 \le 0 \tag{56}$$

$$g_2(\mathbf{x}) = x_1^2 - 8x_2 \le 0 \tag{57}$$

$$g_3(\mathbf{x}) = -x_1 \le 0 \tag{58}$$

The KKT conditions for this problem can be written as follows:

$$x_1 + 4x_2 - 40 \le 0 \tag{59}$$

$$x_1^2 - 8x_2 \le 0 \tag{60}$$

$$-x_1 \le 0 \tag{61}$$

$$\begin{pmatrix} 2x_1 - 28 \\ 2x_2 - 24 \end{pmatrix} + u_1 \begin{pmatrix} 1 \\ 4 \end{pmatrix} + u_2 \begin{pmatrix} 2x_1 \\ -8 \end{pmatrix} + u_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{62}$$

$$u_i \ge 0 \text{ for } i = 1, 2, 3 \tag{63}$$

$$u_1(x_1 + 4x_2 - 40) = 0 \tag{64}$$

$$u_2(x_1^2 - 8x_2) = 0 \tag{65}$$

$$u_3(-x_1) = 0 \tag{66}$$

The reader can verify that $\mathbf{x}^* = (8, 8)^t$ and $\mathbf{u}^* = (28/9, 5/9, 0)^t$ satisfy (59) through (66) and thus, as illustrated in Figure 11, the negative gradient of the objective at $\mathbf{x}^*$, $-\nabla f(\mathbf{x}^*)$, lies in the cone spanned by $\nabla g_1(\mathbf{x}^*)$ and $\nabla g_2(\mathbf{x}^*)$.

It should also be pointed out that if f and g_i for $i = 1, \ldots, m$ are convex functions and h_j for $j = 1, \ldots, p$ are linear functions, then the KKT conditions are also sufficient for optimality. This is the case in Example 8. In fact, these assumptions can be somewhat relaxed using the concepts of generalized convexity. For a detailed discussion of first- and second-order KKT conditions, see, for example, Fiacco and McCormick (1990, pp. 17–34).

The following section discusses a solution procedure that is a direct application of the KKT conditions.

4.3. Quadratic Programming

Quadratic programming problems are an important class of linearly constrained problems having the following form:

$$\text{(QP) Minimize } \mathbf{c}^t\mathbf{x} + 1/2\mathbf{x}^t\mathbf{H}\mathbf{x} \tag{67}$$

$$\text{Subject to } \mathbf{A}\mathbf{x} \le \mathbf{b} \tag{68}$$

$$\mathbf{x} \ge \mathbf{0} \tag{69}$$

where **H** is an $n \times n$ symmetric matrix, **A** is an $m \times n$ matrix, $\mathbf{c} = (c_1, \ldots, c_n)^t$, $\mathbf{b} = (b_1, \ldots, b_m)^t$ and the decision vector is $\mathbf{x} = (x_1, \ldots, x_n)^t$. The following method is based on the Karush–Kuhn–Tucker conditions and reduces the quadratic programming problem to what is referred to as a linear complementarity problem.

Let $\mathbf{u} = (u_1, \ldots, u_m)^t$ and $\mathbf{v} = (v_1, \ldots, v_n)^t$ be the dual multipliers for constraints $\mathbf{Ax} \le \mathbf{b}$ and $-\mathbf{x} \le \mathbf{0}$, respectively. Then, using matrix notation, the KKT conditions for (QP) can be expressed as follows:

$$\mathbf{Ax} + \mathbf{s} = \mathbf{b} \tag{70}$$

$$-\mathbf{A}^t\mathbf{u} - \mathbf{Hx} + \mathbf{v} = \mathbf{c} \tag{71}$$

$$\mathbf{x}, \mathbf{v} \ge \mathbf{0}, \mathbf{s}, \mathbf{u} \ge \mathbf{0} \tag{72}$$

$$\mathbf{x}^t\mathbf{v} = 0, \mathbf{u}^t\mathbf{s} = 0 \tag{73}$$

If **H** is positive semidefinite, then problem (QP) is a convex program. Thus, the KKT conditions are sufficient in this case, and any solution to this system will yield a global optimal solution to (QP). When **H** is indefinite, then local optimal solutions which are not global optimal solutions may occur.

The system (70)–(73) can be expressed in the form

$$\mathbf{w} - \mathbf{Mz} = \mathbf{q} \tag{74}$$

$$\mathbf{w} \ge \mathbf{0}, \mathbf{z} \ge \mathbf{0} \tag{75}$$

$$\mathbf{w}^t\mathbf{z} = 0 \tag{76}$$

where $\mathbf{M} = \begin{pmatrix} \mathbf{0} & -\mathbf{A} \\ \mathbf{A}^t & \mathbf{H} \end{pmatrix}$, $\mathbf{q} = \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix}$, $\mathbf{w} = \begin{pmatrix} \mathbf{s} \\ \mathbf{v} \end{pmatrix}$, $\mathbf{Z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}$.

Written in this form, (74)–(76) are an example of a *linear complementarity problem,* which also has applications in game theory. In this context, (76) is referred to as the *complementarity condition,* and all w_i, z_i pairs are said to be *complementary variables*. A method for finding a solution to this system is the complementary pivoting algorithm credited to Lemke (1968). Under certain assumptions on the matrix **M**, the algorithm determines a solution or finds a direction indicating unboundedness in a finite number of iterations.

Other solution procedures for quadratic programming problems include conjugate gradient methods and the Dantzig–Wolfe method (see Dantzig 1963), which uses a modification of the simplex algorithm for linear programming.

4.4. Separable Programming

Separable programming is a procedure for obtaining an approximate solution to a nonlinear programming problem in which the objective function and constraint functions can be expressed as the sum of univariate functions. A separable programming problem has the following form:

$$\text{(SP) Minimize } f(\mathbf{x}) = \sum_{j=1}^{n} f_j(x_j) \tag{77}$$

$$\text{Subject to } \sum_{j=1}^{n} g_{ij}(x_j) \le b_i \text{ for } i, \ldots, m \tag{78}$$

$$x_j \ge 0 \text{ for } j = 1, \ldots, n \tag{79}$$

Here, for example, $f(\mathbf{x})$ may represent the total cost whereas $f_j(x_j)$ represents the cost contribution of variable x_j.

The solution technique involves approximating each of the functions $f_j(x_j)$ and each of the functions $g_{ij}(x_j)$ by piecewise linear functions on their respective intervals of interest. Before presenting the resulting mathematical programming formulation, it is informative to review the idea of a linear approximation.

A *piecewise linear approximation* of a univariate function $h(x)$ on an interval of interest $[a, b]$ is illustrated in Figure 12. The interval $[a, b]$ is partitioned using t grid points, $a = x_1, x_2, \ldots, x_t = b$. Then the linear approximation, $l_k(x)$, on subinterval $[x_k, x_{k+1}]$ can be written as follows using the concept of a convex combination.

$$l_k(x) = \lambda h(x_k) + (1 - \lambda)h(x_{k+1}) \tag{80}$$

where

$$x = \lambda x_k + (1 - \lambda)x_{k+1} \tag{81}$$

$$\lambda \in [0, 1] \tag{82}$$

Generalizing, the piecewise linear approximation of $h(x)$ on the interval $[a, b]$ is given by

$$l(x) = \sum_{k=1}^{t} \lambda_k h(x_k) \tag{83}$$

$$\sum_{k=1}^{t} \lambda_k = 1 \tag{84}$$

$$\lambda_k \geq 0 \quad \text{for } k = 1, \ldots, t \tag{85}$$

where, at most, two adjacent λ_ks are positive. This last restriction arises because if nonadjacent λ_ks are positive, the resulting approximation will not lie on the piecewise linear approximating function. The accuracy of this approximation improves as the number of grid points increases, but, increasing the number of grid points increases the number of variables in the approximating problem.

Denoting the grid points for variable x_j by x_{kj} for $k = 1, \ldots, t_j$, the approximating problem is

$$\text{Minimize} \sum_{j=1}^{n} \sum_{k=1}^{t_j} \lambda_{kj} f_j(x_{kj}) \tag{86}$$

$$\text{Subject to} \sum_{j=1}^{n} \sum_{k=1}^{t_j} \lambda_{kj} g_{ij}(x_{kj}) = b_i \quad \text{for } i = 1, \ldots, m \tag{87}$$

$$\sum_{k=1}^{t_j} \lambda_{kj} = 1 \quad \text{for } j = 1, \ldots, n \tag{88}$$

$$\lambda_{kj} \geq 0 \quad \text{for } k = 1, \ldots, t_j; \quad j = 1, \ldots, n \tag{89}$$

$$\text{At most, two adjacent } \lambda_{kj}\text{s are positive for } j = 1, \ldots, n \tag{90}$$

Except for constraint (90), this is a linear programming problem in the variables λ_{kj}, which can be solved by the simplex method provided a restricted basis entry rule is used which enforces (90). However, in the case when each f_j is strictly convex and each g_{ij} is convex, constraint (90) can be

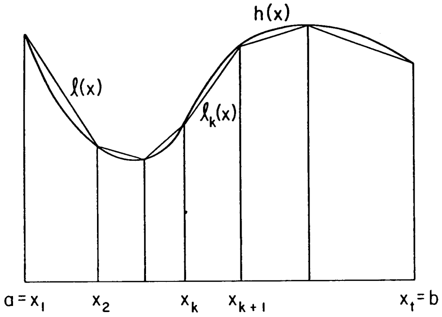


Figure 12 Piecewise Linear Approximation, $l(x)$, of a Function, $h(x)$.

neglected since it will be satisfied automatically by a solution generated by the simplex method. For additional details, see Miller (1963), and Wolfe (1963).

4.5. Geometric Programming

As has been shown in the two previous sections, it is sometimes possible to find the solution of a problem by transforming it into a simpler problem. This is also the case with geometric programming problems, another important class of nonlinear programming problems. The motivation for addressing this class of problems originated with work in engineering design. Geometric programming problems have the following general form:

$$\text{(GP) Minimize } g_0(\mathbf{x}) = \sum_{k=1}^{t_0} c_{0k} \prod_{j=1}^{n} (x_j)^{a_{0jk}} \tag{91}$$

$$\text{Subject to } g_i(\mathbf{x}) = \sum_{k=1}^{t_i} c_{ik} \prod_{j=1}^{n} (x_j)^{a_{ijk}} \leq 1 \text{ for } i = 1, \ldots, m \tag{92}$$

$$x_j > 0 \text{ for } j = 1, \ldots, n \tag{93}$$

where the coefficients c_{ik}, $i = 0, \ldots, m$ must be positive. Although each of the functions $g_i(\mathbf{x})$ for $i = 0, \ldots, m$ is not a true polynomial since the exponents a_{ijk} are not restricted to be integers, they are generally referred to as *posynomials* or positive generalized polynomials.

Problem (GP) has a nonlinear objective function and nonlinear constraints and, as such, is quite difficult to solve. However, geometric programming problems belong to a class of problems whose dual problems involve only linear constraints. Let δ_{ik} be the dual variable associated with the term $c_{ik}\prod_{j=1}^{n} (x_j)^{a_{ijk}}$. Then the dual problem for problem (GP) can be written as follows:

$$\text{Minimize } v(\boldsymbol{\delta}, \boldsymbol{\lambda}) = \prod_{i=0}^{m} \prod_{k=1}^{t_i} \left(\frac{c_{ik}}{\delta_{ik}}\right)^{\delta_{ik}} \prod_{i=1}^{m} (\lambda_i)^{\lambda_i} \tag{94}$$

$$\text{Subject to } \sum_{i=0}^{m} \sum_{k=1}^{t_i} a_{ijk}\delta_{ik} = 0 \text{ for } j = 1, \ldots, n \tag{95}$$

$$\sum_{k=1}^{t_0} \delta_{0k} = 1 \tag{96}$$

$$\sum_{k=1}^{t_i} \delta_{0k} = \lambda_i \text{ for } i = 1, \ldots, m \tag{97}$$

$$\boldsymbol{\delta} \geq \mathbf{0} \tag{98}$$

This dual problem has $t = \sum_{i=0}^{m} t_i$ variables, and at best, in the case when $t - n - 1 = 0$, determining the solution simply involves solving a square system of linear equations in nonnegative variables. The quantity $t - n - 1$ is referred to as the *degree of difficulty*. Although the dual objective function as written in (94) is neither convex nor concave, by using a natural logarithm transformation, the resulting function can be shown to be a concave function. In fact, by taking the logarithm, the objective function becomes separable and the dual problem can be cast as a separable programming problem. Thus, at worst, solving the dual problem involves maximizing a nonlinear concave objective function subject to linear constraints. As such, any local solution to the dual problem is a global solution. Even so, the advantage of solving the dual problem is offset by the fact that it can be difficult to find the optimal primal variables given the optimal dual variables and the optimal objective value.

Example 9

$$\text{Minimize } g_0(\mathbf{x}) = 4x_1x_2^3 + 20x_1^{-2} \tag{99}$$

$$\text{Subject to } g_1(\mathbf{x}) = x_1x_2^{-1} \leq 1 \tag{100}$$

$$x_1, x_2 > 0 \tag{101}$$

Letting δ_{01}, δ_{02}, and δ_{11} be the dual variables for the terms $4x_1x_2^3$, $20x_1^{-2}$, and $x_1x_2^{-1}$, respectively, then the dual problem becomes

$$\text{Maximize } v(\boldsymbol{\delta}, \boldsymbol{\lambda}) = \left(\frac{4}{\delta_{01}}\right)^{\delta_{01}} \left(\frac{20}{\delta_{02}}\right)^{\delta_{02}} \left(\frac{1}{\delta_{11}}\right)^{\delta_{11}} (\lambda_1)^{\lambda_1} \tag{102}$$

$$\text{Subject to } \delta_{01} - 2\delta_{02} + \delta_{11} = 0 \tag{103}$$

$$3\delta_{01} - \delta_{11} = 0 \tag{104}$$

$$\delta_{01} + \delta_{02} = 1 \tag{105}$$

$$\delta_{11} = \lambda_1 \tag{106}$$

$$\delta_{01}, \delta_{02}, \delta_{11} > 0 \tag{107}$$

Condition (103) is derived from the exponents of x_1 in the respective terms of the primal. Similarly, (104) is derived from the exponents of x_2. These are referred to as the *orthogonality conditions,* whereas (105) is termed the *normality condition*. In this case $t - n - 1 = 3 - 2 - 1 = 0$ and the dual problem can be solved by finding the solution of the linear system of equations, (103)–(105). Thus, $\delta^*_{01} = 1/3$, $\delta^*_{02} = 2/3$, $\lambda^*_1 = \delta^*_{01} = 1$, with $g_0(\mathbf{x}^*) = v(\boldsymbol{\delta}^*, \boldsymbol{\lambda}^*) = 12^{1/3}30^{2/3}$.

The optimal values of the primal variables can be recovered by utilizing the following condition:

$$\delta^*_{ok} = \frac{c_{0k}}{g_0(\mathbf{x}^*)} \prod_{j=1}^{n} (x^*_j)^{a_{0jk}} \text{ for } k = 1, \ldots, t_0 \tag{108}$$

Substituting into (108) yields

$$\frac{1}{3} = \frac{4}{12^{1/3}30^{2/3}} x_1 x_2^3 \tag{109}$$

$$\frac{2}{3} = \frac{20}{12^{1/3}30^{2/3}} x_1^{-2} \tag{110}$$

Now solving (109) and (110) gives $x^*_1 = x^*_2 = (5/2)^{1/6}$.

For a complete discussion of geometric programming, the interested reader is referred to Duffin et al. (1967), and Beighter and Phillips (1976). Reviews of software for solving geometric programming problems are provided in Dembo (1976) and Rijckaert and Walraven (1985).

4.6. Methods of Feasible Directions

Consider the following nonlinear programming problem:

$$\text{Minimize } f(\mathbf{x}) \tag{111}$$

$$\text{Subject to } \mathbf{x} \in S \subseteq E_n \tag{112}$$

Conceptually, methods of feasible directions operate in a manner similar to unconstrained multidimensional search techniques. That is, the basic idea is, given a feasible point $\mathbf{x}_k$, determine a direction $\mathbf{d}_k$ and a step length α_k that yield the new point $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$. In the constrained case, however, care must to taken to choose a direction $\mathbf{d}_k$ that not only produces a point $\mathbf{x}_{k+1}$ that improves the objective function, but also maintains the feasibility of $\mathbf{x}_{k+1}$. For a differentiable objective function $f(\mathbf{x})$, an *improving feasible direction* $\mathbf{d}_k$ at the point $\mathbf{x}_k \in S$, has the following two properties (see Figure 13):

1. $\nabla f(\mathbf{x}_k)^t \mathbf{d}_k < 0$, that is, the *directional derivative* at $\mathbf{x}_k$ in the direction $\mathbf{d}_k$ is negative, resulting in a reduction in objective value. Geometrically, this means that $\mathbf{d}_k$ forms an acute angle with $-\nabla f(\mathbf{x}_k)$.
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{d}_k \in S$ for some $\alpha > 0$, that is, a feasible move is possible in the direction $\mathbf{d}_k$.

Thus, methods of feasible directions operate by determining an improving feasible direction and then solving a line search problem to determine the step length in that direction. This process is repeated until some stopping rule is satisfied. Since the sequence of points generated is feasible to the primal problem, these are often called *primal methods*. The way in which the directions are generated and the step sizes are computed determines the various methods.

One such method that can be applied to a nonlinear programming problem with a linear constraint set is the method of reduced gradient, originally proposed by Wolfe (1963). It operates in a manner

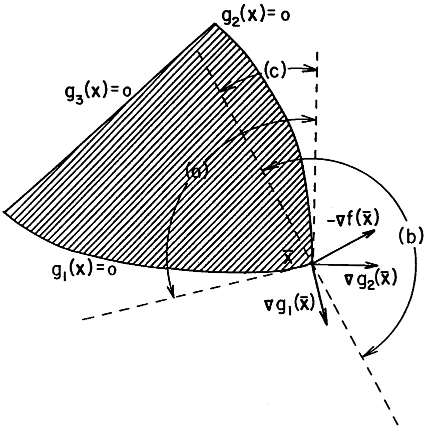


Figure 13 Illustration of (*a*) Feasible direction set at $\bar{\mathbf{x}}$, (*b*) Improving direction set at $\bar{\mathbf{x}}$, and (*c*) Improving feasible direction set at $\bar{\mathbf{x}}$.

similar to the simplex method for linear programming by using a set of independent variables to reduce the dimensionality of the problem. The reduced gradient is the gradient with respect to these independent variables. This method was generalized to handle nonlinear constraints by Abadie and Carpentier (1969) and is called the generalized reduced gradient method (GRG). There are several other methods of feasible directions, including those proposed by Zoutendijk (1960), the gradient projection method of Rosen (1961), and the convex simplex method of Zangwill (1967).

4.7. Sequential Unconstrained Minimization Techniques

In this section, methods for converting a general constrained nonlinear programming problem into an equivalent unconstrained problem are discussed. Once this conversion has been made, algorithms for unconstrained optimization can be applied. Unfortunately, there are computational difficulties associated with this process, and instead of solving a single unconstrained problem, it is usually necessary to solve a sequence of unconstrained problems. Although penalty functions methods were originally introduced by Courant (1943), the sequential unconstrained minimization technique (SUMT) was primarily developed by Fiacco and McCormick (1964). There are two basic approaches, both of which add a penalty term to the objective function. In the *penalty function method* (or exterior penalty function method), the optimal solution is approached by a sequence of infeasible points. That is, the optimum is approached from the exterior of the feasible region. In the second approach, known as the *barrier function method* (or interior penalty function method), the sequence of points generated converges to the optimum from within the feasible region.

4.7.1. Penalty Function Methods

Consider the following problem:

$$\text{Minimize } f(\mathbf{x}) \tag{113}$$

$$\text{Subject to } \mathbf{x} \in S \subseteq E_n \tag{114}$$

The basic idea is to approximate this problem with an unconstrained problem by adding a penalty function to the objective function that prescribes a high cost for violation of the constraint set S. This new unconstrained *auxiliary problem* is of the form

$$\text{Minimize } f(\mathbf{x}) + rP(\mathbf{x}) \tag{115}$$

where r is a positive constant, and $P(\mathbf{x})$ is chosen as a continuous penalty function such that (1) $P(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, and (2) $P(\mathbf{x}) = 0$ if and only if $\mathbf{x} \in S$. For large r, the optimal solution of (115) will be in a region where $P(\mathbf{x})$ is small. However, if r is chosen too large, then the unconstrained problem becomes ill-conditioned and difficult to solve. Thus, a sequence of problems is solved in which r is increased from problem to problem. A sequence of problems that could be used for solving the general nonlinear programming problem (NLP), (46)–(48), is

$$\text{Minimize } f(\mathbf{x}) + r_k \left(\sum_{i=1}^{m} [\max\{0, g_i(\mathbf{x})\}]^q + \sum_{j=1}^{p} |h_j(\mathbf{x})|^q \right) \tag{116}$$

where q is a positive integer and r_k represents a strictly increasing sequence of positive numbers such that $r_k \rightarrow \infty$.

4.7.2. Barrier Function Methods

Barrier function methods are very similar to penalty function methods except that they start at an interior point of the feasible region and set a barrier against leaving the feasible region. In this case, the feasible region must have an interior, so this method is generally restricted to inequality constraints. Consider the nonlinear problem with inequality constraints,

$$\text{(NLPI) Minimize } f(\mathbf{x}) \tag{117}$$

$$\text{Subject to } g_i(\mathbf{x}) \leq 0 \text{ for } i = 1, \ldots, m \tag{118}$$

Ideally, a barrier function, $B(\mathbf{x})$, would assume the value zero for $\mathbf{x} \in \{\mathbf{x}: g_i(\mathbf{x}) < 0 \text{ for } i = 1, \ldots, m\}$ and the value ∞ on the boundary of the feasible region. $B(\mathbf{x})$ is usually defined such that (1) $B(\mathbf{x})$ is continuous and nonnegative on the interior of the feasible region, and (2) $B(\mathbf{x}) \rightarrow \infty$ on the boundary of the feasible region. A typical barrier function is

$$B(\mathbf{x}) = \sum_{i=1}^{m} [-1/g_i(\mathbf{x})] \tag{119}$$

This would result in the sequence of auxiliary problems

$$\text{Minimize } f(\mathbf{x}) + t_k \sum_{i=1}^{m} [-1/g_i(\mathbf{x})] \tag{120}$$

where t_k is a strictly decreasing sequence of positive numbers such that $t_k \rightarrow 0$.

4.7.3. Augmented Lagrangian Methods

In an attempt to avoid the ill-conditioning that occurs in the regular penalty and barrier function methods, Hestenes (1969) and Powell (1969) independently developed a multiplier method for solving nonlinearly constrained problems. This multiplier method was originally developed for equality constraints and involves optimizing a sequence of unconstrained augmented Lagrangian functions. It was later extended to handle inequality constraints by Rockafellar (1973).

Consider the mathematical programming problem

$$\text{(NLPE) Minimize } f(\mathbf{x}) \tag{121}$$

$$\text{Subject to } h_j(\mathbf{x}) = 0 \text{ for } j = 1, \ldots, p \tag{122}$$

The *augmented Lagrangian* function,

$$L_r(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \sum_{j=1}^{p} v_j h_j(\mathbf{x}) + r \sum_{j=1}^{p} (h_j(\mathbf{x}))^2 \tag{123}$$

is formed by introducing a multiplier vector, $\mathbf{v} = (v_1, \ldots, v_p)^t$ and a positive penalty parameter r. Note that $L_r(\mathbf{x}, \mathbf{v})$ in (123) is the Lagrangian function (37) augmented with the term $r \sum_{j=1}^{p} (h_j(\mathbf{x}))^2$. Thus, instead of a single penalty parameter, as in regular penalty function methods, the augmented Lagrangian requires estimates of the Lagrange multipliers. Using these multiplier estimates, which are updated from iteration to iteration, reduces the ill-conditioning of the unconstrained problems. Assuming that at iteration k, estimates $\mathbf{x}_k$, $\mathbf{v}_k$, and penalty parameter r_k are available, the problem minimize $L_{r_k}(\mathbf{x}, \mathbf{v}_k)$ is solved to find the local unconstrained solution $\mathbf{x}_{k+1}$. The updated vector $\mathbf{v}_{k+1}$ is then found using the relationship

$$\mathbf{v}_{k+1} = \mathbf{v}_k + 2r_k \mathbf{h}(\mathbf{x}_{k+1}) \tag{124}$$

where $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_p(\mathbf{x}))^t$. Upon selecting $r_{k+1} \geq r_k$, the process is repeated until $\mathbf{x}_{k+1}$ is sufficiently close to a local solution of (NLPE).

A summary of basic penalty function techniques, as well as exact penalty functions and multiplier methods are discussed in Fletcher (1987, pp. 277–304). For a complete discussion of multiplier methods, the interested reader is referred to Bertsekas (1982).

4.8. Successive Linear Programming

Successive (or sequential) linear programming (SLP) algorithms were introduced by Griffith and Stewart (1961) and have been used in a number of application areas, especially the oil and gas industry. SLP algorithms solve nonlinear optimization problems by using a sequence of linear programs, and computational results have shown they are particularly efficient on problems that are highly constrained. Consider the following nonlinear programming problem:

$$\text{(NLP1) Minimize } f(\mathbf{x}) + \mathbf{c}^t\mathbf{y} \tag{125}$$

$$\text{Subject to } \mathbf{g}(\mathbf{x}) + \mathbf{A}\mathbf{y} = \mathbf{b} \tag{126}$$

$$\mathbf{D}\mathbf{x} + \mathbf{E}\mathbf{y} = \mathbf{e} \tag{127}$$

$$\mathbf{l} \le \mathbf{x} \le \mathbf{u} \tag{128}$$

$$\mathbf{s} \le \mathbf{y} \le \mathbf{t} \tag{129}$$

where $\mathbf{A}$ is an $m \times p$ matrix, $\mathbf{D}$ is $q \times n$, $\mathbf{E}$ is $q \times p$, $\mathbf{c}, \mathbf{s}, \mathbf{t} \in E_p$, $\mathbf{l}, \mathbf{u} \in E_n$, $\mathbf{b} \in E_m$, $\mathbf{e} \in E_q$, $f : E_n \rightarrow E_1$, and $\mathbf{g} : E_n \rightarrow E_m$, that is, $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^t$. Note that the decision variables have been partitioned into the nonlinear variables $\mathbf{x} \in E_n$, and the variables $\mathbf{y} \in E_p$, which appear only linearly. Similarly, the constraints are divided into nonlinear constraints (126) involving the vector of nonlinear differentiable functions $\mathbf{g}(\mathbf{x})$, and the linear constraints (127). The nonlinear variables $\mathbf{x}$ appear in the objective function (125) via the nonlinear differentiable function f.

The basic idea is, given a base point $\bar{\mathbf{x}}$, the nonlinear functions, f, $\mathbf{g}$, are linearized using first order Taylor series approximations. That is, $f(\mathbf{x})$ is replaced by $f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t\mathbf{d}$ where $\mathbf{d} = \mathbf{x} - \bar{\mathbf{x}}$. Similarly, $g_i(\mathbf{x})$ for $i = 1, \ldots, m$, is replaced by $g_i(\bar{\mathbf{x}}) + \nabla g_i(\bar{\mathbf{x}})^t\mathbf{d}$. It is assumed that these linear approximations are accurate on some interval $-\boldsymbol{\delta} \le \mathbf{d} \le \boldsymbol{\delta}$ where $\boldsymbol{\delta} \in E_n$, $\boldsymbol{\delta} > \mathbf{0}$. Substituting these results into (NLP1) yields the linear program

$$\text{(LP}(\bar{\mathbf{x}}, \boldsymbol{\delta})\text{) Minimize } \nabla f(\bar{\mathbf{x}})^t\mathbf{d} + \mathbf{c}^t\mathbf{y} \tag{130}$$

$$\text{Subject to } \mathbf{J}(\bar{\mathbf{x}})\mathbf{d} + \mathbf{A}\mathbf{y} = \mathbf{b} - \mathbf{g}(\bar{\mathbf{x}}) \tag{131}$$

$$\mathbf{D}\mathbf{d} + \mathbf{E}\mathbf{y} = \mathbf{e} - \mathbf{D}\bar{\mathbf{x}} \tag{132}$$

$$\max(\mathbf{l} - \bar{\mathbf{x}}, -\delta) \le \mathbf{d} \le \max(\mathbf{u} - \bar{\mathbf{x}}, \boldsymbol{\delta}) \tag{133}$$

$$\mathbf{s} \le \mathbf{y} \le \mathbf{t} \tag{134}$$

where $\mathbf{J}(\bar{\mathbf{x}})$ is the Jacobian matrix whose jth column is $\nabla g_j(\bar{\mathbf{x}})$.

Assuming a feasible solution exists, LP($\bar{\mathbf{x}}$, $\boldsymbol{\delta}$) is solved to find the solution $\bar{\mathbf{y}}$, $\bar{\mathbf{d}}$ that results in the candidate solution ($\bar{\mathbf{x}} + \bar{\mathbf{d}}$, $\bar{\mathbf{y}}$). If this solution is acceptable, then the bounds $\boldsymbol{\delta}$ may be increased and the process repeated. Otherwise, the bounds $\boldsymbol{\delta}$ are decreased and LP($\bar{\mathbf{x}}$, $\boldsymbol{\delta}$) is resolved. The process terminates when $\|\bar{\mathbf{d}}\|$ is sufficiently small.

Details of SLP algorithms along with computational results are contained in Palacios-Gomez et al. (1982), Baker and Lasdon (1985), and Zhang et al. (1985).

4.9. Successive Quadratic Programming

Successive quadratic programming (SQP) algorithms are an important class of methods that has shown much promise in solving general nonlinear programming problems. The methods are also referred to as Wilson–Han–Powell-type methods (Wilson 1963; Han 1976; Powell 1978) as well as Lagrange–Newton methods. SQP algorithms essentially determine a Karush–Kuhn–Tucker point by applying Newton's method to find a stationary point of the Lagrangian function. For a discussion of SQP algorithms, see, for example, Stoer (1985) and Fletcher (1987, pp. 304–317).

4.10. Nonsmooth Optimization

Nonsmooth or nondifferentiable optimization plays an important role in large-scale programming and addresses mathematical programming problems in which the functions involved have discontinuous first derivatives. Thus, classical methods that rely on gradient information fail to solve these problems, and alternative nonstandard approaches must be used. These alternative methods include subgradient methods and bundle methods. The interested reader is referred to Shor (1985), Zowe (1985), and Fletcher (1987, pp. 357–414).

5. ONLINE SOURCES OF INFORMATION ON OPTIMIZATION

There is a wealth of information on mathematical programming available on the World Wide Web. The resources range from electronic books on optimization to libraries of source code for optimization problems to test data archives for mathematical programming. The following is a listing of a few of the more comprehensive optimization sites. Many of the sites also provide links to other related sites.

- Center for Advanced Modeling and Optimization (CAMO)
  http://www.ici.ro/camo/
- Mathematical Optimization (Computational Science Education Project)
  http://csep1.phy.ornl.gov/mo/mo.html
- The Optimization Technology Center and The Network Enabled Optimization System (NEOS)
  http://www-fp.mcs.anl.gov/otc/
- Nonlinear Programming FAQ (Optimization Technology Center)
  http://www-unix.mcs.anl.gov/otc/Guide/faq/nonlinear-programming-faq.html
- Mathematical Programming Glossary (Harvey J. Greenberg)
  http://www.cudenver.edu/~hgreenbe/glossary/glossary.html

6. NONLINEAR PROGRAMMING CODES

Due to advances in computer technology and algorithmic techniques, mathematical programming codes have made significant progress in recent years. Below is a partial listing of some of the nonlinear programming software that is available to the industrial engineering practitioner. For a detailed discussion of software packages, the interested reader is referred to Moré and Wright (1993). Several online sites also provide listings and reviews of available optimization software. See, for example,

- Decision Tree for Optimization Software (H. D. Mittelmann and P. Spellucci)
  http://plato.la.asu.edu/topics/problems/nlores.html
- Guide to Available Mathematical Software (National Institute of Standards and Technology)
  http://gams.cam.nist.gov/
- Nonlinear Programming FAQ (Optimization Technology Center)
  http://www-unix.mcs.anl.gov/otc/Guide/faq/nonlinear-programming-faq.html
- Nonlinear Programming Packages (Center for Advanced Modeling and Optimization)
  http://www.ici.ro/camo/hnp.htm
- Optimization Software (Optimization Technology Center)
  http://www-fp.mcs.anl.gov/otc/Guide/SoftwareGuide/

6.1. Optimization Software

CONOPT
Problem Type: Nonlinear programs with sparse nonlinear constraints
Method: Generalized reduced gradient
Author: Arne S. Drud, ARKI Consulting and Development A/S, Denmark
Contact: ARKI Consulting and Development A/S, Email: info@arki.dk

DONLP2
Problem Type: Smooth nonlinear functions subject to smooth constraints
Method: Sequential quadratic programming
Author: Peter Spellucci, Technical University Darmstadt, Germany
Contact: http://www.mathematik.tu-darmstadt.de/ags/ag8/spellucci/

EA3
Problem Type: Nonlinear programs
Method: Ellipsoid algorithm
Authors: J. G. Ecker, M. Kupferschmid, Rensselaer Polytechnic Institute, NY
Contact: J. G. Ecker, Department of Mathematical Sciences; M. Kupferschmid, Alan M. Voorhees Computing Center, Rensselaer Polytechnic Institute, Troy, NY 12181

FSQP
Problem Type: Multiple linear/nonlinear objective functions with linear/nonlinear constraints
Method: Sequential quadratic programming
Authors: Eliane R. Panier, Andre Tits, Jian Zhou, Craig Lawrence, University of Maryland

Contact: http://www.isr.umd.edu/Labs/CACSE/FSQP/fsqp.html

GRG2
Problem Type: Nonlinear programs
Method: Generalized reduced gradient
Author: Prof. Leon Lasdon, The University of Texas at Austin
Contact: http://www.optimalmethods.com/

LANCELOT
Problem Type: Large-scale optimization problems
Method: Penalty method
Authors: Andy Conn, IBM T. J. Watson Research Center, NY, Nick Gould, Rutherford Appleton Laboratory, UK, Philippe Toint, Facultés Universitaires Notre Dame de la Paix, Belgium
Contact: http://www.cse.clrc.ac.uk/Activity/LANCELOT+165

LSGRG2
Problem Type: Large-scale nonlinear programs
Method: Generalized reduced gradient
Author: Prof. Leon Lasdon, The University of Texas at Austin
Contact: http://www.optimalmethods.com/

MINOPT
Problem Type: Linear, mixed-integer, nonlinear, dynamic, and mixed-integer nonlinear programs
Method: Generalized benders decomposition, outer approximation and variants, generalized cross decomposition
Authors: C. Schweiger, Christodoulos A. Floudas, Princeton University
Contact: http://titan.princeton.edu/MINOPT/minopt.html

MINOS
Problem Type: Large-scale linear and nonlinear programs
Method: Projected Lagrangian
Authors: Bruce A. Murtagh, University of New South Wales, Australia, Michael A. Saunders, Stanford University
Contact: http://www.stanford.edu/~saunders/brochure/brochure.html

NIMBUS
Problem Type: Differentiable/nondifferentiable multiobjective/single objective optimization problems with nonlinear/linear constraints
Method: Nondifferentiable interactive multiobjective bundle-based optimization
Authors: Kaisa Miettinen, University of Jyväskylä, Finland
Contact: http://nimbus.mit.jyu.fi/

NLPQL
Problem Type: Nonlinear programs
Method: Sequential quadratic programming
Authors: K. Schittkowski, University of Bayreuth, Germany
Contact: http://www.uni-bayreuth.de/departments/math/~kschittkowski/nlpql.htm

NLPQLB
Problem Type: Smooth nonlinear programming with many constraints
Method: Sequential quadratic programming
Authors: Schittkowski, University of Bayreuth, Germany
Contact: http://www.uni-bayreuth.de/departments/math/~kschittkowski/nlpqlb.htm

NPSOL
Problem Type: Dense linear and nonlinear programs
Method: Sequential quadratic programming
Authors: Philip Gill, University of California, San Diego; Walter Murray, Michael A. Saunders, Stanford University; Margaret H. Wright, AT&T Bell Laboratories
Contact: http://www.stanford.edu/~saunders/brochure/brochure.html

OPTIMA Library
Problem Type: Unconstrained and constrained nonlinear optimization
Method: Various methods
Authors: M. C. Bartholomew-Biggs, University of Hertfordshire, United Kingdom
Contact: Dr. M. C. Bartholomew-Biggs, Numerical Optimisation Center, Hatfield, Hertfordshire AL10 9AB, United Kingdom

SNOPT
Problem Type: Large-scale linear and nonlinear programs
Method: Sparse sequential quadratic programming
Authors: Philip Gill, University of California, San Diego; Walter Murray, Michael A. Saunders, Stanford University
Contact: http://www.stanford.edu/~saunders/brochure/brochure.html

SOLVOPT
Problem Type: Nonlinear programs
Method: Exact penalty method
Authors: Alexei V. Kuntsevich, Karl-Franzens Universität Graz, Austria, Franz Kappel
Contact: http://bedvgm.kfunigraz.ac.at:8001/alex/solvopt/

SPENBAR
Problem Type: Nonlinear programs
Method: Modified penalty method
Authors: Neculai Andrei, Research Institute for Informatics, Romania
Contact: Neculai Andrei, Research Institute for Informatics, 8-10, Bdl. Maresal Averescu, 71316 Bucharest, Romania, E-mail: nandrei@u3.ici.ro

TRON
Problem Type: Large bound-constrained optimization problems
Method: Trust region Newton method
Authors: Chih-Jen Lin,National Taiwan University; Jorge Moré, Argonne National Laboratory
Contact: http://www-unix.mcs.anl.gov/~more/tron/

REFERENCES

Abadie, J., and Carpentier, J. (1969), "Generalization of the Wolfe Reduced Gradient Method to the Case of Nonlinear Constraints," in *Optimization,* R. Fletcher, Ed., Academic Press, New York.

Armijo, L. (1966), "Minimization of Functions having Lipschitz Continuous First-Partial Derivatives," *Pacific J. Mathematics,* Vol. 16, No. 1, pp. 1–3.

Avriel, M. (1976), *Nonlinear Programming: Analysis and Methods,* Prentice-Hall, Englewood Cliffs, NJ, pp. 155–183.

Baker, T. E., and Lasdon, L. S. (1985), "Successive Linear Programming at Exxon," *Management Science,* Vol. 31, No. 3, pp. 264–274.

Bazaraa, M. S., Sherali, H. D, and Shetty, C. M. (1994), *Nonlinear Programming: Theory and Applications,* John Wiley & Sons, New York.

Beightler, C. S., and Phillips, D. T. (1976), *Applied Geometric Programming,* John Wiley & Sons, New York.

Bertsekas, D. P. (1982), *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, New York, 1982.

Bracken, J., and McCormick, G. P. (1968), *Selected Applications of Nonlinear Programming,* John Wiley & Sons, New York.

Broyden, C. G. (1970), "The Convergence of a Class of Double Rank Minimization Algorithms 2. The New Algorithm," *Journal of Institute of Mathematics and Its Applications,* Vol. 6, pp. 222–231.

Courant, R. (1943), "Variational Methods for the Solution of Problems of Equilibrium and Vibration," *Bulletin of the American Mathematical Society,* Vol. 49, pp. 1–23.

Dantzig, G. B. (1963), *Linear Programming and Extensions,* Princeton University Press, Princeton, NJ, pp. 490–496.

Davidon, W. C. (1959), "Variable Metric Method for Minimization," *AEC Research Development Report,* ANL-5990.

Dembo, R. S. (1976), "The Current State of the Art of Algorithms and Computer Software for Geometric Programming," Working Paper No. 88, School of Organization and Management, Yale University, New Haven, CT.

Duffin, R. J., Peterson, E. L., and Zener, C. (1963), *Geometric Programming: Theory and Application,* John Wiley & Sons, New York.

Fiacco, A. V., and McCormick, G. P. (1964), "The Sequential Unconstrained Minimization Technique for Nonlinear Programming, A Primal-Dual Method," *Management Science,* Vol. 10, pp. 360–366.

Fiacco, A. V., and McCormick, G. P. (1990), *Nonlinear Programming: Sequential Unconstrained Minimization Techniques,* SIAM, Philadelphia.

Fletcher, R. (1970), "A New Approach to Variable Metric Algorithms," *Computer Journal,* Vol. 13, pp. 317–322.

Fletcher, R. (1987), *Practical Methods of Optimization,* 2nd Ed., John Wiley & Sons, New York.

Fletcher, R., and Powell, M. (1963), "A Rapidly Convergent Descent Method for Minimization," *Computer Journal,* Vol. 6, pp. 163–168.

Fletcher, R., and Reeves, C. M. (1964), "Function Minimization by Conjugate Gradients," *Computer Journal,* Vol. 7, pp. 149–154.

Goldfarb, D. (1970), "A Family of Variable Metric Methods Derived by Variational Means," *Mathematics of Computation,* Vol. 24, pp. 23–26.

Griffith, R. E., and Stewart, R. A. (1961), "A Nonlinear Programming Technique for the Optimization of Continuous Processing Systems," *Management Science,* Vol. 7, pp. 379–392.

Han, S. P. (1976), "Superlinearly Convergent Variable Metric Algorithms for General Nonlinear Programming Problems," *Mathematical Programming,* Vol. 11, pp. 263–282.

Hestenes, M. R. (1969), "Multiplier and Gradient Methods," *Journal of Optimization Theory and Applications,* Vol. 4, pp. 303–320.

Hooke, R., and Jeeves, T. A. (1961), "Direct Search Solution of Numerical and Statistical Problems," *J. Association Computer Machinery*, Vol. 8, pp. 212–229.

Karmarkar, N. (1984), "A New Polynomial-Time Algorithm for Linear Programming," *Combinatorics,* Vol. 4, pp. 373–395.

Karush, W. (1939), "Minima of Functions of Several Variables with Inequalities as Side Conditions," M.S. Thesis, Department of Mathematics, University of Chicago.

Kiefer, J. (1957), "Optimal Sequential Search and Approximation Methods under Minimum Regularity Conditions," *SIAM Journal of Applied Mathematics,* Vol. 5, pp. 105–136.

Kuhn, H. W., and Tucker, A. W. (1951), "Nonlinear Programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,* J. Neyman, Ed., University of California Press, Berkeley, pp. 481–492.

Lemke, C. E. (1968), "On Complementary Pivot Theory," in *Mathematics of the Decision Sciences,* G. B. Dantzig and A. F. Veinott, Eds., American Mathematical Society, Providence, RI.

Levenberg, K. (1944), "A Method for the Solution of Certain Nonlinear Problems in Least Squares," *Quart. Appl. Math.*, Vol. 2, pp. 164–168.

Marqardt, D. W. (1963), "An Algorithm for the Least Squares Estimation of Nonlinear Parameters," *SIAM Journal,* Vol. 11, pp. 431–441.

McCormick, G. P. (1983), *Nonlinear Programming: Theory, Algorithms, and Applications,* John Wiley & Sons, New York, pp. 143–167.

Miller, C. E. (1963), "The Simplex Method for Local Separable Programming," in *Recent Advances in Mathematical Programming,* R. L. Graves and P. Wolfe, Eds., McGraw-Hill, New York.

Moré, J. J., and Wright, S. J. (1993), *Optimization Software Guide,* SIAM, Philadelphia.

Nelder, J. A., and Mead, R. (1964), "A Simplex Method for Function Minimization," *Computer Journal,* Vol. 7, pp. 308–313.

Palacios-Gomez, F., Lasdon, L., and Engquist, M. (1982), "Nonlinear Optimization by Successive Linear Programming," *Management Science,* Vol. 28, No. 10, pp. 1106–1120.

Powell, M. J. D. (1969), "A Method for Nonlinear Constraints in Minimization Problems," in *Optimization,* R. Fletcher, Ed., Academic Press, London.

Powell, M. J. D. (1978), "A Fast Algorithm for Nonlinearly Constrained Optimization Calculation," in *Numerical Analysis,* Lecture Notes in Mathematics, Vol. 630, G. A. Watson, Ed., Springer, Berlin.

Rijckaert, M. J., and Walraven, E. J. C. (1985), "Reflections on Geometric Programming," in *Computational Mathematical Programming,* K. Schittkowski, Ed., Springer, Berlin.

Rockafellar, R. T. (1973), "The Multiplier Method of Hestenes and Powell Applied to Convex Programming," *Journal of Optimization Theory and Applications,* Vol. 12, pp. 555–562.

Rosen, J. B. (1961), "The Gradient Projection Method for Nonlinear Programming Part II: Nonlinear Constraints," *SIAM Journal of Applied Mathematics,* Vol. 9, pp. 514–553.

Rosenbrock, H. H. (1960), "An Automatic Method for Finding the Greatest or Least Value of a Function," *Computer Journal,* Vol. 3, pp. 175–184.

Shanno, D. F. (1970), "Conditioning of Quasi-Newton Methods for Function Minimization," *Mathematics of Computation,* Vol. 24, pp. 647–656.

Shor, N. Z. (1985), *Minimization Methods for Non-differentiable Functions,* Springer, Berlin.

Spendley, W., Hext, G. R., and Himsworth, F. R. (1962), "Sequential Application of Simplex Designs of Optimization and Evolutionary Operations," *Techometrics,* Vol. 4, pp. 441–461.

Stoer, J. (1985), "Principals of Sequential Quadratic Programming Methods for Solving Nonlinear Programs," in *Computational Mathematical Programming,* K. Schittkowski, Ed., Springer, Berlin.

Wilson, R. B. (1963), "A Simplicial Algorithm for Concave Programming," Ph.D. Thesis, Graduate School of Business Administration, Harvard University, Cambridge, MA.

Wolfe, P. (1963), "Methods of Nonlinear Programming," in *Recent Advances in Mathematical Programming,* R. L. Graves and P. Wolfe, Eds., McGraw-Hill, New York.

Zangwill, W. I. (1967), "The Convex Simplex Method," *Management Science,* Vol. 14, pp. 221–283.

Zhang, J., Kim, N., and Lasdon, L. (1985), "An Improved Successive Linear Programming Algorithm," *Management Science,* Vol. 31, No. 10, pp. 1312–1331.

Zoutendijk, G. (1960), *Methods of Feasible Directions,* Elsevier, Amsterdam.

Zowe, J. (1985), "Nondifferentiable Optimization," in *Computational Mathematical Programming,* K. Schittkowski, Ed., Springer, Berlin.

ADDITIONAL READING

Avriel, M., *Nonlinear Programming: Analysis and Methods,* Prentice-Hall, Englewood Cliffs, NJ, 1976.

Bazarra, M. S., Sherali, H. D., and Shetty, C. M., *Nonlinear Programming: Theory & Applications,* John Wiley & Sons, New York, 1994.

Beightler, C. S., Phillips, D. T., and Wilde, D. J., *Foundations of Optimization,* 2nd Ed., Prentice-Hall, Englewood Cliffs, NJ, 1976.

Bertsekas, D. P., *Nonlinear Programming,* 2nd Ed., Athena Scientific, Belmont, MA.

Coleman, T. F., and Li, Y., *Large Scale Numerical Optimization,* SIAM, Philadelphia, 1990.

Dantzig, G. B., *Linear Programming and Extensions,* Princeton University Press, Princeton, NJ, 1963.

Du, D., and Sun, J., *Advances in Optimization and Approximation,* Kluwer, Dordrecht, 1994.

Duffin, R. J., Peterson, E. L., and Zener, C., *Geometric Programming,* John Wiley & Sons, New York.

Ecker, J. G., and Kupferschmid, M., *Introduction to Operations Research,* John Wiley & Sons, New York, 1988.

Fiacco, A. V., and McCormick, G. P., *Nonlinear Programming: Sequential Unconstrained Minimization Techniques,* SIAM, Philadelphia, 1990.

Fletcher, R., *Practical Methods of Optimization,* 2nd Ed., John Wiley & Sons, New York, 1987.

Floudas, C. A., *Deterministic Global Optimization: Theory, Algorithms and Applications,* Kluwer, Dordrecht, 1999.

Gill, P. E., Murray, W., and Wright, M. H., *Practical Optimization,* Academic Press, London, 1981.

Hillier, F. S., and Lieberman, G. J., *Introduction to Operations Research,* 6th Ed., McGraw-Hill, New York, 1995.

Lasdon, L. S., *Optimization Theory for Large Systems,* Macmillan, New York, 1970.

Luenberger, D. G., *Introduction to Linear and Nonlinear Programming,* 2nd Ed., Addison-Wesley, Reading, MA, 1984.

Mangasarian, O. L., *Nonlinear Programming,* McGraw-Hill, New York, 1969.

McCormick, G. P., *Nonlinear Programming: Theory, Algorithms, and Applications,* John Wiley & Sons, New York, 1983.

Minoux, M., *Mathematical Programming: Theory and Algorithms,* John Wiley & Sons, New York, 1986.

Nash, S. G., and Sofer, A., *Linear and Nonlinear Programming,* McGraw-Hill, New York, 1996.

Nocedal, J., and Wright, S. J., *Numerical Optimization,* Springer, New York, 1999.

Shapiro, J. F., *Mathematical Programming: Structures and Algorithms,* John Wiley & Sons, New York, 1979.

Zangwill, W. I., *Nonlinear Programming: A Unified Approach,* Prentice-Hall, Englewood Cliffs, NJ, 1969.

CHAPTER 99

Network Optimization

RICHARD T. WONG
Telcordia Technologies

1. NETWORK MODELS: INTRODUCTION 2568
2. CLASSES OF NETWORK MODELS 2569
  2.1. Minimum Cost Flow Problem 2569
  2.2. Transportation Problem 2570
  2.3. Assignment Problem 2572
  2.4. Shortest Path Problem 2572
  2.5. Longest Path Problem 2572
  2.6. Maximum Flow Problem 2572
  2.7. Additional Models 2573
3. NETWORK MODEL SOLUTION TECHNIQUES 2573
  3.1. Minimum Cost Flow and Transportation Problems 2574
  3.2. Shortest Path Problem 2574
  3.3. Maximum Flow Problem 2574
4. COMPUTER SOFTWARE FOR NETWORK FLOW MODELS 2575
5. ADDITIONAL NETWORK FLOW EXAMPLES 2575
  5.1. An Example of Dynamic Network Flow: A Material-Handling System 2575
  5.2. Personnel Assignment 2576
  5.3. Equipment Replacement 2578
  5.4. Reliability 2579
REFERENCES 2580
ADDITIONAL READING 2581

1. NETWORK MODELS: INTRODUCTION

Network flow models constitute a special class of linear programming problems. Historically, the useful special structure of networks sparked much of the initial interest and enabled the design of special-purpose algorithms that made network models much easier to solve than general linear programming problems. In addition, networks have a number of representational features that enhance their attractiveness for application in areas of industrial engineering such as manufacturing, production, and supply chain management or logistics and distribution systems. Network models can be especially useful in analyzing systems with spatial or temporal relationships. In network models, the nodes can represent entities such as geographic locations or space-time locations and the arcs can represent allowable ways of going from one node to another. One application scenario might be where the goal is to find the shortest travel route (set of roads) between a dispatching depot and a customer in need of a special delivery. The nodes correspond to intersections of roads, and the arcs correspond to sections of roads connecting the intersections. Another example of networks arises in modeling a flexible manufacturing system (Kumar and Kroll 1987) where a node might represent the location of a machine or a storage area. The arcs would correspond to connections between various system locations via a conveyor system or an automated guided vehicle system.

Network models offer a number of desirable characteristics, including:

Flexibility: Network models can accommodate a large number of different situations.

Versatility as a subproblem: Even in situations where linear network flow models do not apply, such as when there are nonlinear elements in the system, network models often arise usefully as subproblems for the more general model.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

Ease of explanation: Since many network models can be depicted in a very easy-to-understand way, network models and their results are generally easy to explain and motivate to nontechnical managers.

Ease of Solution: There are a variety of computer codes (running on computers ranging from mainframes to personal computers) capable of solving large-scale models that might arise in practice; these limited computational requirements also facilitate ‘‘what-if’’ analysis that can be useful in performing sensitivity and scenario analyses on a model.

The remainder of this chapter gives more information concerning the formulation, solution and application of network flow models. Section 2 details some of the principal network models. Sections 3 and 4 discuss solution procedure concepts and give information concerning available software for solving network flow models. Section 5 discusses some examples of network applications.

2. CLASSES OF NETWORK MODELS

All of the network flow models presented in this section can be viewed in terms of a single general model—the minimum cost flow problem. However, there are many important problem subclasses whose structures are particularly suggestive for potential applications and that can be solved with special-purpose solution algorithms. This section gives an overview of various types of network flow models.

2.1. Minimum Cost Flow Problem

The minimum cost flow problem is the most general network flow model. It arises in applications such as the efficient distribution of a single product from a set of source (production) sites to a set of demand sites over a given distribution network. The product may be shipped directly from a source site to a demand site, or it may pass through one or more intermediate points in the distribution network before reaching its final destination.

In the network model, the nodes represent source sites, demand sites, or other (intermediate) nodes that are neither source nor demand sites. The net supply at node i is denoted as b_i (for source sites b_i is positive, for demand sites b_i is negative, and for all other nodes b_i is zero). To simplify our discussion, it is assumed that the total amount produced at the source sites is exactly equal to the total amount required at the demand sites. An arc connecting two nodes represents a transportation link. Associated with each arc (i, j) is an upper bound (arc capacity) u_{ij} limiting the total amount of goods that can flow on it. (In more complex models, there may also be a specific lower bound ij on the total flow on the arc. However, to simplify the discussion in this chapter, it is assumed that all lower bounds are zero.) An uncapacitated arc is one that has no upper bound on the amount of flow it can carry. There is a unit transportation flow cost c_{ij}. The minimum cost flow problem can be represented as the following linear program:

$$\text{Minimize} \sum_{(i,j)} c_{ij}x_{ij} \tag{1}$$

$$\text{Subject to:} \sum_{j} x_{ij} - \sum_{j} x_{ji} = b_i \text{ for all nodes } i \tag{2}$$

$$x_{ij} \leq u_{ij} \tag{3}$$

$$x_{ij} \geq 0 \text{ for all arcs } (i, j) \tag{4}$$

The objective is to find the minimum cost routing of the flows through the network such that all source and demand node constraints (2) and all arc capacity constraints (3) are satisfied.

Figure 1 gives an example of an eight node network that represents a distribution system between factories and retailers. Nodes 1 and 2 are factories whose production capacities (net supplies) are 6 and 9. Nodes 3, 4, and 5 are warehouses (intermediate nodes) whose net supplies are zero. Nodes 6, 7, and 8 correspond to the retailers, whose net supplies are −3, −5, and −7. Each arc is labeled with its unit transportation cost and its upper bound. In this example, there are only arcs between source nodes and intermediate nodes and between intermediate nodes and sink nodes. In a general minimum cost flow model there can be arcs between any two nodes in the network.

In many application situations we may wish to restrict the arc flows to be integer valued. The minimum cost flow problem has the interesting property that if all of the net supplies b_i, and arc upper bounds u_{ij} are integer valued, then whenever there is an optimal solution, there is always an integer-valued optimal solution.

The minimum cost flow problem, besides having many useful applications, also contains a number of special cases that arise in a variety of application scenarios. We will now discuss some of these special cases.

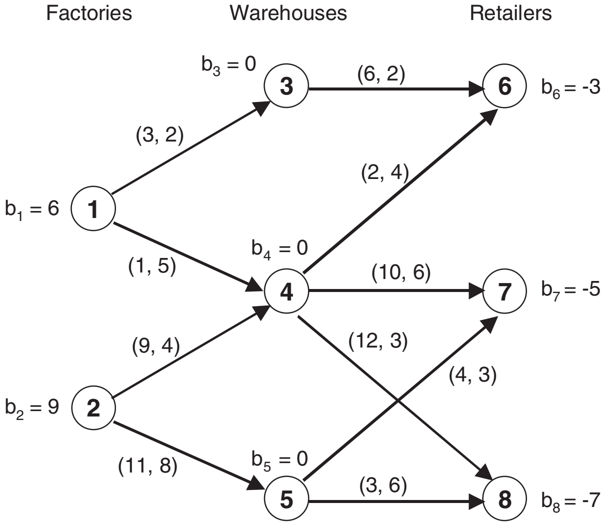


Figure 1 Example of Minimum Cost Flow Problem.

2.2. Transportation Problem

In this subclass of problems there are only source and demand sites (no intermediate nodes are included), and shipments can only be made directly between source and demand sites. The source node i has a supply of s_i and demand node j has demand of d_j (or a net supply of $-d_j$). The transportation problem can be formulated as the following linear programming model:

$$\text{Minimize} \quad \sum_{(i,j)} c_{ij}x_{ij} \tag{5}$$

$$\text{Subject to} \quad \sum_{j} x_{ij} = s_i \text{ for all source nodes } i \tag{6}$$

$$\sum_{i} - x_{ij} = d_j \text{ for all destination nodes } j \tag{7}$$

$$x_{ij} \leq u_{ij} \tag{8}$$

$$x_{ij} \geq 0 \text{ for all arcs } (i,j). \tag{9}$$

Constraints (6) and (7) regulate the net supply for the nodes, and constraint (8) enforces the upper bounds on arc flows. Transportation problems can arise in areas such as logistics inventory control and production planning.

Consider the following example of a production planning problem (Bowman 1956). Production must be scheduled over the next three quarters where the demands are predicted to be 40, 30, and 60, respectively. In each quarter, items may be produced on a regular shift at a cost of $10 per item. During the regular shift there is a capacity restriction limiting production to 45 items per quarter. A maximum of 20 items can also be produced during an overtime shift, at a cost of $12 per item. Finally, items can be held in inventory at a cost of $0.50 per unit per month.

This type of planning problem can be formulated as a transportation model, and the corresponding network is given in Figure 2. Each production shift corresponds to a source node, and the node supply represents the production shift capacity. There are six source nodes since each quarter i can have a regular shift (denoted by node r_i) and an overtime shift (denoted by o_i). For each quarter j there is a demand node j, and finally, there is a dummy demand node DUMMY, which is included to ensure that the total demand equals the total supply. The cost coefficient for arc (g, h) represents the unit cost of supplying demand node h from supply node g. For example, the unit cost of supplying demand node 3 from node o_1, the overtime shift in quarter 1, is ($12.00 + (2 × $0.50) = $13.00). Note that the costs of all arcs incident to the DUMMY node are zero since these flows do not represent any actual production.

This transportation model can also be formulated as the following linear programming problem:

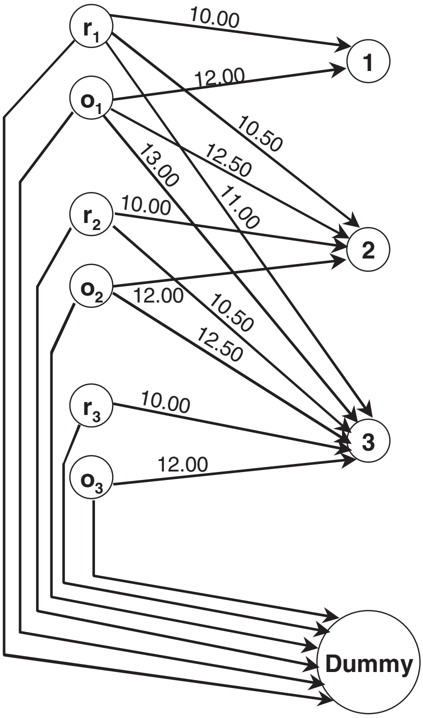


Figure 2 Transportation Problem Representation of Production Planning Example.

$$\begin{aligned}\text{Minimize } & 10.00\,x_{r1,1} + 10.50\,x_{r1,2} + 11.00\,x_{r1,3} + 12.00\,x_{o1,1} + 12.50\,x_{o1,2} \\ & + 13.00\,x_{o1,3} + 10.00\,x_{r2,2} + 10.50\,x_{r2,3} + 12.00\,x_{o2,2} \\ & + 12.50\,x_{o2,3} + 10.00\,x_{r3,3} + 12.00\,x_{o3,3}\end{aligned}$$

$$\begin{aligned}\text{Subject to: } & x_{r1,1} + x_{r1,2} + x_{r1,3} + x_{r1,dummy} && = 45 \\ & x_{o1,1} + x_{o1,2} + x_{o1,3} + x_{o1,dummy} && = 20 \\ & x_{r2,2} + x_{r2,3} + x_{r2,dummy} && = 45 \\ & x_{o2,2} + x_{o2,3} + x_{o2,dummy} && = 20 \\ & x_{r3,3} + x_{r3,dummy} && = 45 \\ & x_{o3,3} + x_{o3,dummy} && = 20 \\ & -x_{r1,1} - x_{o1,1} && = -40 \\ & -x_{r1,2} - x_{r2,2} - x_{o1,2} - x_{o2,2} && = -30 \\ & -x_{r1,3} - x_{r2,3} - x_{r3,3} - x_{o1,3} - x_{o2,3} - x_{o3,3} && = -60 \\ & -x_{r1,dummy} - x_{o1,dummy} - x_{r2,dummy} - x_{o2,dummy} - x_{r3,dummy} - x_{o3,dummy} && = -65 \\ & x_{ij} \geq 0 \text{ for all arcs } (i,j).\end{aligned}$$

Notice that every variable appears in exactly two equations—once with a coefficient of +1 and once with a coefficient of −1. For example, the variable $x_{r1,1}$ appears in the first equation with a +1 coefficient and in the seventh equation with a −1 coefficient. This special property for the variable coefficients holds for any transportation problem and more generally for ANY general minimum cost flow problem. This special structure is also the key to many of the specialized solution approaches for network flow problems.

2.3. Assignment Problem

A special case of the transportation problem is the assignment problem where there are exactly n source nodes that all have a supply of 1 and n demand nodes that all have a demand of 1 (net supply of -1). Thus, each source node must be assigned to a unique demand node. The cost of assigning source node i to demand node j is c_{ij} and all arcs are uncapacitated. The assignment problem can be represented in the following way:

$$\text{minimize} \quad \sum_{(i,j)} c_{ij} x_{ij} \tag{10}$$

$$\text{subject to:} \quad \sum_{j} x_{ij} = 1 \text{ for all sources nodes } i \tag{11}$$

$$\sum_{i} x_{ij} = 1 \text{ for all demand nodes } j \tag{12}$$

$$x_{ij} \geq 0 \text{ for all arcs } (i,j) \tag{13}$$

Constraints (11) and (12) enforce the supply and demand restrictions.

The assignment problem can arise in a manufacturing system when we view the source nodes as jobs and the demand nodes as machines that can perform the jobs. In the next time cycle, each job must be assigned to a unique machine. The coefficient c_{ij} represents the cost of assigning job i to machine j during this time cycle. The optimal solution assigns the jobs to machines so that the total cost during the next time cycle is minimized.

2.4. Shortest Path Problem

Another special case of the uncapacitated minimum cost flow problem is when there is a single source site and a single demand site and all arcs are uncapacitated. The problem reduces to finding the shortest (minimum cost) path from the source node s to the sink node (demand site) t. As discussed in Section 1, one application of this model might be where we wish to find the best travel route for a truck traveling between a depot and a special customer. Another important application is the routing of information packets in the Internet. A shortest path-routing model is used to determine the path traversed by information packets sent over the Internet (Sackett and Metz 1997). The shortest path problem can be formulated as the following linear programming model:

$$\text{minimize} \quad \sum_{(i,j)} c_{ij} x_{ij} \tag{14}$$

$$\text{subject to:} \quad \sum_{j} x_{ij} - \sum_{j} x_{ji} = \begin{cases} 1 & i = s \\ -1 & i = t \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

$$x_{ij} \geq 0 \text{ for all arcs } (i,j) \tag{16}$$

The objective function minimizes the travel costs and constraint (14) ensures that there is a path from node s to node t.

2.5. Longest Path Problem

The longest path problem is identical to the shortest path problem except that the goal is to find the longest path from a source node to a sink node. An important application of this problem is in the area of project scheduling (Elmaghraby 1977), where there are precedence constraints that specify some tasks cannot be completed until other tasks are done. Each arc represents a task and its length is the task duration. Each node indicates a precedence relationship. The tasks corresponding to outgoing nodes cannot be started until all of the tasks corresponding to the incoming nodes are completed. The source and sink nodes represent the overall start and completion of the entire project. The length of the longest path between the source and sink nodes represents the time required to complete the entire project even if unlimited resources are available to perform tasks that could be performed in parallel. The precedence constraints would generally restrict what tasks could be performed in parallel. The arcs on the longest path indicate the "bottleneck" tasks whose reduction in task duration would result in an overall reduction of the project duration.

The linear programming formulation of the longest path problem is identical to that of the shortest path problem except that the objective function must be maximized instead of minimized.

2.6. Maximum Flow Problem

The typical maximum flow problem has a network where there are arc upper bounds and a designated source node s and a designated sink node t. The objective is to find the flow pattern that maximizes the total flow v from the source to the sink node. The maximum flow problem can be formulated as the following linear programming model:

$$\text{minimize} \quad v \tag{17}$$

$$\text{subject to:} \quad \sum_j x_{ij} - \sum_j x_{ji} = \begin{cases} v & i = s \\ -v & i = t \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

$$x_{ij} \leq u_{ij} \tag{19}$$

$$x_{ij} \geq 0 \text{ for all arcs } (i,j) \tag{20}$$

One application of this model is in the area of emergency evacuation of a facility (such as a building or subway station). The source node represents the location of workers in the facility and the sink node represents a safety area. The arcs can correspond to the various links from one part of the facility to another (stairways, corridors, etc.) and the arc capacity indicates the maximum number of people who could traverse a link per unit time. The maximum flow represents the maximum rate at which people could be evacuated from the facility (see Chalmet et al. 1982 for a more elaborate model).

2.7. Additional Models

There are a number of more advanced models that are beyond the scope of this chapter. For example, there are many network flow models that have additional side constraints. The general linear constraints could correspond to tarriff constraints or other restrictions that cannot be directly incorporated into the basic network flow model. There are some advanced network-based solution procedures that can accommodate linear side constraints (Glover and Klingman 1985). These procedures have been incorporated into available computer software. Another useful generalization of the minimum cost flow problem is when there is more than one type of flow in the network. These situations arise frequently in applications where there may be more than one type of product that is being processed by a distribution network or more than one type of material that is flowing through a materials handling facility or supply chain system. Multicommodity flow models, although they can model a wide range of situations, are usually considerably more complicated to solve than comparably sized (single-commodity) network flow models (McBride 1998; Ali et al. 1984).

Other network models incorporate features such as fixed setup costs or nonlinear flow costs. For example, in a production scheduling environment, the amount of flow on an arc could correspond to the number of units assigned to be processed on a machine. There could be a setup cost incurred if there is a nonzero number of units scheduled on the machine. In a transportation network, there could be a fixed construction cost for opening up a transportation link and a variable operating cost that would depend on the traffic flow through the link. The usual modeling procedure for setup and fixed costs is to introduce 0–1 (binary integer) variables and to create a mixed integer programming model.

Another advanced mixed integer programming network model is the traveling salesman problem, which determines the minimum cost or length tour that passes through each node of the network exactly once. Note that such a tour can also be viewed as the sequence of nodes to visit. From this perspective, the traveling salesman problem has applications in machine scheduling where the sequencing of a set of jobs to be run on a machine must be determined. The arc cost for arc (i,j), c_{ij}, corresponds to the setup or changeover cost on the machine from job i to job j. The tour can also be viewed as the travel route of a delivery vehicle that starts out at a depot and then must visit every other node to make deliveries and then return to the depot. In this context, the cost c_{ij} represents the travel cost of going from the customer at node i to the customer at node j. Finally, the traveling salesman problem also arises in the manufacture of printed circuit boards with a robotic assembly unit (Chan and Mercier 1989). There is a given set of locations at which components must be inserted. The term c_{ij} now corresponds to the cost of inserting the component at location j immediately after inserting the component at location i.

Nonlinear costs could arise when the production cost function exhibits economies of scale and is concave. Another example of nonlinear flow costs is in the area of transportation and distribution planning. Models for analyzing the traffic congestion on a road network usually have arc travel costs that are convex functions of the arc flow in order to represent congestion effects on the network. These models also generally have multiple commodities to represent the various origin–destination characteristics of the network users.

3. NETWORK MODEL SOLUTION TECHNIQUES

The current solution technology for network models is quite powerful and is capable of solving large-scale models that can arise in practice. Also, the continually increasing computational power of computers guarantees that the size of network models that can be solved will continue to grow. Specialized solution procedures are available on a wide spectrum of machines, ranging from personal computers and workstations to mainframe machines. Some tests have indicated that specialized network codes can be one to two orders of magnitude faster than general linear programming codes in solving network flow models. This section gives a brief overview of some basic concepts for network flow problem-solution procedures. The next section discusses some available computer software for solving network flow models.

Over the past 50 years, there have been several periods of evolutionary development in the field of network algorithms. In the 1950s and 1960s, many classical results and algorithms were developed (Ford and Fulkerson 1962). In the 1970s, many researchers began to recognize and emphasize the role of computer data structures in the efficient implementation of algorithms. With these new and powerful implementations, a number of large-scale applications of network models were successfully completed.

Starting in the 1970s and continuing into the 1980s and 1990s, researchers began to design and evaluate algorithms according to a worst-case computation time criterion (Ahuja et al 1993). That is, the speed of a solution procedure was evaluated according to the maximum number of solution procedure steps required as a function of the problem size. For example, a shortest path algorithm A with worst-case computation time of $4n^2$ is a procedure that when applied to a problem with n nodes requires no more than $4n^2$ steps. It is widely believed that the rate of worst-case computation growth (for algorithm A, the n^2 term) is more important in evaluating algorithms than the constant (for algorithm A, the factor of 4). So the worst-case computation time of algorithm A is usually described as $O(n^2)$ (pronounced "order n squared"). That is, the worst-case computation time grows as the square of the number of nodes in the network and the exact rate of growth with the constant is not specified. Although the worst-case criterion is mainly a theoretical measure, there is some empirical evidence that indicates that some of the recently developed network solution methods with good worst-case computational performance also perform well relative to other solution procedures.

The following subsections highlight some useful concepts for solving various classes of network flow problems.

3.1. Minimum Cost Flow and Transportation Problems

One of the most effective algorithms for these classes of network problems has been a specialized implementation of the simplex algorithm for linear programming. This type of approach uses special data structures to exploit the special properties of the network models and accelerate the steps of the simplex algorithm. For example, for these network flow models (and all of the other related subclasses discussed in this chapter), the set of basic variables corresponds to a set of arcs that form a spanning tree for the underlying network. Computing such items as the current values for the dual multipliers is easily done with a specialized procedure that exploits the basis tree structure (Ahuja et al. 1993).

3.2. Shortest Path Problem

Dynamic programming and a specialized algorithm due to Dijkstra (Ahuja et al. 1993) are the two most widely used shortest path procedures. Dijkstra's algorithm iteratively labels the nodes with their shortest path distance from the origin node. All nodes are initially unlabeled except for the origin node, which has a label of 0. At each iteration the algorithm labels the unlabeled node that has the minimum shortest path distance from the origin node. The procedure terminates when the destination node is labeled. Dijkstra's algorithm is an $O(n^2)$ algorithm, where n is the number of nodes in the network.

The fastest shortest path codes use either dynamic programming or Dijkstra's algorithm in conjunction with sophisticated data structures that reduce the amount of time spent searching for required problem information.

3.3. Maximum Flow Problem

For the maximum flow problem, the classical solution concept has been the augmenting path. Given a feasible flow pattern, an augmenting path procedure then attempts to find a path from the source node to the sink node such that we can increase the flow along this path and thus increase the total flow from the source node to the sink node. The general augmenting path algorithm iteratively performs a series of path augmentations where each augmentation increases the flow on the selected path as much as possible subject to the arc capacity constraints.

The example in Figure 3*a* shows how a judicious selection of augmenting paths can accelerate the performance of the augmenting path algorithm. In the example, the arc labels are the arc capacities. Nodes s and t are the source and sink nodes, respectively. The initial solution starts with zero flow on all of the arcs. The procedure starts by augmenting the flow on path s–1–2–t by one unit to obtain the flow pattern in Figure 3*b*. Next the procedure augments the flow on path s–2–1–t by one unit to obtain the flow pattern in Figure 3*c*. Iteratively augmenting the flows on these two paths would require a total of 1000 path augmentations. Notice that the optimal solution can be reached

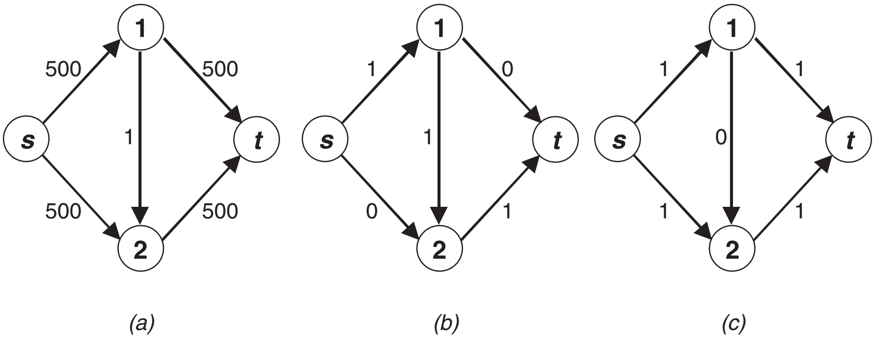


Figure 3 (*a*) Four-Node Example of Maximum Flow Problem (*b*) Example Flow Pattern after One Flow Augmentation (*c*) Example Flow Pattern after Two Flow Augmentations.

from the initial solution in just two augmentations (augment paths *s*–1–*t* and *s*–2–*t* by 500 units each)! Research in the 1970s highlighted the value of judiciously selecting the augmenting paths used and demonstrated that the augmenting path with the minimum number of arcs, the shortest augmenting path, should be chosen (in our example, the paths *s*–1–*t* and *s*–2–*t* would be selected before the path *s*-1-2-*t* since they have fewer arcs). Thus, the maximum flow problem can be approached by solving a series of shortest path problems. This procedure can be improved by saving information about the shortest path computation from one iteration to the next in order to reduce the work required to solve each shortest path problem (Goldfarb and Grigoriadis 1988).

4. COMPUTER SOFTWARE FOR NETWORK FLOW MODELS

A considerable amount of computer software is available for solving network flow problems. During the latter part of the 1990s, the use of the World Wide Web on the Internet has greatly facilitated the dissemination of information and software for solving network flow models.

For personal computers, a number of packages are available which have modules for solving various classes of network flow problems (see Oberstone 1990, chaps. 7, 8 for additional information). The CPLEX (www.cplex.com), OSL (www.6.software.ibm.com/es/oslvZ/features/lib.htm), SAS/OR (www.sas.com), and LINDO (www.lindo.com) commercial systems have network flow modules and run on machines ranging from mainframes to workstations and personal computers. Also, Michael Trick (mat.gsia.cmu.edu/companies.html) has assembled a set of World Wide Web links to a variety of companies that offer OR software including network optimization routines.

There are also a number of codes available for solving network flow models. See Kennington and Helgason (1980), Simeone et al. (1988), and Bertsekas (1991), which all give listings of some network flow codes. More and Wright (1993) contains some discussion of network optimization and available software routines. Also, Michael Trick (see mat.gsia.cmu.edu/resource.html, under Software Packages and Descriptions) has assembled a set of World Wide Web links to a variety of OR software packages (including network optimization routines) that are available for little or no cost.

The field of project management, which includes applications of the longest path problem, has spawned a number of software packages for both commercial and educational use. See Oberstone (1990, chaps. 13, 16) and Wasil and Assad (1988) for a discussion and comparison of some of these packages. It has been estimated that at one time there were over 200 different products on the market (see Scheduling Software, www.buildersnet.org/cpmtutor/html/schedulingsoftware.html). See Microsoft Project (www.microsoft.com/office/project/default.htm) and Job Boss (engineering.software-directory.com/software-2.cdprod1/009/683.Job.Boss.sthml) for two examples of such systems.

5. ADDITIONAL NETWORK FLOW EXAMPLES

This section gives further examples of applications that can be cast as network flow models. These examples also help illustrate the types of situations that can be formulated as network flow problems. In many other applications, the network model structure is often hidden and can require considerable ingenuity to identify and formulate.

5.1. An Example of Dynamic Network Flow: A Material-Handling System

Network flow models can represent temporal as well as spatial relationships. Consider the four-node dynamic maximum flow network example depicted in Figure 4. For each arc the two labels represent

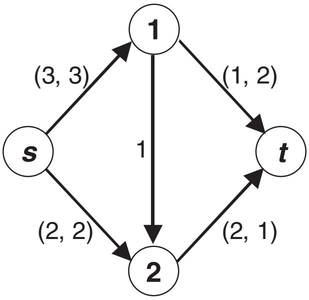


Figure 4 Four-Node Example of Dynamic Network Flow Model.

the arc capacity and the time required to traverse the arc. The dynamic network model can be expanded and represented by an equivalent static model as depicted in Figure 5. In this expanded network, each node represents a specific location at a particular point in time. Such a network could be used, for example, to compute the maximum flow possible between the source and sink node within a given interval of time. By adding arc costs and node demands, the network model could be reformulated as a minimum cost flow problem.

This type of dynamic network flow model can be used to model certain types of material-handling systems (Maxwell and Wilson 1981). The necessary modeling assumptions include that both time and space can be represented by discrete points and that there be only a single type of material which flows through the system. These dynamic network flow models could be useful in evaluating and optimizing preliminary material-handling system configurations and analyzing where system bottlenecks could occur due to storage or processing rate limitations. Maxwell and Wilson (1981) also discuss the role of other methodologies such as simulation (see Chapters 93–96 of this Handbook) in the overall design and evaluation of material-handling systems.

5.2. Personnel Assignment

The assignment of personnel to jobs is an often-cited application area for network flow models. Consider the classical assignment problem where the model represents the assignment of *n* people to *n* jobs. There is one source node for each person available and one demand node for each job. One possible objective function would be to assign people to jobs in order to minimize the relocation costs.

More elaborate network flow models for the personnel assignment problem are also possible (Liang and Thompson 1987). Suppose there are more people available then there are jobs but each

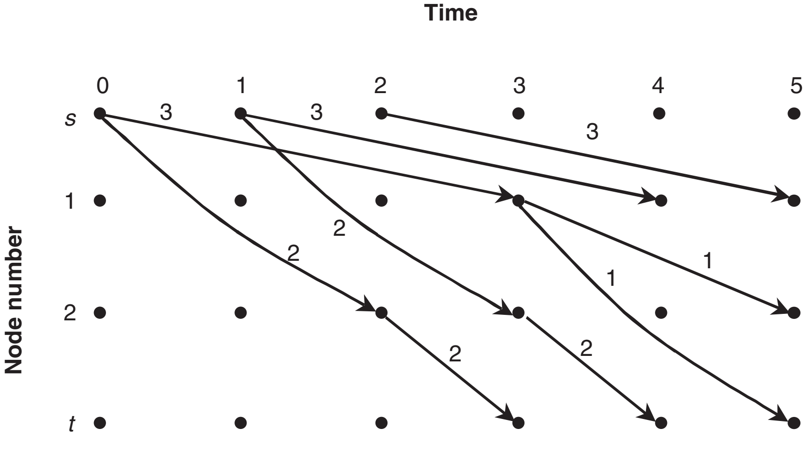


Figure 5 Static Equivalent of Four-Node Dynamic Network Flow Model.

job can be filled by only one person. In particular, consider an example where six people and four different jobs are available. This type of personnel assignment problem can be represented by the minimum cost network flow model in Figure 6. Every node in the layer of nodes (P_1, P_2, . . . , P_6) represents an available person and has a supply of 1. Node t has a net supply of -6. Every node in the layer of nodes (J_1, J_2, J_3, J_4) corresponds to an available job and along with node s are all intermediate nodes whose net supply is zero. All arcs in the network have a capacity of one. In the network flow problem solution, person P_i is assigned to job J_j if there is a flow of one unit on the arcs (P_i, J_j) and (J_j, t). Since arc (J_j, t) has a flow capacity of one, the total flow into node J_j can be at most one and so job j can be assigned to at most one person. Person P_i is not assigned to a job if there is a flow on the arcs (P_i, s) and (s, t). Note that for each personnel node P_i, all of the possible arcs to the job nodes are not included, since a person may only be qualified for a subset of the available jobs.

This type of network flow model can be further expanded by adding another layer of nodes, as depicted in Figure 7. Every job is associated with a particular branch location. In this example there are two branch locations represented by the nodes B_1 and B_2. The total flow between a location node and node t is the total number of jobs at that location that are filled. This information can be useful in enforcing preferences or restrictions concerning the staffing levels at various locations. The staffing preferences can be incorporated by attaching the appropriate arc costs to the arcs incident to node t. As for the restrictions, suppose that company policy specifies a restriction that at most k new jobs at location B_1 will be filled. The model can incorporate this constraint by setting the arc flow capacity on arc (B_1, t) to be k.

This additional layer of nodes illustrates how other preferences or constraints can be added to a network flow model without destroying the special network flow structure. Such modeling techniques are quite useful since network models can be solved much more efficiently than general linear programs.

There is another personnel assignment question for which a network model may be useful. There is a set of jobs and a set of people where for each person there is a list of the jobs that the person is qualified for. The objective is to Maximize the total number of jobs that can be filled with qualified people. This question can be modeled as a maximum flow problem, as shown in Figure 8. All arcs in the network have a capacity of one. The objective function is to maximize the flow between nodes s and t. An arc connecting nodes P_i and J_j indicates that person i is qualified for job j. In the optimal flow pattern, person i is assigned to job j if the flows on arcs (P_i, J_j), (s, P_i) and (J_j, t) are all one. Since every arc (J_j, t) has a flow capacity of one, the total flow into node J_j can be at most one and so job j can be assigned to at most one person. For this model, there are no direct considerations of cost but instead the objective is to maximize the number of jobs filled by qualified people.

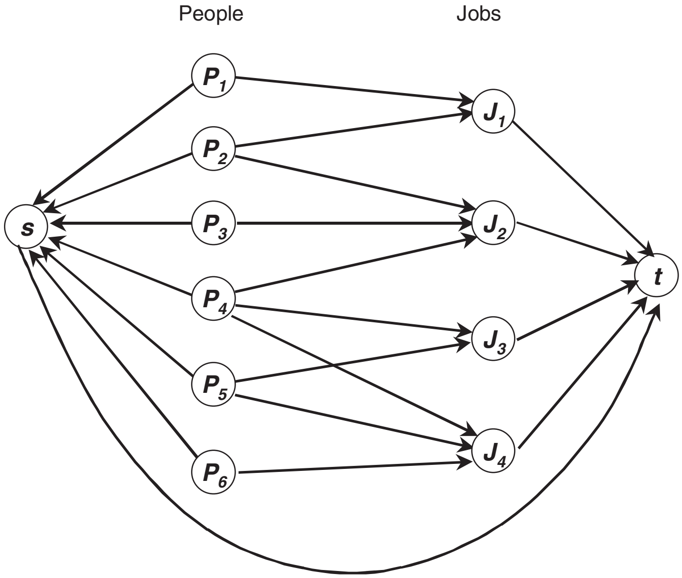


Figure 6 Minimum Cost Network Flow Model of Personnel Assignment Example.

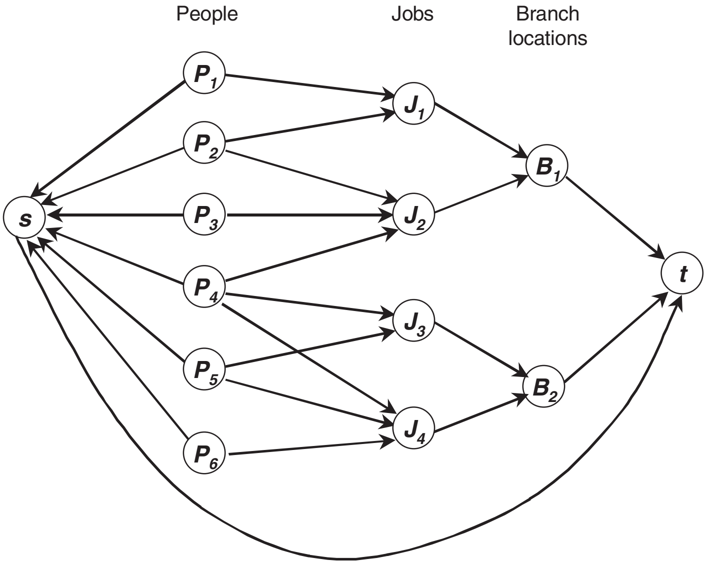

People
Jobs
Branch locations
P_1
P_2
P_3
P_4
P_5
P_6
J_1
J_2
J_3
J_4
B_1
B_2
s
t

Figure 7 Expanded Minimum Cost Network Flow Model of Personnel Assignment Example.

5.3. Equipment Replacement

Network models can also be useful for the decisions related to replacing equipment over a fixed time horizon. For example, a manufacturing center has just installed a new $20,000 cutting machine at the start of year 1. At the start of each of the next four years, the firm has the option to continue operating the current machine for an additional year or to trade in the current machine for a new machine. The operating costs for the machine grow with the age of the machine, while the trade-in value of a machine decreases with age. The operating costs and trade in values are given below:

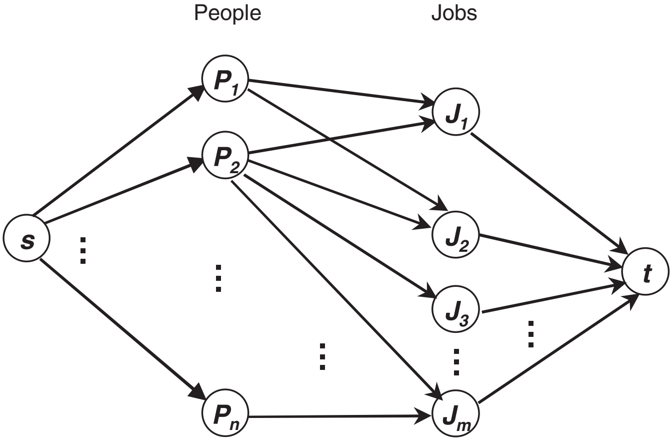

People
Jobs
P_1
P_2
P_n
J_1
J_2
J_3
J_m
s
t

Figure 8 Maximum Flow Model of Alternative Version of the Personnel Assignment Example.

Number of Years Used	Salvage Value	Operating Cost During Last Year
1	$15,000	$1,000
2	$10,000	$2,000
3	$8,000	$4,000
4	$7,000	$8,000

The firm must decide the optimal equipment replacement strategy over the next four years. Figure 9 gives a shortest path network model for determining the optimal replacement strategy. Each of the first five layers of nodes represents the start of a new year. The last layer, which consists of the sink node *t*, represents the end of the planning horizon. The nodes within each layer represent the status (the age or the number of years the machine has been in use) of the current machine at the start of the year. Each arc represents a decision that can be made at the start of a year. For example, the arc from node Year 3–Age 2 to node Year 4–Age 1 corresponds to having a two-year-old cutting machine at the start of year 3 and then trading it in to buy a new machine so that at the start of year 4 the current cutting machine is one year old. The arc cost of $11,000 corresponds to the cost of buying the new cutting machine ($20,000) minus the trade in value of the two-year-old machine ($10,000) plus the operating cost during year 3 ($1,000). The arcs incident to node *t* correspond to the final action of trading in the current machine at the end of the planning horizon. The negative arc cost corresponds to the trade-in (salvage) value of the cutting machine. The arcs on the shortest path from nodes *s* to node *t* will represent the least-cost machine-replacement strategy over the four-year time horizon.

5.4. Reliability

In many engineering systems for areas such as telecommunications and transportation, the issue of system reliability can arise. A telecommunications system can be modeled as a network where the arcs correspond to communication links and the nodes correspond to switching machines. Figure 10 gives an example of such a representation. The system is designed to send messages from node *s* to node *t* where failures may remove an arc (communication link) from service. Assume that the nodes (switching machines) do not fail (or are much less likely to fail than the links). One question of interest is to compute the number of arc disjoint paths between nodes *s* and *t*. The answer provides some measure of the redundancy of the network and its resilience to communication link failures. This question can be answered with a maximum flow network model where each arc has a capacity

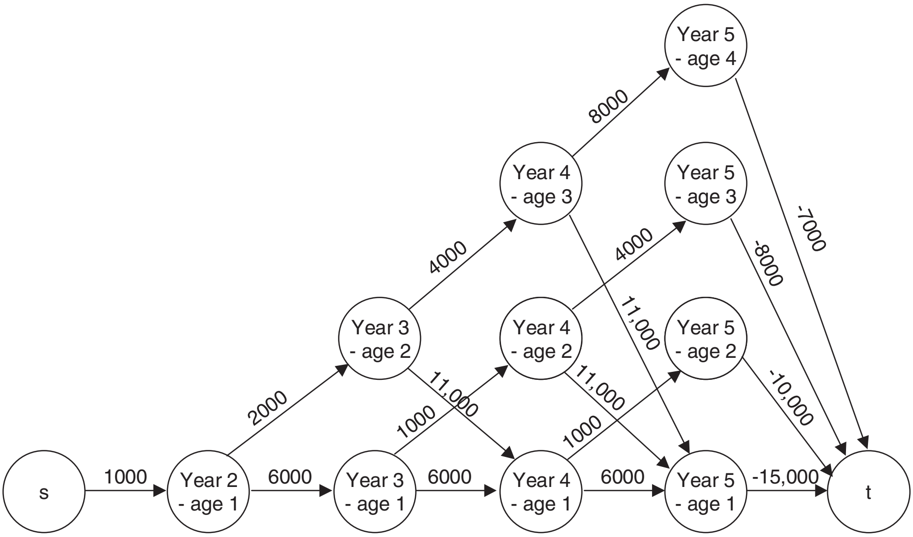


Figure 9 Shortest Path Model of Equipment Replacement Example.

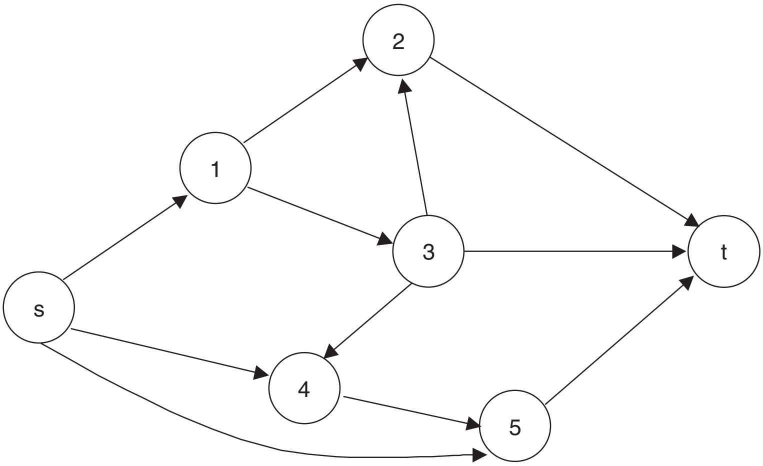


Figure 10 Network Reliability Example.

of one and the maximum flow between *s* and *t* is the number of arc disjoint paths between *s* and *t*. Since each arc has a capacity of one, every arc disjoint path between *s* and *t* can carry one unit of flow and every arc can only belong to at most one flow carrying path, so the maximum flow between these two nodes will equal the (maximum) number of arc disjoint paths between *s* and *t*.

Acknowledgement

Much of the work for the original version of this chapter for the second edition of this volume was performed while the author was at Purdue University. Much of the work done in revising this chapter for the third edition of this volume was performed while the author was at AT&T Labs, Middletown, NJ.

REFERENCES

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993), *Network Flows: Theory, Algorithms, and Applications,* Prentice Hall, Upper Saddle River, NJ.

Ali, I., Barnett, D., Farhangian, K., Kennington, J., Patty, B. Shetty, B., McCarl, B., and Wong, P. (1984), “Multicommodity Network Problems: Applications and Computations,” *AIIE Trans.*, Vol. 16, pp. 127–134.

Bertsekas, D. P. (1991), *Linear Network Optimization: Algorithms and Codes,* MIT Press, Cambridge, MA.

Bowman, E. H. (1956), “Production Scheduling by the Transportation Method of Linear Programming,” *Operations Research,* Vol. 3, pp. 100–103.

Chalmet, L. G., Francis, R. L., and Saunders, P. B. (1982), “Network Models for Building Evacuation,” *Management Science,* Vol. 28, pp. 86–105.

Chan, D., and Mercier, D. (1989), “IC Insertion: An Application of the Travelling Salesman Problem,” *International Journal of Production Research,* Vol. 27, pp. 1837–1841.

Elmaghraby, S. E. (1977), *Activity Networks,* Wiley-Interscience, New York.

Ford, L. R., and Fulkerson, D. R. (1962), *Flows in Networks,* Princeton University Press, Princeton, NJ.

Glover, F., and Klingman, D. (1985), “Basis Exchange Characterizations for the Simplex SON Algorithm for LP/Embedded Networks,” *Mathematical Programming Study,* Vol. 24, pp. 141–157.

Goldfarb, D., and Grigoriadis, M. D. (1988), “A Comparison of the Dinic and Network Simplex Methods for Maximum Flow,” in *Fortran Codes for Network Optimization,* B. Simeone, P. Toth, G. Gallo, F. Maffioli, and S. Pallantino, Eds., J. C. Baltzer, Basel also published as Annals of Operations Research, Vol. 13, pp. 83–123.

Hillier, F. S., and Lieberman, G. J. (1980), *Introduction to Operations Research,* Holden-Day, San Francisco.

Kennington, J. L., and Helgason, R. L. (1980), *Algorithms for Network Programming,* Wiley-Interscience, New York.

Kumar, K. R. and Kroll, D. E. (1987), "Dynamic Network Modeling of an FMS," in *Modern Production Management Systems,* A. Kusiak, Ed., North-Holland, Amsterdam, pp. 19–30.

Liang, T. J., and Thompson, T. J. (1987), "A Large-Scale Personnel Assignment for the Navy," *Decision Sciences,* Vol. 18, pp. 234-249.

Maxwell, W. L., and Wilson, R. C. (1981), "Dynamic Network Flow Modeling of Fixed Path Material Handling Systems," *AIIE Transactions,* Vol. 13, pp. 12–21.

McBride, R. D. (1998), "Advances in Solving the Multicommodity-Flow Problem," *Interfaces,* Vol. 28, pp. 32–41.

More, J. J., and Wright, S. J. (1993), *Optimization Software Guide,* SIAM, Philadelphia.

Oberstone, J. (1990), *Management Science: Concepts, Insights and Applications,* West, St. Paul, MN.

Sackett, G. C., and Metz, C. (1997), *ATM and Multiprotocol Networking* (*Computer Communications*), McGraw-Hill, New York.

Simeone, B., Toth, P., Gallo, G., Maffioli, F., and Pallantino, S., Eds. (1988), *Fortran Codes for Network Optimization,* Annals of Operations Research, Vol. 13.

Syslo, M. M., Deo, N., and Kawalik, J. S. (1983), *Discrete Optimization Algorithms,* Prentice Hall, Englewood Cliffs, NJ.

Wasil, E. A., and Assad, A. A. (1988), "Project Management on the PC: Software, Applications and Trends," *Interfaces,* Vol. 18, pp. 75–84.

ADDITIONAL READING

Bradley, S. P., Hax, A. C., and Magnanti, T. L., *Applied Mathematical Programming,* Addison-Wesley, Reading, MA, 1977.

Ozan, T. M., *Applied Mathematical Programming for Production and Engineering Management,* Prentice-Hall, Englewood Cliffs, NJ, 1986.

Phillips, D. T., and Garcia-Diaz, A., *Fundamentals of Network Analysis,* Prentice-Hall, Englewood Cliffs, NJ, 1981.

CHAPTER 100

Discrete Optimization

RONALD L. RARDIN
Purdue University

1. MODELING **2582**
2. SOLUTIONS **2583**
3. TOTAL ENUMERATION **2584**
4. RELAXATION **2584**
  4.1. Linear Programming Relaxations 2585
  4.2. Lagrangean Relaxations 2587
5. HEURISTIC SEARCH **2589**
  5.1. Constructive or Solution-Building Search 2589
  5.2. Improving or Solution-Enhancing Search 2590
  5.3. Local Optima 2590
  5.4. Tabu, Simulated Annealing, and Genetic Algorithms 2590
6. BACKTRACKING SEARCH AND BRANCH AND BOUND **2591**
  6.1. Tree Representation 2591
  6.2. Branch and Bound 2592
7. GUIDELINES AND LIMITS **2593**
  7.1. Computational Complexity Theory 2594
  7.2. Choosing a Strategy 2595
APPENDIX **2596**
REFERENCES **2600**

Optimization is the process of selecting a solution from among available decision alternatives so that it conforms to all problem constraints and (at least approximately) maximizes or minimizes one or more objective/criterion functions. *Discrete optimization* is the branch confronting the vast array of problems having decisions of a logical or countable nature. Instead of, say, selecting an operating temperature, which is a decision that can pick any value in a *continuous* interval, discrete decisions are those with only specified list of options: turn left or turn right, build or do not build a plant, undertake job *A* before job *B* or *B* before *A*. If all decisions of a problem are discrete, the problem is termed *pure;* otherwise (e.g., if decisions include both whether to setup a process and what temperature to operate it at), the problem is *mixed*. The Appendix to this chapter presents many specific discrete optimization models. Other names for discrete optimization are *combinatorial optimization, integer programming,* and *mixed-integer programming*. A good introduction is provided in Rardin (1998, chaps. 9–12). More advanced books include Wolsey (1988), Parker and Rardin (1988), and Nemhauser and Wolsey (1988). Other standard sources are Schrijver (1986), Papadimitriou and Steiglitz (1982), and Lawler (1976).

1. MODELING

Modeling is the process of mathematically representing a problem in a form conducive to analysis and solution. The logical nature of discrete optimization—especially pure discrete optimization—invites a variety of quite different representations. Many problems can usefully be modeled in terms of logical predicates, objects and sets, graphs, or numerous other constructs.

The format that has proved most useful, and the focus of this chapter, is *numerical representation*—formulation of the problem in terms of numerically valued decision variables. Discrete decision options are encoded as specific numerical variable values. For example, variable $y_j = 1$ may mean plant j is selected for construction and $y_j = 0$ that it is not.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

TABLE 1 Representing Discrete Phenomena with Numerical Variables

Phenomena	Representation
At least K decisions j of subset J must be taken.	$y_j \triangleq 1$ if j is in the solution, 0 otherwise $\Sigma_{j \in J}\, y_j \geq K$
Exactly K decisions j of subset J must be taken.	$y_j \triangleq 1$ if j is in the solution, 0 otherwise $\Sigma_{j \in J}\, y_j = K$
At most K decisions j of subset J can be taken.	$y_j \triangleq 1$ if j is in the solution, 0 otherwise $\Sigma_{j \in J}\, y_j \leq K$
At most K variables z_j with j in subset J can be positive in a solution.	$y_j \triangleq 1$ if $z_j > 0$, 0 otherwise $\Sigma_{j \in J}\, y_j \leq K$ $z_j \leq M\, y_j$ for all $j \in J$
Decision j is allowed only if all decisions $i_1, \ldots, i_n$ are taken.	$y_{ik} \triangleq 1$ if decision i_k is taken, 0 otherwise $y_j \triangleq 1$ if decision j is taken, 0 otherwise $y_j \leq y_{i_k}$ for all $k = 1, \ldots, n$
Decision j is allowed only if some decision $i_1, \ldots, i_n$ is taken.	$y_{i_k} \triangleq 1$ if decision i_k is taken, 0 otherwise $y_j \triangleq 1$ if decision j is taken, 0 otherwise $y_j \leq \Sigma_{k=1}^{n}\, y_{i_k}$
Decision j is implied if all decisions $i_1, \ldots, i_n$ are taken.	$y_{i_k} \triangleq 1$ if decision i_k is taken, 0 otherwise $y_j \triangleq 1$ if decision j is taken, 0 otherwise $\Sigma_{k=1}^{n}\, y_{i_k} \leq (n - 1) + y_j$
Decision j is implied if any decision $i_1, \ldots, i_n$ is taken.	$y_{i_k} \triangleq 1$ if decision i_k is taken, 0 otherwise $y_j \triangleq 1$ if decision j is taken, 0 otherwise $y_{i_k} \leq y_j$ for all $k = 1, \ldots, n$
Nonnegative fixed cost F_j is incurred whenever variable z_j is positive.	$y_j \triangleq 1$ if z_j positive, 0 otherwise $\min \ldots + F_j y_j + \ldots$ $z_j \leq M\, y_j$
If decision i is taken and decision j is taken, cost C_{ij} is incurred.	$y_i \triangleq 1$ if decision i is taken, 0 otherwise $y_j \triangleq 1$ if decision j is taken, 0 otherwise $\min \ldots + C_{ij} y_i y_j + \ldots$
Either task i of duration T_i done before task j of duration T_j or vice versa.	$\mathrm{z}_i \triangleq$ task i start time; $z_j \triangleq$ task j start time $y_{ij} \triangleq 1$ if task i before j, 0 otherwise $0 \leq z_i + T_i \leq z_j - M\,(1 - y_{ij})$ $0 \leq z_j + T_j \leq z_i - M\, y_{ij}$
Function $\theta(p) = \Theta_1, \ldots, \Theta_n$ at $p = P_1 < \ldots < P_n$, with linear interpolation between P_j's.	$y_j \triangleq 1$ if j left interpolation point, 0 otherwise $\theta \triangleq$ interpolated value $z_j \triangleq$ weight on point j $p = \Sigma_{j=1}^{n}\, P_j z_j;\ \theta = \Sigma_{j=1}^{n}\, \Theta_j z_j$ $\Sigma_{j=1}^{n-1}\, y_j = 1;\ \Sigma_{j=1}^{n}\, z_j = 1$ $0 \leq z_1 \leq y_1;\ 0 \leq z_n \leq y_{n-1}$ $0 \leq z_j \leq y_j + y_{j-1}$ for all $j = 2, \ldots, n - 1$
Quantity q is a nonnegative integer variable with value less than or equal to positive integer U.	$y_j \triangleq 1$ if the 2^j bit in the binary representation of q is "on," 0 otherwise $N \triangleq \lfloor \log_2 U \rfloor$ $q = \Sigma_{j=0}^{N}\, 2^j y_j$ $q \leq U$

Although numerical modeling of discrete problems requires more abstraction than some other approaches, it offers a host of advantages. First, objective and constraint functions involving weighted sums and the like are easily expressed in terms of numerically valued variables. Second, methods and encodings evolved for pure problems extend naturally to mixed cases. Most important, however, is the fact that the available methods for continuous optimization are generally more effective than those for discrete problems (see Chapters 97 and 98). Embedding a discrete problem in a continuous environment makes it possible to exploit continuous approximations in the analysis.

Table 1 shows the standard modeling of a variety of discrete notions in terms of numerical decision variables. Throughout, y denotes discrete variables restricted to 0 and 1 values, z represents continuous variables, and M is a large positive constant.

2. SOLUTIONS

In optimization, a solution (choice of values for decision variables) is termed *feasible* if it satisfies all problem constraints and *optimal* if it is feasible and as good as any other feasible solution in objective function value. Models that have no feasible solutions are *infeasible*.

The goal of all discrete optimization analysis is to find feasible solutions with good objective function values. It is rarely important to know a mathematically optimal solution to a model, since the model is itself only an approximation to the underlying problem. However, it is desirable to have a sharp bound on the objective function value that might be obtained by any feasible solution. Then, for example, if a feasible solution to a maximize problem is known with objective function value \$92, and other analysis establishes that no feasible solution can produce better that \$100, one can accept the \$92 solution with confidence that it is no more $(100 - 92)/100 = 8\%$ suboptimal. The attraction of mathematically optimal solutions is that they provide good feasible solutions with zero-error bounds.

Many search methods for discrete optimization move through a sequence of *partial solutions* that assign specific values to some decision variables in a model, leaving others *free* or undetermined. A *completion* of a partial solution is an assignment of specific values to any remaining free variables. Of course, the definition of a partial solution includes the possibility that there are no free variables, in which case the solution is *complete.*

Each change in solution as a search proceeds is termed a *move,* and the collection of partial solutions reachable in one move from the current solution constitutes its *neighborhood.* Moves may change the value of a decision variable already assigned a value in the partial solution, or they may give a value to a previously free decision variable.

To illustrate, consider a pure discrete model with four decision variables y_1, y_2, y_3, y_4. One partial solution is $(y_1, y_2, y_3, y_4) = (1, 0, \#, \#)$, where # denotes a free value. Here only y_1 and y_2 have been assigned a specific value. Assume moves may either increase a single fixed variable from 0 to 1; or decrease a single fixed variable from 1 to 0; or assign the value 0 to a single free variable; or assign the value 1 to a single free variable. The neighborhood of the present solution is then the six partial solutions $(y_1, y_2, y_3, y_4) = (1, 1, \#, \#)$, $(0, 0, \#, \#)$, $(1, 0, 0, \#)$, $(1, 0, 1, \#)$ $(1, 0, \#, 0)$, and $(1, 0, \#, 1)$, reachable in one move.

3. TOTAL ENUMERATION

When there are only a few discrete-valued decision variables in a model, the most effective method of analysis is usually the most direct one: *total enumeration* of all the possibilities. For example, a model with only eight 0–1 variables could be enumerated by trying all $2^8 = 256$ combinations of values for the different variables. If the model is pure discrete, it is only necessary to check whether each possible assignment of values to discrete variables is feasible and to keep track of the feasible solution with best objective function value. For mixed models the process is more complicated because each choice of discrete values yields a residual optimization problem over the continuous variables. Each such continuous problem must be solved or shown infeasible to establish an optimal solution for the full mixed problem.

Although attractive for problems with only a few discrete decisions, enumeration becomes impractical as the number of discrete variables grows to even modest size. Each new 0–1 variable doubles the number of cases that must be considered. In the range of 100–150 discrete decisions, one can compute that this explosive number of cases could occupy the fastest imaginable computer longer than the estimated life of the universe.

4. RELAXATION

When dealing with difficult discrete optimization problems, it is natural to search for related, but easier optimization models that can aid in the analysis. *Relaxations* are auxiliary optimization problems of this sort formed by weakening either the constraints or the objective function of the main problem. Specifically, an optimization problem $(\tilde{P})$ is said to be a *constraint relaxation* of another optimization problem (P) if every solution feasible to (P) is also feasible for $(\tilde{P})$. Similarly, maximize problem (respectively minimize problem) $(\tilde{P})$ is an *objective relaxation* of another maximize (respectively minimize) problem (P) if the two problems have the same feasible solutions and the objective function value in $(\tilde{P})$ of any feasible solution is $\geq$ (respectively $\leq$) the objective function value of the same solution in (P).

To illustrate, consider the discrete optimization problem

$$\begin{aligned} &\min\ -2y_1 + 3y_2 + 6z_1 \\ (P)\ &\text{s.t. } y_1 + y_2 = 1 \\ &z_1 - 10y_1 \leq 0 \\ &y_1, y_2 = 0 \text{ or } 1,\ z_1 \geq 0 \end{aligned}$$

Table 2 shows a variety of relaxations.

TABLE 2 Examples of Relaxations

Relaxation ($\tilde{P}$)	Reason
min $-2y_1$ s.t. $y_1 + y_2 = 1$ $z_1 - 10y_1 \le 0$ $y_1, y_2 = 0$ or 1, $z_1 \ge 0$	Objective relaxation. New objective underestimates at feasible y_1, y_2, z_1.
min $-2y_1 + 3y_2 + 6z_1$ s.t. $y_1 + y_2 = 1$ $y_1, y_2 = 0$ or 1, $z_1 \ge 0$	Constraint relaxation. Second main constraint deleted.
min $-2y_1 + 3y_2 + 6z_1$ s.t. $y_1 + y_2 = 1$ $z_1 - 10y_1 \le 0$ $1 \ge y_1, y_2 \ge 0$, $z_1 \ge 0$	Constraint relaxation. Feasible 0–1 values satisfy $1 \ge y_1, y_2 \ge 0$.
min $-2y_1 + 3y_2 + 6z_1$ $+100(z_1 - 10y_1)$ s.t. $y_1 + y_2 = 1$ $y_1, y_2 = 0$ or 1, $z_1 \ge 0$	Objective relaxation by term $100(z_1 - 10y_1) \le 0$ at feasible solutions. Then constraint relaxation dropping $z_1 - 10y_1 \le 0$.

Relaxations of discrete optimization problems aid in both the good solution finding and the bounding tasks of model analysis.

- An optimal solution to a relaxation can often be rounded or otherwise manipulated in a straightforward way to obtain a satisfactory solution for the main problem.
- The objective function value of an optimal solution to a relaxation bounds the optimal objective function value of the main problem. Specifically, relaxation optima provide lower bounds for minimize problems and upper bounds for maximize problems.
- An optimal solution to a relaxation is an optimal solution to the main problem if (1) it is feasible in the main problem and (2) its objective value in the main problem is the same as that in the relaxation.
- If a relaxation is infeasible, the main problem is infeasible.

4.1. Linear Programming Relaxations

An expression is *linear* if it consists of a weighted sum of variables + or − a constant. Most of the discrete modeling illustrated in Table 1 consists of linear objective functions and constraints. In follows that a great many discrete optimization problems can be effectively modeled as *integer linear programs* of the general form

$$
\begin{array}{lll}
\text{min or max} & \sum_{j=1}^{n} C_j x_j & \\
(ILP) \text{ s.t.} & \sum_{j=1}^{n} A_{ij} x_j \begin{Bmatrix} \ge B_i \\ = B_i \\ \le B_i \end{Bmatrix} & \text{for all } i = 1, \ldots, m \\
& x_j \ge 0 & \text{for all } j \notin J \\
& x_j = 0 \text{ or } 1 & \text{for all } j \in J
\end{array}
$$

Here the C_j are given objective function coefficients, the A_{ij} are given coefficients of the m linear constraints, the B_i are constant terms of the linear constraints, and J is the subset of subscripts $j = 1, \ldots n$ indexing the 0–1 variables.

The *linear programming relaxation* of problem (*ILP*), denoted ($\overline{ILP}$), is the linear program obtained when the last 0–1 system of constraints is replaced by

$$1 \ge x_j \ge 0 \text{ for all } j \in J$$

The third relaxation of Table 2 provides an example.

Because linear programs are the best solved of all optimization problems (see Chapter 97), linear programming relaxations can be optimized for very large models. For this reason, $(\overline{ILP})$s are by far the most commonly employed relaxations, and they form the core of most commercial software for discrete optimization.

Not all linear programming relaxations are close approximations of the discrete problem of interest. Because very complex criterion functions and constraints can be modeled by the linear part of an *ILP,* however, it is often the case that an optimal solution to the relaxation $(\overline{ILP})$ provides a sound starting point for construction of a good feasible solution to the discrete problem. Such constructions may loosely be termed *rounding*.

Let an optimal solution to LP relaxation $(\overline{ILP})$ be denoted by $\bar{x}_j$, $j = 1, \ldots, n$. The easiest case of rounding, yet one that still applies in many models, is where the model permits fractional $\bar{x}_j$ that should be integer (i.e., $j \in J$) to be simply rounded up to the next integer (often denoted $\lceil \bar{x}_j \rceil$) or rounded down to the next integer ($\lfloor \bar{x}_j \rfloor$). General-purpose rounding algorithms are also available, notably Balas and Martin's (1980), pivot and complement procedure. For most cases, however, the methods are dependent on the form of the model.

In a number of very important cases, (*ILP*)s have properties that guarantee there will always be an optimal solution to the corresponding $(\overline{ILP})$ with integer values for variable x_j with $j \in J$. That is, the full model (*ILP*) can be solved optimally by solving the linear relaxation $(\overline{ILP})$ because its only relaxed constraints (x_j integer for $j \in J$) are automatically satisfied by an LP optimum.

The most common of these exact cases are optimization problems that can be modeled as single-commodity *network flows* (see Chapter 99). Equivalently, these are the (*ILP*)s that can be written so that for each variable x_j, at most one constraint coefficient A_{ij} equals 1, at most one A_{ij} equals -1, and all other A_{ij} equal 0. Such (*ILP*)'s are *totally unimodular* in that any submatrix formed by the $\{A_{ij}\}$ associated with a collection of rows i and a like-sized collection of variables j has determinant 0, 1 or -1. This is enough to ensure optimal basic solutions to $(\overline{ILP})$ (produced, for example, by the simplex algorithm for linear programming) are integer whenever right-hand-side coefficients B_i are all integer.

Another important class of (*ILP*)'s with integer optimal solutions to their linear relaxations is those that are *totally dual integer* (TDI). An $(\overline{ILP})$ with integer constraint coefficients A_{ij} is TDI if its linear programming dual (see Chapter 97) has an integer optimal solution for every integer choice of objective function coefficients C_j. Models that are TDI will have integer optimal solutions to their linear programming relaxations if all right-hand-side constants B_i are integer.

When linear programming relaxations are not exact, it is important to make them as sharp an approximation as possible. Different formulations of discrete models as (*ILP*)s can produce quite different linear programming relaxations, even though the models have the same discrete solutions.

To illustrate, consider a problem to

$$\begin{aligned} \min\quad & 30z_1 + 60z_2 - 100z_3 \\ \text{s.t.}\quad & (z_3 = 1 \textit{ only if } z_1 = z_2 = 1) \\ & z_1, z_2, z_3 = 0 \text{ or } 1 \end{aligned}$$

If the specified logical requirement is modeled as $z_1 + z_2 \geq 2z_3$ the optimal solution to the linear programming relaxation is $z_1 = 1$, $z_2 = 0$, $z_3 = 1/2$ with objective function value -20. An equivalent (*ILP*) with a sharper $(\overline{ILP})$ comes from modeling the requirement as $z_1 \geq z_3$, $z_2 \geq z_3$. In this format, an optimal solution to the linear programming relaxation is $z_1 = z_2 = z_3 = 1$ with objective value -10. The latter form is stronger because the relaxation solution obtained is both more nearly feasible for the discrete problem and the source of a more exact bound. (In fact, the second modeling yields an optimal solution for this instance.)

Because of the importance of using formulations with sharp linear programming relaxations, a great deal of research in the recent decades has been addressed to various aspects of that issue. One broad area of opportunity for sharpening linear programming relaxation comes in choosing constraint coefficients more carefully. Assume that each linear inequality of the formulation is rearranged to $\geq$ format with only the constant term on the right-hand side. Each inequality is then of the form $\Sigma_{j=1}^{n} A_{ij}x_j \geq B_i$. Recalling that we also assume all variables are subject to nonnegativity constraints, it is easy to see that this constraint will cut off more LP-feasible points if either B_i can be increased or at least one of the A_{ij} can be decreased.

A classic example relates to large-integer "big M" constants in (*ILP*) formulations. For example, suppose continuous variable z_1 is subject to constraints $0 \leq z_1 \leq 10$ and corresponding 0–1 variable y_1 is to be 1 whenever z_1 is positive. The latter requirement is modeled $My_1 - z_1 \geq 0$, where M is any positive constant at least 10. However, the modeling with the strongest linear programming relaxation will be the one that makes M as small as possible (i.e., $M = 10$).

A second common method of improving linear programming relaxations is to add new valid inequality constraints. An inequality is *valid* for an (*ILP*) if it is satisfied by every (integer) feasible solution to (*ILP*). Technically all the original constraints of (*ILP*) are valid inequalities. However, the term usually refers to added constraints that are not needed for a correct integer linear programming formulation but do sharpen the corresponding linear programming relaxation. The strongest such valid inequalities—ones that are unavoidable if the sharpened LP relaxation is to be exact—are called *facet-inducing, facetial,* or simply *facets*.

It is easy to find valid inequalities for discrete optimization formulations but difficult to find ones strong enough to materially improve the LP relaxation. Most of the families of such inequalities that have proved useful are highly problem specific.

One fairly general case is the family of valid inequalities employed by Crowder et al. (1983), when the original (*ILP*) formulation includes a constraint $\Sigma_{j \in L} A_{ij} x_j \leq B_i$ where nonzero coefficients occur only on integer subscripts in $L \subseteq J$, all coefficients A_{ij} and B_i are positive, and the A_{ij} are not just ones. For any $K \subseteq L$ with $\Sigma_{j \in K} A_{ij} > B_i$, it is easy to see the following inequality is valid $\Sigma_{j \in K} x_j \leq |K| - 1$ ($|K|$ denotes the number of subscripts in K). Furthermore, the smaller we can make K while satisfying its defining requirement, the sharper the revised linear programming relaxation. Choosing minimal K, that is, K that cannot be further reduced, produces a useful family of valid inequalities.

One difficulty with using families of strong valid inequalities to sharpen linear programming relaxations is that the number of inequalities in such families is usually exponentially large. It would be impossible actually to enumerate all such inequalities and add them to the formulation submitted to a linear programming code. Instead, successful applications have employed *separation* subroutines. Such subroutines heuristically select a small subfamily of inequalities that are likely to improve the current formulation. After solving the linear programming relaxation with the selected inequalities added, if the approximation is still not sharp enough, the separation routine may be reinvoked to generate further inequalities.

A third approach to sharpening linear programming relaxations has been to introduce an extended (larger) set of variables (see Martin 1999). Such extra variables are not necessary for a correct (*ILP*) formulation. Still, their presence in the model makes it possible to write new constraints that sharpen the linear programming relaxation.

To illustrate, consider a fixed charge network flow problem with two candidate locations for $50 thousand facilities, each with a demand for 10 thousand units of the commodity to be provided by the facilities. If there is a $1 per unit shipping charge between the facilities, a textbook formulation is

$$\begin{aligned}
\min \quad & z_{12} + z_{21} + 50y_1 + 50y_2 \\
\text{s.t.} \quad & z_1 = z_{12} + 10,\ z_2 = z_{21} + 10 \\
& z_1 \leq 20y_1,\ z_2 \leq 20y_2 \\
& z_1, z_2, z_{12}, z_{21} \geq 0;\ y_1, y_2 = 0 \text{ or } 1
\end{aligned}$$

Here z_i is the total supplied at facility i, z_{ij} represents the number of thousands of units shipped to the other facility, and y_i decides whether facility i is built. This model has LP relaxation optimal solution $\bar{z}_1 = \bar{z}_2 = 10$, $\bar{z}_{12} = \bar{z}_{21} = 0$, $\bar{y} = \bar{y}_2 = 1/2$ and value $50 thousand.

The extended form of Rardin and Wolsey (1993) introduces new variables that subdivide flows z_1 and z_2 according to whether they are directed to the facility site or to its companion site. Specifically,

$$\begin{aligned}
z_1 &= w_{11} + w_{12},\ z_2 = w_{21} + w_{22} \\
z_{12} &= w_{12},\ z_{21} = w_{21} \\
0 &\leq w_{11} \leq 10y_1,\ 0 \leq w_{12} \leq 10y_1 \\
0 &\leq w_{21} \leq 10y_2,\ 0 \leq w_{22} \leq 10y_2
\end{aligned}$$

With this extended system, the LP relaxation yields a discrete optimum of value $60 thousand.

4.2. Lagrangean Relaxations

Linear programming relaxations of integer linear programs produce an easier problem by deleting difficult integrality requirements on variables x_j with $j \in J$. This makes them as widely applicable as integer linear programming itself, but they are not always very good approximations.

Lagrangean relaxations are an alternative appropriate where some of the linear constraints are treated as the complications in an otherwise manageable discrete model. Integrality requirements are explicitly retained in relaxations, and complicating linear constraints are *dualized,* that is, taken to the objective function with appropriate Lagrange multipliers.

The fourth relaxation of Table 2 illustrates the principle of Lagrangean relaxation, but most of the success with Lagrangean relaxation has derived from problem-specific structures. Its power is better illustrated by a classic application.

Generalized assignment problems involve optimal arranging of objects $i = 1, \dots, m$ of known size S_i into locations $j = 1, \dots, n$ of known capacity K_j. An (*ILP*) formulation is

$$\begin{array}{lll} & \min & \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}y_{ij} \\ (GA) & \text{s.t.} & \sum_{j=1}^{n} y_{ij} = 1 \qquad \text{for all } i = 1, \dots, m \\ & & \sum_{i=1}^{m} S_i y_{ij} \le K \quad \text{for all } j = 1, \dots, n \\ & & y_{ij} = 0 \text{ or } 1 \quad \text{for all } i = 1, \dots, m;\ j = 1, \dots, n \end{array}$$

where C_{ij} is the cost of assigning object i to location j.

Either of the main systems of constraints in (GA) might be thought of as complicating because deletion of either leaves an easier model where each variable appears in only one constraint. Thus, either system might be first dualized and then dropped. Corresponding Lagrangean relaxations are

$$\begin{array}{lll} & \min & \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}y_{ij} + \sum_{j=1}^{n} u_j\left(\sum_{i=1}^{m} S_i y_{ij} - K_j\right) \\ (GA_{\mathrm{u}}) & \text{s.t.} & \sum_{j=1}^{n} y_{ij} = 1 \text{ for all } i = 1, \dots, m \\ & & y_{ij} = 0 \text{ or } 1 \text{ for all } i = 1, \dots, m;\ j = 1, \dots, n \end{array}$$

and

$$\begin{array}{lll} & \min & \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}y_{ij} + \sum_{i=1}^{m} v_i\left(\sum_{j=1}^{n} y_{ij} - 1\right) \\ (GA_{\mathrm{v}}) & \text{s.t.} & \sum_{i=1}^{m} S_i y_{ij} \le K_j \text{ for all } j = 1, \dots, n \\ & & y_{ij} = 0 \text{ or } 1 \text{ for all } i = 1, \dots, m;\ j = 1, \dots, n \end{array}$$

The effect of dualizing constraints is to enforce them partially by penalizing violating solutions in the objective function. However, care must be taken to ensure that a valid relaxation results. In (GA_{u}), the dualized constraints are inequalities. Terms weighted by u_j will be negative or zero at feasible solutions. Thus, in order to have a proper objective relaxation, multipliers must satisfy

$$u_j \ge 0 \text{ for all } j = 1, \dots, n$$

Furthermore, an optimal solution to a relaxation (GA_{u}) may not be optimal in (GA) even if it satisfies the relaxed constraints. Since the objective function values in (GA) and (GA_{u}) may differ, *complementary slackness* conditions

$$u_j\left(\sum_{i=1}^{m} S_i y_{ij} - K_j\right) = 0 \text{ for all } j = 1, \dots, n$$

must also be satisfied if a relaxation optimum is to be optimal in (GA).

In (GA_{v}), the situation is much easier because dualized constraints are equalities. Terms weighted by v_i will be zero at feasible solutions. Thus, no sign restrictions on Lagrange multipliers are needed for a proper relaxation, and complementary slackness is not an issue for (GA) optimality.

Any valid choice of multipliers on dualized constraints produces a Lagrangean relaxation, and like all relaxations the optimal objective function value in the relaxation bounds the optimal value of the main problem. However, some choices of constraints to dualize may give rather weak bounds; others may yield very strong bounds (see Parker and Rardin 1988, chap. 5 for a full discussion).

For any fixed dualization strategy, good bounds depend on good multipliers. Successful use of Lagrangean relaxation requires a search where a sequence of Lagrangean relaxations is solved with different multipliers. Results of one relaxation suggest ways to change the multipliers for the next.

The simplest, *subgradient search,* is effective in many settings. Assume the given discrete problem has been expressed as a minimize (*ILP*) with all inequalities $\leq$, $I^{=}$ the collection of dualized equality row numbers, $I^{\leq}$ the collection of dualized ($\leq$) inequalities, and $\{u_i : i \in I^{=} \cup I^{\leq}\}$ the Lagrange multipliers. Then subgradient search updates multipliers:

$$u_i \leftarrow u_i + \alpha \left(\sum_{j=1}^{n} A_{ij}\hat{x}_j - B_i \right) \text{ for all } i \in I^{=} \cup I^{\leq}$$

$$u_i \leftarrow \max\{u_i, 0\} \text{ for all } i \in I^{\leq}$$

where $\hat{x}_j$ denotes the most recent relaxation optimum and α is a stepsize.

5. HEURISTIC SEARCH

When a discrete model has too many decisions for total enumeration and no convenient relaxation that is sharp enough to give sound approximations, the principal remaining approach is to organize a search through a series of solutions (or partial solutions) until a satisfactory feasible solution is isolated.

Sometimes a practical method is available for computing an exact optimum (see Section 6). More commonly, however, only a *heuristic* or *approximate optimum* can be obtained in reasonable computation time. Such solutions are feasible, but there is no guarantee they are optimal. Often it is not even possible to bound the error in the heuristic optima produced.

Implementations of heuristic search in discrete optimization are usually classified according to whether they focus on partial or complete solutions. *Constructive searches* begin with an all-free partial solution and fix one or more components at each move until a complete solution is obtained. *Improving searches* work their way through a sequence of complete solutions, striving at each move to find a complete solution in the neighborhood that is either less infeasible, better in objective value than the current solution, or both. Of course, the strategies can be combined by, for example, using a constructive search to build a first complete solution and then applying an improving search to make it better.

5.1. Constructive or Solution-Building Search

Constructive searches are often called *greedy* or *myopic* because they usually choose variables and values to fix on the basis of estimates of immediate or short-term gain. The main issue in design of such procedures is to choose an informative measure of efficiency or gain on which to base selections.

For certain maximizing models on combinatorial structures called *matroids* (see e.g., Parker and Rardin 1988, chap. 3), an exact optimal solution results from the most naive possible greedy choice rule. Moves iteratively make $=1$ the remaining free variable with greatest objective function coefficient, subject only to the requirement that this choice does not produce infeasibility. The *spanning tree problem* (see the Appendix) is the most famous example of this matroid structure where a greedy algorithm is optimal.

Constructive searches based on more complicated efficiency ratios are much more common than the pure greedy notion of considering only objective coefficient magnitudes. One example is Dobson's (1982), heuristic for *generalized covering* problems of the form

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} C_j y_j \\ \text{s.t.} \quad & \sum_{j=1}^{n} A_{ij} y_j \geq B_i \quad && \text{for all } i = 1, \ldots, m \\ & y_j = 0 \text{ or } 1 && \text{for all } j = 1, \ldots, n \end{aligned}$$

where all A_{ij} and B_i are nonnegative integers. Dobson's algorithm starts with the all free solution and iteratively chooses a new $y_{\tilde{j}}$ to make $= 1$ where $\tilde{j}$ is the free index with $\min\{C_j / \Sigma_{i=1}^{m} \min\{A_{ij}, \tilde{B}_i\}\}$ and $\tilde{B}_i = \max\{0, B_i - \Sigma_{j \text{ fixed } 1} A_{ij}\}$. The denominator of this ratio is the total overall constraints of the amount variable j could contribute toward resolving remaining infeasibility. Thus, the algorithm is choosing on the basis of least cost per unit improvement in infeasibility. The process terminates (making any remaining free variables $= 0$) when the current partial solution is feasible in every main constraint.

The fact that constructive heuristics usually terminate as soon as the first feasible solution is found makes them an attractive choice where solution time is critical, such as in near real-time control. Still, results obviously depend critically on the early decisions taken, so the quality of the heuristic optimum produced tends to deteriorate rapidly with the number of discrete variables in the model.

5.2. Improving or Solution-Enhancing Search

Improving searches begin from a complete solution and apply moves that either improve the objective function value or reduce infeasibility. All or a large part of the complete solutions in the neighborhood are evaluated, and the search advances to the one found most attractive. One example is the *parallel processor* problem:

$$\begin{aligned} \min \quad & z \\ \text{s.t.} \quad & y_{i1} + y_{i2} = 1 \quad \text{for all } i = 1, \ldots, m \\ & \sum_{i=1}^{m} T_i y_{ip} \leq z \quad \text{for all } p = 1, 2 \\ & y_{ip} = 0 \text{ or } 1 \quad \text{for all } i = 1, \ldots, m;\ p = 1, 2 \end{aligned}$$

Here tasks $i = 1, \ldots, m$ of time duration T_i must be scheduled on one of two processors p. Decision variables $y_{ip} = 1$ if task i is assigned to processor p and $= 0$ otherwise. Continuous variable z measures the completion time of all tasks.

An improving search would begin from any feasible choice of values for the discrete variables (each task is assigned to some processor). A major algorithm-design question is what other solutions should be considered neighbors, or equivalently, what set of moves to apply.

One obvious family of moves consists of switching one task from its current processor to the other. Whichever of these changes most improved the objective function would provide the move to be taken.

A neighborhood allowing pairs of tasks to be interchanged between the two processors could yield better results. However, the computational effort per step would increase because there are only m reassignments of one task to a different processor, but more like $m^2/4$ pairwise swaps (assuming about half the tasks will be on each processor).

Still another neighborhood might employ moves that delete or add a single task to one of the processors. Such moves could create infeasibility because a task might be assigned to both processors, or it might not be assigned at all. The difficulty in this case is how to balance improvement in solution value with reduction in infeasibility as the next solution is chosen. A common approach is to add a penalty term in the objective function to discourage infeasibility without prohibiting it. In the parallel processor example above, this penalty term would have the form

$$\alpha \sum_{i=1}^{m} |y_{i1} + y_{i2} - 1|$$

where $\alpha > 0$ is the weight applied to infeasibility.

5.3. Local Optima

Although heuristic search has proved useful on some problems, natural definitions of neighborhoods often lead to poor *local optima,* that is, final feasible solutions that cannot be improved in the neighborhood. Of course, a richer family of allowed moves would make it possible to reach stronger local optima, but the cost of examining the neighborhood at each iteration rapidly becomes prohibitive.

One standard solution to this dilemma, known as *multistart,* is to employ a limited neighborhood but restart the search several times. Each time, the search begins with a randomly chosen starting solution and continues to a local optimum. The best of these local optima is kept as an heuristic optimum.

5.4. Tabu, Simulated Annealing, and Genetic Algorithms

An alternative that has generated much recent research interest is to liberate neighborhood search from the obligation to improve at each step. That is, moves are sometimes adopted that do not improve the objective function (or reduce infeasibility). Comprehensive books on the topic include Reeves (1993), Aarts and Lenstra (1997), and Glover and Laguna (1998).

The immediate difficulty with nonimproving moves is that they can make the search loop. For example, suppose that no improving move is available and a nonimproving move is adopted to change y_{27} from $= 1$ to $= 0$. Assuming some symmetry in the move set, there will certainly be an improving move at the next step: changing y_{27} back from $= 0$ to $= 1$. The algorithm could loop infinitely between the two.

Tabu algorithms of Glover and others deal with repeats by keeping list of moves that are temporarily "tabu" or forbidden. The best improving (or least nonimproving) non-tabu move is adopted at each step of the procedure. For example, in the above case where y_{27} is switched from $= 1$ to $= 0$, any move involving y_{27} might be placed on the tabu list for say the next 5–10 steps and then freed. Solutions can still repeat under tabu, but computational experience has shown promise in a number of applications.

From this simple beginning, tabu methods have developed in a variety of directions. Once a data structure must be maintained to limit moves, it can be used to guide other aspects of the search in a variety of creative ways. For example, moves that have proved useful in the past may be given some preference, or moves that have not been used in the most recent part of the search might be tried. Glover and Laguna (1998) develop a host of alternatives.

A second approach is the stochastic one of *simulated annealing* (see, e.g., Kirkpatrick et al. (1983); Aarts and Lenstra (1998)). With simulated annealing, a move is selected randomly from the available neighborhood at each iteration. If the selected move would result in an improvement, it is adopted and the search advances to the indicated solution. If the move would degrade the solution, it may or may not be adopted, depending on a probability that decreases with the magnitude of the degradation.

A common rule is to accept a nonimproving move with probability $e^{-d/T}$, where d is the amount by which the solution degrades the objective function value and T is a control parameter called the *temperature*. Typically T is started relatively large so that the search can range widely in the early stages. Then T is slowly decreased as the procedure settles into the region of a good feasible solution.

As with tabu, solutions can repeat, but simulated annealing's use of probabilities ensures the search will advance to better solutions if any exist, although it may take a long time. Experience has shown simulated annealing to be a reliable and easy-to-implement way to compute good heuristic solutions in a wide variety of applications. However, comparatively long running times are often required.

Genetic algorithms (see Holland 1975; Goldberg 1989) offer still another approach to dealing with local optima. Instead of keeping just a single current solution at each move, these methods retain a whole *population* of solutions. At each update or *generation,* some or all of these solutions will be replaced by improved ones.

Any of the normal manipulations of neighborhood search can be employed to construct the new solutions, but *crossover* moves, which interchange parts of solutions in the current population, are the most popular. For example, crossover of "parent" solutions (0,1,0,0,1,1) and (1,1,0,1,0,1) by cutting after the third component would produce "offspring" (0,1,0,1,0,1) and (1,1,0,0,1,1). These new solutions would be evaluated and the better ones preserved in the next generation.

Genetic algorithms have become the method of choice in difficult engineering design circumstances where complex feasibility limitations and massive nonlinearity make it difficult to employ neighborhood-based methods. However, other methods usually give better performance on classic combinatorial optimization problems—especially (*ILP*)'s and cases with linear constraints.

6. BACKTRACKING SEARCH AND BRANCH AND BOUND

Heuristic searches usually guarantee neither a global optimal solution nor a bound on the error when computation stops because they make *preemptive* moves—moves for which there are viable alternatives that are not explored. In order to carry out a more exhaustive search, such moves must be taken as *provisional.* That is, a record must be retained of the alternatives not yet pursued, and search must *backtrack* to those alternatives, pursuing them until they prove incapable of producing an satisfactory solution. As the search encouters a particular partial solution, it either *terminates* that solution, that is, finds it to admit no improving move, or *branches* it, that is, extends it by one applicable move. The best feasible solution encountered is recorded as the *incumbent solution.* Thus, if the search exhausts all open alternatives, the incumbent solution is a global optimum.

6.1. Tree Representation

One essential element of backtracking search is a record of provisional moves and alternatives not yet explored. The most convenient format for such a record is a tree like the one in Figure 1. *Nodes* of that tree represent states of the search. *Branches* show the selected and alternative moves available in a state. Search proceeds from the first node or *root* of the tree toward the bottom. A node is numbered when it is actually visited by the search; ones still to be considered are left unnumbered. A backtrack occurs thenever the search skips to an open alternative instead of an extension of the current state.

The example of Figure 1 represents a search that has already completed 6 nodes. Successive moves from the root=1 fixed y_5, then y_{11}, then y_3 to 1. The last produced a feasible solution and node 4, which is terminated. The search now backtracks to one of the three unexplored alternatives

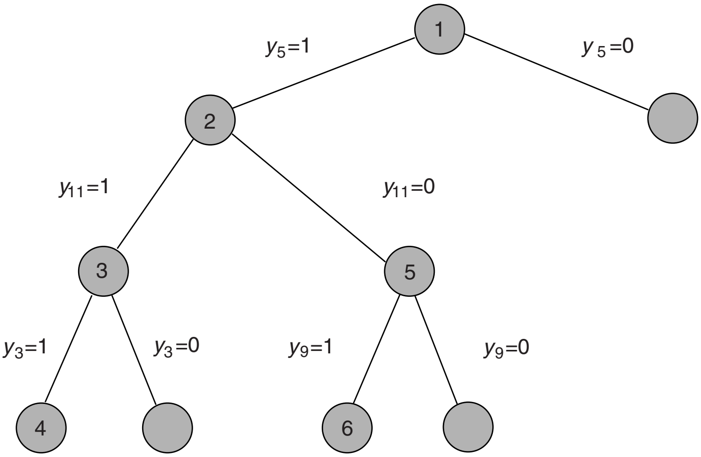


Figure 1 Backtracking Search Tree.

to the moves taken. Here the choice was to adopt move $y_{11} = 0$ to the node marked 5. Moves $y_9 = 1$ and $y_9 = 0$ were now available at this node, so it was branched and the first of these moves was selected.

6.2. Branch and Bound

Branch and bound procedures combine backtracking search with the power of relaxations. Any partial solution in a search that has variables still free defines a *candidate problem,* that is, a discrete optimization problem over the free variables subject to limits imposed by the fixed decisions. Instead of pursuing partial solutions until no further moves are available, branch and bound solves a relaxation of the corresponding candidate problems. If the relaxation optimum satisfies requirements to be optimal for the candidate problem, the partial solution can be terminated immediately; its best completion has been identified. If the relaxation proves infeasible, the partial solution can also be terminated; no completion exists. When neither of these cases occurs, the value of the relaxation optimal solution provides a bound on the value of the candidate problem. That is, it yields a bound on the quality of any completion. If that bound is already worse than the incumbent solution, no completion can improve on the incumbent; the node can be terminated. In any event, the best such bound across all unexplored nodes provides a global bound on the optimal value of the full discrete model.

Any reasonable set of branching moves can be combined with any convenient relaxation to produce a branch and bound procedure. Still, the great majority of successful applications and all commercial codes for discrete optimization use implementations of branch and bound on integer linear programs (*ILP*), with linear programming relaxations. The main ideas of such an algorithm can be outlined as follows:

```
ILP branch and bound (minimize):
initialize $CAND \leftarrow \{(ILP)\}$; $INCUM \leftarrow \infty$; $k \leftarrow 0$
while $CAND \neq \emptyset$ do
    $k \leftarrow k + 1$
    select a member of $CAND$ as $(ILP_k)$
    attempt to solve relaxation $(\overline{ILP}_k)$ for value $VAL_k$
    if $(\overline{ILP}_k)$ is infeasible or $VAL_k \geq INCUM$
        then delete $(ILP_k)$ from $CAND$
        elseif the $(\overline{ILP}_k)$ optimum is feasible in $(ILP)$
            then $INCUM \leftarrow \min\{INCUM, VAL_k\}$
                delete any member of $CAND$ with stored bound $\geq INCUM$
                delete $(ILP_k)$ from $CAND$
        else choose binary free $x_p$ fractional in $(\overline{ILP}_k)$
            replace $(ILP_k)$ in $CAND$ by extensions with $x_p = 0$
                and $x_p = 1$, both with stored bound $VAL_k$
end
```

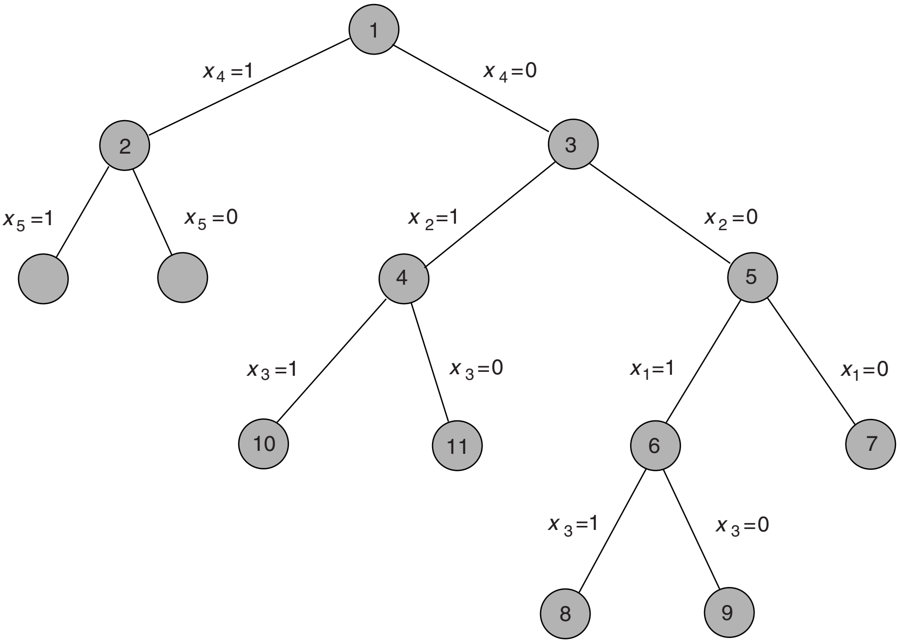


Figure 2 Branch and Bound Example.

In this statement, *CAND* represents a list of the candidate problems associated with active nodes of the search. *Stored bounds* are maintained with members of *CAND* to show the best-known lower bound on the value of an optimal solution to the candidate.

To illustrate this (*ILP*) form of branch and bound, consider the example

$$\begin{aligned} \text{min} \quad & 7x_1 + 12x_2 + 7x_3 + 14x_4 + x_5 + 25x_6 \\ \text{s.t.} \quad & 3x_1 + 6x_2 + 5x_3 + 16x_4 + x_6 \geq 7 \\ & 5x_1 + 4x_2 + 4x_3 + 5x_4 \geq 2 \\ & 3x_1 + 3x_2 + 3x_3 + 10x_5 \leq 3 \\ & x_1, \ldots, x_5 = 0 \text{ or } 1;\ 0 \leq x_6 \leq 3 \end{aligned}$$

Figure 2 shows the search tree for this example, and Table 3 describes processing of each node.

A complete implementation of branch and bound in even the (*ILP*) form given above involves many heuristic rules. The primary ones are (1) which free variable to fix in branching node k and (2) which active node (member of *CAND*) to choose at each iteration k. For the simple example above, corresponding rules were (1) choose the fractional-valued free variable closest to 0.0 and (2) choose the active node with least stored bound, breaking ties in favor of the last fixed variable = 1.

A desirable feature of branch and bound is that a bound on the value of a global optimal solution is always available, so that the algorithm need not be run to termination in order to bound the error of accepting the incumbent solution as approximately optimal. The best stored bound of *CAND* (least for minimize problems, highest for maximize problems) always provides such a bound. Thus, for example, after node 8 is processed in the example of Figure 2, it is certain that any solution to the full (*ILP*) will cost at least $\min\{12.6, 13.4\} = 12.6$. Stopping at that point with the incumbent solution of value 14 would produce at most $(14 - 12.6)/12.6 = 11.1\%$ error.

7. GUIDELINES AND LIMITS

Discrete optimization problems abound in all phases of industrial engineering, and ones with important economic implications easily justify formal modeling and systematic analysis. However, there are no general-purpose methods appropriate for dealing with all or even most models. This concluding

TABLE 3 Processing for Branch and Bound Example

Node	Fixed	LP Relaxation	*VAL*	Processing
1	none	(0,0,0,0.44,0.3,0)	6.42	branch on x_4
2	$x_4 = 1$	(0,0,0,1,0.3,0)	14.30	branch on x_5
3	$x_4 = 0$	(0,0.33,1,0,0,0)	11	branch on x_2
4	$x_4 = 0, x_2 = 1$	(0,1,0.2,0,0,0)	13.4	branch on x_3
5	$x_4 = 0, x_2 = 0$	(0.67,0,1,0,0,0)	11.67	branch on x_1
6	$x_4 = 0, x_2 = 0,$ $x_1 = 1$	(1,0,0.8,0,0,0)	12.6	branch on x_3
7	$x_4 = 0, x_2 = 0,$ $x_1 = 0$	(0,0,1,0,0,2)	57	new incumbent; *INCUM* ← 57; terminate solved
8	$x_4 = 0, x_2 = 0,$ $x_1 = 1, x_3 = 1$	(1,0,1,0,0,0)	14	new incumbent; *INCUM* ← 14; delete extensions of Node 2; terminate solved
9	$x_4 = 0, x_2 = 0,$ $x_1 = 1, x_3 = 0$	infeasible	—	terminate infeasible
10	$x_4 = 0, x_2 = 1,$ $x_3 = 1$	(0,1,1,0,0,0)	19	terminate bound $\geq$ 14
11	$x_4 = 0, x_2 = 1,$ $x_3 = 0$	(0.33,1,0,0,0,0)	14.33	terminate bound $\geq$ 14; incumbent solution optimal

section reviews theory and accumulated wisdom delimiting which of the approaches treated in previous sections are appropriate for a given model.

7.1. Computational Complexity Theory

The formal theory of problem difficulty classification is called *computational complexity theory* (see Garey and Johnson 1979; Papadimitriou 1994; Parker and Rardin 1988, chap. 2). Complexity theory terms a *problem* any collection of related *instances* distinguished only by their size and numerical constants. For example, (*ILP*) is a problem with instances distinguished by counts m and n, discrete variable list J, and constants C_j, A_{ij}, and B_i. The *size* of an instance is the length of the symbol string required to encode it for a computer.

The objective of complexity theory is to classify problems according to how efficiently instances can be solved relative to their size. Bounds on the required computation are expressed as *computational orders,* denoted $O(\ )$. For example, if every instance of a problem can be solved in a number of elementary calculations bounded by the square of its size s, the problem is $O(s^2)$ solvable. More generally, a problem is said to be *polynomially solvable* if there is a constant k such that every instance of size s is solvable in $O(s^k)$ effort. Specifically, $O(s^4)$, $O(s\sqrt{s})$, and $O(s^2 \log s)$ computations are polynomial (the last because $s^2 \log s \leq s^3$). $O(2^s)$ and $O(s!)$ are not polynomial.

One of the most important theoretical achievements of discrete optimization research has been to isolate polynomial solvability as the defining characteristic of truly tractable discrete problems. Every one of the discrete models for which a generally effective algorithm or an exact (polynomially solvable) relaxation has been discovered belongs to the polynomially solvable class.

Across what seems to be a cosmic boundary in mathematics lies the alternative *NP-hard* class, to which almost all discrete optimization models belong that lack such tractable characteristics. *NP*-hard problems are not (yet) provably outside the reach of polynomial solvability, but neither are they just problems for which research has so far failed. Class *NP* is a vast collection of "Does there exist . . ." problems in discrete mathematics and computer science, some of which have been studied for centuries without much progress. No discrete optimization problem actually belongs to *NP* because none is of this existence form. Still, each *NP*-hard discrete optimization problem (*H*) is as difficult as any in *NP* in the sense that a polynomial algorithm for (*H*) would provide one for every member of *NP*. Thus, to seek such an algorithm for any *NP*-hard problem is simultaneously to attack the enormous variety of challenging problems in *NP*. Success is most unlikely. Similar logic establishes that exact relaxations and strong duality theories are also highly improbable for most *NP*-hard problems, although additional technical issues make general statements more difficult.

The fundamental importance of the polynomially solvable vs. *NP*-hard distinction makes it a high-priority matter to try to determine on which side of the boundary any given discrete optimization application falls (theory indicates there may be problems that belong to neither category, but few realistic candidates are known). Polynomial solvability is almost always established by showing the given model is a special case of one of the classic polynomial time models detailed in the Appendix, or inventing some simple enumeration over cases solvable as one of those models. On the other hand, one shows a problem is *NP*-hard by demonstrating that some known *NP*-hard problem can be viewed as a special case (several *NP*-hard models are included in the Appendix, and Garey and Johnson 1979 provides a much bigger list). For example, generalized assignment (*GA*) is known to be *NP*-hard. This is enough to establish that integer linear programming (*ILP*) is *NP*-hard because every instance of (*GA*) is an instance of (*ILP*).

7.2. Choosing a Strategy

When a discrete optimization problem is in the polynomially solvable complexity category, it is usually clear how to proceed with its analysis. A clever and efficient algorithm is at hand. Often an exact linear programming formulation is also known, and a strong duality theory is available for sensitivity studies. Very complete analysis should be possible, unless limits on the time available for solution (e.g., in a real-time setting) mandate quicker methods.

In the far more typical case where the given discrete optimization is *NP*-hard, more care should be exercised in choosing avenues of analysis. Figure 3 offers some very approximate guidelines in terms of two critical characteristics: the number of binary decisions in the model vs. the error of the best available relaxation bound.

For small numbers of decisions, say up to 10–15, total enumeration is recommended, regardless of relaxation quality. As the number of binary decisions increases, models subdivide into three regions. In the outermost, relaxations are too poor to assist in analysis. The only practical strategy is some form of heuristic search lacking even a bound on the suboptimality of results obtained.

The innermost region shows where branch and bound methods may be effective. Relatively strong relaxations are required, and the needed sharpness increases rapidly with the number of binary decisions.

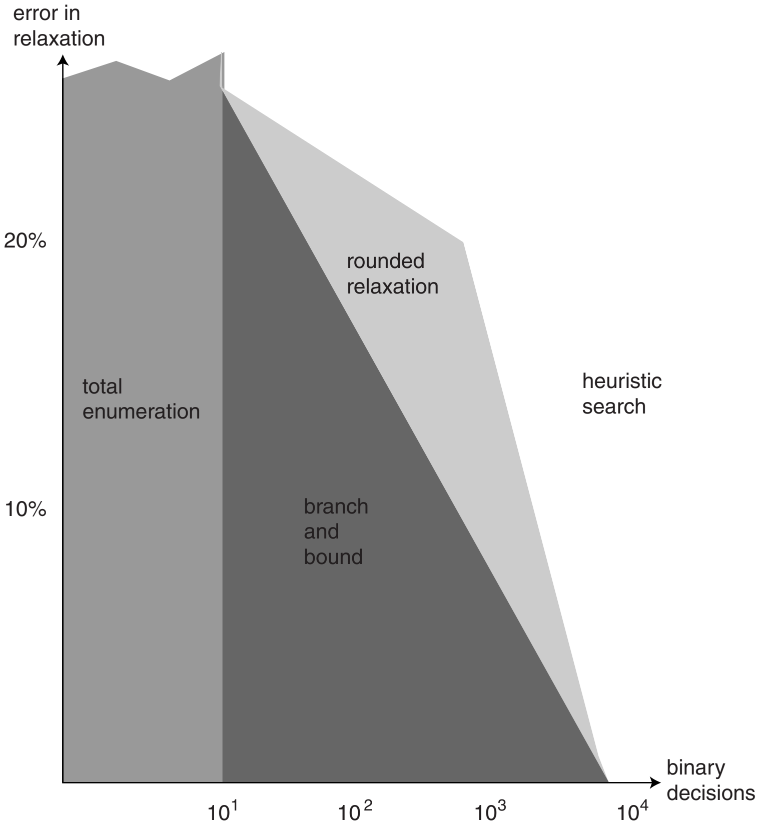


Figure 3 Guidelines for NP-Hard Problems.

Between these two extremes lies a region where relaxations may be helpful but branch and bound is unlikely to be effective. Here, bounds from relaxations at least delimit the suboptimality of solutions found through heuristic search or other means. Relaxation may also provide a good source of feasible discrete solutions when the problem admits easy rounding of relaxation optima.

As with any broad guidelines, the boundaries in Figure 3 are very fuzzy. In particular, it is often too simple to take the number of binary decisions in a model to equal the number of 0–1 variables in an (*ILP*) formulation. A generalized assignment model, for example, with m objects and n locations has mn 0–1 variables. However, each belongs to $\Sigma_j y_{ij} = 1$ multiple-choice set. Thus, there are really only n choices in each of these subsets, or $m \log_2 n$ total binary decisions in the model.

Application of the guidelines in Figure 3 also requires, of course, information about the likely quality of available relaxations. Each family of examples has different behavior, but there are characteristic attributes of relaxations with little promise:

- Nonlinear objective functions or constraints
- Either–or constraints that can only be placed in (*ILP*) format through the use of large positive constants (“big M’s”)
- Massive symmetry introducing numerous feasible alternatives of nearly the same objective function value

Any relaxation possessing such attributes is likely to require strengthening if it is to be of practical use with even moderately large applications.

APPENDIX

This appendix briefly describes and formulates a variety of standard discrete optimization models. Throughout, y denotes 0–1 discrete variables, z represents continuous variables, $D(N, A)$ indicates a directed network or graph with nodes in N and arcs in A, and $G(N, E)$ denotes an undirected network or graph with nodes in N and edges in E. Capital letters are used to denote input and lower case to show decision variables. The size of set S is indicated by $|S|$.

Assignment: Maximum utility pairing of objects from two given sets S and T; $\triangleq$ family of allowed pairs (i, j), $i \in S$, $j \in T$; $C_{ij} \triangleq$ value of pair (i, j).

$$\begin{aligned} y_{ij} &\triangleq 1 \text{ if pair } (i,j) \text{ is chosen, else } 0 \\ \max\ & \Sigma_{(i,j)\in E}\, C_{ij} y_{ij} \\ \text{s.t.}\ & \Sigma_{j\in T}\, y_{ij} = 1,\ i \in S \\ & \Sigma_{i\in S}\, y_{ij} = 1,\ j \in T \\ & y_{ij} = 0 \text{ or } 1,\ (i,j) \in E \end{aligned}$$

Polynomially solvable by combinatorial algorithms. Special case of both Matching and Network Flows. LP relaxation is totally unimodular.

Capital Budgeting: Maximum value subset of objects or projects j to pack within given capacity or resource limits B_i; $A_{ij} \triangleq$ (nonnegative) consumption of resource i by project j; $C_j \triangleq$ value of project j.

$$\begin{aligned} y_j &\triangleq 1 \text{ if project } j \text{ is chosen, else } 0 \\ \max\ & \Sigma_j\, C_j y_j \\ \text{s.t.}\ & \Sigma_j\, A_{ij} \le B_i \quad \text{for all } i \\ & y_j = 0 \text{ or } 1 \quad \text{for all } j \end{aligned}$$

NP-hard. LP relaxation rounds down and can be strengthened with the inequalities discussed in the Valid Inequalities section above. Solution-building heuristics based on ratios of C_j to weighted sums of A_{ij} are common.

Facilities Location: Minimum cost subset of facilities i with capacity U_i, considering both construction, setup, etc. cost $F_i \geq 0$ of the facility; plus variable travel, service cost C_{ij} of serving customers j from facility i; $D_j \triangleq$ demand at j.

$z_{ij} \triangleq$ flow from i to j
$y_i \triangleq$ 1 if i is opened, else 0
min $\Sigma_{i,j}\, C_{ij} z_{ij} + \Sigma_i\, F_i y_i$
s.t. $\Sigma_i\, z_{ij} = D_j$ for all j
$\Sigma_j\, z_{ij} \le U_i y_i$ for all i
$0 \le z_{ij} \le D_j y_i$ for all i, j
$y_i = 0$ or 1 for all i

NP-hard. LP relaxation of the stated form is strong, but weak if the third system of constraints is deleted. Special case of Fixed Charge Network Flows.

Fixed Charge Network Flows: Minimum variable plus fixed cost flow in directed graph $D(N, A)$ with net demands D_k at nodes $k \in N$, and capacities U_{ij} on arc $(i, j) \in A$; $C:_{ij} \triangleq$ unit cost of (i, j) flow; $F_{ij} \triangleq$ fixed cost (nonnegative) of using arc (i, j) at all.

$z_{ij} \triangleq$ flow in arc (i, j)
$y_{ij} \triangleq$ 1 if $z_{ij} > 0$, else 0
min $\Sigma_{(i,j)\in A}\, C_{ij} z_{ij} + \Sigma_{(i,j)\in A}\, F_{ij} y_{ij}$
s.t. $\Sigma_{(i,k)\in A}\, z_{ik} - \Sigma_{(k,j)\in A}\, z_{kj} = D_k,\ k \in N$
$0 \le z_{ij} \le U_{ij} y_{ij},\ (i, j) \in A$
$y_{ij} = 0$ or $1,\ (i, j) \in A$

NP-hard. LP relaxation is poor unless capacities are tight. For other cases much improved *multicommodity* extended LP relaxation is obtained by introducing separate variables z_{ij}^{st} recording the (i, j) flow originating at supply s and bound for demand t.

Generalized Assignment: Minimum cost assignment of objects i to capacitated locations, plants, vehicles, etc.; $S_i \triangleq$ size of object i; $K_j \triangleq$ capacity of location j; C_{ij} cost of assigning i to j.

$y_{ij} \triangleq$ 1 if i assigned to j, else 0
min $\Sigma_{i,j}\, C_{ij} y_{ij}$
s.t. $\Sigma_j\, y_{ij} = 1$ for all i
$\Sigma_i\, S_i y_{ij} \le K_i$ for all j
$y_{ij} = 0$ or 1 for all i, j

NP-hard. Excellent bounds are obtained from Lagrangean relaxations dualizing $=1$ constraints to leave a series of Knapsack problems.

Generalized Covering: Minimum cost subset of patterns j to cover given requirements B_i; $A_{ij} \triangleq$ (nonnegative) contribution to requirement i by pattern j; $C_j \triangleq$ cost of pattern j.

$y_j \triangleq$ 1 if pattern j is chosen, else 0
max $\Sigma_j\, C_j y_j$
s.t. $\Sigma_j\, A_{ij} y_j \ge B_i$ for all i
$y_j = 0$ or 1 for all j

NP-hard. LP relaxation rounds up. Solution building heuristics based on ratios of C_j to weighted sums of A_{ij} are common.

Job Shop Scheduling: Sequence a collection of jobs j with steps $s = 1, \ldots, S_j$ on processors p to minimize the time to complete all jobs; processor sequence for any job is fixed; $T_{js} \triangleq$ duration of step s; $P_{js} \triangleq$ processor of step s.

$z \triangleq$ completion time of all tasks
$z_{js} \triangleq$ start time of job j, step s
min z
s.t. $z_{j1} \ge 0$ for all j
$z_{js} \ge z_{j,s-1} + T_{j,s-1}$,
for all j; $s = 2, \ldots, S_j$
$z \ge z_{jS_j} + T_{jS_j}$ for all j
$z_{js} \ge z_{j's'} + T_{j's'}$ or $z_{j's'} \ge z_{js} + T_{js}$,
for all j, s, j', s' with $P_{js} = P_{j's'}$

NP-hard. LP relaxation formed with big M as in Table 1 is poor. Both solution-building and solution-enhancing heuristics are common.

Knapsack: Maximum value subset of objects or projects j to pack within a given capacity or budget B; $A_j \triangleq$ size of object j; $C_j \triangleq$ value of object j.

$$\begin{aligned} y_j &\triangleq \text{ 1 if object } j \text{ is chosen, else 0} \\ \max \quad & \textstyle\sum_j C_j y_j \\ \text{s.t.} \quad & \textstyle\sum_j A_j y_j \leq B \\ & y_j = 0 \text{ or } 1 \quad \text{for all } j \end{aligned}$$

NP-hard. LP relaxation rounds down and is near optimal in many cases. Heuristics can come arbitrarily close to optimal in polynomial time.

B+ Leontief Flows: Minimum cost flow through state nodes in N of directed hyperarcs or composition operators (I, j) combining positive integer multiples A^i_{Ij} of inputs $i \in I \subseteq N$ to produce one unit at node j; $C_{Ij} \triangleq$ unit cost of (I, j) flow; $B_k \geq 0$ is the nonnegative net requirement at node N.

$$\begin{aligned} z_{Ij} &\triangleq \text{ flow in hyperarc } (I, j) \\ \min \quad & \textstyle\sum_{(I,j)} C_{Ij} z_{Ij} \\ \text{s.t.} \quad & \textstyle\sum_{(I,k)} z_{Ik} - \sum_{(K,j),K \ni k} A^{k\,j}_{K,} \, z_{Kj} \\ & \quad = B_k, \ k \in N \\ & z_{Ij} \geq 0, \quad \text{for all } (I, j) \end{aligned}$$

Polynomially solvable by combinatorial algorithms. LP relaxation is TDI so has integer optima if all B_k are integer.

Matching: Maximum utility nonoverlapping collection of pairs of objects from a given set N; $E \triangleq$ family of allowed pairs (i, j), $i < j$; $C_{ij} \triangleq$ value of pair (i, j).

$$\begin{aligned} y_{ij} &\triangleq \text{ 1 if pair } (i, j) \text{ is chosen, else 0} \\ \max \quad & \textstyle\sum_{(i,j) \in E} C_{ij} y_{ij} \\ \text{s.t.} \quad & \textstyle\sum_{(i,k) \in E} y_{ik} + \sum_{(k,j) \in E} y_{kj} \leq 1, \ k \in N \\ & y_{ij} = 0 \text{ or } 1, \ (i, j) \in E \end{aligned}$$

Polynomially solvable by combinatorial algorithms. An exact (TDI) linear programming relaxation is also known.

Network Flows: Minimum cost flow in directed graph $D(N, A)$ with net demands D_k at nodes $k \in N$, and capacities U_{ij} on arc $(i, j) \in A$; $C_{ij} \triangleq$ unit cost of (i, j) flow.

$$\begin{aligned} z_{ij} &\triangleq \text{ flow in arc } (i, j) \\ \min \quad & \textstyle\sum_{(i,j) \in A} C_{ij} z_{ij} \\ \text{s.t.} \quad & \textstyle\sum_{(i,k) \in A} z_{ik} - \sum_{(k,j) \in A} z_{kj} = D_k, \ k \in N \\ & 0 \leq z_{ij} \leq U_{ij}, \ (i, j) \in A \end{aligned}$$

Polynomially solvable by combinatorial algorithms. LP relaxation is totally unimodular so has integer optima if demands and capacities are integer. Special cases include Assignment, Shortest Path, Maximum Flow, and Minimum Cut (see Chapter 99).

Parallel Processor Scheduling: Assign tasks i of duration T_i to one of several processors p so that the time to complete all tasks (*makespan*) is minimized.

$$\begin{aligned} y_{ip} &\triangleq \text{ 1 if task } i \text{ assigned to } p\text{, else 0} \\ z &\triangleq \text{ completion time of all tasks} \\ \min \quad & z \\ \text{s.t.} \quad & \textstyle\sum_i T_i y_{ip} \leq z \quad \text{for all } p \\ & \textstyle\sum_p y_{ip} = 1 \quad \text{for all } i \\ & y_{ip} = 0 \text{ or } 1 \quad \text{for all } i, p \end{aligned}$$

NP-hard. Closely related to Generalized Assignment. Both solution-building and solution-enhancing heuristics are common.

Quadratic Assignment: Minimum cost assignment of objects $i \in S$ to locations, times $j \in T$, where value is measurable only after pairs of assignments; $E \triangleq$ family of allowed pairs (i, j), $i \in S, j \in T$; $V_{ik} \triangleq$ the shared activity of i and k; $C_{jl} \triangleq$ unit cost or distance of activity between locations j and l.

$$\begin{aligned} y_{ij} &\triangleq 1 \text{ if pair } (i, j) \text{ is chosen, else } 0 \\ \min \quad & \Sigma_{(i,j)\in E}\, \Sigma_{(k,l)\in E}\, (V_{ik}C_{jl})y_{kl} \\ \text{s.t.} \quad & \Sigma_{j\in T}\, y_{ij} = 1,\ i \in S \\ & \Sigma_{i\in S}\, y_{ij} = 1,\ j \in T \\ & y_{ij} = 0 \text{ or } 1.\ (i, j) \in E \end{aligned}$$

NP-hard. No effective relaxations are available for even moderate-sized problems. Local improvement by pairwise exchange of assignments is common.

Set Covering: Finite list of patterns, routes, workers, etc. j that must span a collection of customers, hours, districts, jobs i (duplication allowed); $A_{ij} \triangleq 1$ if pattern j covers i, else 0; $C_j \triangleq$ cost of pattern j.

$$\begin{aligned} y_j &\triangleq 1 \text{ if pattern } j \text{ is chosen, else } 0 \\ \min \quad & \Sigma_j\, C_j y_j \\ \text{s.t.} \quad & \Sigma_j\, A_{ij} y_j \geq 1 \quad \text{for all } i \\ & y_j = 0 \text{ or } 1 \quad \text{for all } j \end{aligned}$$

NP-hard. LP relaxation rounds up and is near optimal in many cases. LP relaxations often numerically difficult.

Set Packing: Finite list of patterns, routes, workers, etc. j that must not overlap in customers, hours, districts, jobs i; $A_{ij} \triangleq 1$ if pattern j uses i, else 0; $C_j \triangleq$ value of pattern j.

$$\begin{aligned} y_j &\triangleq 1 \text{ if pattern } j \text{ is chosen, else } 0 \\ \max \quad & \Sigma_j\, C_j y_j \\ \text{s.t.} \quad & \Sigma_j\, A_{ij} y_j \leq 1 \quad \text{for all } i \\ & y_j = 0 \text{ or } 1 \quad \text{for all } j \end{aligned}$$

NP-hard. LP relaxation rounds down and is near optimal in many cases. For others many valid inequalities are known. LP relaxations often numerically difficult.

Set Partitioning: Finite list of patterns, routes, workers, etc. j that must span a collection of customers, hours, districts, jobs i without duplication; $A_{ij} \triangleq 1$ if pattern j covers i, else 0; $C_j \triangleq$ cost of pattern j.

$$\begin{aligned} y_j &\triangleq 1 \text{ if pattern } j \text{ is chosen, else } 0 \\ \min \quad & \Sigma_j\, C_j y_j \\ \text{s.t.} \quad & \Sigma_j\, A_{ij} y_j = 1 \quad \text{for all } i \\ & y_j = 0 \text{ or } 1 \quad \text{for all } j \end{aligned}$$

NP-hard. LP relaxation gives good bounds in many cases but difficult to round. LP relaxations often numerically difficult.

Spanning Tree: Maximum total weight subset of edges in a connected undirected graph $G(N, E)$ containing exactly one path between each pair of nodes; $C_{ij} \triangleq$ value of edge (i, j), $i < j$.

$$\begin{aligned} y_{ij} &\triangleq 1 \text{ if edge } (i, j) \text{ is chosen, else } 0 \\ \max \quad & \sum_{(i,j)\in E} C_{ij} y_{ij} \\ \text{s.t.} \quad & \sum_{(i,j)\in E} C_{ij} y_{ij} \geq |N| - 1 \\ & \sum_{i,j\in S} y_{ij} \leq |S| - 1,\ S \subseteq N \\ & y_{ij} = 0 \text{ or } 1,\ (i, j) \in E \end{aligned}$$

Polynomially solvable by the greedy algorithm. LP relaxation is Totally Dual Integer.

Steiner Tree: Minimum cost collection of edges of a graph $G(N, E)$ providing a path between vertices in subset $S \subseteq N$; $C_{ij} \triangleq$ (nonnegative) cost of edge $(i, j) \in E$, $i < j$.

$$\begin{array}{rl} y_{ij} \triangleq & 1 \text{ if edge } (i, j) \text{ is chosen, else } 0 \\ \min & \Sigma_{(i,j)\in E}\, C_{ij} y_{ij} \\ \text{s.t.} & \displaystyle\sum_{i\in Q, j\in N\setminus Q} y_{ij} + \sum_{j\in Q, i\in N\setminus Q} y_{ij} \geq 1, \\ & \quad Q \subset N,\ Q \cap S \neq \emptyset,\ S\setminus Q \neq \emptyset \\ & y_{ij} = 0 \text{ or } 1,\ (i, j) \in E \end{array}$$

NP-hard. LP relaxation easily rounded up then enhanced by deleting unnecessary edges. Numerous valid inequalities are known to strengthen the LP relaxation.

Traveling Salesman: Minimum total weight route or sequence visiting each object, job, customer in N exactly once. Represent on a graph with nodes N and edges for allowed i to j transitions; $C_{ij} \triangleq$ cost of transition (i, j), $i < j$.

$$\begin{array}{rl} y_{ij} \triangleq & 1 \text{ if transition } (i, j) \text{ is chosen, else } 0 \\ \max & \Sigma_{i,j}\, C_{ij} y_{ij} \\ \text{s.t.} & \Sigma_{i<k}\, y_{ik} + \Sigma_{j>k}\, y_{kj} = 2,\ k \in N \\ & \Sigma_{i,j\in S}\, y_{ij} \leq |S| - 1,\ S \subset N \\ & y_{ij} = 0 \text{ or } 1,\ (i, j) \in E \end{array}$$

NP-hard. The best-researched problem in discrete optimization. LP relaxation of the given form is strong but requires a separation procedure. Numerous effective solution building and solution enhancing heuristics exist, especially for the *triangular cost* case with $C_{ik} \leq C_{ik} + C_{kj}$.

REFERENCES

Aarts, E., and Lenstra, J. K. (1997), *Local Search in Combinatorial Optimization,* Wiley-Interscience, New York.

Balas, E., and Martin, C. H. (1980), "Pivot and Complement—A Heuristic for 0–1 Programming," *Management Science,. Vol. 26, pp. 86–96.*

Crowder, H., Johnson, E. L., and Padberg, M. (1983), "Solving Large-Scale Zero–One Linear Programming Problems," *Operations Research,* Vol. 31, pp. 803–834.

Dobson, G. (1982), "Worst-Case Analysis of Greedy Heuristics for Integer Programming with Nonnegative Data," *Mathematics of Operations Research,* Vol. 7, pp. 515–531.

Garey, M. R., and Johnson, D. S. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness,* W.H. Freeman, San Francisco.

Glover, F., and Laguna, M. (1998), *Tabu Search,* Kluwer, Norwell, MA.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning,* Addison-Wesley, Reading, MA.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems,* University of Michigan Press, Ann Arbor, MI.

Kirkpatrick, S., Gelatt, J. R., and Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science,* Vol. 220, pp. 671–680.

Lawler, E. L. (1976), *Combinatorial Optimization: Networks and Matroids,* Holt, Rinehart & Winston, New York.

Martin, R. K. (1999), *Large Scale Linear and Integer Optimization: A Unified Approach,* Kluwer, Norwell, MA.

Nemhauser, G. L., and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization,* John Wiley & Sons, New York.

Papadimitriou, C. H. (1994), *Computational Complexity,* Addison-Wesley, Reading, MA.

Papadimitriou, C. H., and Steiglitz, K. (1982), *Combinatorial Optimization: Algorithms and Complexity,* Prentice-Hall, Englewood Cliffs, MJ.

Parker, R. G., and Rardin, R. L. (1988), *Discrete Optimization,* Academic Press, Boston.

Rardin, R. L. (1998), *Optimization in Operations Research,* Prentice Hall, Upper Saddle River, NJ.

Rardin, R. L., and Wolsey, L. A. (1993), "Valid Inequalities and Projecting the Multicommodity Extended Formulation for Uncapacitated Fixed Charge Network Flow Problems," *European Journal of Operational Research,* Vol. 71, 95–109.

Reeves, C. R. (1993), *Modern Heuristic Techniques in Combinatorial Problems,* Halsted Press, New York.

Schrijver, A. (1986), *Theory of Linear and Integer Programming,* John Wiley & Sons, New York.

Wolsey, L. A. (1998), *Integer Programming,* Wiley-Interscience, New York.

CHAPTER 101

Multicriteria Optimization

Po Lung Yu
University of Kansas
National Chiao Tung University

Chin I. Chiang
Gwo Hshiung Tzeng
National Chiao Tung University

1. INTRODUCTION 2602
2. PREFERENCES 2603
3. VALUE FUNCTIONS 2605
  3.1. Revealed Preferences and Value Functions 2605
  3.2. Methods of Constructing Value Functions 2605
4. SATISFICING AND COMPROMISE SOLUTIONS 2608
  4.1. Satisficing Models 2608
  4.2. Compromise and Goal Programming Solutions 2610
  4.3. Computing Compromise Solutions and Goal Programming Solutions 2611
    4.3.1. The Ideal Point as the Target Point 2611
    4.3.2. General Target Points and Goal Programming 2612
    4.3.3. Interactive Methods of Compromise and Goal Programming Models 2612
  4.4. Further Comments 2614
5. DOMINATION STRUCTURES 2614
  5.1. Constant Cone Domination Structures and Solutions 2615
  5.2. Variable Cone Domination Structures and Methods for Seeking Good Solutions 2616
6. MC^2-SIMPLEX METHOD FOR LINEAR CASES 2618
  6.1. Nondominated Solutions 2618
  6.2. Potential Solutions and MC^2-Simplex Method 2618
7. FUZZY MULTICRITERIA OPTIMIZATION 2620
8. EXTENSIONS AND CONCLUDING REMARKS 2620
REFERENCES 2621
ADDITIONAL READING 2623

1. INTRODUCTION

All animals have multiple goals to fulfill in their lives. They want to survive, to perpetuate their own species (including sex), to dominate, and so on. If you kick a dog, the dog will fight back or run away. Humans are no exception. We want to have a good life, which may mean more wealth, power, respect, and time for ourselves, together with good health and a successful next generation, and so on. We want food to taste, smell, and look good and be nutritious. Ever since Adam, human beings have been continuously confronted with multiple-criteria decision problems.

Although there have been abundant records of decision analyses based on multiple criteria in human history, putting the analyses in a formal mathematical setting is fairly new. Though still young, the literature of such mathematical treatments has exploded during the last three decades. The interested reader is referred to Stadler (1979, 1981), Steuer et al (1996) for surveys and historic notes.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

In this chapter, we will report on six basic concepts of multicriteria optimization. Just as the three primary colors (red, yellow, and blue) can produce an infinite number of pictures and the seven basic notes of the musical scale (do, re, mi, etc.) can produce an infinite number of songs, the six basic concepts we will describe should allow the reader to generate an infinite number of models to solve the complex multiple-criteria decision problems. The six basic concepts are:

1. *Preferences as binary relations,* in which mathematical functional relations are generalized to represent revealed preferences and some basic solution concepts of MCDM (multiple-criteria decision making) will be introduced (Section 2).
2. *Value functions,* in which preferences are represented by powerful numerical ordering and multicriteria optimization problems are reduce to single-criterion optimization (Section 3).
3. *Satisficing, goal programming, and compromise solutions,* in which we model the preferences in terms of human goal-seeking behavior and introduce goal programming (which is well known). Again in this framework multicriteria problems are reduced to single-criterion ones (Section 4).
4. *Domination structures,* in which the preferences are represented in terms of multidimensional comparisons (in contrast to 2 and 3, which are of single-dimensional comparison) and solution concepts and methods to obtain the solutions are introduced (Section 5).
5. *MC^2-simplex method for linear cases,* in which powerful computation method for linear cases are introduced (Section 6).
6. *Fuzzy multicriteria optimization,* is sketched in Section 7.

In Section 8, we offer further comments on multicriteria optimization.

2. PREFERENCES

Given any pair of decision outcomes y^1 and y^2, one and only one of the following can occur:

1. We are convinced that y^1 is better than or preferred to y^2, denoted by $y^1 > y^2$.
2. We are convinced that y^1 is worse than or less preferred to y^2, denoted by $y^1 < y^2$.
3. We are convinced that y^1 is equivalent to or equally preferred to y^2, denoted by $y^1 \sim y^2$; or
4. We have no sufficient evidence to say either 1, 2, or 3, denoted by $y^1 \; ? \; y^2$. Thus, the preference relation between y^1 and y^2 is indefinite or not yet clarified.

Note that each of the above statements involves a comparison or relation between a pair of outcomes. Let Y denote the set of all possible outcomes. Any revealed preference information, accumulated or not, can be represented by a subset of the Cartesian product $Y \times Y$, or by a so-called binary relation. We have the following definition.

Definition 2.1. A binary relation R on Y is a subset of the Cartesian product $Y \times Y$. Binary relation R is

1. *Reflexive* if $(y, y) \in R$ for all $y \in y$
2. *Symmetric* if $(y^1, y^2) \in R$ implies that $(y^2, y^1) \in R$ for all $y^1, y^2 \in y$
3. *Transitive* if $(y^1, y^2) \in R$ and $(y^2, y^3) \in R$ imply that $(y^1, y^3) \in R$, for all $y^1, y^2, y^3 \in Y$
4. *C*omplete if $(y^1, y^2) \in R$ or $(y^2, y^1) \in R$ for all $y^1, y^2 \in y$ and $y^1 \neq y^2$
5. An *equivalence* if R is reflexive, symmetric, and transitive

Definition 2.2.

1. A preference based on $>$ (respectively $<$, $\sim$, or ?) is a binary relation on Y, denoted by $\{>\}$ (respectively $\{<\}$, $\{\sim\}$, or $\{?\}$), so that whenever $(y^1, y^2) \in \{>\}$ (respectively $\{<\}$, $\{\sim\}$, or $\{?\}$) $y^1 > y^2$ (respectively $y^1 < y^2$, $y^1 \sim y^2$, or $y^1 ? \; y^2$).
2. For convenience, we also define $\{> \sim\} = \{>\} \cup \{\sim\}; \{>?\} = \{>\} \cup \{?\}$, $\{<\sim?\} = \{<\sim\} \cup \{?\}$, etc.

By a preference (or revealed preference) structure we mean the collection of all the above preferences. Because such structures are uniquely determined by $\{>\}$, $\{\sim\}$ and $\{?\}$, a preference structure will be denoted by $\mathcal{P}(\{>\}, \{\sim\}, \{?\})$, or simply by $\mathcal{P}$.

Remark 2.1. Note that the larger $\{?\}$ is, the less the revealed preference is clarified, which will usually cloud the decision process. Sometimes it may take discipline and/or new information for the decision maker to clarify his or her preference in order to reach the final solution. The concept of *pseudo-orders* (Roy and Vincke 1987) may help in resolving some of the problems.

Example 2.1 (*Pareto preference*). Let greater values be more preferred for each component y_i and assume that no other information on the preference is available or established. Then Pareto preference is defined by $y^1 > y^2$ if and only if $y_i^1 \geq y_i^2$ for all i and $y^1 \neq y^2$. Therefore,

$$\begin{aligned}\{>\} &= \{(y^1, y^2) | \; y_i^1 \geq y_i^2 \text{ for all } i \text{ and } y^1 \neq y^2\}\\ \{\sim\} &= \{(y, y) | y \in Y\}\\ \{?\} &= \{(y^1, y^2) | \text{there exist } i \text{ and } j \text{ such that } y_i^1 > y_i^2, \, y_j^1 < y_j^2\}\end{aligned}$$

Example 2.2 (*value function*). A value function v is a real-valued function which is defined on Y, the set of decision outcomes, such that $v(y^1) > v(y^2)$ if and only if y^1 is preferred to y^2 and $v(y^1) = v(y^2)$ if and only if y^1 is indifferent to y^2. It is easy to see that

$$\begin{aligned}\{>\} &= \{(y^1, y^2) \in Y \times Y | v(y^1) > v(y^2)\}\\ \{\sim\} &= \{(y^1, y^2) \in Y \times Y | v(y^1) = v(y^2)\}\\ \{?\} &= \varnothing\end{aligned}$$

Example 2.3 (*a lexicographic ordering*). Let $y = (y_1, y_2, \ldots, y_q)$ be indexed so that the kth component is overwhelmingly more important than the $(k + 1)$th component for $k = 1, 2, \ldots, q - 1$. A lexicographic ordering preference is defined as follows: the outcome $y^1 = (y_1^1, y_2^1, \ldots, y_q^1)$ is preferred to $y^2 = (y_1^2, y_2^2, \ldots, y_q^2)$ if and only if $y_1^1 > y_1^2$, or there is some k so that $y_k^1 > y_k^2$ and $y_j^1 = y_j^2$ for $j = 1, 2, \ldots, k - 1$. We see that

$$\begin{aligned}\{>\} &= \{(y^1, y^2) \in Y \times Y | y^1 \text{ is lexicographically preferred to } y^2\}\\ \{\sim\} &= \{(y, y) \in Y \times Y\}\\ \{?\} &= \varnothing\end{aligned}$$

Definition 2.3. Given a preference and a point $y^0 \in Y$, we define the *better, worse, equivalent,* and *indefinite* sets with respect to (*w.r.t.*) y^0 as:

1. $\{y^0 <\} = \{y \in Y | y^0 < y\}$ (the better set *w.r.t* y^0)
2. $\{y^0 >\} = \{y \in Y | y^0 > y\}$ (the worse set *w.r.t* y^0)
3. $\{y^0 \sim\} = \{y \in Y | y^0 \sim y\}$ (the equivalent set *w.r.t* y^0)
4. $\{y^0 ?\} = \{y \in Y | y^0 ? y\}$ (the indefinite set *w.r.t* y^0)
5. $\{y^0 > \sim\} = \{y^0 >\} \cup \{y^0 \sim\}$
6. $\{y^0 > ?\} = \{y^0 >\} \cup \{y^0 ?\}$
7. $\{y^0 > \sim ?\} = \{y^0 > \sim\} \cup \{y^0 ?\}$

Now we can define various solution concepts as follows.

Definition 2.4. Given a preference structure $\mathcal{P}(\{>\}, \{\sim\}, \{?\})$ defined on the outcome space Y, we define:

1. $y^0 \in Y$ is an $N[<]$-solution (point) if and only if $\{y^0 <\} \cap Y = \varnothing$; the collection of all such solutions is denoted by $N[<]$.
2. $y^0 \in Y$ is an $N[<\sim]$-solution if and only if $\{y^0 < \sim\} \cap Y = \{y^0\}$; the collection of all such solutions is denoted by $N[<\sim]$.
3. $y^0 \in Y$ is an $N[<?]$-solution if and only if $\{y^0 ? <\} \cap Y = \varnothing$; the collection of all such solutions is denoted by $N[?]$.
4. $y^0 \in Y$ is an $N[<\sim ?]$-solution if and only if $\{y^0 < \sim ?\} \cap Y = \{y^0\}$; the collection of all such solutions is denoted by $N[<\sim ?]$.

The following can be easily established (Chien 1987; Chien et al. 1990):
Theorem 2.1.

1. $N[<\sim?] \subseteq N[<?] \subseteq N[<]$
2. $N[<\sim?] \subseteq N[<\sim] \subseteq N[<]$
3. $N[<\sim?]$ contains at most one point

Remark 2.2. In application, we may first try to locate $N[<]$, then $N[< ?]$ or $N[< \sim]$, and finally $N[< \sim?]$, which, if nonempty, will contain only one clear superior solution, and no alternative optimals would occur.

3. VALUE FUNCTIONS

3.1. Revealed Preferences and Value Functions

Numerical systems have a prevailing impact on our culture and thinking. It is natural and important for us to ask, Is it possible to express our preference over the outcomes in terms of numbers so that the larger the number the stronger the preference, and if it is, how do we do it?

Suppose that our preference structure $\mathcal{P}$ can be represented by a value function $v{:}Y \rightarrow R$ so that $v(y^1) > v(y^2)$ if and only if y^1 is preferred to y^2 and $v(y^1) = v(y^2)$ if and only if y^1 is indifferent to y^2 and that Y is convex and v is continuous on Y. We immediately can obtain the following properties for $\mathcal{P}$ (see Example 2.2).

1. $\{?\} = \varnothing$
2. $\{>\}$ and $\{> \sim\}$ are transitive.
3. Let $\mathcal{Y}$ be the collection of all isovalued curves (or surfaces) of v in Y. Note that different isovalued curves never intersect. Since Y is convex and v is continuous, $V = v[Y]$ is an interval which contains dense countable rational numbers. Their corresponding isovalued curves are thus countable and dense in $\mathcal{Y}$.
4. $\{y <\}$ and $\{y >\}$ are open sets in Y for every $y \in Y$, if Y is open.

We summarize some important results on value functions in the following remark.

Remark 3.1. (a) Property 1 of the above is essential. If $\{?\} \neq \varnothing$, then value function representation cannot offer meaningful numerical ordering; (b) Properties 1 and 2 together mean that $\mathcal{P}$ is a weak order in literature; (c) Properties 1–3 conversely ensure the existence of value function representation; (d) Properties 1–2 and 4 conversely ensure the existence of *continuous* value function representation. The reader is referred to Fishburn (1970), Debreu (1954, 1960), Keeney and Raiffa (1976), Yu (1985), Gorman (1968), Yu and Takeda (1987), and others for further detailed discussion on the existence condition of value functions and their forms.

3.2. Methods of Constructing Value Functions

Assuming that the existence conditions are satisfied, we are interested in methods of constructing value functions to approximate the revealed preferences. Let us first roughly classify the existing methods before we illustrate some popular methods.

One large class of techniques for constructing value functions is a direct application of calculus. These methods are usually based on the construction of approximate indifference curves. This class includes methods using trade-off ratios, tangent planes, gradients, and line integrals. Some of these methods are discussed in Yu (1985).

A second class has specifically been developed for the case of additive value functions. One of the well-known methods in this class is the midvalue method (Keeney and Raiffa 1976; Yu 1985), which is based on the pairwise information of $\{\sim\}$ and $\{>\}$.

A third class takes into account the fact that usually the revealed preference contains conflicting information, making it a virtually impossible task to construct a consistent value function. The conflicting nature of the information may be due to an unclear perception on the part of the DM of his true preference structure or to imperfect and/or incorrectly used interaction techniques. The objective of these methods is to find a value function and/or ideal point that minimizes the inconsistencies. Some of the techniques, such as regression analysis, are based on statistical theory, others on mathematical programming models, such as least distance and minimal inconsistency methods (using some appropriate l_p-norm). Included are methods based on weight ratios, pairwise preference information, or the distance from a (perhaps unknown) idea/target point (see Yu 1985, Section 6.3.4 and the citations therein). Another group of methods in this class is that of eigenweight vectors (see, e.g., Saaty 1980; Cogger and Yu 1985; Yu 1985, Section 6.3.3.). Yet another group uses holistic assessment utilizing the orthogonal design of statistical experiment (see Yu 1985, Section 6.3.3.2 and citations therein).

Each method has its strengths and weaknesses. The selection of the best/correct method is truly an art and poses a challenge for the analyst and decision maker. Assume that there are q criteria under consideration and that more is better for each y_k ($k = 1, 2, \ldots, q$) with y_k taking values in the interval $[a_k, b_k]$. Thus, the outcome space is the rectangle $[a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_q, b_q]$. Assume that the value function is additive as follows:

$$v(y) = w_1 v_1(y_1) + w_2 v_2(y_2) + \ldots + w_q v_q(y_q)$$

where $v_k(y_k)$, $k = 1, 2, \ldots, q$ are the individual components and $(w_1, w_2, \ldots, w_q)$ is the weight distribution. Without loss of generality, we can assume that $v_k(a_k) = 0$ and $v_k(b_k) = 1$. To assess individual $v_k(y_k)$, there are several methods. For details, see Yu (1985, chap. 5).

Assume that we have successfully generated the individual function v_k ($k = 1, 2, \ldots, q$). We want to determine the weights $w_1, w_2, \ldots, w_q$ so that we can obtain the overall value function

$$v(y) = w_1 v_1(y_1) + w_2 v_2(y_2) + \ldots + w_q v_q(y_q)$$

by aggregating the individual components. There are many ways to assess the weights (see, e.g., Hwang and Masu 1979; Hwang and Yoon 1981; Yu 1973). Being limited by space, we shall discuss only two popular methods for different situations.

Method 1. Recall that $v_k(a_k) = v_k(y_k^0) = 0$ and $v_k(b_k) = v_k(y_k^1) = 1$. Set $y^0 = (y_1^0, \ldots, y_{\mathrm{q}}^0)$ and $y^1 = (y_1^1, \ldots, y_q^1)$. Denote

$$(y_k^1, y_{\bar{k}}^0) = (y_1^0, \ldots, y_{k-1}^0, y_k^1, y_{k+1}^0, \ldots, y_q^0)$$

then

$$v((y_k^1, y_{\bar{k}}^0)) = w_k v_k(y_k^1) = w_k$$

This simply means that we need only to know the value of the value function at point $(y_k^1, y_{\bar{k}}^0)$ to know w_k. If this is difficult to accomplish, we can alternatively identify $q - 1$ pairs of indifferent points, then, using the additive form of the value function, we can obtain $q - 1$ equations for the q unknowns $(w_1, w_2, \ldots, w_q)$. After normalizing, we obtain a unique solution.

Example 3.1. Assume that there are three criteria under consideration and that $Y_1 = Y_2 = Y_3 = [0, 2]$. Suppose that we have obtained the individual functions as

$$\begin{aligned} v_1(y_1) &= (5/6)y_1 - (1/6)\, y_1^2 \\ v_2(y_2) &= y_2 - (1/4)\, y_2^2 \\ v_3(y_3) &= (3/4)\, y_3 - (1/8)\, y_3^2 \\ (1, 2, 0) &\sim (1, 0, 2) \\ (1, 0, 1) &\sim (0, 2, 0) \end{aligned}$$

Using the additive form of the value function, we obtain

$$\begin{aligned} v((1, 2, 0)) &= w_1 v_1(1) + w_2 v_2(2) + w_3 v_3(0) = (2/3)w_1 + w_2 \\ v((1, 0, 2) &= w_1 v_1(1) + w_2 v_2(0) + w_3 v_3(2) = (2/3)w_1 + w_3 \end{aligned}$$

Therefore, (1, 2, 0) ~ (1, 0, 2) implies that $w_2 = w_3$. Similarly, from (1, 0, 1) ~ (0, 2, 0), we have $(2/3)w_1 + (5/8)w_3 = w_2$. Combined with $w_1 + w_2 + w_3 = 1$, we obtain the weights

$$w_1 = \frac{9}{41},\ w_2 = \frac{16}{41},\ w_3 = \frac{16}{41}$$

Therefore, the overall value function is given by

$$v(y) = \left(\frac{15}{82}\right) y_1 - \left(\frac{3}{82}\right) y_1^2 + \left(\frac{16}{41}\right) y_2 - \left(\frac{4}{41}\right) y_2^2 + \left(\frac{12}{41}\right) y_3 - \left(\frac{2}{41}\right) y_3^2$$

Method 2. Saaty's Eigenweight Vector Method (Saaty 1980). Suppose that the value function is of the form

$$v(y) = w_1 y_1 + w_2 y_2 + \ldots + w_q y_q$$

Thus, the value function is determined if the weights are known. Without losing generality, we can assume that $w_k > 0$ $(k = 1, 2, \ldots, q)$.

If weights are given, we can define the weight ratio by

$$w_{ij} = \frac{w_i}{w_j}$$

Then the weight ratio matrix $W = [w_{ij}]_{q \times q}$ is *consistent* in the sense that for any i, j, and k,

$$w_{ij} = w_{ji}^{-1}$$

$$w_{ik} = w_{ij} w_{jk}$$

On the other hand, given a consistent matrix, we can find the weight vector that generates it.

Theorem 3.1. Let $w_{q \times q}$ be any consistent matrix. Then

1. The maximum eigenvalue for W is q and all the other eigenvalues are 0.
2. The eigenvector corresponding to the maximum eigenvalue, which is unique after normalizing, is the weight vector $w = (w_1, w_2, \ldots, w_q)$, which generates W by $w_{ij} = w_i / w_j$.

Therefore, to find the weights $w_1, w_2, \ldots, w_q$, we need only to find the weight ratios w_{ij} $(i, j = 1, 2, \ldots, q)$. Saaty proposed the following procedure:

Step 1: For $i < j$, estimate or elicit the weight ratio w_{ij}, by a_{ij}. Let $a_{ii} = 1$ for all i and $a_{ij} = a_{ij}^{-1}$ for $i > j$. Denote $A = [a_{ij}]_{q \times q}$.

Step 2: Since A is found as an approximation for W, when the consistency conditions are "almost" satisfied for A, one would expect that the normalized eigenvector corresponding to the maximum eigenvalue of A, denoted by $\lambda_{\max}$, will be close to w. Thus, w can be approximated by the normalized eigenvector corresponding to $\lambda_{\max}$.

Here the problem is, however, that the estimated weight ratios are generally not consistent, especially when criteria are not few, since human perception and judgment are subject to change when the information input or psychological state of the decision maker change. The using of the above procedure is partially justified by the following theorem.

Theorem 3.2 (Saaty 1980).

1. Let $\overline{w} = (\overline{w_1}, \overline{w_2}, \ldots, \overline{w_q})$ be the normalized eigenvector corresponding to $\lambda_{\max}$ of A. Then $\overline{w_i} > 0$ for all i.
2. Given i and k, if $a_{ij} \geq a_{kj}$ for all j, then $\overline{w_i} \geq \overline{w_k}$.

Example 3.2 (Saaty 1980). Three criteria are under consideration. Assume that the following estimations of weight ratios are obtained:

$$a_{12} = 9,\ a_{13} = 7,\ a_{23} = 1/5$$

We construct the following matrix:

$$A = \begin{bmatrix} 1 & 9 & 7 \\ 1/9 & 1 & 1/5 \\ 1/7 & 5 & 1 \end{bmatrix}$$

To find $\lambda_{\max}$, we solve $\det[A - \lambda\, I] = 0$, or $(1 - \lambda)^3 - (1 - \lambda) + 9/35 + 35/9 = 0$. The maximum solution is $\lambda_{\max} \approx 3.21$. Then we find the corresponding normalized eigenweight vector $\overline{w}$ with

$$\overline{w_1} = 0.77,\ \overline{w_2} = 0.05,\ \overline{w_3} = 0.17$$

The value function is therefore given by

$$v(y) = 0.77 y_1 + 0.05 y_2 + 0.17\, y_3$$

Remark 3.2. In Saaty's method, one needs to solve a nonlinear equation to find the maximum eigenvalue. Cogger and Yu (1985) proposed another eigenweight vector method that can be applied to find the eigenweight vector easily by using the deductive formula. The weights determined by the latter method are different from those determined by the former. It is important to note that there is no need to have all a_{ij}, $i < j$, in order to apply the above method. For such exploration, the reader is referred to Takeda and Yu (1995).

Many other methods have been developed that we cannot discuss here due to limited space. The interested reader is referred to Yu (1985), Yu and Takeda (1987, 1995), Hwang and Masud (1979), Hwang and Yoon (1981), and Yu (1973) and quotes therein. It should also be stressed that all these methods are approximation methods. When inconsistencies occur, we may wish to represent the decision maker's preference in ways other than value functions. This will be discussed shortly.

4. SATISFICING AND COMPROMISE SOLUTIONS

Each human being has a set of ideal goals to achieve and maintain (see Yu 1990, 1995, and 1985, chap. 9 and citations therein). When a perceived state (or value) has unfavorably deviated from its targeted or ideal state (or value), a charge (tension or pressure) will be produced to prompt an action to reduce or eliminate the deviation. The behavior of taking actions, including adjustment of the ideal values to move the perceived states to the targeted ideal states is called *goal-seeking behavior.*

This goal-seeking behavior has an important and pervasive impact on human decision making. In this section, we shall focus on two concepts that model human goal-seeking behavior. These are satisficing goal programming and compromise models.

4.1. Satisficing Models

Example 4.1. Consider the problem of selecting an athlete for a basketball team. Assume that quickness, f_1, and accuracy of shooting, f_2, of the player are the main concerns. Let both f_1 and f_2 be indexed from 0 to 10, where the higher indices, the better the player. Assume that a player, x, will be selected with satisfaction if $f_1(x) \geq 9$ or $f_2(x) \geq 9$, or $f_1(x) + f_2(x) \geq 15$. Let $f = f(x) = (f_1(x), f_2(x))$ be the score vector of x and $G_1(f) = f_1 - 9$, $G_2(f) = f_2 - 9$, and $G_3(f) = f_1 + f_2 - 15$. If an athlete's score vector is in at least one of the sets , $S_i = \{f | G_i(f) \geq 0\}$, $i = 1, 2, 3$, then he or she is to be selected with satisfaction. Note that these sets are defined in terms of the scores (outcomes) instead of the candidates (decision alternatives). Here S_i, $i = 1, 2, 3$, so defined, are called satisficing sets.

Note that the specifying of lower bounds 9, 9, and 15 in the above example is an act of *goal setting*. Goal setting for satisficing models is defined as the procedure of identifying a satisficing set S such that whenever the decision outcome is an element of S, the decision maker will be happy and satisfied and is assumed to have reached the optimal solution. When S contains only a single point, the point is called the goal or target point.

Let our MCDM problem be specified by the criteria $f = (f_1, f_2, \ldots, f_q)$ and the feasible set is $X = \{x \in R^n | g(k) \leq 0\}$, where $g = (g_1, g_2, \ldots, g_m)$. Note that the outcome of a decision x is specified by $y(x) = f(x)$.

To specify S, one can start with each individual f_i and find its satisficing or acceptable intervals. One can then consider two or more criteria simultaneously for their corresponding trade-off. In a general form, in the final satisficing set S can be defined as the union of

$$S_k = \{f | G_k(f) \geq 0\}, k = 1, 2, \ldots, r$$

where G_k is a vector function that reflects the trade-off over f. Note that when the union contains a single set (i.e., $r = 1$), S is defined by a system of inequalities.

Once S, the satisficing set, has been determined, finding a satisficing solution x^0, such that $f(x^0) \in S$, is a mathematical programming problem.

Since $S = \cup \{S_k | k = 1, 2, \ldots, r\}$, $Y \cap S \neq \varnothing$ if and only if there is an S_k such that $Y \cap S_k \neq \varnothing$. Because we need only one point of $Y \cap S$ when it is nonempty, we can verify individually if $Y \cap S_k \neq \varnothing$, $k = 1, 2, \ldots, r$, and find one point from the nonempty intersections.

To verify whether or not $Y \cap S_k \neq \varnothing$ is empty and to find a point in $Y \cap S_k$, if it is nonempty, we can use the following mathematical program. First rewrite each satisficing set as

$$S_k = \{f(x) | G_{kj}(f(x)) \geq 0, j = 1, 2, \ldots, J_k\}$$

where G_{kj}, $j = 1, 2, \ldots, J_k$, are components of G_k, and J_k is the number of components of G_k.

Program 4.1. $V_k = \min d_1 + d_2 + \ldots + d_{J_k}$

$$\text{s.t. } G_{kj}(f(x)) + d_j \geq 0, j = 1, 2, \ldots, J_k$$

$$x \in X, d_j \geq 0, j = 1, 2, \ldots, J_k$$

It is readily verified that $Y \cap S_k \neq \varnothing$ if and only if $V_k = 0$. We thus have:

Theorem 4.1. Let $S = \cup [S_k | k = 1, 2, \ldots, r\}$. Then a satisficing solution exists if and only if there is at least one $k \in \{1, 2, \ldots, r\}$ such that Program 4.1 yields $V_k = 0$.

Observe that when one satisficing solution is found, the decision problem is solved, at least temporarily. Otherwise the decision maker can activate either positive problem solving or negative problem avoidance to restructure the problem. The former will enable careful restudy and restructuring of the problem so as to find a solution, perhaps a new one, that lies in the satisficing set. The latter will try to reduce the aspiration levels or play down the importance of making a good decision, thus, lowering the satisficing set to have a nonempty intersection with Y. While the psychological attitude in problem solving and in problem avoidance may be different (see Yu 1990, 1995, 1985, chap. 9 for further discussion), the consequences are the same. That is, eventually the newly structured problem enables the decision maker to have $Y \cap S \neq \varnothing$. The following provides some helpful information for decision maker for restructing S and Y.

1. Depending on the formation of S, if Program 4.1 is used to identify a satisficing solution, one can produce x^k, $k = 1, 2, \ldots, r$, which solves Program 4.1 with S_k as the satisficing set. One then can compute $f(x^k)$, $G_k(f(x^k))$, the distance between $f(x^k)$ and S_k, and the trade-off among the $G_{kj}(f(x))$, $j = 1, 2, \ldots, J_k$, at x^k. All of these can be helpful for the decision maker to reframe S_k.
2. Suppose that it has been revealed or established that more is better for each criterion and that the revealed preference contains at least the Pareto preference. One can then generate (a) all or some representatives of the $N[<]$ points, the corresponding $f(x)$, and optimal weights $\Lambda(x)$ for the $N[<]$ points (see Sections 5 and 6 for further discussion); (b) the value $f_k^* = \max \{f_k(x) | x \in X\} (k = 1, 2, \ldots, q)$, which gives the best value of $f(x)$ over X. This information can help the decision maker to restructure his or her S. Relaxing of the target goal with reference to the ideal point $(f_1^*, f_2^*, \ldots, f_g^*)$ sequentially and interactively with the decision maker, until a satisficing solution is obtained, is certainly an art. It is especially useful when X and f are fairly fixed and not subject to change.
3. If possible (for instance, in bicriteria cases), displaying Y and S, even if only partially, can help the decision maker conceptualize where Y and S are, to make it easier to restructure the problem.

One must keep in mind that with X, $f(x)$, and S, one can generate as much information as one wishes, just as one can generate as many statistics from sample values as one wants. However, relevant information (e.g., useful statistics) that can positively affect the problem solving may not be too much. Irrelevant information can become a burden for decision analysis and decision making.

Example 4.2. Consider a sample production problem. Let the resource constraints in the amount produced, x_1 and x_2, be given as follows:

$$\begin{aligned} g_1(x) &= 3x_1 + x_2 - 12 \leq 0 \\ g_2(x) &= 2x_1 + x_2 - 9 \leq 0 \\ g_3(x) &= x_1 + 2x_2 - 12 \leq 0 \\ & x_1 \geq 0, x_2 \geq 0 \end{aligned}$$

The following two objective functions are related to gross output and net profit respectively:

$$\begin{aligned} y_1 &= f_1(x) = x_1 + x_2 \\ y_2 &= f_2(x) = 10x_1 - x_1^2 + 4x_2 - x_2^2 \end{aligned}$$

Assume that the decision maker specifies his or her satisficing set as

$$S = \{(y_1, y_2) | y_1 \geq 10, y_2 \geq 30\}$$

Then, using Program 4.1, we solve the problem of

$$\begin{aligned} V = {} & \min d_1 + d_2 \\ \text{s.t.} \quad & x_1 + x_2 + d_1 \geq 10 \\ & 10x_1 - x_1^2 + 4x_2 - x_2^2 + d_2 \geq 30 \\ & 3x_1 + x_2 \leq 12 \\ & 2x_1 + x_2 \leq 9 \\ & x_1 + 2x_2 \leq 12 \\ & x_1 \geq 0, x_2 \geq 0 \end{aligned}$$

It turns out that $V > 0$, thus $S \cap Y = \emptyset$. To find a satisficing solution, we must restructure X, f, and/or S. If X and f are fairly fixed, we may offer the information that $f_1^* = 7$ and $f_2^* = 26.5$, which simply means that any outcome y with $y_1 > 7$ or $y_2 > 26.5$ is unobtainable. Thus, the goals $y_1 > 10$ and $y_2 > 30$ must come down to allow a satisficing solution. The restructuring of S will continue until a satisficing solution is found.

4.2. Compromise and Goal Programming Solutions

Assume that each criterion $f_i (i = 1, 2, \ldots, q)$ is characterized by "more is better." Let $y^* = (y_1^*, y_2^*, \ldots, y_q^*)$ where $y_i^* = \sup\{f_i(x) | x \in X\}$ with X the feasible set. The point y^* is called *ideal* (or *utopia*) *point* because it is usually unattainable even though y_i^* may individually be attainable.

Example 4.3. Consider the following maximization problem with two criteria:

$$\begin{aligned} \max \quad & y_1 = f_1(x) = 6x_1 + 4x_2 \\ \max \quad & y_2 = f_2(x) = x_1 \\ \text{s.t.} \quad & g_1(x) = x_1 + x_2 \leq 100 \\ & g_2(x) = 2x_1 + x_2 \leq 150 \\ & x_1, x_2 \geq 0 \end{aligned}$$

To find y_1^*, we solve

$$\begin{aligned} & \max y_1 = f_1(x) = 6x_1 + 4x_2 \\ & \text{s.t. } x_1 + x_2 \leq 100 \\ & \quad 2x_1 + x_2 \leq 150 \\ & \quad x_1, x_2 \geq 0 \end{aligned}$$

The solution is $x_1^* = x_2^* = 50$. Thus, $y_1^* = f_1(x^*) = 500$. Similarly, we find $y_2^* = 75$. Therefore, the ideal point is $y^* = (500, 75)$. Note that y^* is not attainable.

In group decision problems, if each criterion represents a player's payoff, then y^*, if attainable, would make each player happy because it would simultaneously maximize each player's payoff. Even if one is a dictator, he cannot do better than y^* for himself. As y^* is usually not attainable, a compromise is needed if no other alternative is available to dissolve the group conflict. This offers a natural explanation of why the solution to be introduced is called a *compromise solution* (Yu 1973). The reader can extend this explanation easily to multiple-criteria problems.

Now given $y \in Y$, the outcome space, the *regret* of using y instead of obtaining the *ideal* point y^* may be approximated by the distance between y and y^*. Thus, we define the (group) regret of using y by

$$r(y) = \|y - y^*\|$$

where $\|y - y^*\|$ is the distance between y and y^* according to some specific norm. Typically, the l_p-norm will be used in our discussion because it is easy to understand, unless otherwise specified. To make this more specific, define for $p \geq 1$,

$$r(y; p) = \|y - y^*\|_p = \left[\sum |y_i - y_i^*|^p \right]^{1/p}$$

and

$$r(y; \infty) = \max \{|y_i - y_i^*|, i = 1, 2, \ldots, q\}$$

Then $r(y; p)$ is a measurement of regret from y to y^* according to the l_p-norm.

Definition 4.1. The compromise solution with respect to the l_p-norm is $y^p \in Y$, which minimizes $r(y; p)$ over Y, or is $x^p \in X$, which minimizes $r(f(x); p)$ over X. When the ideal point y^* is replaced by a specific goal or target point, the resulting compromise solution is called the goal programming solution with respect to the goal point.

Remark 4.1. In group decision problems, $r(y; p)$ may be interpreted as group regret and the compromise solution y^p, is the one which minimizes the group regret in order to maintain a cooperative group spirit. As the parameter p varies, the solution y^p can change.

Note that $r(y; p)$ treats each $|y_i^* - y_i|$ as having the same importance in forming the group regret and the multiple criteria problems. If the criteria have different degrees of importance, then a weight vector may be assigned to signal the different degrees of importance. The regret function $r(y; p)$ can be modified in a natural way (Yu 1985). Observe that using weights in the regret function has the effect of changing the scale of each criterion. Conversely, the scale can be adjusted so that the regret function is reduced to that of equal weight. Thus, in studying the properties of compromise solutions, without loss of generality, one may focus on the equal weight case. We shall assume the equal weight case from now on, unless specified otherwise.

Remark 4.2. Observe that the compromise solution is not scale *independent*. Scale independence, an important criterion in group decision problems, can prevent players from artificially changing the scale to obtain a better arbitration for themselves. Note that when the scale of $f_i(x)$ or y_i is changed and the weight of importance of each criterion is changed, so is the compromise solution.

Remark 4.3. Compromise solutions with proper assumptions enjoy a number of properties, such as feasibility, least group regret, no dictatorship, Pareto optimality, uniqueness, symmetry, independence of irrelevant alternatives, continuity, monotonicity, and boundedness. In terms of parameter p (associated with p-norms), we may say that when $p = 1$, the sum of the group utility is mort emphasized, and when $p = \infty$, the individual regret of the group is mort emphasized. For the details and further exploration, see Yu (1985, 1973) and Freimer and Yu (1976).

4.3. Computing Compromise Solutions and Goal Programming Solutions

Now we turn to the computation of compromise solutions. We first consider the case when the target point is the ideal point, then the general case.

4.3.1. The Ideal Point as the Target Point

To find a compromise solution when the ideal point is the target point, we have to solve $q + 1$ mathematical programming problems. The first q problems are to find the utopia or ideal point where $y^* = (y_1^*, y_2^*, \ldots, y_q^*)$, where $y_i^* = \sup\{f_i(x) | x \in X\}$. The last one is to find the compromise solution y^p. Suppose that X contains only countable points. We will have $q + 1$ integer programming problems. Otherwise, if X is a region, we will have $q + 1$ nonlinear programming problems.

If X and $f_i(x)$, $i = 1, 2, \ldots, q$ have some special structure, then more efficient computational techniques are available. For instance, if X is a convex set and each $f_i(x)$ is concave, then we first have q concave programming and then a convex programming (because $r(f(x); p)$ is convex under the assumptions). If X is a polyhedron defined by a system of linear inequalities and each $f_i(x)$ is linear, then the ideal points y^* can be found by q simple linear programming problems. Furthermore, the compromise solutions of y^1 and y^∞ can be found by a linear programming problem (the other compromise solutions, y^p, $1 < p < \infty$, can be found by convex programming) (Yu 1985).

Example 4.4. Reconsider Example 4.3. The ideal point is (500, 75). Now we show how to find compromise solutions y^1, y^2, and y^∞.

1. To find the y^1 compromise solution, we solve

$$\begin{array}{ll} \min & \{[500 - (6x_1 + 4x_2)] + [75 - x_1]\} \\ \text{s.t.} & x_1 + x_2 \leq 100 \\ & 2x_1 + x_2 \leq 150 \\ & x_1, x_2 \geq 0 \end{array}$$

We obtain $x_1^* = 50$, $x_2^* = 50$ and the compromise solution is then $y^1 = (500, 50)$.

2. Similar to (1), we obtain compromise solution $y^2 = (490, 55)$.
3. For the y^∞ compromise solution, we solve

$$\begin{aligned} \min\max\quad & \{500 - (6x_1 + 4x_2),\ 75 - x_1\} \\ \text{s.t.}\quad & x_1 + x_2 \leq 100 \\ & 2x_1 + x_2 \leq 150 \\ & x_1, x_2 \geq 0 \end{aligned}$$

This problem can be simplified as

$$\begin{aligned} \min\quad & v \\ \text{s.t.}\quad & v \geq 500 - (6x_1 + 4x_2) \\ & v \geq 75 - x_1 \\ & x_1 + x_2 \leq 100 \\ & 2x_1 + x_2 \leq 150 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Note that this is a linear program. We can obtain the solution easily as $x_1^* = 58.33$, $x_2^* = 33.33$, and $y^\infty = (483.33, 58.33)$.

4.3.2. General Target Points and Goal Programming

The computation for the compromise solution with general target point can be slightly more complicated. This is due to the possibility that $y_i^* \geq y_i$ for all $y \in Y$ and $i = 1, 2, \ldots, q$ no longer hold. In this case, we need some transformation of variables.

Let d_i^+ and d_i^- be respectively the positive and negative pairs of $y_i - y_i^*$, i.e., $d_i^+ = y_i - y_i^*$ if $y_i > y_i^*$ and 0 if otherwise, and $d_i^- = y_i^* - y_i$ if $y_i \leq y_i^*$ and 0 if otherwise. Now given a target point y^*, the compromise solution with l_p-norm can be found by solving the following (Yu 1985):

Program 4.2.

$$\begin{aligned} \min\quad & \sum (d_i^+ + d_i^-)^p \\ \text{s.t.}\quad & y_i^* - f_i(x) = d_i^- - d_i^+,\ i = 1, 2, \ldots, q \\ & d_i^+, d_i^- \geq 0,\ i = 1, 2, \ldots, q \\ & g(x) \leq 0 \end{aligned}$$

Remark 4.4.

1. Suppose that all $f_i(x)$, $i = 1, 2, \ldots, q$, and $g(x)$ are linear. Then Program 4.2 for l_l-compromise solutions reduces to a linear program typically known as (linear) *goal programming*. By adding weights to or imposing lexicographical ordering on the criteria, one can generalize the concept discussed here to a variety of goal programming formats. Goal programming and compromise solutions are closely related. The major advantages of goal programming and l_l-compromise solutions are that they are easily understood and can be easily computed by linear programs.
2. When all $f_i(x)$, $i = 1, 2, \ldots, q$, and $g(x)$ are linear, Program 4.2 for l_2-compromise solutions becomes a quadratic program. The program can be solved without major difficulty.
3. When all $f_i(x)$, $i = 1, 2, \ldots, q$, and $g(x)$ are linear, Program 4.2 for l_∞-compromise solutions becomes a linear program.

4.3.3. Interactive Methods of Compromise and Goal Programming Models

For nontrivial decision problems, an optimal decision usually cannot be reached by applying the compromise solution only once. The following interactive method, according to Figure 1, may be helpful.

Box (*1*): Identifying X, f, and y^* is an art that requires careful observation and conversation with the decision maker. The target point y^*, weight vector w, and p for l_p-norm may be more difficult to specify precisely. However, one can use the ideal point as the first step of approximation for y^*, select some representative weights for w, and let $p = 1$, 2, or ∞ to begin.

Box (*2*): One can locate compromise solutions according to Program 4.2 for the specified values of y^*, w, and p.

Box (*3*): All relevant information obtained in Box (1) and (2) can be presented to the decision maker. These include X, f, Y, ideal points y_i^*, $i = 1, 2, \ldots, q$, the optimal vector value that maximizes the ith criterion, and the compromise solutions y^1, y^2, and y^* with their weights and norms.

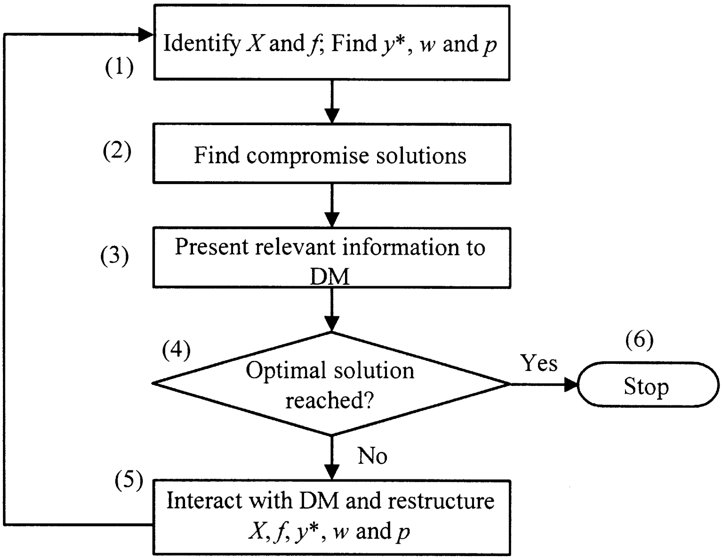

(1)
Identify X and f; Find y*, w and p
(2)
Find compromise solutions
(3)
Present relevant information to DM
(4)
Optimal solution reached?
Yes
(6)
Stop
No
(5)
Interact with DM and restructure X, f, y*, w and p

Figure 1 An Interactive Method for Compromise Solutions.

Boxes (4) and (6): If an optimal solution is reached in Box (3) and the decision maker is satisfied, the process is terminated at Box (6); otherwise, we go to Box (5) to obtain more information.

Box (5): Through conversation, we may obtain new information on X, f, y^*, weight vector w, and norm parameter p. Note that each of these five elements may change with time. To help the decision maker locate or change the target point, the "one-at-a-time method may be used. That is, only one value of f_i, or y_i, $i = 1, 2, \ldots, q$, is to be changed and the rest remain unchanged. One may ask the decision maker, "Since no satisfactory solution is obtained with the current target point y^*, would it be possible to decrease (or increase) the value of y_i^*? And, if so, by how much?" This kind of suggestive question may help the decision maker to think carefully of possible changes in the target points. The other method is called the pairwise trade-off method. We first select two criteria, say f_i and f_j, and then ask: "To maintain the same degree of satisfaction for the target point everything else being equal, how many units of f_j must be increased in order to compensate a one-unit decrease in f_i?" Again, this kind of suggestive question can help the decision maker to think carefully of the possible changes in the target points. Finally, we recall that the target point may be regarded as the satisficing set containing a single satisficing solution. Thus, by locating the satisficing set, we may locate the target point.

Example 4.5. Let us use Example 4.3 to illustrate the interactive method.

Step 1: Identify X, f, y^*, w, and p. The decision space X is given by

$$\begin{aligned} g_1(x) &= x_1 + x_2 \le 100 \\ g_2(x) &= 2x_1 + x_2 \le 150 \\ x_1, x_2 &\ge 0 \end{aligned}$$

and the objective functions are

$$\begin{aligned} y_1 &= f_1(x) = 6x_1 + 4x_2 \\ y_2 &= f_2(x) = x_1 \end{aligned}$$

Through conversation with the decision maker, assume that the ideal point will be used as the approximation for the target point. The ideal point (y_1^*, y_2^*) is (500, 75). We also assume that the two criteria are of the same weights. We shall use $p = 1, 2$, and ∞ for l_p-norm.

Step 2: Find compromise solutions. The computation of compromise solutions for $p = 1, 2$, and ∞ was illustrated in Example 4.4. The compromise solutions are $y^1 = (500, 50)$, $y^2 = (490, 55)$, and $y^\infty = (483.33, 58.33)$.

Step 3: Present relevant information to DM. All relevant information obtained in steps 1 and 2 will be presented to the decision maker. This includes X, f, Y, ideal points and the compromise solutions y^1, y^2, and y^∞ with their weights (w) and norms (p).

Step 4: Optimal solution reached? Suppose the decision maker is not satisfied with the compromise solutions presented. Assume that he or she is not satisfied with the target point (y^*) and the weight vector (w). The decision maker prefers to have the target point a little bit closer to the set Y and also feels that more weight should be given to the first criterion (f_1) rather than having equal weights.

Step 5: Through conservation with the decision maker, new relevant information on the target point (y^*) and the weight vector (w) will be obtained. Assume that the information on X, f, and p remains the same and that the new y^* is (450, 75) and the new w is (2/3, 1/3). With this new information compromise solutions y^1, y^2, and y^∞ can be recalculated and the new solutions will be presented to the decision maker. If the decision maker is satisfied with the new solutions, the procedure will be terminated; otherwise, more information from the decision maker on X, f, y^*, w, and p will be needed. We go to Box 5 and the process continues.

4.4. Further Comments

Utilizing human goal-seeking behavior, we have described the main concepts of satisficing and compromise solutions. In the real world, although specifying the satisficing set or the target point is not trivial, we can always start with the ideal point and then interact with the decision maker to gradually reach the final decision. For more details see Yu (1985, chap. 4).

There is a rich literature related to satisficing and compromise models. For instance, for goal programming see Charnes (1975), Ignizio (1976), Lee (1972) and those quoted therein; for different kinds of norms or penalty functions see Hwang and Yoon (1981), Gearhart (1979), Pascoletti and Serafini (1984), and the citations therein; for restructuring ideal points see Zeleny (1975).

Finally, observe that methods of this and the previous section are called methods of one-dimensional comparison, which tries to convert the preferences into a single-criterion numerical ordering. Other methods, such as Elimination et Choice Translating Reality (ELECTRE) (see Roy 1971 for further details) also belong to this class.

5. DOMINATION STRUCTURES

As mentioned in the introduction, preferences may be represented by one-dimensional comparison, which we discussed in the previous two sections, or in terms of multidimensional comparison, which we will discuss in this and the following section. Note that in one-dimensional comparison, we implicitly or explicitly assume that $\{?\} = \varnothing$ and that no ambiguity exists in preference. Once the value function or proper regret function is determined, MCDM becomes a one-dimensional comparison or a mathematical programming problem. In this section we shall tackle the problems with $\{?\} \neq \varnothing$.

Recall Definition 2.3 and $\{y^0<\}$, $\{y^0>\}$, $\{y^0 \sim\}$, and $\{y^0?\}$ are respectively the sets of points in Y that are better (preferred), worse (less preferred), equivalent and indefinite to y^0. For simplicity, we rewrite:

$$\begin{aligned}\{y^0<\} &= y^0 + P(y^0)\\ \{y^0>\} &= y^0 + D(y^0)\\ \{y^0\sim\} &= y^0 + I(y^0)\\ \{y^0?\} &= y^0 + U(\mathrm{y}^0)\end{aligned}$$

where $P(y^0)$, $D(y^0)$, $I(y^0)$, and $U(y^0)$ are respectively the sets of preferred, dominated, indifferent (equivalent), and unclarified (indefinite) factors.

Example 5.1

1. In Pareto preference (Example 2.1), for each y, $P(y) = \Lambda^{\geqq} = \{d \in R^q | d_i \geq 0$ for all i and $d \neq 0\}$, $D(y) = \Lambda^{\leqq} = \{d \in R^q | d_i \leq 0$ for all i and $d \neq 0\} = -P(y)$, $I(y) = \{0\}$, $U(y) = R^q \backslash (\Lambda^{\geqq} \cup \Lambda^{\leqq} \cup \{0\})$.
2. If the preference is represented by a concave differentiable value function $v(y)$, then $D(y)$ contains $\{d | \nabla v(y) \cdot d < 0\}$ which is a "half" space no matter what the dimensionality of Y is. Except linear $v(y)$, $D(y)$ is a function of y and in general $D(y) \neq -P(y)$. Note that in general $D(y)$ and $P(y)$ are not convex cones. Nevertheless, if we use the *tangent* cones of $\{y>\}$ and

$\{y<\}$ at y to represent the local sets of the preferred and dominated factors, denoted by $LD(y)$ and $LP(y)$, respectively, then $LD(y)$ and $LP(y)$ can be convex cones and $LD(y) = -LP(y)$. For a detail of such treatment, see Yu (1985, chap. 7).

In order to simplify our presentation, we shall assume the following throughout this section.

Assumption 5.1. For each $y \in Y$, $D(y)$ and $P(y)$ are convex cones. Furthermore, $D(y) = -P(y)$.

Remark 5.1. Pareto preference (Example 5.l number 1) satisfies Assumption 5.l, so does the preference represented by a linear value function. As indicated in Example 5.l, number 2, one can use $LD(y)$ and $LP(y)$ to replace $D(y)$ and $P(y)$ respectively when the assumption does not hold. If we do so, our results stated in this section are still valid, but only in a local sense (local optimal vs. global optimal). For the details of such treatment and conditions for the local results to be valid as the global results refer to Yu (1985, chap. 7).

Definition 5.1. A point $y^0 \in Y$ is a *nondominated point* or N-points if $y^0 N[<]$, where $N[<]$ is as defined in Definition 2.4.

In Section 5.1, we discuss some basic properties of N-points when $D(y)$ is constant. In Section 5.2, we explore the case when $D(y)$ varies with y.

5.1. Constant Cone Domination Structures and Solutions

For simplicity, when $D(y) = \Lambda$ for all y, we shall call Λ the dominated cone and denote the set of all nondominated points or (N-points) by $N(Y, \Lambda)$. In this subsection we shall discuss two sets of conditions that characterize N-points. To facilitate our discussion, let us introduce the following concepts:

Definition 5.2

1. Y is Λ-convex if $Y + \Lambda$ is a convex set.
2. The polar cone of Λ is defined by $\Lambda^* = \{\lambda | \lambda \cdot d \leq 0 \text{ for all } d \in \Lambda\}$; interior is denoted by int Λ^*.
3. Given $\lambda \in \Lambda^*$, $\lambda \neq 0$, the set of all maximum points in Y with respect to linear functional $\lambda \cdot y$ is denoted by $Y^0(\lambda)$.

For further discussion on cone convexity, see Yu (1974, 1985).

Theorem 5.1

1. $\cup \{Y^0(\lambda) | \lambda \in \text{int } \Lambda^*, \lambda \neq 0\} \subseteq N(Y, \Lambda)$
2. If Y is Λ-convex and Λ is pointed (that is, $\Lambda \cap (-\Lambda) = \{0\}$), then

$$N(Y, \Lambda) \subseteq \cup \{Y^0(\lambda) | \lambda \in \Lambda^*, \lambda \neq 0\}$$

Note that the theorem states that conditions for N-points to be found by maximizing a linear function over Y; 1 is a sufficient condition and 2 a necessary condition. If one is interested in the entire set of N-points, then the theorem serves as an approximation for the set; 1 is the inner approximation and 2 the outer approximation for $N(Y, \Lambda)$. The results are derived in Yu (1974, 1985); their further refinement can be found in Hartley (1978).

A cone that is a closed polyhedron is called a *polyhedral cone*. If Λ is a polyhedral cone, so is Λ^*. In this case, there are a finite number of vectors $\{H^1, H^2, \ldots, H^p\}$ so that Λ^* is the cone generated by $\{H^1, H^2, \ldots, H^p\}$. That is,

$$\Lambda^* = \{a_1 H^1 + a_2 H^2 + \ldots + a_p H^p | a_i \in R^1, a_1 \geq 0, i = 1, 2, \ldots, p\}$$

$\{H^1, H^2, \ldots, H^p\}$ will be called a *generator* of Λ^*. From now on, unless otherwise specified, whenever Λ is a polyhedral cone, we shall assume that $\{H^1, H^2, \ldots, H^p\}$ is a generator for Λ^*.

Let $r(j)$ be the vector in R_k^{p-1} representing $\{r_k \in R^1 | k \in \{1, 2, \ldots, p\} \backslash \{j\}\}$. Define $Y(r(j)) = \{y \in Y | H^k \cdot y \geq r_k, k \in \{1, 2, \ldots, p\} \backslash \{j\}\}$. Note that r_k may be regard as an aspiration level or satisfying level for $H^k \cdot y$.

Theorem 5.2 (Yu 1974). Let $\Lambda' = \Lambda \cup \{0\}$ be a polyhedral cone. Then $y^0 \in N(Y, \Lambda)$ if and only if for any arbitrary $j \in \{1, 2, \ldots, p\}$, there is $r(j)$ such that y^0 uniquely maximizes $H^j \cdot y$ over $Y(r(j))$.

This theorem essentially says that in search of $N(Y, \Lambda)$ by mathematical programming, constraints and objective functions are interchangeable.

5.2. Variable Cone Domination Structures and Methods for Seeking Good Solutions

In this subsection we describe a convergent method and a heuristic method to locate $N(Y, D(\cdot))$. To begin, let us assume that each $D(y)$ contains $\Lambda^{\leqq}$. Define $\Lambda^0 = \cap \{D(y)|y \in Y\}$ and set $Y^0 = Y$. Note that $\Lambda^{\leqq} \subseteq \Lambda^0$. Recursively, let us define, for $n = 0, 1, 2, \ldots$

$$\Lambda^n = \cap \{D(y)|y \in Y^n\} \text{ and } Y^{n+1} = N(Y^n, \Lambda^n)$$

Then

$$Y^{n+1} \subseteq Y^n \text{ and } \Lambda^n \subseteq \Lambda^{n+1}$$

It follows that the sequences converge. Denote by $\lim Y^n$ and $\lim \Lambda^n$ the limits. It can be shown (Yu 1985) that for each n, $N(Y, D(\cdot)) \subseteq Y^n$ and $N(Y, D(\cdot)) \subseteq \lim Y^n$. This observation allows us to locate the N-points according to Flowchart 1 (method l) of Figure 2.

Flowchart 1 is self-explanatory. Although the method guarantees a convergent set containing the N-points, it may not be effective and efficient. The following heuristic method according to Flowchart 2 (method 2) can be more efficient in reaching the final decision at the risk of missing some N-points. Let us explain Flowchart 2.

Box (*0*): We first explore the relationship among the criteria and locate some plausible aspiration levels and trade-offs among the criteria. Then we use Theorem 5.1 or 5.2 to locate a set of initial "good" alternatives corresponding to different aspiration levels and trade-offs, using mathematical programming if necessary. One can also start with the points that respectively maximize the individual criteria. This initial set is denoted by Z^0.

Box (*1*): We should first estimate $D(y)$ for some representative points in Z^n and find their intersection for Λ^n. The more representative points considered, the smaller is Λ^n and the less chance there is to miss N-points; but it may take longer to find the final solutions. In estimate $D(y)$, we can use the definition to locate it directly, or we can use the bounds of trade-off ratios to locate it indirectly. For further discussion, see Yu (1985).

Box (*2*): Here we can use Theorem 5.1 or 5.2 to locate the entire set of Z^{n+1} or to locate a representative set for Z^{n+1}, depending on how much we want to avoid missing some N-points.

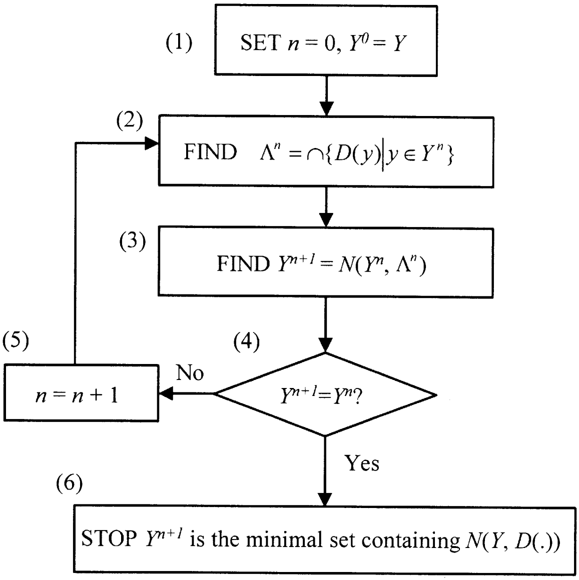

(1)
SET $n = 0$, $Y^0 = Y$
(2)
FIND $\Lambda^n = \cap\{D(y)|y \in Y^n\}$
(3)
FIND $Y^{n+1} = N(Y^n, \Lambda^n)$
(4)
$Y^{n+1} = Y^n$?
No
(5)
$n = n + 1$
Yes
(6)
STOP Y^{n+1} is the minimal set containing $N(Y, D(.))$

Figure 2 Flowchart 1: The First Method for Nondominated Solutions.

Boxes (3) and (4): Box (3) is a comparison to see if the process has reached a steady state. If not, the process will continue to Box (4), which is to eliminate those dominated points in Z^{n+1}. Here $\hat{D}(Z^{n+1})$ denotes estimated dominated points in Z^{n+1}, and W^{n+1} will be the remaining "good" alternatives after the elimination.

Boxes (5) and (6): Box (5) is a comparison to see if the elimination in Box (4) is effective if $W^{n+1} = Z^{n+1}$, then Z^{n+1} can be a set of N-points and the process can reach its steady state. Box (6) is to replace Z^{n+1} by W^{n+1} for the next iteration.

Boxes (7)–(10): Z^{n+1} is stable if no element in Z^{n+1} is dominated by any other element in Z^{n+1} and every element outside of Z^{n+1} is dominated by some element in $Z^{n+1} = N(Z^{n+1}, D(\cdot))$. If this condition is satisfied, Z^{n+1} can probably be the set of all N-points. In Box (9), nondominance is verified either by Theorem 5.1 or 5.2. If Z^{n+1} is the set of all N-points, the process is stopped at Box (10) with Z^{n+1} as the set for the final decision. Otherwise, the process should be repeated again from Box (0). If Z^{n+1} is not stable, then Z^{n+1} contains some dominated points or there are N-points not contained in Z^{n+1}. We shall accordingly either eliminate the dominated points from Z^{n+1} or add new N-points into Z^{n+1}. This adjustment on Z^{n+1} is performed at Box (8).

Box (11): This step is obvious for changing the step variable in the process.

For examples of applying the two methods described in this subsection, the interested reader is referred to Yu (1985).

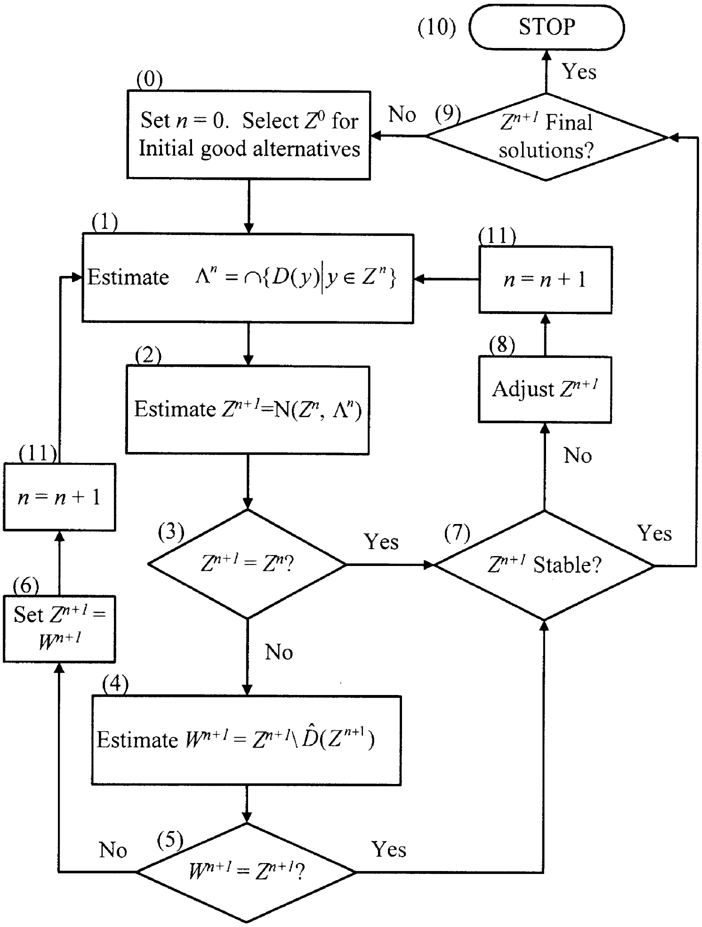

(10) STOP
Yes
(0)
Set $n = 0$. Select Z^0 for Initial good alternatives
No
(9) Z^{n+1} Final solutions?
(1)
Estimate $\Lambda^n = \cap\{D(y) | y \in Z^n\}$
(11)
$n = n + 1$
(8)
Adjust Z^{n+1}
(2)
Estimate Z^{n+1}=N(Z^n, Λ^n)
(11)
$n = n + 1$
No
(3) $Z^{n+1} = Z^n$?
Yes
(7) Z^{n+1} Stable?
Yes
(6)
Set $Z^{n+1} = W^{n+1}$
No
(4)
Estimate $W^{n+1} = Z^{n+1} \backslash \hat{D}(Z^{n+1})$
No
(5) $W^{n+1} = Z^{n+1}$?
Yes

Figure 3 Flowchart 2: The Second Method for Nondominated Solutions.

6. MC^2-SIMPLEX METHOD FOR LINEAR CASES

Linear function is the best function in analysis and application because it is easy to understand and compute. Because of these traits of linear functions, linear programming is popular and powerful. In this section we shall extend the linear programming with single criterion to that with multiple criteria and multilevel constraints of the resource. Instead of the simplex method with single criterion, we shall discuss the multicriteria (MC) simplex method and the multicriteria multiconstraint-level (MC^2) simplex method to help resolve the difficulty of decision problems.

Algebraically, the MC^2-simplex method can be represented by

$$\begin{array}{ll} \max & Cx \\ \text{s.t.} & Ax \leq D,\ x \geq 0 \end{array} \tag{1}$$

where $C = C_{q\times n}$, $A = A_{m\times n}$, and $D = D_{m\times k}$ are matrices. Note that if $k = 1$, then D is a vector and (1) becomes the MC simplex problem; if $q = 1$, then C is a vector and (1) becomes the ordinary simplex problem.

The MC^2-simplex formulation can arise in many ways. Recall that Theorem 5.2 implies that to find N-points, the constraints and the objective functions in mathematical programming formulation are interchangeable. One may put some criteria in constraints with multiple levels, and thus the result will be a MC^2-simplex form. Another case is in the design of optimal systems, in which one wishes to design a system that is optimal over all contingencies rather than find an optimal point within a given system (see Seiford and Yu 1979; Yu 1985, chap. 8 for further discussion). Alternatively, one may view the constraint level as occurring according to a random rule or influenced by some uncertain factor but contained within a set. In multiple-person decision making, the resource levels may reflect the views of different coalitions of the players.

6.1. Nondominated Solutions

The following result (Yu and Zeleny 1974) connects the relationship between N-points and the MC-simplex method.

Theorem 6.1. When D is a column vector, x^0 is an N-point in X with respect to $D(y) = \Lambda^{\leq}$ (i.e., $Cx^0 \in N(Y, \Lambda^{\leq})$) if and only if there exists some $\lambda > 0$ such that x^0 solves

$$\begin{array}{ll} \max & \lambda Cx \\ \text{s.t.} & Ax \leq D_{m\times 1},\ x \geq 0 \end{array} \tag{2}$$

Thus, by varying over λ over $\Lambda^{>} = \{\lambda|\lambda > 0\}$, we can locate all possible nondominated extreme points (N_{ex}-points) of X. Indeed, there are only a finite number of N_{ex}-points in X, and they are connected. (That is, by simplex pivoting, one can always arrive at any N_{ex}-point from other N_{ex}-point without leaving the bases of N_{ex}-points.) Locating all N_{ex}-points is in fact computationally feasible. Once all N_{ex}-points are located, they can be used to locate the entire N-points. (See Yu 1985, Yu and Zeleny 1974 for further discussion.)

To find N_{ex}-points, one can use the MC-simplex method, which appends multiple-criteria rows to the simplex tableaus. Because the MC-simplex method is a special case of the MC^2-simplex method, we shall only describe the MC^2 simplex method in the next subsection.

Now suppose $\Lambda' = \Lambda \cup \{0\}$ is a general pointed polyhedral cone with (H^1, H^2, . . . , H^p) as a generator for Λ^*. One can define $C' = HC$, where H is the matrix with H^j as its jth row. Then $Cx^0 \in N(Y, \Lambda)$ if and only if there is $\lambda_{1\times p} > 0$ so that x^0 is a solution of (2) with C' replacing C. (See Yu 1985; Yu and Zeleny 1974 for further discussion.) This result suggests that in locating N_{ex}-points for a general polyhedral dominated cone Λ one can first convert the objective coefficient C into C', and then it becomes a problem of locating N_{ex}-points with dominated cone $\Lambda^{\leq}$. Thus, unless specified, we shall assume $D(y) = \Lambda^{\leq}$ for all y throughout the remainder of this subsection.

6.2. Potential Solutions and MC^2-Simplex Method

Returning to (1), we generalize the concept of nondominated solutions into that of a potential solution by defining: A basis J is a potential basis (without confusion we also call J a potential solution) for the MC^2 problem (1) if there are $\lambda > 0$ and $\gamma > 0$ such that J is an optimal basis for

$$\begin{array}{ll} \max & \lambda Cx \\ \text{s.t.} & Ax \leq D\gamma,\ x \geq 0 \end{array} \tag{3}$$

Note that if D is a vector, $\gamma \in R^1$. By normalization, we can set $\gamma = 1$ and (3) reduces to (2). Then Theorem 6.1 ensures that a potential solution with D being a vector is indeed an N_{ex}-point. Thus, the concept of potential solutions is a generalization of that of nondominated solutions.

Note that the simplex tableau of (3) can be written as:

$$\begin{array}{cc|cr} \hline A & I & D\gamma & \qquad (4) \\ \hline -\lambda C & 0 & 0 & \qquad (5) \\ \hline \end{array}$$

Let B be the basis matrix associated with J. Since each set of basic vectors J is uniquely associated with a column index set, we shall, without confusion, let J be this set of indices and J' the set of non-basic columns. The simplest tableau associated with J is

$$\begin{array}{cc|cr} \hline B^{-1}A & B^{-1} & B^{-1}D\gamma & \qquad (6) \\ \hline \lambda C_B B^{-1}A - \lambda C & \lambda C_B B^{-1} & \lambda C_B B^{-1}D\gamma & \qquad (7) \\ \hline \end{array}$$

where (6) = $B^{-1} \cdot$ (4) (i.e., premultiply (4) by B^{-1} on both sides of the equation), (7) = $\lambda C_B \cdot$ (6) + (5), and C_B is the submatrix of criteria columns associated with the basis vectors.

Dropping γ and λ, we obtain the MC^2 tableau associated with basis J:

$$\begin{array}{cc|cr} \hline B^{-1}A & B^{-1} & B^{-1}D & \qquad (8) \\ \hline C_B B^{-1}A - C & C_B B^{-1} & C_B B^{-1}D & \qquad (9) \\ \hline \end{array}$$

which we write as

$$\begin{array}{c|c} \hline Y & W \\ \hline Z & V \\ \hline \end{array}$$

where $Y = [B^{-1}A, B^{-1}]$, $W = B^{-1}D$, $Z = [C_B B^{-1}A - C, C_B B^{-1}]$ and $V = C_B B^{-1}D$.

Let $W(J)$ and $Z(J)$ be the submatrices of the tableau associated with basis J Define:

$$\Gamma(J) = \{\gamma > 0 | W(J)\gamma \geq 0\},\ \Lambda(J) = \{\lambda > 0 | \lambda Z(J) \geq 0\}$$

Note that, because of (6) and (8), the basis is feasible for all $\gamma \in \Gamma(J)$; and because of (7) and (9), the basis is also optimal for all $\lambda \in \Lambda(J)$. Immediately, we have (Yu 1985; Seiford and Yu 1979):

Theorem 6.2.

1. J is a potential solution if and only if $\Lambda(J) \times \Gamma(J)\varnothing$.

2. Given basis J, $\Gamma(J) \neq \varnothing$ if and only if $w_{\max} = 0$ for

$$\begin{array}{ll} \max & w = e_1 + e_2 + \ldots + e_q \\ \text{s.t.} & Z(J)x + e = 0 \\ & x \geq 0,\ e \geq 0, \text{ where } e \in R^q \end{array}$$

3. Given basis J, $\Gamma(J) \neq \varnothing$ if and only if $w_{\max} = 0$ for

$$\begin{array}{ll} \max & w = d_1 + d_2 + \ldots + d_k \\ \text{s.t.} & yW(J)x + d = 0 \\ & y \geq 0,\ x \geq 0, \text{ where } d \in R^k \end{array}$$

There are practical ways to verify whether a basis is a potential solution. It can be shown (Yu 1985; Seiford and Yu 1979) that the set of all potential bases is connected (that is, using pivoting, one can arrive at any potential solution without leaving the potential bases). This result makes locating all potential solutions feasible.

Given a potential basis, its corresponding sets of weights $\Lambda(J)$ and $\Gamma(J)$ on criteria and constraint levels can be specified. Even if we do not know the precise weights λ and γ, as long as $\lambda \in \Lambda(J)$ and $\gamma \in \Gamma(J)$), we know that J is the optimal basis and the decision process terminates at J for the solution. Since there are only a finite number of potential bases, finding them and identifying their corresponding weight sets can greatly simplify the difficulty in finding the final solution.

Suppose that we have $N = \{1, \ldots, \mathrm{n}\}$ opportunities or products, from which we want to choose a subset to undertake or produce as to maximize profit. We can formulate the problem as in (3). We want to maximize the objective $\lambda\ Cx$ with constraint and $Ax \leq D\lambda$ and $x \geq 0$. Here λ and γ are uncertain. In this way we have a system design problem with multiple-criteria and multiple-resource availability levels. We can use MC^2-simplex method to identify a set of potentially good systems as candidates for the optimal system. Here, each potentially good system is a subset of a given opportunity set, each of which optimizes the objective when the parameters of contribution coefficients and that of resource availability levels fall in certain region. For further discussion along this line and applications, see Lee et al. (1990), Shi (1998a, b).

7. FUZZY MULTICRITERIA OPTIMIZATION

In addition to the above methods, we could imbed fuzzy set theory into multiple-criteria linear programming. Bellman and Zadeh (1970) propose the concept of fuzzy set, and Zimmermann (1978) first propose this idea. In fuzzy set theory, there is a membership function $\mu(x)$ indicating each element x the degree of membership for x to belong to a set. Fuzzy multiple-objective linear programming formulates the objectives and the constraints as fuzzy sets, characterized by their individual linear membership functions. The decision set is defined as the intersection of all fuzzy sets and the set defined by relevant hard constraints. A crisp (nonfuzzy) solution is generated by selecting the solution that has the highest degree of membership in the decision set. For further discussions, the reader is referred to Zimmermann (1978), Werners (1987), Martinson (1993), and Lee and Li (1993).

The fuzzy multiple objectives linear programming (f-MOLP) usually has the following format:

$$\begin{aligned}
&\max\ \tilde{z}_k = \sum_{j=1}^{n} \tilde{c}_{kj} x_k,\ k = 1, 2, \ldots, q_1 \\
&\min\ \tilde{w}_k = \sum_{j=1}^{n} \tilde{c}_{kj} x_j,\ k = q_1 + 1, \ldots, q \\
&\text{s.t.} \sum_{j=1}^{n} \tilde{a}_{ij} x_j \leq \tilde{b}_i,\ i = 1, 2, \ldots, m_1 \\
&\sum_{j=1}^{n} \tilde{a}_{ij} x_j \geq \tilde{b}_i,\ i = m_1 + 1, \ldots, m_2 \\
&\sum_{j=1}^{n} \tilde{a}_{ij} x_j = \tilde{b}_i,\ i = m_2 + 1, \ldots, m \\
&x_j \geq 0, j = 1, 2, \ldots, n
\end{aligned} \tag{10}$$

where $\tilde{c}_{kj}$ is the jth coefficient of kth objective, $\tilde{a}_{ij}$ is jth coefficient of ith constraint, and $\tilde{b}_i$ is the right-hand side (RHS) of ith constraint. Note that $\tilde{c}_{kj}$, $\tilde{a}_{ij}$, and $\tilde{b}_i$ are fuzzy numbers. The f-MOLP problem (10) can be solved by transferring it into crisp MOLP (c-MOLP), shown as (11).

$$\begin{aligned}
&\max\ (z_k)_a = \sum_{j=1}^{n} (c_{kj})_a^U x_j,\ k = 1, 2, \ldots, q_1 \\
&\min\ (w_k)_a = \sum_{j=1}^{n} (c_{kj})_a^L x_j,\ k = q_1 + 1, \ldots, q \\
&\text{s.t.} \sum_{j=1}^{n} (a_{ij})_a^L x_j \leq (b_i)_a^U,\ i = 1, 2, \ldots, m_1, m_2 + 1, \ldots, m \\
&\sum_{j=1}^{n} (a_{ij})_a^U x_j \geq (b_i)_a^L,\ i = m_1 + 1, \ldots, m \\
&x_j \geq 0, j = 1, 2, \ldots, n
\end{aligned} \tag{11}$$

where $(c_{kj})_a^U$ and $(c_{kj})_a^L$, $(a_{ij})_a^U$ and $(a_{ij})_a^L$ and $(b_i)_a^U$ and $(b_i)_a^L$ are upper and lower bound of fuzzy number $\tilde{c}_{kj}$, $\tilde{a}_{ij}$ and $\tilde{b}_i$, respectively, by taking α-level cut. Problem (11) can be solved by a fuzzy algorithm interactively. For details, see Zimmermann (1978) and Lee and Li (1993). For applications and extensions along this line, see Sakawa et al. (1994, 1995), Shibano et al. 1996), Shih et al. (1996), Ida and Gen (1997), and Shih and Lee (1999) and those quoted therein.

8. EXTENSIONS AND CONCLUDING REMARKS

We have briefly sketched six important topics of MCDM problems. Many more topics, such as interactive methods including adapted gradient search method, surrogate-worth trade-off methods (Haimes and Hall 1974), the Zionts–Wallenius method (Zionts and Wallenius 1976), the paired comparison simplex method (Malakooti and Ravindran (1986), the bireference procedure (Michalowski and Szapiro 1992), the paired comparison method (Shin and Allen 1994), preference over uncertain outcomes, and second-order games can be found in Yu (1985, chap. 10 and the citations therein). For multiple-criteria dynamic optimization problems, refer to Li and Haimes (1989), Yu and Seiford (1981), Yu and Leitmann (1974) and those quoted therein. The reader who is interested in decision aid and MCDM in abstract spaces is referred to Roy (1977) and Dauer and Stadler (1986) respectively, and the citations therein. For other mathematical analysis on multiple criteria optimization, see Metev and Yordanova-Markova (1997), Ehrgott and Klamroth (1997), and Gal and Hanne (1999).

There are a number of computer programs to solve multiple-criteria decision making (MCDM) problems. Because many MCDM problems are reformulated and solved by converting them into single-criterion optimization problems (as discussed in the previous sections), many computer programs for mathematical programming are adopted in MCDM computer software. Since the problem domains of MCDM are rich and complex, so are the computer software programs. Space limitations make it difficult to list and discuss them here. However, interested readers can refer to www.cba.uga.edu/mcdm.html, the website of the International Society on Multiple Criteria Decision Making, for MCDM software listing.

With creative thinking and practicing, the basic concepts of MCDM described above can generate an infinite number of concepts and be applied to an infinite number of multicriteria problems. Rigid habitual ways of thinking usually are the stumbling blocks to creativity. In solving nontrivial problems, we need not only the mastery of the mathematical tools but also a good understanding of human behavior and the habitual domains of the decision makers (see Yu 1990, 1995 for a detailed discussion). At the least, we should not become so overenthusiastic about one particular concept or method that, so to speak, we cut our feet to fit the already-made shoes.

REFERENCES

Bellman, R. E., and Zadeh, L. A. (1970), "Decision-Making in a Fuzzy Environment," *Management Science*, Vol. 17, pp. 140–164.

Charnes, A., and Cooper, W. W. (1975), "Goal Programming and Constrained Regression—A Comment," *Omega,* Vol. 3, pp. 403–409.

Chien, I. S. (1987), "Restructuring and Computer Support in Multiple Criteria Decision Making," Ph.D. dissertation, University of Kansas, Lawrence, KS.

Chien, I. S., Yu, P. L., and Zhand, D. (1990), "Indefinite Preference Structures and Decision Analysis," *Journal of Optimization Theory and Application*, Vol. 64, pp. 71–85.

Cogger, K. O., and Yu, P. L. (1985), "Eigen Weight Vectors and Least Distance Approximation for Revealed Preference in Pairwise Weight Ratios," *Journal of Optimization Theory and Application,* Vol. 46, pp. 483–491.

Dauer, J. P., and Stadler, W. (1986), "A Survey of Vector Optimization in Infinite-Dimensional Space, Part 2," *Journal of Optimization Theory and Application,* Vol. 51, pp. 205–241.

Debreu, G. (1954), "Representation of a Preference Ordering by a Numerical Function," in *Decision Processes*, R. M. Thrall, C. H. Coombs, and R. L. Davis, Eds., John Wiley & Sons, New York.

Debreu, G. (1960), "Topological Methods in Cardinal Utility Theory," in *Mathematical Methods in Social Sciences*, K. J. Arrow, S. Karlin, and P. Suppes, Eds., Stanford University Press, Stanford, CA.

Ehrgott, M., and Klamorth, K. (1997), "Connectedness of Efficient Solutions in Multiple Criteria Combinatorial Optimization," *European Journal of Operational Research,* Vol. 97, pp. 159–166.

Freimer, M., and Yu, P. L. (1976), "Some New Results on Compromise Solutions for Group Decision Problem," *Management Science*, Vol. 22, pp. 688–693.

Fishburn, P. C. (1970), *Utility Theory for Decision Making*, John Wiley & Sons, New York.

Gal, T., and Hanne T. (1999), "Consequences of Dropping Nonessential Objectives for the Application of MCDM Methods," *European Journal of Operational Research,* Vol. 119, pp. 373–378.

Gearhart, W. B. (1979), "Compromise Solutions and Estimation of the Noninferior Set," *Journal of Optimization Theory and Application,* Vol. 28, pp. 29–47.

Gorman, W. M. (1968), "The Structure of Utility Functions," *Review of Economic Studies,* Vol. 35, pp. 367–390.

Haimes, Y. Y., and Hall, W. A. (1974), "Multiobjectives in Water Resources System Analysis: The Surrogate Worth Trade off Method," *Water Resources Research*, Vol. 10, pp. 615–623.

Hartley, R. (1978), "On Cone-Efficiency, Cone-Convexity and Cone-Compactness," *SIAM Journal of Applied Mathematics*, Vol. 34, pp. 211–222.

Hwang, C. L., and Masud, A. S. M. (1979), *Multiple Objective Decision Making—Methods and Applications: A State-of-the-Art Survey*, Springer, New York.

Hwang, C. L., and Yoon, K. (1981), *Multiple Attribute Decision Making—Methods and Applications: A State-of-the-Art Survey,* Springer, New York.

Ida, K. and Gen, M. (1997), "Improvement of Two-Phase Approach for Solving Fuzzy Multiple Objective Linear Programming," *Journal of Japan Society for Fuzzy Theory and System,* Vol. 19, pp. 115–121 (in Japanese).

Ignizio, J. P. (1976), *Goal Programming and Extensions,* D.C. Heath, Lexington, MA.

Keeney, R. L., and Raiffa, H. (1976), *Decisions with Multiple Objectives: Preferences and Value Tradeoffs,* John Wiley & Sons, New York.

Lee, E. S., and Li, R. J. (1993), "Fuzzy Multiple Objective Programming and Compromise Programming with Pareto Optimum," *Fuzzy Sets and Systems,* Vol. 53, pp. 275–288.

Lee, S. M. (1972), *Goal Programming for Decision Analysis*, Auerback, Philadelphia.

Lee, Y. R., Shi, Y., and Yu, P. L. (1990), "Linear Optimal Designs and Optimal Contingency Plans," *Management Science*, Vol. 36, pp. 1106–1119.

Li, D., and Haimes, Y. Y. (1989), "Multiobjective Dynamic Programming: The State of the Art," *Control Theory and Advanced Technology,* Special Issue on Multiobjective Discrete Dynamic System Control Theory and Advanced Technology, Y. Y. Haimes and D. Li, Eds., Vol. 5, pp. 471–483.

Malakooti, B. and Ravindran, A. (1986), "Experiments with an Interactive Paired Comparison Simplex Method for MOLP Problems," *Annals of Operations Research,* Vol. 5 pp. 575–597.

Martinson, F. K. (1993), "Fuzzy vs. Minmax Weighted Multiobjective Linear Programming Illustrative Comparisons," *Decision Sciences,* Vol. 24, pp. 809–824.

Metev, B. S., and Yordanova-Markova, I. T. (1997), "Multi-objective Optimization over Convex Disjunctive Feasible Sets Using Reference Points," *European Journal of Operational Research,* Vol. 98, pp. 124–137.

Michalowski W. and Szapiro T. (1992), "A Bi-reference Procedure for Interactive Multiple Criteria Programming," *Operations Research*, Vol. 40, pp. 247–258.

Pascoletti, A., and Serafini, P. (1984), "Scalarizing Vector Optimization Problems," *Journal of Optimization Theory and Application,* Vol. 42, pp. 499–524.

Roy, B. (1971), "Problems and Methods with Multiple Objective Functions," *Math Programming,* Vol. 1, pp. 239–266.

Roy, B. (1977), "A Conceptual Framework for a Prescriptive Theory of Decision-Aid," in *Multiple Criteria Decision Making,* M. K. Starr and M. Zeleny, Eds., TIMS Studies in Management Sciences, Vol. 6, North-Holland, Amsterdam.

Roy, B. and Vincke, P. H. (1987), "Pseudo-Orders: Definition, Properties, and Numerical Representation," *Mathematical Social Sciences,* Vol. 14, pp. 263–274.

Saaty, T. L. (1980), *The Analytic Hierarchy Process*, McGraw-Hill, New York.

Sakawa, M., Inuiguchi, M., Sundad, H., and Sawada, K. (1994), "Fuzzy Multiobjective Combinatorial Optimization through Revised Genetic Algorithms," *Journal of Japan Society for Fuzzy Theory and System,* Vol. 6, pp. 177–186 (in Japanese).

Sakawa, M., Kato, K., Sundad, H., and Enda, Y. (1995), "An Interactive Fuzzy Satificing Method for Multiobjective 0-1 Programming Problems through Revised Genetic Algorithms," *Journal of Japan Society for Fuzzy Theory and System,* Vol. 17, pp. 361–370 (in Japanese).

Seiford, L., and Yu, P. L. (1979), "Potential Solutions of Linear Systems: The Multi-Criteria Multiple Constrain Levels Program," *Journal of Mathematical Analysis and Applications,* Vol. 69, pp. 283–303.

Shi, Y. (1998a), "Finding Dual Flexible Contingency Plans for Optimal Generalized Linear Systems," *International Transactions in Operational Research,* Vol. 5, pp. 303–315.

Shi, Y. (1998b), "Optimal System Design with MC^2 Linear Programming: A Dual Contingency Plan Approach," *European Journal of Operational Research,* Vol. 107, pp. 692–709.

Shibano, T., Sakawa M., and Obata H. (1996), "Interactive Decision Making for Multiobjective 0–1 Programming Problems with Fuzzy Parameters through Genetic Algorithms," *Journal of Japan Society for Fuzzy Theory and System,* Vol. 18, pp. 1144–1153 (in Japanese).

Shih, H. S., and Lee, E. S. (1999), "Fuzzy Multi-Level Minimum Cost Flow Problems," *Fuzzy Sets and Systems*, Vol. 107, pp. 159–176.

Shih, H. S., Lai, Y. J., and Lee, E. S. (1996), "Fuzzy Approach for Multiple-Level Programming Problems," *Computers and Operation Researches,* Vol. 23, pp. 73–91.

Shin, W. S., and Allen, D. B. (1994), "An Interactive Paired Comparison Method for Bicriterion Integer Mathematical Programming," *Naval Research Logistics,* Vol. 41, pp. 423–434.

Stadler, W. (1979), "A Survey of Multicritena Optimization or the Vector Maximization Problem, Part I: 1776–1960," *Journal of Optimization Theory and Application,* Vol. 29, pp. l–52.

Stadler, W. (1981), "A Comprehensive Bibliography on Multicriteria Decision Making and Related Areas," Working Paper, University of California, Berkeley, CA.

Steuer, R. E., Gardiner, I. R., and Gray, J. (1996), "A Bibliographic Survey of the Activities and International Nature of Multiple Criteria Decision Making," *Journal of Multi-Criteria Decision Analysis,* Vol. 5, No. 3, pp. 195–217.

Takeda, E., and Yu, P. L. (1995), "Assessing Priority Weights from Subsets of Pairwise Comparisons in Multiple Criteria Optimization Problems," *European Journal of Operational Research*, Vol. 86.

Werners, B. (1987), "Interactive Multiple Objective Programming Subject to Flexible Constraints," *European Journal of Operational Research,* Vol. 31, pp. 342–349.

Yu, P. L. (1973), "A Class of Solutions for Group Decision Problems," *Management Science,* Vol. 19, pp. 936–946.

Yu, P. L. (1974), "Cone Convexity, Cone Extreme Points and nondominated Solutions in Decision Problems with Multiobjectives," *Journal of Optimization Theory and Applications*, Vol. 14, pp. 319–377.

Yu, P. L. (1985), *Multiple Criteria Decision Making Concepts, Techniques and Extensions*, Plenum Press, New York.

Yu, P. L. (1990), *Forming Winning Strategies—An Integrated Theory of Habitual Domains,* Springer, Berlin.

Yu, P. L. (1995), *Habitual Domains: Freeing Yourself from the Limits on Your Life,* Highwater Editions, Shawnee Mission, KS.

Yu, P. L., and Leitmann, G. (1974), "Nondominated Decisions and Cone Convexity in Dynamic Multicriteria Decision Problems," *Journal of Optimization Theory and Applications,* Vol. 14, pp. 573–584.

Yu, P. L., and Seiford, L. (1981), "Multistage Decision Problems with Multicriteria," in *Multicriteria Analysis: Practical Methods*, P. NijKamp and J. Spronk, Eds, Gower Press, London.

Yu, P. L., and Takeda, E. (1987), "A Verification Theorem of Preference Separability for Additive Value Functions," *Journal of Mathematical Analysis and Applications,* Vol. 126, pp. 382–396.

Yu, P. L. and Zeleny, M. (1974), "The Set of All Nondominated Solutions in the Linear Case and a Multicriteria Simplex Method," *Journal of Mathematical Analysis and Applications,* Vol. 49, pp. 430–468.

Zeleny, M. (1975), "The Theory of the Displaced Ideal," in M. Zeleny, Ed., Springer, New York, pp. 151–205.

Zimmermann, H. J. (1978), "Fuzzy Programming and Linear Programming with Several Objective Functions," *Fuzzy Sets and Systems,* Vol. 1, pp. 45–55.

Zionts, S. and Wallenius, J. (1976), "An Interactive Programming Method for Solving the Multiple Criteria Problem," *Management Science*, Vol. 22, pp. 652–663.

ADDITIONAL READING

Changkong, V., and Haimes, Y. Y., *Multiobjective Decision Making Theory and Methodology,* North-Holland, New York, 1983.

Dyer, J. S., "Interactive Goal Programming," *Management Science,* Vol. 19, 1970, pp. 62–70.

Fandel, G., Grauer, M., Kurzhanski, A., and Wierzbicki, A. P., Eds., *Large-Scale Modelling and Interactive Decision Makings, Proceedings, Eisenbach, GRD, 1985,* Lecture Notes in Economics and Mathematic No. 273, Springer, New York, 1986.

Gal, T., *Postoptimal Analysis Mathematic Programming and Related Topics,* McGraw-Hill, New York, 1979.

Haimes, Y. Y., and Changkong, V., Eds., *Decision Making with Multiple Objectives. Proceedings, Cleveland, Ohio, 1984,* Lecture Notes in Economics and Mathematical Systems No. 242, Springer, New York, 1985.

Hansen, P., Ed., *Essays and Surveys on Multiple Criteria Decision Making, Proceedings, Mons, 1982,* Lecture Notes in Economics and Mathematical Systems No. 209, Springer, New York, 1983.

Hazen, G. B., and Morin, T. L., "Optimality Conditions for Non-conical Multiple-Objective Programming," *Journal of Optimization Theory and Application,* Vol. 40, 1983, pp. 25–59.

Morse, J. N., Ed., *Organization: Multiple Agents with Multiple Criteria. Proceedings, University of Delaware, Newark, 1980,* Lecture Notes in Economics and Mathematical Systems No. 190, Springer, New York, 1981.

Pareto, V., 1906; Piccola *Manual of Political Economy*, A. S. Schwier, Trans., MacMillan, New York, 1971.

Sawaragi, Y., Nakayama, H., and Tamno, T. (1985), *Theory of Multiobjective Optimization,* Academic Press, Orlando, FL.

Steuer, R. E., *Multiple Criteria Optimization,* John Wiley & Sons, New York, 1985.

Tanino, T., and Sawaragi, Y., ‘‘Stability of Nondominated Solutions in Multicriteria Decision Making,’’ *Journal of Optimization Theory and Application,* Vol. 30, 1980, pp. 229–253.

Wang, P. Z., *Fuzzy Sets and the Fall-Shadow of Random Sets,* Beijing Normal University Press, Beijing, 1985 (in Chinese).

White, D. J., *Optimality and Efficiency,* John Wiley & Sons, New York.

Zeleny, M. (1982), *Multiple Criteria Decision Making*, McGraw-Hill, New York.

CHAPTER 102

Stochastic Optimization

ANTON J. KLEYWEGT
ALEXANDER SHAPIRO
Georgia Institute of Technology

1. INTRODUCTION 2625

2. OPTIMIZATION UNDER UNCERTAINTY 2625

3. STOCHASTIC PROGRAMMING 2628

- 3.1. Stochastic Programming with Recourse 2629
- 3.2. Sampling Methods 2631
- 3.3. Perturbation Analysis 2632
- 3.4. Likelihood Ratio Method 2633
- 3.5. Simulation-Based Optimization Methods 2634

4. DYNAMIC PROGRAMMING 2636

- 4.1. Basic Concepts in Dynamic Programming 2636
  - 4.1.1. Decision Times 2636
  - 4.1.2. States 2637
  - 4.1.3. Decisions 2637
  - 4.1.4. Transition Probabilities 2638
  - 4.1.5. Rewards and Costs 2638
  - 4.1.6. Policies 2639
  - 4.1.7. Example 2640
- 4.2. Finite Horizon Dynamic Programs 2641
  - 4.2.1. Optimality Results 2641
  - 4.2.2. Finite Horizon Algorithm 2641
  - 4.2.3. Structural Properties 2642
- 4.3. Infinite Horizon Dynamic Programs 2643
- 4.4. Infinite Horizon Discounted Dynamic Programs 2643
  - 4.4.1. Optimality Results 2643
  - 4.4.2. Infinite Horizon Algorithms 2644
- 4.5. Approximation Methods 2645

REFERENCES 2646

1. INTRODUCTION

Decision makers often have to make decisions in the presence of uncertainty. Decision problems are often formulated as optimization problems, and thus in many situations decision makers wish to solve optimization problems that depend on parameters which are unknown. Typically, it is quite difficult to formulate and solve such problems, both conceptually and numerically. The difficulty already starts at the conceptual stage of modeling. Often there are a variety of ways in which the uncertainty can be formalized. In the formulation of optimization problems, one usually attempts to find a good trade-off between the realism of the optimization model, which affects the usefulness and quality of the obtained decisions, and the tractability of the problem, so that it can be solved analytically or numerically. As a result of these considerations, there are a large number of different approaches for formulating and solving optimization problems under uncertainty. It is impossible to give a complete survey of all such methods in one article. Therefore, this chapter aims only to give a flavor of prominent approaches to optimization under uncertainty.

2. OPTIMIZATION UNDER UNCERTAINTY

To describe some issues involved in optimization under uncertainty, we start with a static optimization problem. Suppose we want to maximize an objective function $G(x, \omega)$, where x denotes the decision

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

to be made, χ denotes the set of all feasible decisions, ω denotes an outcome that is unknown at the time the decision has to be made, and Ω denotes the set of all possible outcomes.

There are several approaches for dealing with optimization under uncertainty. Some of these approaches are illustrated next in the context of an example.

Example 1 (Newsvendor problem). Many companies sell seasonal products, such as fashion articles, airline seats, Christmas decorations, magazines, and newspapers. These products are characterized by a relatively short selling season, after which the value of the products decreases substantially. Often a decision has to be made how much of such a product to manufacture or purchase before the selling season starts. Once the selling season has started, there is not enough time remaining in the season to change this decision and implement the change, so that at this stage the quantity of the product is given. During the season the decision maker may be able to make other types of decisions to pursue desirable results, such as to change the price of the product as the season progresses and sales of the product take place. Such behavior is familiar in many industries. Another characteristic of such a situation is that the decisions have to be made before the eventual outcomes become known to the decision maker. For example, the decision maker has to decide how much of the product to manufacture or purchase before the demand for the product becomes known. Thus, decisions have to be made without knowing which outcome will take place.

Suppose that a manager has to decide how much of a seasonal product to order. Thus, the decision variable x is a nonnegative number representing the order quantity. The cost of the product to the company is c per unit of the product. During the selling season the product can be sold at a price (revenue) of r per unit of the product. After the selling season, any remaining product can be disposed of at a salvage value of s per unit of the product, where typically $s < r$. The demand D for the product is unkown at the time the order decision x has to be made. If the demand D turns out to be greater than the order quantity x, then the whole quantity x of the product is sold during the season, and no product remains at the end of the season, so that the total revenue and the profit turn out to be rx and $rx - cx = (r - c)x$, respectively. If the demand D turns out to be less than the order quantity x, then quantity D of the product is sold during the season, and the remaining amount of product at the end of the season is $x - D$, so that the total revenue and the profit turn out to be $rD + s(x - D)$ and $rD + s(x - D) - cx = (s - c)x + (r - s)D$, respectively. Thus, the profit is given by

$$G(x, D) = \begin{cases} (s - c)x + (r - s)D & \text{if } x \geq D \\ (r - c)x & \text{if } x < D \end{cases} \tag{1}$$

The manager would like to choose x to maximize the profit $G(x, D)$, but the dilemma is that D is unknown, or in other words is uncertain, at the time the decision should be made.

Note that if $r \leq c$ and $s \leq c$, then the company can make no profit from buying and selling the product, so that the optimal order quantity is $x^* = 0$, irrespective of what the demand D turns out to be. Also, if $s \geq c$, then any unsold product at the end of the season can be disposed of at a value at least equal to the cost of the product, so that it is optimal to order as much as possible, irrespective of what the demand D turns out to be. These, of course, are obvious cases. Therefore, we assume in the remainder of this example that $s < c < r$. Under this assumption, for any given $D \geq 0$, the function $G(\cdot, D)$ is a piecewise linear function with positive slope $r - c$ for $x < D$ and negative slope $s - c$ for $x > D$. Therefore, if the demand D is known at the time the order decision has to be made, then the best decision is to choose order quantity $x^* = D$.

However, if D is not known, then the problem becomes more difficult. Sometimes a manager may want to hedge against the worst possible outcome. Suppose the manager thinks that the demand D will turn out to be some number in the interval $[a, b]$ with $a < b$, that is, the lower and upper bounds for the demand are known to the manager. In that case, in order to hedge against the worst possible scenario, the manager will choose the value of x that gives the best profit under the worst possible outcome. That is, the manager will maximize the function $g(x) \equiv \min_{D \in [a,b]} G(x, D)$ over $x \geq 0$. This leads to the following max-min problem:

$$\max_{x \geq 0} \min_{D \in [a,b]} G(x, D) \tag{2}$$

It is not difficult to see that $g(x) = G(x, a)$, and hence $x^* = a$ is the optimal solution from the point of view of the worst-case scenario. Clearly, in many cases this will be an overly conservative decision.

Sometimes a manager may want to make the decision that under the worst possible outcome will still appear as good as possible compared with what would have been the best decision with hindsight, that is, after the outcome becomes known. For any outcome of the demand D, let

$$g^*(D) \equiv \max_{x \geq 0} G(x, D) = (r - c)D$$

denote the optimal profit with hindsight, also called the optimal value with perfect information. The optimal decision with perfect information, $x^* = D$, is sometimes called the wait-and-see solution. Suppose the manager chose to order quantity x, so that the actual profit turned out to be $G(x, D)$. The amount of profit that the company missed out on because of a suboptimal decision is given by $g^*(D) - G(x, D)$. This quantity is often called the absolute regret. The manager may want to choose the value of x that minimizes the absolute regret under the worst possible outcome. For any decision x, the worst possible outcome is given by $\max_{D \in [a,b]} [g^*(D) - G(x, D)]$. Since the manager wants to choose the value of x that minimizes the absolute regret under the worst possible outcome, this leads to the following min-max problem:

$$\min_{x \geq 0} \max_{D \in [a,b]} [g^*(D) - G(x, D)] \tag{3}$$

The optimal solution of this problem is $x^* = [(c - s)a + (r - c)b]/(r - s)$. Note that x^* is a convex combination of a and b, and thus $a < x^* < b$. The larger the salvage loss per unit $c - s$, the closer x^* is to a, and the larger the profit per unit $r - c$, the closer x^* is to b. That seems to be a more reasonable decision than $x^* = a$.

It was assumed in both variants of the worst-case approach discussed above that no a priori information about the demand D was available to the manager except the lower and upper bounds for the demand. In some situations this may be a reasonable assumption and the worst-case approach could make sense if the range of the demand is known and is not too large.

Another approach to decision making under uncertainty, different from the worst-case approaches described above, is the stochastic optimization approach, on which we focus in the remainder of this article. Suppose that the demand D can be viewed as a *random variable*. This means that the probability distribution of D is known, or at least can be estimated, by using historical data and/or a priori information available to the manager. Let $F(w) \equiv \mathbb{P}(D \leq w)$ be the corresponding cumulative distribution function (cdf) of D. Then one can try to optimize the objective function on *average*, that is, to maximize the expected profit $\mathbb{E}[G(x, D)] = \int_0^\infty G(x, w)\, dF(w)$. This leads to the stochastic program

$$\max_{x \geq 0} \{g(x) \equiv \mathbb{E}[G(x, D)]\} \tag{4}$$

In the present case it is not difficult to solve the above optimization problem in a closed form. For any $D \geq 0$, the function $G(\cdot, D)$ is concave (and piecewise linear). Therefore, the expected value function $g(\cdot)$ is also concave. Suppose for a moment that $F(\cdot)$ is continuous at a point $x > 0$. Then

$$g(x) = \int_0^x [(s - c)x + (r - s)w]\, dF(w) + \int_x^\infty (r - c)x dF(w)$$

Using integration by parts it is possible to calculate that

$$g(x) = (r - c)x - (r - s) \int_0^x F(w)\, dw \tag{5}$$

The function $g(\cdot)$ is concave, and hence continuous, and therefore Eq. (5) holds even if $F(\cdot)$ is discontinuous at x. It follows that $g(\cdot)$ is differentiable at x iff $F(\cdot)$ is continuous at x, in which case

$$g'(x) = r - c - (r - s)F(x) \tag{6}$$

Consider the inverse $F^{-1}(\alpha) \equiv \min\{x : F(x) \geq \alpha\}$ function of the cdf F, which is defined for $\alpha \in (0, 1)$. ($F^{-1}(\alpha)$ is called the α-quantile of the cdf F.) Since $g(\cdot)$ is concave, a necessary and sufficient condition for $x^* > 0$ to be an optimal solution of problem (4), is that $g'(x^*) = 0$, provided that $g(\cdot)$ is differentiable at x^*. Note that because $s < c < r$, it follows that $0 < (r - c)/(r - s) < 1$. Consequently, an optimal solution of (4) is given by

$$x^* = F^{-1}\left(\frac{r - c}{r - s}\right) \tag{7}$$

This holds even if $F(\cdot)$ is discontinuous at x^*.

Clearly the above approach explicitly depends on knowledge of the probability distribution of the demand D. In practice, the corresponding cdf $F(\cdot)$ is never known exactly and can be approximated (estimated) at best. Nevertheless, often the obtained optimal solution is robust with respect to perturbations of the cdf $F(\cdot)$.

Another point worth mentioning is that by solving (4), the manager tries to optimize the profit on average. However, the realized profit $G(x^*, D)$ could be very different from the corresponding expected value $g(x^*)$, depending on the particular realization of the demand D. This may happen if $G(x^*, D)$, considered as a random variable, has a large variability that could be measured by its variance $\mathrm{Var}[G(x^*, D)]$. Therefore, if the manager wants to hedge against such variability, he may consider the following optimization problem

$$\max_{x \geq 0} \{g_\beta(x) \equiv \mathbb{E}[G(x, D)] - \beta \, \mathrm{Var}[G(x, D)]\} \tag{8}$$

The coefficient $\beta \geq 0$ represents the weight given to the conservative part of the decision. If β is large, then the above optimization problem tries to find a solution with minimal profit variance, while if $\beta = 0$, then problem (8) coincides with problem (4). Note that since the variance $\mathrm{Var}[G(x, D)] \equiv \mathbb{E}[(G(x, D) - \mathbb{E}[G(x, D)])^2]$ is itself an expected value, from a mathematical point of view, problem (8) is similar to the expected value problem (4). Thus, the problem of optimizing the expected value of an objective function $G(x, D)$ is very general—it could include the means, variances, quantiles, and almost any other aspects of random variables of interest.

The following deterministic optimization approach is also often used for decision making under uncertainty. The random variable D is replaced by its mean $\mu = \mathbb{E}[D]$, and then the following deterministic optimization problem is solved:

$$\max_{x \geq 0} G(x, \mu) \tag{9}$$

A resulting optimal solution $\bar{x}$ is sometimes called an expected value solution. Of course, this approach requires that the mean of the random variable D be known to the decision maker. In the present example, the optimal solution of this deterministic optimization problem is $\bar{x} = \mu$. Note that the mean solution $\bar{x}$ can be very different from the solution x^* given in (7). It is well known that the quantiles are much more stable to variations of the cdf F than the corresponding mean value. Therefore, the optimal solution x^* of the stochastic optimization problem is more robust with respect to variations of the probability distributions than an optimal solution $\bar{x}$ of the corresponding deterministic optimization problem. This should be not surprising, since the deterministic problem (9) can be formulated in the framework of the stochastic optimization problem (4) by considering the trivial distribution of D being identically equal to μ.

Also note that, for any x, $G(x, D)$ is concave in D. As a result, it follows from Jensen's inequality that $G(x, \mu) \geq \mathbb{E}[G(x, D)]$, and hence

$$\max_{x \geq 0} G(x, \mu) \geq \max_{x \geq 0} \mathbb{E}[G(x, D)]$$

Thus, the optimal value of the deterministic optimization problem is biased upward relative to the optimal value of the stochastic optimization problem.

One can also try to solve the optimization problem

$$\max_{x \geq 0} G(x, D) \tag{10}$$

for different realizations of D, and then take the expected value of the obtained solutions as the final solution. In the present example, for any realization D, the optimal solution of (10) is $x = D$, and hence the expected value of these solutions, and final solution, is $\bar{x} = \mathbb{E}[D]$. Note that this approach does not have a clear rationale, and moreover, in many optimization problems it may not make sense to take the expected value of the obtained solutions. This is usually the case in optimization problems with discrete solutions, for example, when a solution is a path in a network, there does not seem to be a useful way to take the average of several different paths. Therefore, we do not discuss this approach further.

3. STOCHASTIC PROGRAMMING

The discussion of the above example motivates us to introduce the following model optimization problem, referred to as a *stochastic programming* problem:

$$\min_{x \in \chi} \{g(x) \equiv \mathbb{E}[G(x, \omega)]\} \tag{11}$$

(We consider a minimization rather than a maximization problem for the sake of notational convenience.) Here $\chi \subset \mathbb{R}^n$ is a set of permissible values of the vector x of decision variables and is referred to as the feasible set of problem (11). Often χ is defined by a (finite) number of smooth (or even linear) constraints. In some other situations the set χ is finite. In that case problem (11) is called a *discrete* stochastic optimization problem (this should not be confused with the case of discrete probability distributions). Variable ω represents random (or stochastic) aspects of the problem. Often ω can be modeled as a finite dimensional random vector, or in more involved cases as a random process. In the abstract framework we can view ω as an element of the probability space $(\Omega, \mathfrak{F}, P)$ with the known probability measure (distribution) P.

It is also possible to consider the following extensions of the basic problem (11).

- One may need to optimize a function of the expected value function $g(x)$. This happened, for example, in problem (8), where the manager wanted to optimize a linear combination of the expected value and the variance of the profit. Although important from a modeling point of view, such an extension usually does not introduce additional technical difficulties into the problem.
- The feasible set can also be defined by constraints given in a form of expected value functions. For example, suppose that we want to optimize an objective function subject to the constraint that the event $\{h(x, W) \geq 0\}$, where W is a random vector with a known probability distribution and $h(\cdot, \cdot)$ is a given function, should happen with a probability not bigger than a given number $p \in (0, 1)$. Probability of this event can be represented as the expected value $\mathbb{E}[\psi(x, W)]$, where

$$\psi(x, w) \equiv \begin{cases} 1 & \text{if} \quad h(x, w) \geq 0 \\ 0 & \text{if} \quad h(x, w) < 0 \end{cases}$$

Therefore, this constraint can be written in the form $\mathbb{E}[\psi(x, W)] \leq p$. Problems with such probabilistic constraints are called *chance constrained* problems. Note that even if the function $h(\cdot, \cdot)$ is continuous, the corresponding indicator function $\psi(\cdot, \cdot)$ is discontinuous unless it is identically equal to zero or one. Because of that, it may be technically difficult to handle such a problem.

- In some cases the involved probability distribution P_θ depends on parameter vector θ, whose components also represent decision variables. That is, the expected value objective function is given in the form

$$g(x, \theta) \equiv \mathbb{E}_\theta[G(x, \omega)] = \int_\Omega G(x, \omega)\, dP_\theta(\omega) \tag{12}$$

By using a transformation it is sometimes possible to represent the above function $g(\cdot)$ as the expected value of a function, depending on x and θ, with respect to a probability distribution that is independent of θ. We shall discuss such likelihood ratio transformations in Section 3.4

The above formulation of stochastic programs is somewhat too general and abstract. In order to proceed with a useful analysis we need to identify particular classes of such problems that on one hand are interesting from the point of view of applications and on the other hand are computationally tractable. In the following sections we introduce several classes of such problems and discuss various techniques for their solution.

3.1. Stochastic Programming with Recourse

Consider again problem (4) of the newsvendor example. We may view that problem as a two-stage problem. At the first stage a decision should be made about the quantity x to order. At this stage the demand D is not known. At the second stage a realization of the demand D becomes known and, given the first stage decision x, the manager makes a decision about the quantities y and z to sell at prices r and s, respectively. Clearly the manager would like to choose y and z in such a way as to maximize the profit. It is possible to formulate the second stage problem as the simple linear program

$$\max_{y,z}\ ry + sz \quad \text{subject to } y \leq D,\ y + z \leq x,\ y \geq 0,\ z \geq 0 \tag{13}$$

The optimal solution of the above problem (13) is $y^* = \min\{x, D\}$, $z^* = \max\{x - D, 0\}$, and its optimal value is the profit $G(x, D)$ defined in (1). Now at the first stage, before a realization of the demand D becomes known, the manager chooses a value for the first-stage decision variable x by maximizing the expected value of the second-stage optimal profit $G(x, D)$.

This is the basic idea of a two-stage stochastic program with recourse. At the first stage, before a realization of the random variables ω becomes known, one chooses the first-stage decision variables x to optimize the expected value $g(x) \equiv \mathbb{E}[G(x, \omega)]$ of an objective function $G(x, \omega)$ that depends on the optimal second stage objective function.

A *two-stage stochastic linear program with fixed recourse* is a two-stage stochastic program of the form

$$\begin{aligned} &\min_x c^T x + \mathbb{E}[Q(x, \xi)] \\ &\text{s.t. } Ax = b,\ x \geq 0 \end{aligned} \tag{14}$$

where $Q(x, \xi)$ is the optimal value of the second-stage problem

$$\begin{aligned} &\min_y q(\omega)^T y \\ &\text{s.t. } T(\omega)x + Wy = h(\omega),\ y \geq 0 \end{aligned} \tag{15}$$

The second-stage problem depends on the data $\xi(\omega) \equiv (q(\omega), h(\omega), T(\omega))$, elements of which can be random, while the matrix W is assumed to be known beforehand. The matrices $T(\omega)$ and W are called the *technology* and *recourse* matrices, respectively. The expectation $\mathbb{E}[Q(x, \xi)]$ is taken with respect to the random vector $\xi = \xi(\omega)$, whose probability distribution is assumed to be known. The above formulation originated in the works of Dantzig (1955) and Beale (1955).

Note that the optimal solution $y^* = y^*(\omega)$ of the second-stage problem (15) depends on the random data $\xi = \xi(\omega)$ and therefore is random. One can write $Q(x, \xi(\omega)) = q(\omega)^T y^*(\omega)$.

The next question is how one can solve the above two-stage problem numerically. Suppose that the random data have a *discrete* distribution with a finite number K of possible realizations $\xi_k = (q_k, h_k, T_k)$, $k = 1, \ldots, K$, (sometimes called *scenarios*), with the corresponding probabilities p_k. In that case, $\mathbb{E}[Q(x, \xi)] = \sum_{k=1}^{K} p_k Q(x, \xi_k)$, where

$$Q(x, \xi_k) = \min \{q_k^T y_k : T_k x + W y_k = h_k,\ y_k \geq 0\}$$

Therefore, the above two-stage problem can be formulated as one large linear program:

$$\begin{aligned} \min\ & c^T x + \textstyle\sum_{k=1}^{K} p_k q_k^T y_k \\ \text{s.t. } & Ax = b \\ & T_k x + W y_k = h_k \\ & x \geq 0,\ y_k \geq 0,\ k = 1, \ldots, K \end{aligned} \tag{16}$$

The linear program (16) has a certain block structure that makes it amenable to various decomposition methods. One such decomposition method is the popular L-shaped method developed by Van Slyke and Wets (1969). We refer the interested reader to the recent books by Kall and Wallace (1994) and Birge and Louveaux (1997) for a thorough discussion of stochastic programming with recourse.

The above numerical approach works reasonably well if the number K of scenarios is not too large. Suppose, however, that the random vector ξ has m independently distributed components, each having just three possible realizations. Then the total number of different scenarios is $K = 3^m$. That is, the number of scenarios grows exponentially fast in the number m of random variables. In that case, even for a moderate number of random variables, say $m = 100$, the number of scenarios becomes so large that even modern computers cannot cope with the required calculations. It seems that the only way to deal with such exponential growth of the number of scenarios is to use sampling. Such approaches are discussed in Section 3.2.

It may also happen that some of the decision variables at the first or second stage are integers, such as binary variables representing "yes" or "no" decisions. Such integer (or discrete) stochastic programs are especially difficult to solve, and only very moderate progress has been reported so far. A discussion of two-stage stochastic integer programs with recourse can be found in Birge and Louveaux (1997). A branch and bound approach for solving stochastic discrete optimization problems was suggested by Norkin et al. (1998). Schultz et al. (1998) suggested an algebraic approach for solving stochastic programs with integer recourse by using a framework of Gröbner basis reductions. For a recent survey of mainly theoretical results on stochastic integer programming see Klein Haneveld and Van der Vlerk (1999).

Conceptually the idea of two-stage programming with recourse can be readily extended to *multistage* programming with recourse. Such an approach tries to model the situation where decisions are made periodically (in stages) based on currently known realizations of some of the random variables. An H-stage stochastic linear program with fixed recourse can be written in the form

$$
\begin{aligned}
\min\ & c^1x^1 + \mathbb{E}\{\min c^2(\omega)x^2(\omega) + \ldots + \mathbb{E}[\min c^H(\omega)x^H(\omega)]\} \\
\text{s.t.}\ & W^1x^1 = h^1 \\
& T^1(\omega)x^1 + W^2x^2(\omega) = h^2(\omega) \\
& \ldots\ldots \\
& T^{H-1}(\omega)x^{H-1}(\omega) + W^Hx^H(\omega) = h^H(\omega) \\
& x^1 \geq 0, x^2(\omega) \geq 0, \ldots, x^H(\omega) \geq 0
\end{aligned}
\tag{17}
$$

The decision variables $x^2(\omega), \ldots, x^H(\omega)$ are allowed to depend on the random data ω. However, the decision $x^t(\omega)$ at time t can only depend on the part of the random data that is known at time t (these restrictions are often called nonanticipativity constraints). The expectations are taken with respect to the distribution of the random variables whose realizations are not yet known.

Again, if the distribution of the random data is discrete with a finite number of possible realizations, then problem (17) can be written as one large linear program. However, it is clear that even for a small number of stages and a moderate number of random variables the total number of possible scenarios will be astronomical. Therefore, a current approach to such problems is to generate a reasonable number of scenarios and solve the corresponding (deterministic) linear program, hoping to catch at least the flavor of the stochastic aspect of the problem. The argument is that the solution obtained in this way is more robust than the solution obtained by replacing the random variables with their means.

Often the same practical problem can be modeled in different ways. For instance, one can model a problem as a two-stage stochastic program with recourse, putting all random variables whose realizations are not yet known at the second stage of the problem. Then, as realizations of some of the random variables become known, the solutions are periodically updated in a two-stage rolling horizon fashion, every time by solving an updated two-stage problem. Such an approach is different from a multistage program with recourse, where every time a decision is to be made, the modeler tries to take into account that decisions will be made at several stages in the future.

3.2. Sampling Methods

In this section we discuss a different approach that uses Monte Carlo sampling techniques to solve stochastic optimization problems.

Example 2. Consider a stochastic process I_t, $t = 1, 2, \ldots$, governed by the recursive equation

$$I_t = [I_{t-1} + x_t - D_t]^+ \tag{18}$$

with initial value I_0. Here D_t are random variables and x_t represent decision variables. (Note that $[a]^+ \equiv \max\{a, 0\}$.) The above process I_t can describe the waiting time of the tth customer in a $G/G/1$ queue, where D_t is the interarrival time between the $(t - 1)$th and tth customers and x_t is the service time of $(t - 1)$th customer. Alternatively, I_t may represent the inventory of a certain product at time t, with D_t and x_t representing the demand and production (or ordering) quantities, respectively, of the product at time t.

Suppose that the process is considered over a finite horizon with time periods $t = 1, \ldots, T$. Our goal is to minimize (or maximize) the expected value of an objective function involving $I_1, \ldots, I_T$. For instance, one may be interested in maximizing the expected value of a profit given (Albritton et al. 1999);

$$
\begin{aligned}
G(x, W) &\equiv \textstyle\sum_{t=1}^{T} \{\pi_t \min[I_{t-1} + x_t, D_t] - h_tI_t\} \\
&= \textstyle\sum_{t=1}^{T} \pi_tx_t + \sum_{t=1}^{T-1} (\pi_{t+1} - \pi_t - h_t)I_t + \pi_1I_0 - (\pi_T + h_T)I_T
\end{aligned}
\tag{19}
$$

Here $x = (x_1, \ldots, x_T)$ is a vector of decision variables, $W = (D_1, \ldots, D_T)$ is a random vector of the demands at periods $t = 1, \ldots, T$, and π_t and h_t are nonnegative parameters representing the marginal profit and the holding cost, respectively, of the product at period t.

If the initial value I_0 is sufficiently large, then with probability close to one, variables $I_1, \ldots, I_T$ stay above zero. In that case $I_1, \ldots, I_T$ become linear functions of the random data vector W, and hence components of the random vector W can be replaced by their means. However, in many practical situations the process I_t hits zero with high probability over the considered horizon T. In such cases the corresponding expected value function $g(x) \equiv \mathbb{E}[G(x, W)]$ cannot be written in a closed form. One can use a Monte Carlo simulation procedure to evaluate $g(x)$. Note that for any given realization of D_t, the corresponding values of I_t, and hence the value of $G(x, W)$, can be easily calculated using the iterative formula (18).

That is, let $W^i = (D^i_1, \ldots, D^i_T)$, $i = 1, \ldots, N$, be a random (or pseudorandom) sample of N independent realizations of the random vector W generated by computer, that is, there are N generated realizations of the demand process D_t, $t = 1, 2, \ldots, T$, over the horizon t. Then for any given x the corresponding expected value $g(x)$ can be approximated (estimated) by the sample average

$$\hat{g}_N(x) \equiv \frac{1}{N} \sum_{i=1}^{N} G(x, W^i) \tag{20}$$

We have that $\mathbb{E}[\hat{g}_N(x)] = g(x)$, and by the law of large numbers, that $\hat{g}_N(x)$ converges to $g(x)$ with probability one (w.p.1) as $N \rightarrow \infty$. That is, $\hat{g}_N(x)$ is an *unbiased* and *consistent* estimator of $g(x)$.

Any reasonably efficient method for optimizing the expected value function $g(x)$, say by using its sample average approximations, is based on estimation of its first (and maybe second) order derivatives. This has an independent interest and is called *sensitivity* or *perturbation* analysis. We will discuss that in Section 3.3. Recall that $\nabla g(x) \equiv (\partial g(x)/\partial x_1, \ldots, \partial g(x)/\partial x_T)$ is called the gradient vector of $g(\cdot)$ at x.

It is possible to consider a stationary distribution of the process I_t (if it exists) and to optimize the expected value of an objective function with respect to the stationary distribution. Typically, such a stationary distribution cannot be written in a closed form and is difficult to compute accurately. This introduces additional technical difficulties into the problem. Also, in some situations the probability distribution of the random variables D_t is given in a parametric form whose parameters are decision variables. We will discuss dealing with such cases later.

3.3. Perturbation Analysis

Consider the expected value function $g(x) \equiv \mathbb{E}[G(x, \omega)]$. An important question is under which conditions the first order derivatives of $g(x)$ can be taken inside the expected value, that is, under which conditions the equation

$$\nabla g(x) \equiv \nabla \mathbb{E}[G(x, \omega)] = \mathbb{E}[\nabla_x G(x, \omega)] \tag{21}$$

is correct. One reason why this question is important is the following. Let $\omega^1, \ldots, \omega^N$ denote a random sample of N independent realizations of the random variable with common probability distribution P, and let

$$\hat{g}_N(x) \equiv \frac{1}{\mathrm{N}} \sum_{i=1}^{N} G(x, \omega^i) \tag{22}$$

be the corresponding sample average function. If the interchangeability equation (21) holds, then

$$\mathbb{E}[\nabla \hat{g}_N(x)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\nabla_x G(x, \omega^i)] = \frac{1}{N} \sum_{i=1}^{N} \nabla \mathbb{E}[G(x, \omega^i)] = \nabla g(x) \tag{23}$$

and hence $\nabla \hat{g}_N(x)$ is an unbiased and consistent estimator of $\nabla g(x)$.

Let us observe that in both examples 1 and 2 the function $G(\cdot, \omega)$ is piecewise linear for any realization of ω, and hence is not everywhere differentiable. The same holds for the optimal value function $Q(\cdot, \xi)$ of the second-stage problem (15). If the distribution of the corresponding random variables is discrete, then the resulting expected value function is also piecewise linear and hence is not everywhere differentiable.

On the other hand, expectation with respect to a continuous distribution typically smoothes the corresponding function and in such cases Eq. (21) often is applicable. It is possible to show that if the following two conditions hold at a point x, then $g(\cdot)$ is differentiable at x and Eq. (21) holds:

1. The function $G(\cdot, \omega)$ is differentiable at x w.p.1.
2. There exists a positive valued random variable $K(\omega)$ such that $\mathbb{E}[K(\omega)]$ is finite and the inequality

$$|G(x_1, \omega) - G(x_2, \omega)| \leq K(\omega)\|x_1 - x_2\| \tag{24}$$

holds w.p.1 for all x_1, x_2 in a neighborhood of x.

If the function $G(\cdot, \omega)$ is not differentiable at x w.p.1 (i.e., for P-almost every $\omega \in \Omega$), then the right-hand side of Eq. (21) does not make sense. Therefore, clearly the above condition (1) is necessary for (21) to hold. Note that condition (1) requires $G(\cdot, \omega)$ to be differentiable w.p.1 at the given (fixed) point x and does not require differentiability of $G(\cdot, \omega)$ everywhere. The second condition (ii) requires $G(\cdot, \omega)$ to be continuous (in fact Lipschitz continuous) w.p.1 in a neighborhood of x.

Consider, for instance, function $G(x, D)$ of example 1 defined in 1. For any given D, the function $G(\cdot, D)$ is piecewise linear and differentiable at every point x except at $x = D$. If the cdf $F(\cdot)$ of D is continuous at x, then the probability of the event $\{D = x\}$ is zero and hence the interchangeability equation (21) holds. Then $\partial G(x, D)/\partial x$ is equal to $s - c$ if $x > D$, and is equal to $r - c$ if $x < D$. Therefore, if $F(\cdot)$ is continuous at x, then $G(\cdot, D)$ is differentiable at x and

$$g'(x) = (s - c)\mathbb{P}(D < x) + (r - c)\mathbb{P}(D > x)$$

which gives the same equation as (6). Note that the function $\partial G(\cdot, D)/\partial x$ is discontinuous at $x = D$. Therefore, the second order derivative of $\mathbb{E}[G(\cdot, D)]$ cannot be taken inside the expected value. Indeed, the second order derivative of $G(\cdot, D)$ is zero whenever it exists. Such behavior is typical in many interesting applications.

Let us calculate the derivatives of the process I_t, defined by the recursive equation (18), for a particular realization of the random variables D_t. Let τ_1 denote the first time that the process I_t hits zero, that is, $\tau_1 \geq 1$ is the first time $I_{\tau_1 - 1} + x_{\tau_1} - D_{\tau_1}$ becomes less than or equal to zero, and hence $I_{\tau_1} = 0$. Let $\tau_2 > \tau_1$ be the second time that I_t hits zero, etc. Note that if $I_{\tau_1 + 1} = 0$, then $\tau_2 = \tau_1 + 1$, and so on. Let $1 \leq \tau_1 < \ldots < \tau_n \leq T$ be the sequence of hitting times. (In queueing terminology, τ_i represents the starting time of a new busy cycle of the corresponding queue.) For a given time $t \in \{1, \ldots, T\}$, let $\tau_{i-1} \leq t < \tau_i$ Suppose that the events $\{I_{\tau-1} + x_\tau - D_\tau = 0\}$, $\tau = 1, \ldots, T$, occur with probability zero. Then, for almost every W, the gradient of I_s with respect to the components of vector x_t can be written as follows:

$$\nabla_{x_t} I_s = \begin{cases} 1 & \text{if } t \leq s < \tau_i \text{ and } t \neq \tau_{i-1} \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

Thus, by using Eqs. (19) and (25), one can calculate the gradient of the sample average function $\hat{g}_N(\cdot)$ of example (2), and hence one can consistently estimate the gradient of the expected value function $g(\cdot)$.

Consider the process I_t defined by the recursive equation (18) again. Suppose now that variables x_t do not depend on t, and let x denote their common value. Suppose further that D_t, $t = 1, \ldots,$ are independently and identically distributed with mean $\mu > 0$. Then for $x < \mu$ the process I_t is stable and has a stationary (steady-state) distribution. Let $g(x)$ be the steady-state mean (the expected value with respect to the stationary distribution) of the process $I_t = I_t(x)$. By the theory of regenerative processes it follows that for every $x \in (0, \mu)$ and any realization (called sample path) of the process D_t, $t = 1, \ldots,$ the-long run average $\hat{g}_T(x) \equiv \Sigma_{t=1}^T I_t(x)/T$ converges w.p.1 to $g(x)$ as $T \to \infty$. It is possible to show that $\nabla \hat{g}_T(x)$ also converges w.p.1 to $\nabla g(x)$ as $T \to \infty$. That is, by differentiating the long-run average of a sample path of the process I_t we obtain a consistent estimate of the corresponding derivative of the steady-state mean $g(x)$. Note that $\nabla I_t(x) = t - \tau_{i-1}$ for $\tau_{i-1} \leq t < \tau_i$, and hence the derivative of the long-run average of a sample path of the process I_t can be easily calculated.

The idea of differentiation of a sample path of a process in order to estimate the corresponding derivative of the steady-state mean function by a single simulation run is at the heart of *infinitesimal perturbation analysis*. We refer the interested reader to Glasserman (1991) and Ho and Cao (1991) for a thorough discussion of that topic.

3.4. Likelihood Ratio Method

The Monte Carlo sampling approach to derivative estimation introduced in Section 3.3 does not work if the function $G(\cdot, \omega)$ is discontinuous or if the corresponding probability distribution also depends on decision variables. In this section we discuss an alternative approach to derivative estimation known as the *likelihood ratio* (or *score function*) method.

Suppose that the expected value function is given in the form $g(\theta) \equiv \mathbb{E}_\theta[G(W)]$, where W is a random vector whose distribution depends on the parameter vector θ. Suppose further that the distribution of W has a probability density function (pdf) $f(\theta, w)$. Then for a chosen pdf $\phi(w)$ we can write

$$\mathbb{E}_\theta[G(W)] = \int G(w) f(\theta, w)\, dw = \int G(w) \frac{f(\theta, w)}{\phi(w)} \phi(w)\, dw$$

and hence

$$g(\theta) = \mathbb{E}_{\phi}[G(Z)L(\theta, Z)] \tag{26}$$

where $L(\theta, z) \equiv f(\theta, z)/\phi(z)$ is the so-called likelihood ratio function, $Z \sim \phi(\cdot)$ and $\mathbb{E}_{\phi}[\cdot]$ means that the expectation is taken with respect to the pdf ϕ. We assume in the definition of the likelihood ratio function that $0/0 = 0$ and that the pdf ϕ is such that if $\phi(w)$ is zero for some w, then $f(\theta, w)$ is also zero, that is, we do not divide a positive number by zero.

The expected value in the right-hand side of (26) is taken with respect to the distribution ϕ, which does not depend on the vector θ. Therefore, under appropriate conditions ensuring interchangeability of the differentiation and integration operators, we can write

$$\nabla g(\theta) = \mathbb{E}_{\phi}[G(Z)\nabla_{\theta}L(\theta, Z)] \tag{27}$$

In particular, if for a given θ_0 we choose $\phi(\cdot) \equiv f(\theta_0, \cdot)$, then $\nabla_{\theta}L(\theta, z) = \nabla_{\theta}f(\theta, z)/f(\theta_0, z)$, and hence $\nabla_{\theta}L(\theta_0, z) = \nabla_{\theta}\ln[f(\theta_0, z)]$. The function $\nabla_{\theta}\ln[f(\theta, z)]$ is called the score function; thus the name of this technique.

Now by generating a random sample $Z^1, \ldots, Z^N$ from the pdf $\phi(\cdot)$, one can estimate $g(\theta)$ and $\nabla g(\theta)$ by the respective sample averages

$$\tilde{g}_N(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N} G(Z^i)L(\theta, Z^i) \tag{28}$$

$$\nabla\tilde{g}_N(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N} G(Z^i)\nabla_{\theta}L(\theta, Z^i) \tag{29}$$

This can be readily extended to situations where function $G(x, W)$ also depends on decision variables.

Typically, the density functions used in applications depend on the decision variables in a smooth and even analytic way. Therefore, usually there is no problem in taking derivatives inside the expected value in the right-hand side of (26). When applicable, the likelihood ratio method often also allows estimation of second and higher order derivatives. However, note that the likelihood ratio method is notoriously unstable and a bad choice of the pdf ϕ may result in huge variances of the corresponding estimators. This should not be surprising since the likelihood ratio function may involve divisions by very small numbers, which of course is a very unstable procedure. We refer to Glynn (1990) and Rubinstein and Shapiro (1993) for a further discussion of the likelihood ratio method.

As an example consider the optimal value function of the second stage problem (15). Suppose that only the right-hand-side vector $h = h(\omega)$ of the second-stage problem is random. Then $Q(x, h) = G(h - Tx)$, where $G(\chi) \equiv \min\{q^T y{:}Wy = \chi, y \geq 0\}$. Suppose that the random vector h has a pdf $f(\cdot)$. By using the transformation $z = h - Tx$, we obtain

$$\mathbb{E}_f[Q(x, h)] = \int G(\eta - Tx)f(\eta)\,d\eta = \int G(z)f(z + Tx)\,dz = \mathbb{E}_{\phi}[G(Z)L(x, Z)] \tag{30}$$

Here ϕ is a chosen pdf, z is a random vector having pdf ϕ, and $L(x, z) \equiv f(z + Tx)/\phi(z)$ is the corresponding likelihood ratio function. It can be shown by duality arguments of linear programming that $G(\cdot)$ is a piecewise linear convex function. Therefore, $\nabla_x Q(x, h)$ is piecewise constant and discontinuous, and hence second order derivatives of $\mathbb{E}_f[Q(x, h)]$ cannot be taken inside the expected value. On the other hand, the likelihood ratio function is as smooth as the pdf $f(\cdot)$. Therefore, if $f(\cdot)$ is twice differentiable, then the second order derivatives can be taken inside the expected value in the right-hand side of (30), and consequently the second order derivatives of $\mathbb{E}_f[Q(x, h)]$ can be consistently estimated by a sample average.

3.5. Simulation-Based Optimization Methods

There are basically two approaches to the numerical solution of stochastic optimization problems by using Monte Carlo sampling techniques. One approach is known as the *stochastic approximation* method and originated in Robbins and Monro (1951). The other method was discovered and rediscovered by different researchers and is known under various names.

Suppose that the feasible set χ is convex and that at any point $x \in \chi$ an estimate $\hat{\gamma}(x)$ of the gradient $\nabla g(x)$ can be computed, say by a Monte Carlo simulation method. The stochastic approximation method generates the iterates by the recursive equation

$$x_{\nu+1} = \Pi_{\chi}(x_{\nu} - \alpha_{\nu}\hat{\gamma}(x_{\nu})) \tag{31}$$

where $\alpha_{\nu} > 0$ are chosen step sizes and Π_{χ} denotes the projection onto the set χ, that is, $\Pi_{\chi}(x)$ is the point in χ closest to x. Under certain regularity conditions the iterates x_{ν} converge to a locally optimal solution of the corresponding stochastic optimization problem, that is, to a local minimizer x^* of $g(x)$ over χ. Typically, in order to guarantee this convergence the following two conditions are imposed on the step sizes: (1) $\Sigma_{\nu=1}^{\infty} \alpha_\nu = \infty$, and (2) $\Sigma_{\nu=1}^{\infty} \alpha_\nu^2 < \infty$. For example, one can take $\alpha_\upsilon \equiv c/\nu$ for some $c > 0$.

If the exact value $\gamma_\nu \equiv \nabla g(x_\nu)$ of the gradient is known, then $-\gamma_\nu$ gives the direction of steepest descent at the point x_ν. This guarantees that if $\gamma_\nu \neq 0$, then moving along the direction $-\gamma_\nu$ the value of the objective function decreases, that is, $g(x_\nu - \alpha\gamma_\nu) < g(x_\nu)$ for $\alpha > 0$ small enough. The iterative procedure (31) tries to mimic that idea by using the estimates $\hat{\gamma}(x_\nu)$ of the corresponding "true" gradients. The projection Π_χ is needed in order to enforce feasibility of the generated iterates. If the problem is unconstrained, that is, the feasible set χ coincides with the whole space, then this projection is the identity mapping and can be omitted from (31). Note that $\hat{\gamma}(x_\nu)$ does not need to be an accurate estimator of $\nabla g(x_\nu)$.

Kushner and Clark (1978) and Benveniste et al. (1990) contain expositions of the theory of stochastic approximation. Applications of the stochastic approximation method, combined with the infinitesimal perturbation analysis technique for gradient estimation, to the optimization of the steady-state means of single-server queues were studied by Chong and Ramadge (1992) and L'Ecuyer and Glynn (1994).

An attractive feature of the stochastic approximation method is its simplicity and ease of implementation in those cases in which the projection $\Pi_\chi(\cdot)$ can be easily computed. However, it also has severe shortcomings. The crucial question in implementations is the choice of the step sizes α_ν. Small step sizes result in very slow progress towards the optimum while large step sizes make the iterates zigzag. Also, a few wrong steps in the beginning of the procedure may require many iterations to correct. For instance, the algorithm is extremely sensitive to the choice of the constant c in the step size rule $\alpha_\nu = c/\nu$. Therefore, various step size rules were suggested in which the step sizes are chosen adaptively (see Ruppert 1991 for a discussion of that topic).

Another drawback of the stochastic approximation method is that it lacks good stopping criteria and often has difficulties with handling even relatively simple linear constraints.

Another simulation-based approach to stochastic optimization is based on the following idea. Let $\hat{g}_N(x)$ be the sample average function defined in (22), based on a sample of size N. Consider the optimization problem

$$\min_{x \in \chi} \hat{g}_N(x) \tag{32}$$

We can view the above problem as the sample average approximation of the "true" (or expected value) problem (11). The function $\hat{g}_N(x)$ is random in the sense that it depends on the corresponding sample. However, note that once the sample is generated, $\hat{g}_N(x)$ becomes a deterministic function whose values and derivatives can be computed for a given value of the argument x. Consequently, problem (32) becomes a deterministic optimization problem and one can solve it with an appropriate deterministic optimization algorithm.

Let $\hat{v}_N$ and $\hat{x}_N$ denote the optimal objective value and an optimal solution of the sample average problem (32), respectively. By the law of large numbers we have that $\hat{g}_N(x)$ converges to $g(x)$ w.p.1 as $N \to \infty$. It is possible to show that under mild additional conditions, $\hat{v}_N$ and $\hat{x}_N$ converge w.p.1 to the optimal objective value and an optimal solution of the true problem (11), respectively. That is, $\hat{v}_N$ and $\hat{x}_N$ are consistent estimators of their "true" counterparts.

This approach to the numerical solution of stochastic optimization problems is a natural outgrowth of the Monte Carlo method of estimation of the expected value of a random function. The method is known by various names, and it is difficult to point out who was the first to suggest this approach. In the recent literature a variant of this method, based on the likelihood ratio estimator $\tilde{g}_N(x)$, was suggested in Rubinstein and Shapiro (1990) under the name *stochastic counterpart method* (also see Rubinstein and Shapiro 1993 for a thorough discussion of such a likelihood ratio–sample approximation approach). In Robinson (1996) such an approach is called the *sample path method*. This idea can also be applied to cases in which the set χ is finite, that is, to stochastic discrete optimization problems (Kleywegt and Shapiro 1999).

Of course, in a practical implementation of such a method, one has to choose a specific algorithm for solving the sample average approximation problem (32). For example, in the unconstrained case, one can use the steepest descent method. That is, iterates are computed by the procedure

$$x_{\nu+1} = x_\nu - \alpha_\nu \nabla \hat{g}_N(x_\nu) \tag{33}$$

where the step size α_ν is obtained by a line search, such as $\alpha_\nu \equiv \arg\min_\alpha \hat{g}_N(x_\nu - \alpha \nabla \hat{g}_N(x_\nu))$. Note that this procedure is different from the stochastic approximation method (31) in two respects. Typically a reasonably large sample size N is used in this procedure, and, more importantly, the step sizes are calculated by a line search instead of being defined a priori. In many interesting cases

$\hat{g}_N(x)$ is a piecewise smooth (and even piecewise linear) function and the feasible set is defined by linear constraints. In such cases bundle-type optimization algorithms are quite efficient (see Hiriart-Urruty and Lemarechal 1993 for a discussion of the bundle method).

A well-developed statistical inference of the estimators $\hat{v}_N$ and $\hat{x}_N$ exists (Rubinstein and Shapiro 1993). That inference aids in the construction of stopping rules, validation analysis, and error bounds for obtained solutions and, furthermore, suggests variance reduction methods that may substantially enhance the rate of convergence of the numerical procedure. For a discussion of this topic and an application to two-stage stochastic programming with recourse, we refer to Shapiro and Homem-de-Mello (1998).

If the function $g(x)$ is twice differentiable, then the above sample path method produces estimators that converge to an optimal solution of the true problem at the same asymptotic rate as the stochastic approximation method, provided that the stochastic approximation method is applied with the *asymptotically optimal* step sizes (Shapiro 1996). On the other hand, if the underlying probability distribution is discrete and $g(x)$ is piecewise linear and convex, then w.p.1 the sample path method provides an exact optimal solution of the true problem for N large enough, and moreover the probability of that event approaches one exponentially fast as $N \to \infty$ (Shapiro and Homem-de-Mello 1999).

4. DYNAMIC PROGRAMMING

Dynamic programming (DP) is an approach for the modeling of dynamic and stochastic decision problems, the analysis of the structural properties of these problems, and the solution of these problems. Dynamic programs are also referred to as Markov decision processes (MDP). Slight distinctions can be made between DP and MDP, such as that in the case of some deterministic problems the term *dynamic programming* is used rather than *Markov decision processes*. The term *stochastic optimal control* is also often used for these types of problems. We shall use these terms synonymously.

Dynamic programs and multistage stochastic programs deal with essentially the same types of problems, namely dynamic and stochastic decision problems. The major distinction between dynamic programming and stochastic programming is in the structures that are used to formulate the models. For example, in DP, the so-called state of the process, as well as the value function, that depends on the state, are two structures that play a central role, but these concepts are usually not used in stochastic programs. Section 4.1 provides an introduction to concepts that are important in dynamic programming.

Much has been written about dynamic programming. Some books in this area are Bellman (1957), Bellman (1961), Bellman and Dreyfus (1962), Nemhauser (1966), Hinderer (1970), Bertsekas and Shreve (1978), Denardo (1982), Ross (1983), Puterman (1994), Bertsekas (1995), and Sennott (1999).

The dynamic programming modeling concepts presented in this chapter are illustrated with an example, which is both a multiperiod extension of the single-period newsvendor problem of Section 2 as well as an example of a dynamic pricing problem.

Example 3 (Revenue management problem). Managers often have to make decisions repeatedly over time regarding how much inventory to obtain for future sales as well as how to determine the selling prices. This may involve inventory of one or more products, and the inventory may be located at one or more locations, such as warehouses and retail stores. The inventory may be obtained from a production operation that is part of the same company as the decision maker, and such a production operation may be a manufacturing operation or a service operation, such as an airline, hotel, or car rental company, or the inventory may be purchased from independent suppliers. The decision maker may also have the option to move inventory between locations, such as from warehouses to retail stores. Often the prices of the products can be varied over time to attempt to find the most favorable balance between the supply of the products and the dynamically evolving demand for the products. Such a decision maker can have several objectives, such as to maximize the expected profit over the long run. The profit involves both revenue, which is affected by the pricing decisions, as well as cost, which is affected by the inventory replenishment decisions.

In Section 4.1 examples are given of the formulation of such a revenue management problem with a single product at a single location as a dynamic program.

4.1. Basic Concepts in Dynamic Programming

In this section the basic concepts used in dynamic programming models are introduced.

4.1.1. Decision Times

A dynamic programming model should distinguish between the decisions made at different points in time. The major reason for this is that the information available to the decision maker is different at different points in time—typically more information is available at later points in time (in fact, many people hold this to be the definition of time).

A second reason why distinguishing decision points is useful is that for many types of DP models it facilitates the computation of solutions. This seems to be the major reason why dynamic programming is used for deterministic decision problems. In this context, the time parameter in the model does not need to correspond to the notion of time in the application. The important feature is that a solution is decomposed into a sequence of distinct decisions. This facilitates computation of the solution if it is easier to compute the individual decisions and then put them together to form a solution than it is to compute a solution in a more direct way.

The following are examples of ways in which the decision points can be determined in a DP model:

- Decisions can be made at predetermined discrete points in time. In the revenue management example, the decision maker may make a decision once per day regarding what prices to set during the day, as well as how much to order on that day.
- Decisions can be made continuously in time. In the revenue management example, the decision maker may change prices continuously in time (which is likely to require a sophisticated way of communicating the continuously changing prices).
- Decisions can be made at random points in time when specific events take place. In the revenue management example, the decision maker may decide on prices at the random points in time when customer requests are received and may decide whether to order and how much to order at the random points in time when the inventory changes.

A well-formulated DP model specifies the way in which the decision points in time are determined.

Most of the results presented in this article are for DP models where decisions are made at predetermined discrete points in time, denoted by $t = 0, 1, \ldots, T$, where T denotes the length of the time horizon. DP models with infinite time horizons are also considered. DP models such as these are often called discrete-time DP models.

4.1.2. States

A fundamental concept in DP is that of a state, denoted by s. The set $\mathbb{S}$ of all possible states is called the state space. The decision problem is often described as a controlled stochastic process that occupies a state $S(t)$ at each point in time t.

Describing the stochastic process for a given decision problem is an exercise in modeling. The modeler has to determine an appropriate choice of state description for the problem. The basic idea is that the state should be a sufficient, and efficient, summary of the available information that affects the future of the stochastic process. For example, for the revenue management problem, choosing the state to be the amount of the product in inventory may be an appropriate choice. If there is a cost involved in changing the price, then the previous price should also form part of the state. Also, if competitors' prices affect the demand for the product, then additional information about competitors' prices and behavior should be included in the state.

Several considerations should be taken into account when choosing the state description, some of which are described in more detail in later sections. A brief overview is as follows. The state should be a sufficient summary of the available information that affects the future of the stochastic process in the following sense. The state at a point in time should not contain information that is not available to the decision maker at that time, because the decision is based on the state at that point in time. (There are also problems, called partially observed Markov decision processes, in which what is also called the state contains information that is not available to the decision maker. These problems are often handled by converting them to Markov decision processes with observable states. This topic is discussed in Bertsekas [1995].) The set of feasible decisions at a point in time should depend only on the state at that point in time, and maybe on the time itself, and not on any additional information. Also, the costs and transition probabilities at a point in time should depend only on the state at that point in time, the decision made at that point in time, and maybe on the time itself, and not on any additional information. Another consideration is that often one would like to choose the number of states to be as small as possible since the computational effort of many algorithms increase with the size of the state space. However, the number of states is not the only factor that affects the computational effort. Sometimes it may be more efficient to choose a state description that leads to a larger state space. In this sense the state should be an efficient summary of the available information.

The state space $\mathbb{S}$ can be a finite, countably infinite, or uncountable set. This article addresses dynamic programs with finite or countably infinite, also called discrete, state spaces $\mathbb{S}$.

4.1.3. Decisions

At each decision point in time, the decision maker has to choose a decision, also called an action or control. At any point in time t, the state s at time t, and the time t, should be sufficient to determine the set $\mathcal{A}(s, t)$ of feasible decisions, that is, no additional information is needed to determine the admissible decisions. (Note that the definition of the state of the process should be chosen in such a way that this holds for the decision problem under consideration.) Sometimes the set of feasible decisions depends only on the current state s, in which case the set of feasible decisions is denoted by $\mathcal{A}(s)$. Although most examples have finite sets $\mathcal{A}(s, t)$ or $\mathcal{A}(s)$, these sets may also be countably or uncountably infinite.

In the revenue management example, the decisions involve how much of the product to order, as well as how to set the price. Thus, decision $a = (q, r)$ denotes that quantity q is ordered and that the price is set at r. Suppose the supplier requires that an integer amount between a and b be ordered at a time. Also suppose that the state s denotes the current inventory, and that the inventory may not exceed capacity Q at any time. Then the order quantity may be no more than $Q - s$. Also suppose that the price can be set to be any real number between r_1 and r_2. Then the set of feasible decisions is $\mathcal{A}(s) = \{a, a + 1, a + 2, \ldots, \min\{Q - s, b\}\} \times [r_1, r_2]$.

The decision maker may randomly select a decision. For example, the decision maker may roll a die and base the decision on the outcome of the die roll. This type of decision is called a randomized decision, as opposed to a nonrandomized, or deterministic, decision. A randomized decision for state s at time t can be represented by a probability distribution on $\mathcal{A}(s, t)$ or $\mathcal{A}(s)$. The decision at time t is denoted by $A(t)$.

4.1.4. Transition Probabilities

The dynamic process changes from state to state over time. The transitions between states may be deterministic or random. The presentation here is for a dynamic program with discrete time parameter $t = 0, 1, \ldots$, and with random transitions.

The transitions have a memoryless, or Markovian, property, in the following sense. Given the history $H(t) \equiv (S(0), A(0), S(1), A(1), \ldots, S(t))$ of the process up to time t, as well as the decision $A(t) \in \mathcal{A}(S(t), t)$ at time t, the probability distribution of the state that the process is in at time $t + 1$ depends only on $S(t)$, $A(t)$, and t, that is, the additional information in the history $H(t)$ of the process up to time t provides no additional information for the probability distribution of the state at time $t + 1$. (Note that the definition of the state of the process should be chosen in such a way that the probability distribution has this memoryless property.)

Such memoryless random transitions can be represented in several ways. One representation is by transition probabilities, which are denoted by $p[s'|s, a, t] \equiv \mathbb{P}[S(t + 1) = s'|H(t), S(t) = s, A(t) = a]$. Another representation is by a transition function f, such that given $H(\mathrm{t})$, $S(t) = s$, and $A(\mathrm{t}) = a$, the state at time $t + 1$ is $S(t + 1) = f(s, a, t, \omega)$, where ω is a random variable with a known probability distribution. The two representations are equivalent, and in this article we use mostly transition probabilities. When the transition probabilities do not depend on the time t beside depending on the state s and decision a at time t, they are denoted by $p[s'|s, a]$.

In the revenue management example, suppose the demand has probability mass function $\tilde{p}(r, d) \equiv \mathbb{P}[D = d|\text{price} = r]$ with $d \in \{0, 1, 2, \ldots\}$. Also suppose that a quantity q that is ordered at time t is received before time $t + 1$, and that unsatisfied demand is back-ordered. Then $\mathcal{S} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, and the transition probabilities are as follows.

$$p[s'|s, (q, r)] = \begin{cases} \tilde{p}(r, s + q - s') & \text{if } s' \leq s + q \\ 0 & \text{if } s' > s + q \end{cases}$$

If a quantity q that is ordered at time t is received after the demand at time t, and unsatisfied demand is lost, then $\mathcal{S} = \{0, 1, 2, \ldots\}$, and the transition probabilities are as follows:

$$p[s'|s, (q, r)] = \begin{cases} \tilde{p}(r, s + q - s') & \text{if } q < s' \leq s + q \\ \Sigma_{d=s}^{\infty} \tilde{p}(r, d) & \text{if } s' = q \\ 0 & \text{if } s' < q \text{ or } s' > s + q \end{cases}$$

4.1.5. Rewards and Costs

Dynamic decision problems often have as objective to maximize the sum of the rewards obtained in each time period, or equivalently, to minimize the sum of the costs incurred in each time period. Other types of objectives sometimes encountered are to maximize or minimize the product of a sequence of numbers resulting from a sequence of decisions, or to maximize or minimize the maximum or minimum of a sequence of resulting numbers.

In this article we focus mainly on the objective of maximizing the expected sum of the rewards obtained in each time period. At any point in time t, the state s at time t, the decision $a \in \mathcal{A}(s, t)$ at time t, and the time t, should be sufficient to determine the expected reward $r(s, a, t)$ at time t.

(Again, the definition of the state should be chosen so that this holds for the decision problem under consideration.) When the rewards do not depend on the time t beside depending on the state s and decision a at time t, they are denoted by $r(s, a)$.

Note that even if in the application the reward $\tilde{r}(s, a, t, s')$ at time t depends on the state s' at time $t + 1$, in addition to the state s and decision a at time t, and the time t, the expected reward at time t can still be found as a function of only s, a, and t, because

$$r(s, a, t) = \mathbb{E}[\tilde{r}(s, a, t, s')] = \sum_{s' \in \mathbb{S}} \tilde{r}(s, a, t, s') p[s' | s, a, t]$$

In the revenue management example, suppose unsatisfied demand is back-ordered and that an inventory cost/shortage penalty of $h(s)$ is incurred when the inventory level is s at the beginning of the time period. Then $\tilde{r}(s, (q, r'), s') = r' \, (s + q - s') - h(s)$ with $s' \leq s + q$. Thus,

$$r(s, (q, r')) = \sum_{d=0}^{\infty} \tilde{p}(r', d) r' d - h(s)$$

If unsatisfied demand is lost, then $\tilde{r}(s, (q, r'), s') = r' \, (s + q - s') - h(s)$ with $q \leq s' \leq s + q$. Thus,

$$r(s, (q, r')) = \sum_{d=0}^{s-1} \tilde{p}(r', d) r' d + \sum_{d=s}^{\infty} \tilde{p}(r', d) r' s - h(s)$$

In finite horizon problems, there may be a salvage value $v(s)$ if the process terminates in state s at the end of the time horizon t. Such a feature can be incorporated in the previous notation by letting $\mathcal{A}(s, T) = \{0\}$, and $r(s, 0, T) = v(s)$ for all $s \in \mathbb{S}$.

Often the rewards are discounted with a discount factor $\alpha \in [0, 1]$, so that the discounted expected value of the reward at time t is $\alpha^t r(s, a, t)$. Such a feature can again be incorporated in the previous notation by letting $r(s, a, t) = \alpha^t \bar{r}(s, a, t)$ for all s, a, and t, where $\bar{r}$ denotes the undiscounted reward function. When the undiscounted reward does not depend on time, it is convenient to denote explicitly the discounted reward by $\alpha^t r(s, a)$.

4.1.6. Policies

A policy, sometimes called a strategy, prescribes the way a decision is to be made at each point in time, given the information available to the decision maker at the point in time. Therefore, a policy is a solution for a dynamic program.

There are different classes of policies of interest, depending on which of the available information the decisions are based on. A policy can base decisions on all the information in the history of the process up to the time the decision is to be made. Such policies are called history-dependent policies. Given the memoryless nature of the transition probabilities, as well as the fact that the sets of feasible decisions and the expected rewards depend on the history of the process only through the current state, it seems intuitive that it should be sufficient to consider policies that base decisions only on the current state and time, and not on any additional information in the history of the process. Such policies are called memoryless, or Markovian, policies. If the transition probabilities, sets of feasible decisions, and rewards do not depend on the current time, then it also seems intuitive that it should be sufficient to consider policies that base decisions only on the current state, and not on any additional information in the history of the process or on the current time. (However, this intuition may be wrong, as shown by counterexample in Section 4.1.7). Under such policies, decisions are made in the same way each time the process is in the same state. Such policies are called stationary policies.

The decision maker may also choose to use some irrelevant information to make a decision. For example, the decision maker may randomly select a decision by rolling a die or drawing a card from a deck of cards. Policies that allow such randomized decisions are called randomized policies, and policies that do not allow randomized decisions are called nonrandomized or deterministic policies.

Combining the above types of information that policies can base decisions on, the following types of policies are obtained: the class Π^{HR} of history dependent randomized policies, the class Π^{HD} of history dependent deterministic policies, the class Π^{MR} of memoryless randomized policies, the class Π^{MD} of memoryless deterministic policies, the class Π^{SR} of stationary randomized policies, and the class Π^{SD} of stationary deterministic policies. The classes of policies are related as follows: $\Pi^{SD} \subset \Pi^{MD} \subset \Pi^{HD} \subset \Pi^{HR}$, $\Pi^{SD} \subset \Pi^{MD} \subset \Pi^{MR} \subset \Pi^{HR}$, $\Pi^{SD} \subset \Pi^{SR} \subset \Pi^{MR} \subset \Pi^{HR}$.

For the revenue management problem, an example of a stationary deterministic policy is to order quantity $q = s_2 - s$ if the inventory level $s < s_1$, for chosen constants $s_1 \leq s_2$, and to set the price at level $r = \check{r}(s)$ for a chosen function $\check{r}(s)$ of the current state s. An example of a stationary randomized policy is to set the price at level $r = \check{r}_1(s)$ with probability $p_1(s)$ and at level $r = \check{r}_2(s)$

with probability $1 - p_1(s)$ for chosen functions $\check{r}_1(s)$, $\check{r}_2(s)$, and $p_1(s)$ of the current state s. An example of a memoryless deterministic policy is to order quantity $q = s_2(t) - s$ if the inventory level $s < s_1(t)$, for chosen functions $s_1(t) \leq s_2(t)$ of the current time t, and to set the price at level $r = \check{r}(s, t)$ for a chosen function $\check{r}(s, t)$ of the current state s and time t.

4.1.7. Example

In this section an example is presented that illustrates why it is sometimes desirable to consider more general classes of policies, such as memoryless and/or randomized policies, instead of stationary deterministic policies, even if the sets of feasible solutions, transition probabilities, and rewards are stationary. More such examples may be found in Ross (1970), Ross (1983), Puterman (1994), and Sennott (1999).

The examples are for dynamic programs with stationary input data and objective to minimize the long-run average cost per unit time, $\limsup_{T\to\infty} \mathbb{E}[\sum_{t=0}^{T-1} r(S(t), A(t)) | S(0)]/T$. For any policy π, let

$$V^{\pi}(s) \equiv \limsup_{T\to\infty} \frac{1}{T}\, \mathbb{E}^{\pi}\left[\sum_{t=0}^{T-1} r(S(t), A(t)) \,\middle|\, S(0) = s\right]$$

denote the long-run average cost per unit time under policy π if the process starts in state s, where $\mathbb{E}^{\pi}[\cdot]$ denotes the expected value if policy π is followed.

A policy π^* is called optimal if $V^{\pi^*}(s) = \inf_{\pi\in\Pi^{HR}} V^{\pi}(s)$ for all states s.

Example 4. This example shows that even if the dynamic program has stationary input data, it does not always hold that for any policy $\tilde{\pi}$, and any $\epsilon > 0$, there exists a stationary deterministic policy π that has value function V^{π} within ϵ of the value function $V^{\tilde{\pi}}$ of policy $\tilde{\pi}$.

The state space $\mathbb{S} = \{0, 1, 1', 2, 2', 3, 3', \ldots\}$. Feasible decision sets are $\mathcal{A}(0) = \{a\}$, $\mathcal{A}(i) = \{a, b\}$, and $\mathcal{A}(i') = \{a\}$ for each $i \in \{1, 2, 3, \ldots\}$. When in state $i \in \{0, 1, 2, \ldots\}$, a cost of 2 is incurred, otherwise there is no cost. That is, the costs are $r(0, a) = 2$, $r(i, a) = r(i, b) = 2$, and $r(i', a) = 0$. The transition probabilities are as follows:

$$p[0|0, a] = 1,\ p[i+1|i, a] = 1,\ p[i'|i, b] = 1 - p[0|i, b] = p_i \qquad \text{for all } i$$

$$p[1|1', a] = 1, \text{ and } p[(i-1)'|i', a] = 1 \qquad \text{for all } i \geq 2$$

The values p_i can be chosen to satisfy

$$p_i < 1 \text{ for all } i, \text{ and } \prod_{i=1}^{\infty} p_i = \frac{3}{4}$$

Suppose the process starts in state 1. The idea is simple: we would like to go down the chain i', $(i-1)'$, $\ldots$, $1'$ as much as possible. To do that, we also need to go up the chain $1, 2, \ldots, i$, and then go from state i to state i' by making decision b. When we make decision b in state i, there is a risk $1 - p_i > 0$ of making a transition to state 0, which is very bad.

A stationary deterministic policy π that chooses decision a for each state i, has long-run average cost per unit time of $V^{\pi}(1) = 2$, which is as bad as can be. The only other possibility for a stationary deterministic policy π is to choose decision b for the first time in state j. In that case, each time state j is visited, there is a positive probability $1 - p_j > 0$ of making a transition to state 0. It follows that the mean time until a transition to state 0 is made is less than $2j/(1 - p_j) < \infty$, and the long-run average cost per unit time is $V^{\pi}(1) = 2$. Thus, $V^{\pi}(1) = 2$ for all stationary deterministic policies π.

Consider the memoryless deterministic policy $\tilde{\pi}$ that on its jth visit to state 1, chooses decision a, $j - 1$ times, and then chooses decision b. With probability $\Pi_{i=1}^{\infty} p_i = 3/4$, the process never makes a transition to state 0 and the long-run average cost per unit time is 1. Otherwise, with probability $1 - \Pi_{i=1}^{\infty} p_i = 1/4$, the process makes a transition to state 0 and the long-run average cost per unit time is 2. Hence, the expected long-run average cost per unit time is $V^{\tilde{\pi}}(1) = 3/4 \times 1 + 1/4 \times 2 = 5/4$. Thus, there is no ε-optimal stationary deterministic policy for $\varepsilon \in (0, 3/4)$. In fact, by considering memoryless deterministic policies $\tilde{\pi}_k$ that on their jth visit to state 1, choose decision a, $j + k$ times and then choose decision b, one obtains policies with expected long-run average cost per unit time $V^{\tilde{\pi}_k}(1)$ arbitrarily close to 1 for sufficiently large values of k. It is clear that $V^{\pi}(1) \geq 1$ for all policies π, and thus $V^*(1) = 1$, and there is no ε-optimal stationary deterministic policy for $\varepsilon \in (0, 1)$.

4.2. Finite Horizon Dynamic Programs

In this section we investigate dynamic programming models for optimization problems with the form

$$\max_{(A(0),A(1),\ldots,A(T))} \mathbb{E}\left[\sum_{t=0}^{T} r(S(t), A(t), t)\right] \tag{34}$$

where $T < \infty$ is the known finite horizon length and decisions $A(t)$, $t = 0, 1, \ldots, T$, have to be feasible and may depend only on the information available to the decision maker at each time t, that is, the history H(t) of the process up to time t, and possibly some randomization. For the presentation we assume that $\mathbb{S}$ is countable and r is bounded. Similar results hold in more general cases, subject to regularity conditions.

4.2.1. Optimality Results

From the memoryless properties of the feasible sets, transition probabilities, and rewards, it is intuitive that it should be sufficient to consider memoryless deterministic policies. This can be shown to be true for finite horizon problems of the form (34).

The value function V^π of a memoryless policy π is defined by

$$V^\pi(s, t) \equiv \mathbb{E}^\pi\left[\sum_{\tau=t}^{T} r(S(\tau), A(\tau), \tau) \,\middle|\, S(t) = s\right] \tag{35}$$

Then, because it is sufficient to consider memoryless deterministic policies, the optimal value function V^* is given by

$$V^*(s, t) \equiv \sup_{\pi\in\Pi^{MD}} V^\pi(s, t) \tag{36}$$

It is easy to see that the value function V^π of a memoryless policy π satisfies the following inductive equation:

$$V^\pi(s, t) = r(s, \pi(s, t), t) + \sum_{s'\in\mathbb{S}} p[s'|s, \pi(s, t), t)]V^\pi(s', t + 1) \tag{37}$$

(Recall that $\pi(s, t)$ denotes the decision under policy π if the process is in state s at time t. If π is a randomized policy, then the understanding is that the expected value is computed with the decision distributed according to probability distribution $\pi(s, t)$. Also, even history-dependent policies satisfy a similar inductive equation, except that the value function depends on the history up to time t.) Similarly, the optimal value function V^* satisfies the following inductive optimality equation:

$$V^*(s, t) = \sup_{a\in\mathbb{A}(s,t)} \left\{ r(s, a, t) + \sum_{s'\in\mathbb{S}} p[s'|s, a, t]V^*(s', t + 1) \right\} \tag{38}$$

4.2.2. Finite Horizon Algorithm

Solving a finite horizon dynamic program usually involves using (38) to compute V^* with the following backward induction algorithm. An optimal policy $\pi^* \in \Pi^{MD}$ is then obtained using (40), or an ε-optimal policy $\pi^*_\varepsilon \in \Pi^{MD}$ is obtained using (41).

Finite horizon backward induction algorithm.

0. Set $V^*(s, T + 1) = 0$ for all $s \in \mathbb{S}$.

1. For $t = T, \ldots, 1$, repeat steps 2 and 3.

2. For each $s \in \mathbb{S}$, compute

$$V^*(s, t) = \sup_{a\in\mathbb{A}(s,t)} \left\{ r(s, a, t) + \sum_{s'\in\mathbb{S}} p[s'|s, a, t]V^*(s', t + 1) \right\} \tag{39}$$

3. For each $s \in \mathbb{S}$, choose a decision

$$\pi^*(s, t) \in \arg \max_{a \in \mathcal{A}(s,t)} \left\{ r(s, a, t) + \sum_{s' \in \mathcal{S}} p[s'|s, a, t] V^*(s', t + 1) \right\} \tag{40}$$

if the maximum on the right hand side is attained. Otherwise, for any chosen $\varepsilon > 0$, choose a decision $\pi_\varepsilon^*(s, t)$ such that

$$\begin{aligned} & r(s, \pi_\varepsilon^*(s, t), t) + \sum_{s' \in \mathcal{S}} p[s'|s, \pi_\varepsilon^*(s, t), t] V^*(s', t + 1) + \frac{\varepsilon}{T + 1} \\ & > \sup_{a \in \mathcal{A}(s,t)} \left\{ r(s, a, t) + \sum_{s' \in \mathcal{S}} p[s'|s, a, t] V^*(s', t + 1) \right\} \end{aligned} \tag{41}$$

The value function V^π of a policy π can be calculated with a similar algorithm, except that (37) is used instead of (39), that is, the maximization on the right-hand side of (39) is replaced by the decision under policy π, and step 3 is omitted.

4.2.3. Structural Properties

Dynamic programming is useful not only for the computation of optimal policies and optimal expected values, but also for determining insightful structural characteristics of optimal policies. In fact, for many interesting applications the state space is too big to compute optimal policies and optimal expected values exactly, but dynamic programming can still be used to establish qualitative characteristics of optimal quantities. Some such structural properties are illustrated with examples.

Example 5 (inventory replenishment). A business purchases and sells a particular product. A decision maker has to decide regularly, say once every day, how much of the product to buy. The business does not have to wait to receive the purchased product. In contrast to the newsvendor problem, here product that is not sold on a particular day can be kept in inventory for the future. The business pays a fixed cost K plus a variable cost c per unit of product each time product is purchased. Thus, if a units of product are purchased, then the purchasing cost is $K + ca$ if $a > 0$, and it is 0 if $a = 0$. In addition, if the inventory level at the beginning of the day is s, and a units of product is purchased, then an inventory cost of $h(s + a)$ is incurred, where h is a convex function. The demand for the product on different days are independent and identically distributed. If the demand D is greater than the available inventory $s + a$, then the excess demand is backlogged until additional inventory is obtained, at which time the backlogged demand is filled immediately. Inventory remaining at the end of the time horizon has no value. The objective is to minimize the expected total cost over the time horizon. This problem can be formulated as a discrete-time dynamic program. The state $S(t)$ is the inventory at the beginning of day t. The decision $A(t)$ is the quantity purchased on day t, and the single-stage cost $r(s, a) = (K + ca)\, I_{\{a>0\}} + h(s + a)$. The transitions are given by $S(t + 1) = S(t) + A(t) - D(t)$. Dynamic programming can be used to show that the following policy is optimal. If the inventory level $S(t) < \sigma^*(t)$, where $\sigma^*(t)$ is called the optimal reorder point at time t, then it is optimal to purchase $\Sigma^*(t) - S(t)$ units of product at time t, where $\Sigma^*(t)$ is called the optimal order-up-to point at time t. If the inventory level $S(t) \geq \sigma^*(t)$, then it is optimal not to purchase any product. Such a policy is often called an (s, S)-policy, or a (σ, Σ)-policy. Similar results hold in the infinite horizon case, except that σ^* and Σ^* do not depend on time t anymore.

Example 6 (resource allocation). A decision maker has an amount of resource that can be allocated over some time horizon. At each discrete point in time, a request for some amount of resource is received. If the request is for more resource than the decision maker has available, then the request has to be rejected. Otherwise, the request can be accepted or rejected. A request must be accepted or rejected as a whole—the decision maker cannot allocate a fraction of the amount of resource requested. Rejected requests cannot be recalled later. If the request is accepted, the amount of resource available to the decision maker is reduced by the amount of resource requested and the decision maker receives an associated reward in return. The amounts of resource and the rewards of future requests are unknown to the decision maker, but the decision maker knows the probability distribution of these. At the end of the time horizon, the decision maker receives a salvage reward for the remaining amount of resource. The objective is to maximize the expected total reward over the time horizon. Problems of this type are encountered in revenue management and the selling of assets such as real estate and vehicles. This resource-allocation problem can be formulated as a dynamic program. The state $S(t)$ is the amount of resource available to the decision maker at the beginning of time period t. The decision $A(t)$ is the rule that will be used for accepting or rejecting requests during time period t. If a request for amount Q of resource with an associated reward R is accepted in time period t, then the single-stage reward is R and the next state is $S(t + 1) = S(t) - Q$. If the request is rejected, then the next state is $S(t+1) = S(t)$. It is easy to see that the optimal value function $V^*(s, t)$ is increasing in s and decreasing in t. The following threshold policy, with reward threshold function $x^*(q, s, t) = V^*(s, t+1) - V^*(s-q, t+1)$, is optimal. Accept a request for amount Q of resource with an associated reward R if $Q \leq S(t)$ and $R \geq x^*(Q, S(t), t)$, and reject the request otherwise. If each request is for the same amount of resource (say 1 unit of resource), and the salvage reward is concave in the remaining amount of resource, then the optimal value function $V^*(s, t)$ is concave in s and t, and the optimal reward threshold $x^*(1, s, t) = V^*(s, t+1) - V^*(s-1, t+1)$ is decreasing in s and t. These intuitive properties do not hold in general if the requests are for different amounts of resource.

Structural properties of the optimal value functions and optimal policies of dynamic programs have been investigated for many different applications. Some general structural results are given in Serfozo (1976), Topkis (1978), and Heyman and Sobel (1984).

4.3. Infinite Horizon Dynamic Programs

In this section we present dynamic programming models with an infinite time horizon. Although an infinite time horizon is a figment of the imagination, these models often are useful for decision problems with many decision points. Many infinite horizon models also have the desirable feature that there exist stationary deterministic optimal policies. Thus, optimal decisions depend only on the current state of the process and not on the sometimes artificial notion of time, as in finite horizon problems. This characteristic makes optimal policies easier to understand, compute, and implement, which is desirable in applications.

We again assume that $\mathbb{S}$ is countable and r is bounded. Similar results hold in more general cases, subject to regularity conditions. We also assume that the sets $\mathcal{A}(s)$ of feasible decisions depend only on the states s, the transition probabilities $p[s'|s, a]$ depend only on the states s, s', and decisions a, and the rewards $r(s, a)$ depend only on the states s and decisions a, and not on time, as in the finite horizon case.

In this article we focus on dynamic programs with total discounted reward objectives. As illustrated in the example of Section 4.1.7, infinite horizon dynamic programs with other types of objectives, such as long-run average reward objectives, may exhibit undesirable behavior. A proper treatment of dynamic programs with these types of objectives requires more space than we have available here, and therefore we refer the interested reader to the references. Besides, in most practical applications, rewards and costs in the near future are valued more than rewards and costs in the more distant future, and hence total discounted reward objectives are preferred for applications.

4.4. Infinite Horizon Discounted Dynamic Programs

In this section we investigate dynamic programming models for optimization problems with the form

$$\max_{(A(0), A(1), \ldots)} \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r(S(t), A(t))\right] \tag{42}$$

where $\alpha \in (0, 1)$ is a known discount factor. Again, decisions $A(t)$, $t = 0, 1, \ldots$ have to be feasible and may depend only on the information available to the decision maker at each time t, that is, the history H(t) of the process up to time t, and possibly some randomization.

4.4.1. Optimality Results

From the stationary properties of the feasible sets, transition probabilities, and rewards, one would expect that it should be sufficient to consider stationary deterministic policies. This can be shown to be true for infinite horizon discounted problems of the form (42).

The value function V^π of a stationary policy π is defined by

$$V^\pi(s) \equiv \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \alpha^t r(S(t), A(t)) \,\middle|\, S(0) = s\right] \tag{43}$$

Then, because it is sufficient to consider stationary deterministic policies, the optimal value function V^* is given by

$$V^*(s) \equiv \sup_{\pi \in \Pi^{SD}} V^\pi(s) \tag{44}$$

Again motivated by the stationary input data, it is intuitive, and can be shown to be true, that the value function V^π of a stationary policy π satisfies an equation similar to (37) for the finite horizon case, that is,

$$V^{\pi}(s) = r(s, \pi(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, \pi(s)]V^{\pi}(s') \tag{45}$$

Similarly, the optimal value function V^* satisfies the following optimality equation:

$$V^*(s) = \sup_{a \in \mathcal{A}(s)} \left\{ r(s, a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, a]V^*(s') \right\} \tag{46}$$

4.4.2. Infinite Horizon Algorithms

Solving an infinite horizon discounted dynamic program usually involves computing V^*. An optimal policy $\pi^* \in \Pi^{SD}$ or an ε-optimal policy $\pi^*_{\varepsilon} \in \Pi^{SD}$ can then be obtained, as shown in this section.

Unlike the finite horizon case, V^* is not computed directly using backward induction. An approach that is often used is to compute a sequence of approximating functions V_i, $i = 0, 1, 2, \ldots$, such that $V_i \to V^*$ as $i \to \infty$.

Approximating value functions provide good policies, as motivated by the following result. Suppose V^* is approximated by $\hat{V}$ such that $\|V^* - \hat{V}\|_{\infty} \le \varepsilon$. Consider any policy $\hat{\pi} \in \Pi^{SD}$ such that

$$r(s, \hat{\pi}(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, \hat{\pi}(s)]\hat{V}(s') + \delta \ge \sup_{a \in \mathcal{A}(s)} \left\{ r(s, a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, a]\hat{V}(s') \right\}$$

for all $s \in \mathcal{S}$, that is, decision $\hat{\pi}(s)$ is within δ of the optimal decision using approximating function $\hat{V}$ on the right-hand side of the optimality equation (46). Then

$$V^{\hat{\pi}}(s) \ge V^*(s) - \frac{2\alpha\varepsilon + \delta}{1 - \alpha} \tag{47}$$

for all $s \in \mathcal{S}$, that is, policy $\hat{\pi}$ has value function within $(2\alpha\varepsilon + \delta)/(1 - \alpha)$ of the optimal value function.

4.4.2.1. Value Iteration One algorithm based on a sequence of approximating functions V_i is called value iteration, or successive approximation. The iterates V_i of value iteration correspond to the value function $V^*(s, T + 1 - i)$ of the finite horizon dynamic program with the same problem parameters. Specifically, starting with initial approximation $V_0(s) = 0 = V^*(s, T + 1)$ for all s, the ith approximating function $V_i(s)$ is the same as the value function $V^*(s, T + 1 - i)$ of the corresponding finite horizon dynamic program, that is, the value function for time $T + 1 - i$ that is obtained after i steps of the backward induction algorithm.

Value iteration algorithm

0. Choose initial approximation V_0 and stopping tolerance ε. Set $i \leftarrow 0$.

1. For each $s \in \mathcal{S}$, compute

$$V_{i+1}(s) = \sup_{a \in \mathcal{A}(s)} \left\{ r(s, a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, a]V_i(s') \right\} \tag{48}$$

2. If $\|V_{i+1} - V_i\|_{\infty} < (1 - \alpha)\varepsilon/2\alpha$, then go to step 3. Otherwise, set $i \leftarrow i + 1$ and go to step 1.

3. For each $s \in \mathcal{S}$, choose a decision

$$\pi^*_{\varepsilon}(s) \in \arg\max_{a \in \mathcal{A}(s)} \left\{ r(s, a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, a]V_{i+1}(s') \right\}$$

if the maximum on the right-hand side is attained. Otherwise, for any chosen $\delta > 0$, choose a decision $\pi^*_{\delta}(s)$ such that

$$r(s, \pi^*_{\delta}(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, \pi^*_{\delta}(s)]V_{i+1}(s') + (1 - \alpha)\delta > \sup_{a \in \mathcal{A}(s)} \left\{ r(s, a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, a]V_{i+1}(s') \right\}$$

It can be shown, using the contraction property provided by the discount factor α, that $V_i \to V^*$ as $i \to \infty$ for any initial approximation V_0. Also, the convergence is geometric with rate α. Specifically, for any V_0, $\|V_i - V^*\|_\infty \leq \alpha^i \|V_0 - V^*\|_\infty$. That implies that the convergence rate is faster if the discount factor α is smaller.

When the value iteration algorithm stops, the final approximation V_{i+1} satisfies $\|V_{i+1} - V^*\|_\infty < \varepsilon/2$. Furthermore, the chosen policy π^*_ε is an ε-optimal policy, and the chosen policy π^*_δ is an $(\varepsilon + \delta)$-optimal policy.

There are several versions of the value iteration algorithm. One example is Gauss–Seidel value iteration, which uses the most up-to-date approximation $V_{i+1}(s)$ on the right-hand side of (48) as soon as it becomes available, instead of using the previous approximation $V_i(s')$ as shown in (48). Gauss–Seidel value iteration has the same convergence properties and performance guarantees given above, but in practice it usually converges faster.

There are several other algorithms for solving infinite horizon discounted dynamic programs. One of these is policy iteration, which computes a sequence of policies π_i, and their value functions V^{π_i}. Another algorithm is modified policy iteration, which is a generalization of both value iteration and policy iteration. With correct choice of algorithm parameters, modified policy iteration often performs much better than value iteration and policy iteration. There are also several variations on these algorithms, obtained with different choices of algorithm control methods, such as adaptive control, as well as parallel versions. Most books on dynamic programming in the References discuss one or more of these algorithms.

4.5. Approximation Methods

For many interesting applications the state space $\mathcal{S}$ is too big for any of the algorithms discussed so far to be used. This is usually due to the "curse of dimensionality"—the phenomenon that the number of states grows exponentially in the number of dimensions of the state space. When the state space is too large, not only is the computational effort required by these algorithms excessive, but storing the value function and policy values for each state is impossible with current technology.

Recall that solving a dynamic program usually involves using (38) in the finite horizon case or (46) in the infinite horizon case to compute the optimal value function V^*, and an optimal policy π^*. To accomplish this, the following major computational tasks are performed:

1. Estimation of the optimal value function V^* on the right-hand side of (38) or (46).
2. Estimation of the expected value on the right-hand side of (38) or (46). For many applications, this is a high-dimensional integral that requires a lot of computational effort to compute accurately.
3. The maximization problem on the right hand side of (38) or (46) has to be solved to determine the optimal decision for each state. This maximization problem may be easy or hard, depending on the application. The first part of this article discusses several methods for solving such stochastic optimization problems.

Approximation methods usually involve approaches to perform one or more of these computational tasks efficiently, sometimes by sacrificing optimality.

For many applications, the state space is uncountable and the transition and cost functions are too complex for closed form solutions to be obtained. To compute solutions for such problems, the state space is often discretized. Discretization methods and convergence results are discussed in Wong (1970a), Fox (1973), Bertsekas (1975), Chow and Tsitsiklis (1991), and Kushner and Dupuis (1992).

For many other applications, such as queueing systems, the state space is countably infinite. Computing solutions for such problems usually involves solving smaller dynamic programs with finite state spaces, often obtained by truncating the state space of the original DP, and then using the solutions of the smaller DPs to obtain good solutions for the original DP. Such approaches and their convergence are discussed in Fox (1971), White (1980a, b, 1982), White (1982), Cavazos-Cadena (1986), Van Dijk (1991), and Sennott (1997).

Even if the state space is not infinite, the number of states may be very large. A natural approach is to aggregate states, usually by collecting similar states into subsets, and then to solve a related DP with the aggregated state space. Aggregation and aggregation/disaggregation methods are discussed in Mendelssohn (1982), Chatelin (1984), Schweitzer et al. (1985), Bean et al. (1987), and Bertsekas and Castanon (1989).

Another natural approach for dealing with a large-scale DP is to decompose the DP into smaller related DPs, which are easier to solve, and then to use the solutions of the smaller DPs to obtain a good solution for the original DP. Decomposition methods are presented in Wong (1970b), Collins and Lew (1970), Courtois (1977), and Kleywegt et al. (1999).

Some general state space-reduction methods that include many of the methods mentioned above are analyzed in Whitt (1978, 1979a, b), Hinderer (1976, 1978), Hinderer and Hübner (1977), and Haurie and L'Ecuyer (1986). Surveys are given in Morin (1978) and Rogers et al. (1991).

Another natural and quite different approach for dealing with DPs with large state spaces is to approximate the optimal value function V^* with an approximating function $\hat{V}$. It was shown in Section

4.4.2 that good approximations $\hat{V}$ to the optimal value function V^* lead to good policies $\hat{\pi}$. Polynomial approximations, often using orthogonal polynomials such as Legendre and Chebychev polynomials, have been suggested by Bellman and Dreyfus (1959), Chang (1966), and Schweitzer and Seidman (1985). Approximations using splines have been suggested by Daniel (1976), and approximations using regression splines by Chen et al. (1999). Estimation of the parameters of approximating functions for infinite horizon discounted DPs have been studied in Tsitsiklis and Van Roy (1996), Van Roy and Tsitsiklis (1996), and Bertsekas and Tsitsiklis (1996). Some of this work was motivated by methods proposed for reinforcement learning; see Sutton and Barto (1998) for an overview.

REFERENCES

Albritton, M., Shapiro, A., and Spearman, M. L. (1999), "Finite Capacity Production Planning with Random Demand and Limited Information," Preprint.

Beale, E. M. L. (1955), "On Minimizing a Convex Function Subject to Linear Inequalities," *Journal of the Royal Statistical Society, Series B,* Vol. 17, pp. 173—184.

Bean, J. C., Birge, J. R., and Smith, R. L. (1987), "Aggregation in Dynamic Programming," *Operations Research,* Vol. 35, pp. 215–220.

Bellman, R., and Dreyfus, S. (1959), "Functional Approximations and Dynamic Programming," *Mathematical Tables and Other Aids to Computation,* Vol. 13, pp. 247–251.

Bellman, R. E. (1957), *Dynamic Programming,* Princeton University Press, Princeton, NJ.

Bellman, R. E. (1961), *Adaptive Control Processes: A Guided Tour,* Princeton University Press, Princeton, NJ.

Bellman, R. E., and Dreyfus, S. (1962), *Applied Dynamic Programming,* Princeton University Press, Princeton, NJ.

Benveniste, A., Métivier, M., and Priouret, P. (1990), *Adaptive Algorithms and Stochastic Approximations,* Springer, Berlin.

Bertsekas, D. P. (1975), "Convergence of Discretization Procedures in Dynamic Programming," *IEEE Transactions on Automatic Control,* Vol. AC-20, pp. 415–419.

Bertsekas, D. P. (1995), *Dynamic Programming and Optimal Control,* Athena Scientific, Belmont, MA.

Bertsekas, D. P., and Castanon, D. A. (1989), "Adaptive Aggregation Methods for Infinite Horizon Dynamic Programming," *IEEE Transactions on Automatic Control,* Vol. AC-34, pp. 589–598.

Bertsekas, D. P., and Shreve, S. E. (1978), *Stochastic Optimal Control: The Discrete Time Case,* Academic Press, New York.

Bertsekas, D. P., and Tsitsiklis, J. N. (1996), *Neuro-Dynamic Programming,* Athena Scientific, Belmont, MA.

Birge, J. R., and Louveaux, F. (1997), *Introduction to Stochastic Programming,* Springer Series in Operations Research, Springer, New York.

Cavazos-Cadena, R. (1986), "Finite-State Approximations for Denumerable State Discounted Markov Deci sion Processes," *Applied Mathematics and Optimization,* Vol. 14, pp. 1–26.

Chang, C. S. (1966), "Discrete-Sample Curve Fitting Using Chebyshev Polynomials and the Approximate Determination of Optimal Trajectories via Dynamic Programming," *IEEE Transactions on Automatic Control,* Vol. AC-11, pp. 116–118.

Chatelin, F. (1984), "Iterative Aggregation/Disaggregation Methods," in *Mathematical Computer Performance and Reliability,* G. Iazeolla, P. J. Courtois, and A. Hordijk, Eds., Elsevier, Science Publishers Amsterdam, pp. 199–207.

Chen, V. C. P., Ruppert, D., and Shoemaker, C. A. (1999), "Applying Experimental Design and Regression Splines to High-Dimensional Continuous-State Stochastic Dynamic Programming," *Operations Research,* Vol. 47, pp. 38–53.

Chong, E. K. P., and Ramadge, P. J. (1992), "Convergence of Recursive Optimization Algorithms Using Infinitesimal Perturbation Analysis Estimates," *Discrete Event Dynamic Systems: Theory and Applications,* Vol. 1, pp. 339–372.

Chow, C. S., and Tsitsiklis, J. N. (1991), "An Optimal One-Way Multigrid Algorithm for Discrete-Time Stochastic Control," *IEEE Transactions on Automatic Control,* Vol. AC-36, pp. 898–914.

Collins, D. C., and Lew, A. (1970), "A Dimensional Approximation in Dynamic Programming by Structural Decomposition," *Journal of Mathematical Analysis and Applications,* Vol. 30, pp. 375–384.

Courtois, P. J. (1977), *Decomposability: Queueing and Computer System Applications,* Academic Press, New York.

Daniel, J. W. (1976), "Splines and Efficiency in Dynamic Programming," *Journal of Mathematical Analysis and Applications,* Vol. 54, pp. 402–407.

Dantzig, G. B. (1955), "Linear Programming under Uncertainty," *Management Science,* Vol. 1, pp. 197–206.

Denardo, E. V. (1982), *Dynamic Programming Models and Applications,* Prentice-Hall, Englewood Cliffs, NJ.

Fox, B. L. (1971), "Finite-State Approximations to Denumerable-State Dynamic Programs," *Journal of Mathematical Analysis and Applications,* Vol. 34, pp. 665–670.

Fox, B. L. (1973), "Discretizing Dynamic Programs," *Journal of Optimization Theory and Applications,* Vol. 11, pp. 228–234.

Glasserman, P. (1991), *Gradient Estimation via Perturbation Analysis,* Kluwer, Norwell, MA.

Glynn, P. W. (1990), "Likelihood Ratio Gradient Estimation for Stochastic Systems," *Communications of the ACM,* Vol. 33, pp. 75–84.

Haurie, A., and L'Ecuyer, P. (1986), "Approximation and Bounds in Discrete Event Dynamic Programming," *IEEE Transactions on Automatic Control,* Vol. AC-31, pp. 227–235.

Heyman, D. P., and Sobel, M. J. (1984), *Stochastic Models in Operations Research,* Vol. 2, McGraw-Hill, New York.

Hinderer, K. (1970), *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter.* Springer, Berlin.

Hinderer, K. (1976), "Estimates for Finite-Stage Dynamic Programs," *Journal of Mathematical Analysis and Applications,* Vol. 55, pp. 207–238.

Hinderer, K. (1978), "On Approximate Solutions of Finite-Stage Dynamic Programs," in *Dynamic Programmming and Its Applications,* M. L. Puterman, Ed., Academic Press, New York, pp. 289–317.

Hinderer, K., and Hübner, G. (1977), "On Exact and Approximate Solutions of Unstructured Finite-Stage Dynamic Programs," in *Markov Decision Theory: Proceedings of the Advanced Seminar on Markov Decision Theory* (Amsterdam, September 13–17, 1976), H. C. Tijms and J. Wessels, Eds., Mathematisch Centrum, Amsterdam. pp. 57–76.

Hiriart-Urruty, J. B., and Lemarechal, C. (1993), *Convex Analysis and Minimization Algorithms,* Springer, Berlin.

Ho, Y. C., and Cao, X. R. (1991), *Perturbation Analysis of Discrete Event Dynamic Systems,* Kluwer, Norwell, MA.

Kall, P., and Wallace, S. W. (1994), *Stochastic Programming,* John Wiley & Sons, Chichester.

Klein Haneveld, W. K., and Van der Vlerk, M. H. (1999), "Stochastic Integer Programming: General Models and Algorithms," *Annals of Operations Research,* Vol. 85, pp. 39–57.

Kleywegt, A. J., and Shapiro, A. (1999), "The Sample Average Approximation Method for Sto chastic Discrete Optimization," Preprint, available at Stochastic Programming E-Print Series, http://dochost.rz.hu-berlin.de/speps/.

Kleywegt, A. J., Nori, V. S., and Savelsbergh, M. W. P. (1999), "The Stochastic Inventory Routing Problem with Direct Deliveries," Technical Report TLI99-01, The Logistics Institute, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

Kushner, H. J., and Clark, D. S. (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems,* Springer, Berlin.

Kushner, H. J., and Dupuis, P. (1992), *Numerical Methods for Stochastic Control Problems in Continuous Time,* Springer, New York.

L'Ecuyer, P., and Glynn, P. W. (1994), "Stochastic Optimization by Simulation: Convergence Proofs for the GI/G/1 Queue in Steady-State," *Management Science,* Vol. 11, pp. 1562–1578.

Mendelssohn, R. (1982), "An Iterative Aggregation Procedure for Markov Decision Processes," *Operations Research,* Vol. 30, pp. 62–73.

Morin, T. (1978), "Computational Advances in Dynamic Programming," in *Dynamic Programmming and its Applications,* M. L. Puterman, Ed., Academic Press, New York. pp. 53–90.

Nemhauser, G. L. (1966), *Introduction to Dynamic Programming,* John Wiley & Sons, New York.

Norkin, V. I., Pflug, G. C., and Ruszczyński, A. (1998), "A Branch and Bound Method for Stochastic Global Optimization," *Mathematical Programming,* Vol. 83, pp. 425–450.

Puterman, M. L. (1994), *Markov Decision Processes,* John Wiley & Sons, New York.

Robbins, H., and Monro, S. (1951), "On a Stochastic Approximation Method," *Annals of Mathematical Statistics,* Vol. 22, pp. 400–407.

Robinson, S. M. (1996), “Analysis of Sample-Path Optimization,” *Mathematics of Operations Research,* Vol. 21, pp. 513–528.

Rogers, D. F., Plante, R. D., Wong, R. T., and Evans, J. R. (1991), “Aggregation and Disaggregation Techniques and Methodology in Optimization,” *Operations Research,* Vol. 39, pp. 553–582.

Ross, S. M. (1970), *Applied Probability Models with Optimization Applications,* Dover, New York.

Ross, S. M. (1983), *Introduction to Stochastic Dynamic Programming,* Academic Press, New York.

Rubinstein, R. Y. and Shapiro, A. (1990), “Optimization of Static Simulation Models by the Score Function Method,” *Mathematics and Computers in Simulation,* Vol. 32, pp. 373–392.

Rubinstein, R. Y., and Shapiro, A. (1993), *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method,* John Wiley & Sons, Chichester.

Ruppert, D. (1991), “Stochastic Approximation,” in *Handbook of Sequential Analysis,* B. K. Ghosh and P. K. Sen, Eds., Marcel Dekker, New York, pp. 503–529.

Schultz, R., Stougie, L., and Van der Vlerk, M. H. (1998), “Solving Stochastic Programs with Integer Recourse by Enumeration: A Framework Using Gröbner Basis Reductions,” *Mathematical Programming,* Vol. 83, pp. 229–252.

Schweitzer, P. J., and Seidman, A. (1985), “Generalized Polynomial Approximations in Markovian Decision Processes,” *Journal of Mathematical Analysis and Applications,* Vol. 110, pp. 568–582.

Schweitzer, P. J., Puterman, M. L., and Kindle, K. W. (1985), “Iterative Aggregation-Disaggregation Procedures for Discounted Semi-Markov Reward Processes,” *Operations Research,* Vol. 33, pp. 589–605.

Sennott, L. I. (1997), “The Computation of Average Optimal Policies in Denumerable State Markov Decision Chains,” *Advances in Applied Probability,* Vol. 29, pp. 114–137.

Sennott, L. I. (1999), *Stochastic Dynamic Programming and the Control of Queueing Systems,* John Wiley & Sons, New York.

Serfozo, R. F. (1976), Monotone Optimal Policies for Markov Decision Processes. *Mathematical Programming Study,* Vol. 6, pp. 202–215.

Shapiro, A. (1996), “Simulation-Based Optimization: Convergence Analysis and Statistical Inference,” *Stochastic Models,* Vol. 12, pp. 425–454.

Shapiro, A., and Homem-de-Mello, T. (1998), “A Simulation-Based Approach to Two-Stage Stochastic Programming with Recourse,” *Mathematical Programming,* Vol. 81, pp. 301–325.

Shapiro, A., and Homem-de-Mello, T. (1999), “On the Rate of Convergence of Optimal Solutions of Monte Carlo Approximations of Stochastic Programs,” *SIAM Journal on Optimization,* Vol. 11, No. 1, pp. 70–86, 2000.

Sutton, R. S., and Barto, A. G. (1998), *Reinforcement Learning: An Introduction,* MIT Press, Cambridge, MA.

Topkis, D. M. (1978), “Minimizing a Submodular Function on a Lattice,” *Operations Research,* Vol. 26, pp. 305–321.

Tsitsiklis, J. N., and Van Roy, B. (1996), “Feature-Based Methods for Large-Scale Dynamic Programming,” *Machine Learning,* Vol. 22, pp. 59–94.

Van Dijk, N. (1991), “On Truncations and Perturbations of Markov Decision Problems with an Application to Queueing Network Overflow Control,” *Annals of Operations Research,* Vol. 29, pp. 515–536.

Van Roy, B., and Tsitsiklis, J. N. (1996), “Stable Linear Approximations to Dynamic Programming for Stochastic Control Problems with Local Transitions,” *Advances in Neural Information Processing Systems 8,* MIT Press, Cambridge, MA, pp. 1045–1051.

Van Slyke, R., and Wets, R. J. B. (1969), “L-Shaped Linear Programs with Application to Optimal Control and Stochastic Programming,” *SIAM Journal of Applied Mathematics,* Vol. 17, pp. 638–663.

White, D. J. (1980a), “Finite-State Approximations for Denumerable-State Infinite-Horizon Discounted Markov Decision Processes: The Method of Successive Approximations,” in *Recent Developments in Markov Decision Processes,* R. Hatley, L. C. Thomas, and D. J. White, Eds., Academic Press, New York, pp. 57–72.

White, D. J. (1980b), “Finite-State Approximations for Denumerable-State Infinite-Horizon Discounted Markov Decision Processes,” *Journal of Mathematical Analysis and Applications,* Vol. 74, pp. 292–295.

White, D. J. (1982), “Finite-State Approximations for Denumerable-State Infinite Horizon Discounted Markov Decision Processes with Unbounded Rewards,” *Journal of Mathematical Analysis and Applications,* Vol. 86, pp. 292–306.

Whitt, W. (1978), ‘‘Approximations of Dynamic Programs, I,’’ *Mathematics of Operations Research,* Vol. 3, pp. 231–243.

Whitt, W. (1979a), ‘‘A-Priori Bounds for Approximations of Markov Programs,’’ *Journal of Mathematical Analysis and Applications,* Vol. 71, pp. 297–302.

Whitt, W. (1979b), ‘‘Approximations of Dynamic Programs, II,’’ *Mathematics of Operations Research,* Vol. 4, pp. 179–185.

Wong, P. J. (1970a), ‘‘An Approach to Reducing the Computing Time for Dynamic Programming,’’ *Operations Research,* Vol. 18, pp. 181–185.

Wong, P. J. (1970b), ‘‘A New Decomposition Procedure for Dynamic Programming,’’ *Operations Research,* Vol. 18, pp. 119–131.

AUTHOR INDEX

Aaras, A., *1108*
Aarts, E., 2590, 2591, *2600*
Abadie, J., 2560, *2565*
Abernathy, R. B., 1945, *1954*
Ackerman, K. B., *1547*, 2079, *2081*
Acton, F. S., 2274, *2292*
Adachi, T., 1777, *1787*
Adamopoulos, J., *972*
Adams, J., 602, *617*, 1729, *1739*
Addison, J. L., 1897, 1898, *1917*
Adelman, L., 145, *149*
Adelsberger, H. H., 698, *706*, 1735, 1736, *1739*
Adelstein, B. D., *2520*
Adiputra, N., *993*
Adkins, C. L., *866*
Adkins, M., *2216*
Adkins, R., 2424, *2441*, *2442*
Adler, L., 1733, *1739*
Adler, P. S., *560*
Adolf, W. W., 583, *587*
Agarwal, M. K., *707*
Agarwal, V., 759, *770*
Agee, M. H., *2351*, *2405*
Agha, G., *2444*
Agnew, P. W., 229, *257*
Agrawal, S. C., 503, *528*
Aguiar, M. W. C., 520, 522, 523, *528*
Ahire, S. L., *1806*
Ahlberg-Hulten, G., *1236*
Aho, A. V., 816, *822*
Ahuja, R. K., 808–811, *822*, 2574, *2580*
Ainslie, G., 2204, *2215*
Ainsworth, L., 1025, 1028, *1038*
Ainsworth, L. K., 1209, 1211, *1233*
Aiyer, A., *1918*
Akao, Y., 13, *24*
Akella, R., 1890, *1919*
Akiba, M., 552, *559*
Akkermans, H., *226*
Akturk, M. S., 545, *559*
Alabiso, B., 173, *175*
Albanese, R., 882, *894*
Albert, S., 147, *149*
Alberts, D. S., *149*
Albritton, M., 2631, *2646*
Alderson, W., 2113, *2138*, *2139*
Aldrich, T., *2442*
Aldrich, T. B., 2410, 2423, *2443*
Aldridge, A., *1129*
Alemi, F., *990*
Alessi, S. M., 932, *941*
Alexander, C., 761, *770*
Alexander, D. C., 1097, *1100*, 1131, *1152*
Alexander, F., 539, *540*
Algeo, James M., Sr., 352
Algeo, M. E. A., *352*
Alhaery, M., 1056, *1103*
Ali, I., 2573, *2580*
Allais, M., 2202, *2215*
Allee, V., 147, *149*
Allemang, R., 1404, *1407*
Allen, D., *483*
Allen, D. B., 2621, *2623*
Allen, D. K., 462, *482*
Allen, D. M., 2284, *2292*, *2293*
Allen, D. S., *352*
Allen, D. T., 531, 533, 534, *540*
Allen, J. A., *946*
Allen, J. F., 1287, *1294*
Allen, J. S., *864*
Allen, R. B., 1210, *1230*
Allen, W. H., 928, *941*
Allenby, B. R., 530, 531, 533, 537, *540*, *541*
Allender, L., 2424, 2426, *2441*, *2442*, *2444*
Allport, D. A., 1016, *1037*
Allread, W. G., *1105*
Alluisi, E. A., 929, 930, *941*, *943*
Allweyer, T., 218, *223*
Almanza, B. A., 826, 827, 833, 834, *836*
Almaraz, J., 1794, 1795, *1805*
Almeida, V. A. F., *736*
Al-Shedi, A., *1108*
Altan, T., *587*
Alting, L., 537, *540*
Altiok, K., 1668, *1668*
Amason, A. C., 982, *989*
AMA Survey on Quality and Customer Satisfaction Programs, *1806*
Ambegaonkar, P., 228, *257*
Ambler, Bruce, 352
American Educational Research Association (AERA), 921, *941*
American National Standards Institute (ANSI), 1164, 1165, 1176, 1177, *1188*, *1190*, 1195, 1197, 1198, 1200, 1201, 1203, 1204, *1230*
American Productivity Center (APC), 1562, *1582*
American Society for Quality (ASQ), *1974*
American Society for Quality Control (ASQC), *1855*
American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), 1133, 1139, 1147, *1152*

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

Amick, B. C., 1222, *1235*
Amidon, D. M., 213, *226*
Amirhosseini, M. M., 2093, 2095, *2108, 2109*
Ammons, J. C., *542, 543*
Amussen, S., 2163, *2171*
Anderl, R., 191–193, 204, *223, 224*
Andersen, S. O., 532, *540*
Anderson, C. J., *2223*
Anderson, D., 785, *785*
Anderson, D. J., *540*
Anderson, E. W., 623, 624, 629, *631*
Anderson, J. C., 617, *618*
Anderson, J. R., 929, *941*
Anderson, N. H., 2200, *2215*
Anderson, V. L., 2225, 2229, *2239*
Anderson Consulting, 2125, *2138*
Andersson, G. B. J., 1046, 1071, 1072, *1100, 1101, 1107, 1128*
Andradóttir, S., 2492, *2494*
Andrei, Neculai, 2565
Andren, E., *353*
Andrews, D., 1705, 1707, *1715*
Andrews, D. C., 2126, *2138*
Andrews, D. H., 929, *941*
Andrews, I. R., 2195, *2222*
Andriole, S., 145, *149*
Anesgart, M., *2442*
Angell, D., 229, *257*
Anger, W. K., *1190*
Anjewierden, A., *226*
Ankele, A., *399*
ANSI/ISA 2000, 1769, *1787*
ANSI Z-365 Draft, 1093–1095, *1100*
Anthony, R. N., 111, *149*
Anthony, W. P., 856, 860, *864*
Anton, J., 651, 657, *663*
Antonioni, D., *867*
Aoki, M., 263, *278*
Apple Computer, Inc., 536, *540*
Applegate, L. M., 125, 130, 144, *149*
Apte, U., 165, *175*
Aram, J. D., *989*
Arcangeli, K. K., *1109*
Archer, R., *2444*
Archer, R. D., *2443*
Archer, S., *2442*
Archer, S. G., 2424, *2441, 2443*
Ardekani, S., *541*
Arden, B. W., 116, *149*
Arditi, A., 2434, *2442*
Argote, L., *895*, 2212, *2215*
Argyris, C., 874, *894*, 939, *941*
Arifin, R., 1525, *1525*
Arizono, I., 1778, *1787*
Arkes, H. R., 2197, 2204, *2215*
Armada, M., *398*
Armijo, L., 2549, *2565*
Armistead, C., 1697, 1710, *1716*
Armstrong, A. G., 147, *151*
Armstrong, G., 327, *352*
Armstrong, J. E., 111, 126, 128, *153*
Armstrong, R. D., 2093, *2108*
Armstrong, T., 1061, *1100, 1235*
Armstrong, T. J., 1083, 1084, 1086, 1087, *1100, 1107, 1109*, 1201, 1202, *1230, 1232*, 1365, *1389*
Arnold, S. F., *2495*
Arnott, J. L., *1234*
Arntzen, B. C., 2112, 2127, *2138*
Aronsson, G., *896*, 1220–1222, *1230, 1233*
Arthur, A., *588*
Arthur, J., 951, *969*
Arthur, M. B., *866*
Arthur, W. B., 602, *617*
Arvey, R., 914, 916, *917*
Asakura, T., 1222, 1223, *1230*
Asfour, S. S., 1071, *1100, 1104, 1108*
Ash, Mary Kay, 847, 848
Ashayberi, J., 518, *528*
Ashe, J. T., *1583*
Ashikaga, T., *1102*
Ashworth, S., *944*
Asian, S., *749*
Aspinwall, E. M., 1697, *1716*
Aspinwall, L., 2115, *2138*
Assad, A., 2061, *2068*
Assad, A. A., 2575, *2581*
Associates in Process Improvement (API), 1809–1820, 1823–1826, *1827*, 1829, 1831–1839, 1841, 1843–1854, *1855*
Astfeld, A. M., *1104*
Astrand, P. O., 871, 876, *894*
Atencio, A., 2431, *2442*
Atherton, A., *1582*
AT&T, 1863, 1865, 1867, 1872, 1874, *1876*
Attewell, P., 1225, *1230*
Atwater, D., *945*
Atwater, L. E., 858, 859, *867*
Atwood, M. E., *1232, 2442*
Atzeni, P., 117, 121, *149, 150*
Aukstakalnis, S., 2505, *2519*
Ausubel, J., *541*
Automotive Industry Action Group (AIAG), *1974*
Autosimulations Inc., 2456, *2465*
Avery, A. C., 826–828, 830, 833, 834, *836*
AVISEM, 1145, *1152*
Avolio, B. J., 842–845, 847–852, 854, 855, 857, 859–861, *864–867*
Avriel, M., 2544, *2565, 2567*
Avula, X., *1789*
Axelsson, J., *1188*
Ayas, K., *1407, 1408*
Aydin, C. E., 1217, *1230*
Ayman, R., 2208, *2216*
Ayoub, M. A., 1081, *1100*
Ayoub, M. M., 1050, 1052, 1070, 1071, 1080, *1100, 1104, 1105, 1107–1109, 1152*
Ayres, R. U., 487, *528*
Azadeh, M. A., *972*
Azimi, H., 957, *969*
Azueta, S., 2434, *2442*
Azumi, K., *994*

Ba, S., *972*
Babakus, E., 628, 630, *631*

Babbage, C., 870, *894*
Babu, P. S., 1524, *1526*
Bach, V., 214, 215, 218, *223*
Baddeley, A. D., 2434, 2435, *2442*
Badeau, P., *824*
Badger, D., 1076, *1102*
Badler, N., 2434, *2442*
Badler, N. I., 1112, 1113, 1127, *1127*
Badler, N. L., 1050, *1100*
Bahe, F., 193, *226*
Bahrick, H. P., 930, *941*
Bailyn, L., 1222, *1230*
Bainbridge, L., 1135, *1152,* 1894, *1916*
Baines, R. W., 508, *528*
Bair, J. H., 220, 222, *223*
Baker, A. D., 697, *706*
Baker, C. C., *1154*
Baker, C. R., *2220*
Baker, C. V., 934, *944*
Baker, E., *824*
Baker, K. R., *1740,* 1746, *1766, 2053*
Baker, O., *990*
Baker, R. H., 228, *257*
Baker, T. E., 2562, *2565*
Bakken, G. M., *1100, 1108*
Bakshi, B., *542*
Bal, J., 1696, 1706, 1707, *1715*
Balakrishnan, N., *1955*
Balakrishnan, S., *1788*
Balas, E., *1739,* 2586, *2600*
Balci, O., 2461, *2465*
Baldwin, E. N., *1330*
Baldwin, R. A., 690, 691, *706*
Baldwin, T. R., *1790*
Baldwin, T. T., 933, *941*
Balkin, D. B., *865*
Ball, M., 2061, *2068*
Ballakur, A., *1330*
Ballou, R. H., 2071, *2081*
Baloff, N., 988, *989*
Balzer, W. K., 2200, *2215*
Bammer, F., *1230*
Bammer, G., *1230*, 1224
Banaag, J., 1071, *1102*
Bancroft, N., 86, 90, 93, 94, *107*
Banda, C., *2442*
Bandura, A., 986, *989*
Bane, L. J., 1932, 1945, *1954*
Banker, R. D., 877, *894*
Banks, J., 128, *150,* 2446, 2449, 2455, 2456, 2461, 2465, *2465–2467, 2494*
Bannon, L., *1918*
Bao, S., *1236*
Bapat, V., *2467*
Barash, M. M., 473, *482,* 697, *709*
Barbey, J., *398*
Barbosa, L. C., 130, *150*
Barbuceanu, M., *2019*
Bard, J., 796, 800, 802, *823, 1251, 1278*
Bar-Hillel, M., 2196, 2197, *2215*
Barkawi, H., *1108*
Barker, T. B., 2235, 2237, *2239, 2240*
Barkmeyer, E. J., *352*
Barkmeyer, Tom, 352
Barley, S. R., 952, *969*
Barling, J., 852, 860, *864*
Barlow, R. E., 1932, 1935, *1954*
Barnard, P., 1135, *1155*
Barnes, C. D., 2458, *2465*
Barnes, M. M., *2467*
Barnes, R. M., 874, *894*
Barness, Z. I., *989*
Barnett, C. K., *1806*
Barnett, D., *2580*
Barnett, J., *971*
Barnett, V., 2272, *2292*
Barnthouse L., 537, *540*
Barnum, J., *540*
Baron, J., 2196, 2210–2212, *2215*
Baron, S., 2429, *2442, 2444*
Barquín, R. C., *107*
Barreto, S. M., 1161, 1177, *1188*
Barrett, Colleen, 849
Barrie, D., *1807*
Barry, M., 83, *107*
Bart, V., 686, *706*
Barta, T. M., 2464, *2467*
Bartholdi, J., 802, *822*
Bartholdi, J. B., 2093, 2105, *2108*
Bartholomew, D., 1701, 1706, 1708, 1709, *1715*
Bartholomew, R., *1919*
Bartholomew-Biggs, M. C., 2564
Barto, A. G., 2646, *2648*
Barton, Jeff, 352
Barton, R. R., 2480, *2494*
Barua, A., 260, *278*
Basak, I., 2214, *2215*
Bass, B. M., 842, 843, 845, 847–851, 853, 854, 857–861, *864–866*
Bass, S. F., *1107*
Bässler, R., *399*
Bastress, E. K., 1176, *1188*
Bateman, B. O., 531, *540*
Bateman, T. S., *994*
Bates, M. W., 917, *917*
Bateson, J., 625, *632*
Bateson, J. E. G., 624, *631*
Battié, M. C., 1070, *1100, 1101*
Battig, W. F., 929, *941*
Bauer, W., *2520*
Bauman, R. D., *1189*
Baumgartner, P., 602, 603, *619*
Baumol, W. J., 2074, *2081*
Baveja, A., 1896, *1916*
Bazaraa, M. S., 2530, *2538,* 2545, 2549, *2565, 2567*
Bazerman, M., 2210–2212, *2215*
Beach, L. R., 2205, 2207, 2214, *2215*
Beale, E. M. L., 2630, *2646*
Bean, J. C., 2645, *2646*
Beauchesne, M.-N., *1153*
Beaumariage, T. G., *176*
Bechhofer, R. E., *2239,* 2488, *2494*
Bechtel, C., 2112, *2138*
Beckers, S., 761, *770*

Becket, W. M., *1100*
Beckhard, R., *57*
Beckman, R. J., 2285, *2292*
Beckman, T., 148, *152*
Bednar, D. A., 625, 626, *632*
Beebower, G. L., *770*
Beers, M. C., *224*
Beevis, D., *2443*
Beightler, C. S., 2559, *2565, 2567*
Beishon, R. J., 1135, *1152*
Beitz, W., 207, *225, 587*
Belcher, D. W., 903, 908, *918*
Bell, D., 623, *631,* 2180, 2203, *2215*
Bell, H. H., *941*
Bell, T., 29, 32, 35, 44, 53, *57*
Bellman, R., 2636, 2646, *2646*
Bellman, R. E., 2620, *2621*
Bellovin, S., 229, *257*
Belsley, D. A., 2284, *2292*
Belyavin, A. J., *2442*
Benabdallah, S., *1918*
Bengard, M. L., *1583*
Benge, E. J., 904, *918*
Bennet, J. L., 131, *150*
Bennett, D., 538, *540*
Bennett, K. B., *1039, 1040*
Benneyan, J., 745, *748*
Benson, R. F., *2053*
Bentley, J. P., 1883, *1885*
Benveniste, A., 2635, *2646*
Berchem-Simon, O., 1144, 1145, *1152*
Berens, A. P., 1909, *1918*
Berg, J., *352*
Berg, T., *352*
Bergamasco, M., *2520*
Bergen, J., 2434, *2443*
Bergen, M. E., *2032*
Berger, C. J., 870, 871, 893, *894*
Bergqvist, U., 1202, *1235*
Beringer, D. B., 1186, *1189*
Berkelman, R. L., *1104*
Berkenkamp, J., *542*
Berliner, C., 2318, *2329*
Bermudez, J., *2053*
Bernard, B. P., *989*
Bernard, R., 228, *257*
Berners-Lee, Tim, 244
Bernoulli, D., 2181, *2216*
Berra, P. B., 473, *482*
Berry, L. L., 624, *631–633, 650, 1965*
Berry, W. L., *561,* 1752, *1766, 2053*
Berson, Y., 859, *864*
Bertsekas, D. P., 2562, *2565, 2567,* 2575, *2580,* 2636, 2637, 2645, 2646, *2646*
Bertsimas, D., 801, *822*
Berwick, D. M., *989, 1855*
Besachio, Mary Eileen, 352
Bethea, N. J., *1100, 1108*
Bettman, J. R., *2221, 2443*
Beuter, A., *1102*
Beyer, H., 956, 964, *970*
Beyer, J. M., *969*
Bezjian-Avery, A., *632*
Bhaskaran, K., *1717, 1739*
Bhattacharya, A. K., *2139*
Bhote, K. R., 2233, *2239*
Biazzo, S., 1697, *1715*
Bickley, W., *945*
Bierbaum, C., 2423, *2442*
Bierbaum, C. R., *2442*
Bieser, M., *2221*
Biggemann, M., *1101, 1108*
Bigos, S. J., 1070, 1071, *1100, 1101, 1107*
Bikson, T. K., 1217, *1232*
Billesbach, T., 559, *559*
Billing, C. E., 962, *970*
Billings, C. E., 1297, *1310*
Billington, C., 1669, 1692, *1693,* 2112, 2113, 2127, *2140*
Billington, P. J., 2044, *2052*
Bindiganavale, R., *1127*
Birchfield, J. C., 829, 830, *836*
Birge, J. R., 2630, *2646*
Birnbaum, M. H., 2194, 2197, 2198, *2216, 2223*
Bishko, S. A., *542*
Bishop, R., 1363, *1389*
Bitner, M. J., 623, 624, *631, 633,* 641, *649, 650*
Bito, J. F., *399*
Bittner, A. C., *1109,* 1115, *1127*
Bixby, Robert, 2535
Bixby Cooper, M., *2140*
Bjork, R. A., 929, 937, *942, 945*
Bjørkland, R., 1086, *1108*
Blach, R., 2510, *2519*
Black, A., *399*
Blackburn, R., *1806*
Blackler, F., 1225, 1226, *1230*
Blackman, H. S., 1941, *1954,* 2189, 2191, 2192, *2218*
Blackstone, J. H., *353*
Blaha, M., *1789*
Blanding, W., 652, 653, 655, *663*
Blank, G. E., *2468*
Blank, L. T., *2351*
Blanning, R. W., 145, *150*
Bläßer, S., *225*
Blatner, D., 2505, *2519*
Blattberg, R. C., 679, *683*
Blauner, R., 874, *894*
Blick, A. C., *1101*
Blickensderfer, E. L., *945*
Blignault, I., 1224, *1230*
Block, J., *353*
Block, M. R., 539, *540*
Bloemhof-Ruwaard, J. M., 538, *540, 541*
Blok, K., *542*
Blood, M. R., 886, *896*
Bloom, J., *897*
Bloom, K. C., 931, *941*
Bloor, S., *225, 707*
Blumer, C., 2204, *2215*
Boal, K. B., 845, *864*
Board of Certification in Professional Ergonomics, *1101*
Bobkranz, R., *1104*
Bobo, S., 1908, *1916*

Bobrow, D. G., *154*
Boccardi, S., *1101*
Bochtler, W., *224*
Bock, R. D., 2182, *2216*
Bock, Y., *225*
Boddy, D., 1222, *1230*
Bodker, S., 964, *970*
Boeing Commercial Aircraft Group (BCAG), 957, *970*
Boender, E., 201, *223*
Boerstler, H., *993*
Boettcher, K., 2426, *2442*
Boff, K. R., 2410, *2442*
Boguski, T. L., *543*
Bohr, D., *989*
Bohr, P. C., *990*
Boies, S. J., *1232*
Boisot, M. H., 148, *150*
Bolat, A., 1734, *1740*
Bolijn, A. J., 1145, *1152*
Boller, G. W., 628, 630, *631*
Bolman, L. G., 982, *989*
Bolton, G. E., 2210, 2211, *2216*
Bonczek, R., 107, *107*
Bond, B., *352, 353*
Bond, Bruce, 352
Bond, W., 1700, 1710, 1711, *1716*
Boniface, D. R., *2240*
Bonney, M. C., *1106, 1108*
Bonney, R., 1063, *1101*
Bontadelli, J. A., *2405*
Booch, 292
Booher, H. R., 1297, *1310*
Booker, J. D., *483*
Booker, S., 1216, *1236*
Booms, B. H., *649*
Boothroyd, G., *399,* 403, *445,* 1328, *1330*
Boothroyd Dewhurst, Inc., *399*
Borenstein, J., 1525, *1525*
Borg, G. A. V., 1071, *1101*
Bornstein, G., 2212, *2216*
Borrego, S. A., *1876*
Borsenik, F. D., 835, *836*
Bort, J., 228, *257*
Boss, Bernard M., 844
Botta, E., 207, *224*
Bottomley, A., 220, *223*
Boucher, T. O., 1902, *1916*
Boudreau, M., 94, *107*
Bourne, L. E., Jr., *945*
Bourne, T. A., *258*
Bourret, P., 1779, *1787*
Boustead, I., 537, *540*
Bouzid, N., 931, *941*
Bovair, S., 931, *943*
Bowen, D. E., 559, *559,* 625, *633,* 975, 978, *990*
Bowen, H. K., 16, *24,* 551, *561*
Bowen, T., 95, *108*
Bowers, C. A., *947, 971*
Bowers, M. R., 635, *649*
Bowersox, D. J., 2113, 2123, 2128, *2138, 2140*
Bowersox, Donald J., 2056
Bowker, A. H., 2243, 2248, 2255, *2263*
Bowman, E. H., 2568, *2580*
Box, G. E. P., *2003,* 2225, 2227, 2229, 2232–2235, *2239,* 2273, *2293*
Boyce, P., *1918*
Boyett, J. H., *57*
Boyett, J. T., *57*
Boynton, A. C., 214, *226,* 686, *706*
Bozer, Y., *2109*
Bozer, Y. A., 1509, 1525, *1525, 1526,* 2095, 2106, *2108*
Brace, G., *2139*
Brache, A. P., 925, *945*
Bracken, J., 2540, *2565*
Bradley, K., 147, *149*
Bradley, S. P., 813, 814, *823,* 2074, *2081,* 2537, *2538, 2581*
Bradtmiller, B., *1128*
Braig, B. M., *632*
Bralla, J. G., 1315, *1330*
Bramel, Julien, 2019
Bramson, M., 2167, *2171*
Brand, A., 435, *445*
Brand, R. R., 2113, *2139*
Brandon, J., 1703, 1704, *1716*
Branson, R. K., 926, *941*
Branson, Richard, 849
Bras, B., *708*
Brashers, D. E., 2211, 2214, *2216*
Bratley, P., 2485, *2494*
Brauchler, R., *1104*
Brauchler, W., *1104*
Braun, M., *1231*
Bray, D., 939, *943*
Bray, O. H., 486, *528*
Brazier, D., 1313, *1330*
Brazile, R. P., 1734, *1739*
Breant, F., *2520*
Breaugh, J. A., 857, *864*
Brecht, L., 218, *224*
Brehmer, B., 2200, *2216, 2218*
Breinin, E., *866*
Breining, R., *2520*
Brennan, L., 1050, *1101*
Breslow, L., 1179, *1188*
Bressan, C., 745, *748*
Brett, B., *2442*
Bricken, W., *2520*
Briggs, G. E., 934, *941*
Brimson, J. A., 2318, *2329*
Brinckmann, P., 1072, *1101, 1108*
Brinson, G. P., 752, *770*
Brisley, C. L., 1448, *1462*
BRITE-EURAM, 209, *223*
British Standards Institution, 1133, *1152*
Brnich, M. J., Jr., *2220*
Broadbent, D. E., 1133, *1153*
Brobst, S., *152*
Brobst, S. A., *152*
Brockmann, T., *1501*
Broderick, R. L., *1152*
Brodie, M. L., 122, *152*
Brogmus, G., 1202, *1232*
Brokaw, N., 1080, *1101, 1103*
Bromme, R., 214, *223*

Brooke, A., 2536, *2538*
Brookhouse, J. K., 2195, *2216*
Brooking, A., 147, 148, *150*
Brooks, F. P., 2507, *2519*
Brookshear, J. G., 74, *108*
Brousseau, K. R., *970*
Brouver, M. L., *1104*
Brouwer, M., *1153*, 1366, *1389*
Brouwer, M. L., *1153*
Brown, C., 88, *108*, 1215, 1226, *1230*
Brown, C. M. L., *1230*
Brown, D. R., *2240*
Brown, G. G., *2138*
Brown, M. L., *1105*
Brown, R. G., 2498, *2519*
Brown, S., 83, *108*
Brown, S. W., 630, *633, 650*
Brown, T. J., *631, 632*
Brown, W. S., *2443*
Browne, J., 229, *258*, 1747, *1767*
Brownell, W. D., *1766*
Broyden, C. G., 2552, *2565*
Bruce, D. R., *942*
Brucker, P., *1740*
Brückmann, R., *2520*
Bruderlin, A., *1127*
Brüggemann, K., *587*
Bruni, J. R., *991*
Brunner, D. T., 2455, *2467*
Brunnermeier, S., 350, *352*
Brunswick, E., 2200, *2216*
Bryant, F. B., *2218*
Brynjolfsson, E., 146, *150*
Bryson, J. M., 845, *864*
Bryson, S., 2498, 2503, 2505, *2519, 2520*
Bubb, H., 1112, *1127*
Buch, K., 979, *989*
Buchanan, D., 1697, 1700, 1701, *1715*
Buchanan, D. A., *1230*, 1222
Buck, J. R., 2173, 2191, *2216*, 2360, *2392*
Buckle, P., 1061, *1100, 1107*
Bucklin, L. P., 2113, 2115, *2139*
Budescu, D., 2202, *2216*
Budescu, D. V., *2222*
Budros, A., 1888, *1916*
Buehler, J. W., *1104*
Buell, P., 1179, *1188*
Buffa, E. S., 485, *528*
Buffa, M., 1906, *1916*
Buford, J. F. K., 229, *257*
Buie, E., 1215, 1216, *1231*
Bukowitz, W. R., 148, *150*
Bulkley, J. W., *541*
Buller, B. J., *972*
Bullinger, H.-J., 207, *223*, 634, *649*, 1050, *1101*, 1202, *1231*, 1284, 1285, *1294*
Bulow, J., 277, *279*
Bunting, A. J., *2442*
Burandt, U., 1117, *1127*
Burbridge, J. L., 462, 475, *482*
Burdea, G., *2520*
Burdea, G. C., 169, *175, 400*
Burdick, Dave, 352
Burdorf, A., *1108*
Bureau of Labor Statistics (BLS), *1101*, 1157, 1158, 1164, 1173, 1174, *1188*, 2395, *2405*
Bureau of National Affairs (BNA), 902, *918*, 1070, *1101*
Burgelman, R. A., 955, *970*
Burger, G. C. E., 1137, *1152*
Burgess, B. H., 303, *307*
Burghardt, M., 205, *223*
Burk, S. L. H., *918*
Burke, P., 697, *706*
Burns, J. M., 845, 847, 852, 853, *864*
Burns, L., 1734, *1739*
Burns, M., 912, 913, *918*
Burns, R. N., 1746, 1747, 1750–1753, 1757–1758, 1760, 1763–1766, *1766*
Burr, Donald, 849
Burress, R., *540*
Burri, G., 1373, *1389*
Burrows, W., *1583*
Busalacchi, F. A., 605, *617*
Busemeyer, J. R., 2205, *2216, 2222*
Buttle, F., 629, 630, *631*
Buzacott, J. A., 1669, 1691, 1692, *1693, 1740*, 2146, 2150, 2168, 2170, *2171*
Buzacott, J. S., 1640, 1644, 1648, 1649, 1652, 1655, 1656, 1660, 1664, 1668, *1668*
Buzzell, R. D., 623, *631*
Bycio, P., 851, *864*
Byham, W. C., *989*

CAD-FEM GmbH, 203, *223*
Cady, F. B., *2293*
Caesar, C., *224*
Cai, K. Y., *1955*
Cai, W., *2466*
Cailliau, Robert, 244
Cakir, A., 1131, 1145, *1152*, 1228, *1231*
Cala, F., 536, *540*
Calderwood, R., *1038*
Caldwell, B., 91, *108*
Caldwell, L. S., 1056, *1101*
Calhoun, J. C., *175*
Calichman, M., 746, *749*
Callaway, 1787, *1787*
Callery, C. A., *2467*
Calogero, J. A., *1104*
Calzon, J., 1956, *1965*
Camillerapp, J., *1918*
Cammann, C., *993*
Camp, R., 1703, *1715*
Camp, R. C., 2112, *2139*
Campbell, D. T., 885, *895*
Campbell, J., 921, *941*
Campbell, J. C., *560*
Campbell, R. J., *865*
Campion, M. A., 870, 871, 873, 875, 877, 883, 884, 886–889, 891–894, *894, 895, 898*, 976, 977, 984, 986, 987, *989, 992*, 1217, *1231*
Cana, O., 1569, *1582*
Cannon, J. R., *1039*
Cannon-Bowers, J. A., *897*, 933, 934, *941, 942, 945, 947*
Canter, R. R., *895*

Cao, X. R., 2633, *2647*
Caplan, J. S., *1583*
Caplan, R. D., 874, *895, 989*
Carayon, P., 981, *991,* 1221, 1222, 1224–1226, 1228, 1229, *1231, 1233–1235*
Carayon-Sainfort, P., 1222, 1223, *1231, 1236*
Carbon, M., 214, *223, 225*
Card, K., 1015, *1037*
Card, S. K., 133, *150,* 1208, 1209, 1216, *1231,* 1297, *1310,* 2434, *2442*
Cardy, R. L., *865*
Carlisle, K. E., 926, *941*
Carlson, A. E., 1131, *1152*
Carlson, E. D., 113, 114, 125, *153,* 2079, *2082*
Carlson, W. E., *1129*
Carman, J. M., 628, 630, *631, 993*
Carpentier, J., 2560, *2565*
Carr, D. K., 980–982, *989*
Carrico, T. M., *2466*
Carrie, A., 503, 506, *528*
Carrier, R., *1102*
Carroll, J. M., 1210, 1213, 1214, *1231, 1234*
Carroll, S. J., Jr., 907, *919*
Carrubba, E. R., 1922, *1954*
Carson, A. B., 1131, *1152*
Carson, J. S., *2465, 2494*
Carter, K., *1234*
Carter, M. F., 499, *528*
Carter, M. W., 1746, 1747, 1750, 1752, 1753, 1758, 1760, 1765, *1766*
Cartter, A. M., 901, *918*
Cartwright, D., 880, 882, *895*
Cascio, W. F., *1716*
Case, D., 1217, *1235*
Case, K., *1106, 1108*
Case, K. E., *2351, 2405*
Cassell, J., 1127, *1128*
Castanon, D. A., 2645, *2646*
Castro, E., 77, *108*
Cats-Baril, W., 1119, *1128*
Cats-Baril, W. L., *990*
Cattanach, R. E., 534, *540*
Caudill, M., 1906, *1916*
Caudle, S. L., 1696, 1698, 1702, 1705, 1708, 1712, *1715*
Cavanagh, R. C., *1038*
Cavazos-Cadena, R., 2645, *2646*
Caverni, J. P., *2216*
CCE-CNMA Consortium, *528*
Celi, I., *225*
Celiker, T., *225*
Center for Health Systems Research and Analysis (CHSRA), 984–987, *989*
Centers for Disease Control (CDC), 1163, 1164, 1168, *1188, 1190,* 1195, 1199, 1205, *1231*
Centry Research Corp., 1160, *1188*
Cerf, Vint, 238
Ceri, S., *150*
Ceroni, J. A., 604, 607–609, *617*
Chadha, B., 171, *175*
Chadha, N., 534, *540*
Chaffey, D., 143, *150*
Chaffin, D. B., 1046, 1050–1054, 1058, 1059, 1061, 1068, *1101, 1102, 1108, 1109,* 1112, 1117, 1121, 1126, *1128, 1129*
Chain Store Age, 266, *279*
Chalmet, L. G., 2573, *2580*
Champy, J., *224, 528,* 747, *749,* 950, *970,* 1700, 1706, 1709, 1714, *1716,* 2049, *2052,* 2126, 2132, *2139*
Chan, D., 2573, *2580*
Chan, S., 1697, *1715*
Chandra, J., 1777, *1787*
Chandru, V., *1330*
Chaney, F. B., 1896, 1897, *1916*
Chang, C. S., 2646, *2646*
Chang, D. T., 1524, *1525*
Chang, P., 1777, *1788*
Chang, S., *1790*
Chang, S. H., 1524, *1525*
Chang, T., 1777, 1781, *1788, 1790*
Chang, T. C., 452, 472, *482, 483*
Chang, T.-C., 448, 462, 478, *483*
Changkong, V., *2623*
Chao, E. Y., 1125, 1126, *1128*
Chao, X., *1740*
Chapman, R. L., 1556, *1582*
Charlton, S. G., 1889, *1919*
Charnes, A., 2614, *2621*
Chase, R. B., 559, *559,* 624, *631*
Chase, W. G., 1023, *1037*
Chatelin, F., 2645, *2646*
Chatterjee, K., 2210, 2211, *2216*
Chatterjee, S., *2293*
Chemers, M. M., 2208, *2216*
Chen, C., 1779, *1788*
Chen, D., *528*
Chen, J., 1692, *1693*
Chen, J-M., 1904, *1920*
Chen, M., 171, *176*
Chen, N.-F., 1724, *1739*
Chen, P. P., 121, *149*
Chen, P. P. S., 121, *150,* 510, *528*
Chen, Q. X., *617*
Chen, T. T., *149*
Chen, V. C. P., 2646, *2646*
Chen, X., 602, *617*
Chen, Y., 507, *528*
Cheng, F., 1669, 1692, *1693*
Cheng, P. E., 2200, 2208, *2216*
Cheraskin, L., *894*
Cherkassky, V., *1790*
Cherns, A., 874, *895*
Chervak, S., 1136, 1137, 1141, *1152*
Chesler, D. J., 902, *918*
Cheswick, W., 229, *257*
Chi, C.-F., 1896, 1898, *1916*
Chicago Tribune, 671, *683*
Chien, I. S., 2604, *2621*
Chignell, M., *152*
Ching, C., 107, *108*
Ching, W., *1717*
Chionglo, J. F., *2019*
Chiu, C., 1776, *1788*
Cho, F., *561*
Cho, H., 1777, 1779, *1788*

Cho, M. S., *1526*
Choe, K.-I., 2093, 2095, 2106, *2108*
Choi, C. F., 1697, *1715*
Choi, J. H., *599*
Choi, S. C., 689, *706, 707*
Choi, S.-Y., 267, 278, *279*
Chong, E. K. P., 2635, *2646*
Choo, C. W., 149, *150*
Choong, Y.-Y., 1228, *1231*
Chopra, V., 756, *770*
Chou, Y. M., 1869, 1870, *1876*
Chow, C. S., 2645, *2646*
Chow, W. M., 1511, *1525*
Christensen, D., *1102, 1108, 1109*
Christensen, J., *2520*
Christensen-Szalanski, J. J., 2196–2198, *2216*
Christina, Vivi, 539
Christofides, N., 2062, *2068*
Christopher, M., 2116, *2139*
Christopher, M. G., *2140*
Christopher, Neil, 352
Chu, W. W., *1717*
Chubb, A. P., *2444*
Chui, Y. P., *707*
Chung, M. J., 690, 691, *706*
Church, R. L., 2076, *2081*
Churchill, G. A., *631,* 2277, *2293*
Churchill, G. A., Jr., *632*
Churchill, T., *1128*
Churchill, Winston, 134
Chvatal, V., 538, *540*
Chvátal, V., 811, 814, *823*
Ciborra, C., 951, *970*
Ciesla, M., *225*
Ciriello, V. M., 1055–1058, 1071, 1072, 1080, *1101, 1107,* 1117, 1118, *1128*
Cirillo, J., 745, *749*
Claderwood, R., *2219*
Cladwell, C., *989*
Clark, A. J., 1678, *1693*
Clark, D. S., 2635, *2647*
Clark, G. M., 2488, 2494, *2494*
Clark, H. H., 1023, *1037*
Clark, J., 826, *836*
Clark, K. B., 556, *559, 2329*
Clark, R. E., 928, 935, 940, *941*
Clark, W. C., *542*
Clauser, C. E., *1128*
Clearwater, S. H., 697, *707*
Clegg, C., *1716*
Cleland, D. I., *1350*
Clemen, R. T., 2173, 2177, 2178, 2183, 2187, 2191, 2213, *2216, 2218*
Clement, A., 964, *970,* 1225, *1231*
Clement, C., 1213, *1232*
Clements, J. H., *1102, 1128*
Clendenin, J. A., *2139*
Cleveland, R., *1189*
Cleveland, R. J., 1161, 1183, *1188*
Cleveland, W. S., *2240*
Clifton, T. C., *866*
Clinton-Cirocco, A., *1038*
Closs, D. J., 2113, *2138*
CMS Research Inc., 2458, *2465*
Coad, P., 123, *150*
Coase, R. H., 2123, *2139*
Cobb, G. W., *2240*
Cobb, S., *895, 989*
Cochran, W. G., 2255, *2263*
Cod, E. F., 80
Cofer, C. N., 929, 931, *942*
Coffey, G., *2216*
Coffey, R., 747, *749*
Cogger, K. O., 2605, 2608, *2621*
Cohen, A., 165, *175,* 1158, 1161, 1176, 1180, 1183, 1184, *1188, 1189*
Cohen, A. L., 981, *989*
Cohen, B. G. F., *1235*
Cohen, H. H., *1188, 1189*
Cohen, J., *989*
Cohen, M. D., *152*
Cohen, M. S., 2197, 2198, 2205, *2216*
Cohen, P., *989*
Cohen, P. A., 928, *942, 943*
Cohen, W., *1236*
Cohon, J. L., 2074, *2081*
Coiffet, P., 169, *175*
Colbert, D. N., 2112, *2139*
Cole, L. L., *1236*
Cole, R., 1386, *1389*
Cole, R. E., *1806*
Coleman, E. P., 852, *864*
Coleman, J. L., 2133, *2139*
Coleman, P. J., *1189*
Coleman, T. F., *2567*
Colligan, M., *1189*
Colligan, M. J., 1158, 1176, 1180, 1184, *1188*
Collins, C., *2442*
Collins, D. C., 2645, *2646*
Collins, J., 656, *664*
Collins, J. C., 7, 8, 10, *24*
Collins, J. M., 915, *918*
Colombini, D., 1064, 1067, *1101*
Colquhoun, G. J., 508, *528*
Colson, K., *918*
CommerceNet, 343, *352*
Compton, D., *947*
Compton, W. D., 2168, *2171*
Computerworld, 949–951, *970*
Comrie, P. R., *541, 600*
Conard, R., 1181, *1188*
Conard, R. J., *1188*
Condit, P. M., 956, *970*
Conen, W., 698, *706*
Conger, J. A., 846–849, 854, 855, 859, *864, 865*
Conn, Andy, 2564
Connelly, David, 352
Connelly, M. S., *866*
Conner-Sax, K., 229, *257*
Connolly, T., 2204, *2216*
Connor, G., 761, *770*
Connors, D., 1692, *1693*
Conolley, E. S., *942*
Conrad, R., *1190*
Constanza, M. C., *1102*
Consumer Product Safety Commission (CPSC), *1129*
Consumer Reports, 536, *540*
Contractor, N., 952, *970*

Contreras, L. E., *2393*
Converse, S. A., *945, 993*
Conway, F., *1236*
Conway, R. W., *1740*
Cook, J., *989,* 1229, *1231*
Cook, R. D., 2285, *2292*
Cook, T. D., 885, *895*
Cook, W. D., *2108*
Cooke, D. P., 953, 965, *970*
Cooke, D. S., *399*
Cooke, J. A., 2065, *2068, 2069*
Cookson, B., *1789*
Coombs, B., *2220*
Coombs, R., 1696, 1700, *1715*
Cooper, C. L., *1190,* 1226, *1233*
Cooper, J. C., 671, *683*
Cooper, M. C., 2072, *2082,* 2112–2119, 2123–2126, 2133, *2139, 2140*
Cooper, R., 2319, *2330*
Cooper, R. G., *989*
Cooper, T. D., *1919*
Cooper, W. W., *2621*
Coopers & Lybrand, 780, *785*
Copacino, W. C., 2112, 2115, *2139*
Corbett, C., *989*
Cordata, J. W., 148, *150*
Cordery, J. L., 877, *895*
Corker, K., *2444*
Corker, K. M., 2431, 2434, 2437, *2442, 2444*
Corlett, E., *1037*
Corlett, E. N., 1061, 1063, *1101, 1105,* 1117, *1128*
Corlett, N., 1363, *1389*
Cormier, D., 1940, *1955*
Cormier, D. R., 1906, *1916*
Cormier, S. M., 929, *942*
Cornaby, B. W., *543*
Cornell, J. A., 2239, *2239, 2240*
Coronel, C., 80–82, *109,* 117, *153*
Corrigan, S., 1705, *1715*
Cotton, J. L., 978, 979, *989*
Coughlan, A., *2140*
Coughlan, R., *2216*
Council of Logistics Management, 2113, *2140*
Courant, R., 2560, *2565*
Courtois, P. J., 2645, *2647*
Covey, S. R., 24, *24*
Cowan, D., 487, *529*
Cox, J., 557, *560*
Cox, J. F., 557, *561*
Cox, J. F., III, *353*
Cox, R., 2113, *2139*
Coxx, J. L., *941*
Craig, A. T., 2243, 2248, 2254, *2263*
Craig, J. J., *399*
Craik, F. I. M., 929, *942*
Craik, K. J. W., *2444*
Crain, R. C., 2455, *2465*
Crawshaw, C. M., 931, *941*
Cressman, G. E., 667, *683*
Criswell, H., 950, *970*
Criteria for Performance Excellence (2000), 1961, 1964, *1965*
Crocker, D. C., 2274, 2276, *2292*
Croll, A. A., 229, *257*
Cronbach, L. J., 630, *631, 990*
Cronin, J. J., Jr., 628, 630, *631*
Crookall, P., 851, 860, *865*
Crosby, P. B., 626, *631,* 1794, *1805*
Crosby, Philip, 747
Crossman, E. R. F. W., 886, *895*
Crowcroft, J., *736*
Crowder, H., 809, 811, *823,* 2587, *2600*
Crowe, T. J., 1698, *1715*
Crowley, P., 759, *771*
Crowston, K., 1287, *1295*
Crozier, M., 1225, 1226, *1231*
Cruikshank, G., *1234*
Crutzen, P. J., 537, *541*
Cruz-Neira, C., 2503, 2507, *2519, 2520*
Csikszentmihalyi, M., 4, *24*
Csurgai, G., 172, *175*
Cucuzza, T. G., 672, *683*
Cullinane, T. P., *529*
Cullinan-James, C., 2538, *2539*
Cuneo, J., *1153*
Cunningham, J. B., 914, *918*
Curington, W., *918*
Curphy, G. J., *865*
Curran, M. A., *543*
Curran, M. S., *540*
Curran, T., 86–88, 90, 91, *108*
Curtis, M. A., 473, *483*
Cvan Riel, M. P. J. V., *1107*
Czaja, S. J., 1223, *1231*
Czepiel, J. A., 624, 625, *631, 633*

Daganzo, C. F., 1734, *1739*
Dahl, S. G., 2410, *2442–2444*
Dahlin, L. B., *1104*
Dai, F., *2520*
Dai, J. G., 2167, *2171*
Dailey, A., *353*
Dale, B. G., 1696, 1697, *1716*
Daley, D. J., 1635, *1668*
Damkot, D. K., 1071, *1102*
Damodaran, L., 964, *970*
Dani, T. H., 2498, *2519*
Daniel, C., 2227, *2239*
Daniel, J. W., *2647*
Daniels, E., *749*
Danielsen, N., *1104*
D'Anjou, L. O., 599, *599*
Danner, D. L., 1424, 1425, *1462*
Dantzig, G. B., 2526, 2528, *2538,* 2543, 2556, *2565, 2567,* 2630, *2647*
Danziger, J. N., 1220, *1231*
Daouas, T., 1777, *1788*
Dar-El, E., 1405, *1407, 1408*
Darwen, H., 122, *150*
Daskin, M., 2067, *2068*
Daskin, M. S., 802, *823*
Date, C. J., 117, 122, *150*
Dauer, J. P., 2621, *2621*
Dauzère-Pérès, S., 2044, *2052*
D'Aveni, R. A., 955, *970*
Davenport, T., 86, 88, 92, *108,* 957, 962, *970*
Davenport, T. H., *57,* 148, *150,* 213, 215, 218, *224,* 348, *352,* 1696, 1700, *1715*
Davidon, W. C., 2551, 2552, *2565*

Davilla, D., *2443*
Davis, B., 87, *108*
Davis, G., 1550, *1582*
Davis, G. A., 532, 536, *540*
Davis, G. B., 127, *150*
Davis, J. B., 532, *540*
Davis, J. H., 2212, 2213, *2216*
Davis, K. R., 914, *918*
Davis, L., 1025, *1037*, 1781, *1788*, 2449, 2464, *2465*
Davis, L. E., 869, 870, 874, 884–886, *895*
Davis, Maggie, 352
Davis, R. B., 1125, *1128*
Davis, R. N., 984–986, *990*
Davis, S. M., *57*, 685, *707*
Davis, T., *736*, 2112, 2121, 2127, *2139*
Davis, T. P., *2240*
Davis, W. J., *2465*
Davis-Sacks, M. L., 982, 983, *990*
Dawes, R. M., 2198, 2209, *2216*
Dawson, Jim, 848
Dawson, K., 745, *749*
Day, D., 981, *990*
Dayhoff, J. E., 163, *175*
Dayton, T., 1206, *1233*
Deal, T. E., 982, *989*
Dean, A., 2235, *2239*
Dean, J. W., 975, 978, *990*
Dearden, N. J., *149*
Debenham, J., 122, *150*
Debreu, G., 2605, *2621*
Dechamplain, B., *1102*
Decker, P. J., 921, *945*
DeFanti, T. A., *2519*
Degani, A., *1152*
DeGarmo, E. P., 2401, *2405*
DeGeus, A., 7, 8, *24*
De Hoog, R., 213, *224*, *226*
Deivanayagam, S., *1100*, *1104*
de Jong, J. R., 1137, *1152*
DeKeyser, V., 1023, *1037*
Dekker, R., *541*
Delbecq, A. L., 909, *918*, 2213, *2217*
Delleman, N., *1101*
Delleman, N. J., 1076, *1102*
Delphi-Studie, 311, *322*
Delwiche, M., 1900, 1902, 1906, 1907, *1920*
Demaree, R. G., *991*
Dembo, R. S., 2559, *2565*
Demel, G., 957, *970*
De Meuse, K. P., *897*
Deming, W., 1794, *1805*, *1806*
Deming, W. E., 5, 20, 24, 62, 747, 975, *990*, 1830–1832, 1836, *1855*, 1958
Dempsey, P. G., *1100*, *1103*
Dempster, M. A. H., *1740*
Denardo, E. V., 2636, *2647*
Denison, D. R., 877, *895*
Denison, R., *540*
Denmark Ministry of the Environment, 532, *540*
Dennerlein, J. T., *1235*
Dennis, A. R., *2220*
Denoble, R. A., 740, *749*
Deo, N., *2581*
Derfler, F., 228, 229, *257*
Derks, S., *1189*
Dertien, M. G., 915, *918*
Dertouzos, M., 950, *970*
De Ruyter, J. C., 628, 629, *631*, *632*
De Ruyter, K., *631*
Dervitsiotis, K. N., 1700, 1706, *1715*
Dery, Arie, 842
DeSanctis, G., 144, 145, 147, *150*, *154*, 952, *970*, *973*
de Santos, P. G., *398*
Desarbo, W. S., 689, *706*, *707*
Deschamps, J., *57*
Desrosiers, J., 794, 809, *823*
Dess, G. G., *57*
Dessouky, Y. M., 2464, *2467*
de Treville, S., 2168, *2171*
Dettmer, H. W., 557, *560*
Deutsch, C., 86, *108*
Deutsch, M., 880, *895*
Deutsch, S., 2440, *2444*
Deutschmann, D., 1202, *1236*
Devin, K., 148, *150*
Devlin, S. E., *945*
DeVor, R., 602, *617*
Devore, J. L., 2243, 2248, 2254, *2263*
De Vries, W. R., *399*
Dewhurst, P., *1330*
deWitte, P., *529*
DFG, 444, *445*
Dhillon, B. S., *1583*, *1955*
Diamond, W., 2225, *2239*
Diamond, W. J., *2239*
Diaper, D., 1208, 1209, *1231*
Dickie, B. N., 1699, 1701, *1715*
Dickinson, D. A., 599, *600*
Dickinson, T. L., *945*, 987, *990*, *993*
Dickson, M. W., 984, 986, 987, *990*
Diehl, A., 1900, *1916*
Diehl, M., *895*
Dieter, G. E., *399*
Dietterich, T., 1780, *1790*
Dilger, K. A., 2064, *2068*
Dilla, B. L., *993*
Dillon, P. S., *540*
Dilworth, J. B., *353*
Dincer, M., 402, *446*
DiPasquale, J., *2466*
Dirken, J. M., 1137, *1152*
Dixit, A., 2212, *2217*
Doane, M., 93, *108*
Dobson, G., 2589, *2600*
Dodson, V. N., *1235*
Doebelin, E. O., 1884, *1885*
Doege, E., *587*
DOE OEPA, 594, *600*
Doherty, E. M., 979, 988, *989*, *990*
Doherty, M. E., *2215*, *2216*
Dolk, D. R., 131, *150*
Doller, H. J., 747, *749*
Dologite, D. G., 2214, *2220*
Donabedian, A., 637, 640, *649*, 983, *990*
Donald, D. L., 2460, *2465*

Dong, J., 1228, *1231, 1232*
Dong, Y., *528*
Donnell, M. L., *1920*
Donnelly, J. H., *865*
Doremus, Paul, 352
Dorfman, R., 626, *631*
Dorris, A. L., 2196, 2197, *2217*
Doscher, E., 1772, 1774, *1788*
Doss, B., *1920*
Dossantos, J. R. G. L., *1526*
Dougherty, E. R., 2243, 2254, *2263*
Doumeingts, G., 507, 513, *528*
Douwes, M., *1105*
Dowd, K., 229, *257*
Dowdy, L. W., *736*
Doyle, L. E., *1330*
Dragonetti, N. C., *153*
Drake, G., *2467*
Draper, A., *1330*
Draper, A. B., *1330*
Draper, C. W., *600*
Draper, N. R., *2003,* 2229, *2239, 2293*
Draper, S. W., 1297, *1310*
Drayer, R., 780, *785*
Drews, C., *2444*
Drews, C. W., *2444*
Dreyfus, H., 1023, *1037*
Dreyfus, S., 2636, 2646, *2646*
Dreyfuss, H., 1117, 1120, *1128*
Driskell, J. E., 2208, *2217*
Driver, M. J., 964, *970*
Drost, M. R., *1102*
Drozda, T. J., *483*
Drucker, P., 322, *322*
Drucker, P. F., 954, *970,* 2205, 2206, *2217*
Druckman, D., 937, *942*
Drud, Arne S., 2563
Drury, C., 1025, *1037*
Drury, C. G., 1131, 1132, 1134–1137, 1139, 1141, 1145–1147, 1151, *1152–1154,* 1890, 1891, 1893, 1896–1900, 1909, 1912, 1914, 1916, *1916–1920*
Du, D., *2567*
Du, X., 701, *708*
Du, Y., *823*
Duane, J. T., 1952, *1954*
Duarte, A. M., 767, 768, *771*
Dubinsky, A. J., 844, 851, *867*
Du Charme, W., *2217*
Duclos, L. K., 559, *560*
Duda, R. O., *2217*
Dueser, R., *587*
Duff, I. S., 2534, *2538*
Duffany, B. H., 926, *945*
Duffin, R. J., 2559, *2565, 2567*
Duffy, L., 2212, *2217*
Duffy, V., 1193, *1232*
Duffy, V. G., 983, *990*
Duket, S. D., *2444*
Dul, J., *993, 1105*
Dumas, M., 745, *749*
Dumas, M. B., 745, *749*
Dumas, Y., *823*
Dumdie, D. P., *1583*
Duncan, A. J., *2003*
Duncan, J. H., *1462*
Duncan, W. R., *1350*
Dunkle, D. E., *1231*
Dunleavy, J., *109*
Dunn, M. S., 463, *483*
Dunn, R. L., *1583,* 1586, 1587, *1622*
Dunnette, M. D., *866*
Dupont, P., 193, *224*
Du Pont Co., 2225, 2226, 2235, 2238, *2239*
Dupuis, P., 2645, *2647*
Düring, H., 193, *224*
Durlach, N., 2498, *2519*
Dutkewych, J., *750*
Dutta, P. K., 128, *150*
Dutta, S. P., *1110*
Dvir, T., 852–854, 861, *865*
Dwyer, D., 922, *942*
Dwyer, D. J., *942*
Dyer, J. L., 933, *942*
Dyer, J. S., 2080, *2081, 2623*
Dyer, L., 862, *865, 2215*
Dzida, W., 1228, *1231*
Dzierba, S., *990*

Early, J. F., 984, 986, 987, *990*
Earnshaw, R., *2520*
Easingwood, C., 635, *650*
Eason, K., 964, *970*
Easterby, R. S., 1135, 1137, *1152*
Easton, G. S., *1806*
Ebbesen, E. B., 930, *947*
Ebbinghaus, H., 930, *942*
Ebeling, B. J., *942*
Eberts, R., 1205, 1206, 1212, 1229, *1232, 1233, 1236*
Eberts, R. E., 604, *617,* 1209, *1232*
Eck, C. D., 214, *224*
Ecker, J. G., 2563, *2567*
Eckstrand, G., 1180, *1188*
The Economist, 705, *707,* 951, 969, *970*
Eddy, F., *1789*
Edelstein, H., 83, *107*
Eden, D., 847, 854, *865*
Edland, E., 2209, *2217*
Edosmwan, J. A., *1330*
Edosomwan, J. A., 1793–1801, 1804, *1806, 1917*
Edvinsson, L., 147, 148, *150, 153*
Edwards, C. D., 1794, *1806*
Edwards, F. C., 1069, *1107*
Edwards, F. R., 759, *771*
Edwards, J., *736, 2218*
Edwards, R. H. T., 1086, *1102*
Edwards, W., 2173, 2178, 2182, 2183, 2185, 2187, 2191–2197, 2201, *2217, 2223*
Egan, D. E., 1207, *1232*
Egbelu, P. J., 1511, 1524, 1525, *1525*
Ehrenfeld, J. R., 537, *541*
Ehrgott, M., 2621, *2621*
Eiglier, P., *632*
Einhorn, H. J., 1023, *1037,* 2196, 2197, 2199, 2201, *2217, 2218*
Einstein, W. O., *864*

Eisenberg, E., 952, *970*
Eisenhart, C., 2267, 2269, *2293*
Eisenstein, D. D., 2105, *2108*
Eisler, H., 1071, *1102*
Eklund, J., 1186, *1188*
Eklund, N., *1918*
El-Ansary, A., *2140*
Elders, L., *1108*
Electric Power Research Institute (EPRI), 960, *970*
Electronic Business XML, 343, *352*
Eliason, A., 74, *108*
Elizur, D., 908, *918*
Ellis, S., 2499, *2520*
Ellis, S. R., 2499, *2520*
Ellram, L. M., 2112, 2115, 2125, *2139, 2140*
Ellsberg, D., 2202, *2217*
Elmaghraby, S. E., 2037, *2052,* 2572, *2580*
Elmasri, R., 80, *108*
Elnakhal, A. E., 168, *175*
Elpet, B., 206, *224*
Elsayed, A. E., *529,* 1932, *1954*
Elsayed, E. A., 1524, *1525*
Elster, J., *2217*
El-Taji, S., *2467*
Emami-Naei, A., *1885*
Embrey, D., 1028, *1038*
Embrey, D. E., *2217*
Emerson, S. M., 902, *918*
Emery, F., 964, *970*
Emery, F. E., 1186, *1188*
Emmelhainz, M. A., 2135–2137, *2139, 2140*
Emmons, H., 1727, *1739,* 1764, *1766*
Emory, F. E., 874, *895*
Enda, Y., *2622*
Engel, F. L., 1895, *1917*
Engelhardt, M., 1932, 1945, *1954*
Englebart, Douglas C., 134
English, W. L., 745, *750*
Englund, R. L., *1262*
Engquist, M., *2566*
Enkawa, T., 552, 555, *559, 560*
Entin, E., *1039*
Environmental Protection Agency (EPA), 1164, 1168, *1188*
Enz, C. A., 855, *865*
EPA AP-42, 597, *600*
EPA M204, 598, *600*
EPA OCEPA, 591, *600*
EPA OPA, 594, *600*
EPA OPPT, 594, *600*
EPA OSW, 594, *600*
EPA OW, 595, *600*
Eppinger, S. D., 688, *708*
Erens, F., 694, *707*
Erens, F. J., *707*
Ergun, O., 2059, *2068*
Erhun, F., 545, *559*
Erickson, T. D., 1213, *1232*
Ericsson, K. A., *945,* 1027, *1037,* 1209, *1232,* 2206, *2217*
Ericsson, M., *151*
Erig, M., 1054, *1101,* 1121, *1128*
Eriksen, C. W., 1895, *1917*
Erisman, A. M., *2538*
Erixon, G., 689, *707*
Ernst, A. T., 2458, *2465*
Ernst, J., *1107*
Erwin, P. J., 1160, 1161, *1188*
Esbeck, E. S., *989*
Eschenbach, T. D., *2392*
Eschinger, Chad, 352
Eskigun, E., 538, *540*
Esprit Consortium AMICE, 489, 490, 512, *528*
Estes, C. B., *2405*
Estes, J. H., *2392*
Etter, G., 745, *750*
Ettl, M., 1669, 1690–1692, *1693*
Ettlie, J., 949, *971*
Ettlie, J. E., *560*
Etzioni, A., 2206, *2217*
Eubanks, C. F., *707*
EUREKA, *399*
European Environmental Agency, 531, *540*
Evanoff, B. A., 981, *990*
Evans, J. B. T., 2196–2199, 2201, *2217*
Evans, J. R., 1889, 1900, *1917, 2648*
Evans, P., *57*
Eveland, J. D., 1217, *1232*
Eversheim, W., 206, 207, *224,* 637, 640, 643, 647, *650*

Faard, H. F., 1053, *1105*
Fabre, J. M., *2216*
Fabrycky, W. J., *2351, 2393,* 2395, 2399, *2405*
Facchin, P., *748*
Fahey, L., *57*
Fähnrich, K.-P., 635, 637, 639, *650*
Fahs, M., *1189*
Faley, R. H., 923, *943*
Fallentin, N., *1101*
Fallon, E., *1918*
Fallon, E. F., 1184, *1189*
Fan, M., 277, *279*
Fan, Y., 490, 496, 508, *528, 529*
Fandel, G., *2623*
Fang, L., 128, *150*
Fanger, P. O., 1133, *1153*
Faraway, J., 1126, *1128*
Faraway, J. J., *1128*
Farbo, P. C., 915, *919*
Farebrother, R. W., *2293*
Farfan, H. F., 1076, *1102*
Fargher, H. E., 2044, *2052*
Farhangian, K., *2580*
Farhoodi, F., 1777, *1788*
Farmer, E. W., 2410, *2442*
Farr, J. L., 987, *992*
Farr, M. J., 930, 931, *942*
Farrel, L., *1101*
Fassnacht, M., *683*
Fathallah, F. A., *1105*
Faucett, J., 1224, *1232*
Fava, J., *540*
Fava, J. A., 537, *540*
Feather, J. J., 1705–1707, 1709, 1713, *1715*
Feather, N. T., 2204, *2217*
Federal Aviation Administration (FAA), 961, *971,* 1909, *1917*
Federal Register, 1070, *1102*

Federgruen, A., 1684, *1693*
Fedrizzi, M., 2186, *2217*
Feigenbaum, A. V., 1794, *1806, 2003*
Feigin, G., *1693*
Feigin, G. E., 1692, *1693*
Feitzinger, E., 689, *707*
Feldmann, K., 402, 404–406, 422, 435, 437, 439–441, 443, *445*
Felix, B., 228, *257*
Felten, D. F., 964, *973,* 1889, *1920*
Feng, T., *2441, 2444*
Ferguson, S. A., 1061, *1102, 1105*
Fernandez, J. E., *1108*
Fernie, J., 776, *785*
Ferreira, P. M., 1328, *1330*
Ferrell, O. C., 625, *632*
Ferrie, J., 1714, *1715*
Ferzacca, N., *1918*
Festa, M., 2394, *2405*
Fetter, R., *866*
Fiacco, A. V., 2530–2532, *2538,* 2551, 2555, 2560, *2565, 2567*
Fiedler, J. A., *2218*
Field, F. R., III, 537, *541*
Field, J. M., *894*
Fikes, R. E., *707*
Financial Risk Management, LTD., 759, *771*
Finch, P. D., 2266, *2293*
Fine, L. J., *989, 1100, 1104, 1107, 1129, 1155*
Finkelstein, J., *1101*
Finley, L., 955, *972*
Firby, R. J., 2435, *2442*
Firesmith, D. G., 124, *150*
Fischer, G., 1210, *1232*
Fischhoff, B., 2191, 2196, 2197, 2201, *2217, 2219–2221*
Fishbein, B. K., *540*
Fishburn, P. C., 2179, *2217,* 2605, *2621*
Fisher, C. D., 856, 858, 859, 861, 862, *865*
Fisher, J., 1701, 1708, 1710, *1715*
Fisher, K., *57*
Fisher, L., *1107,* 2335, *2351*
Fisher, L. D., *1100, 1101*
Fisher, M., 796, 803, 810, *823*
Fisher, M. D., *57*
Fisher, M. L., 808, *823,* 2113, *2139*
Fisher, R. A., 1421, *1462*
Fishman, G. S., 2485, *2494*
Fisk, R. P., 636, *650*
Fitzgerald, L., *1668*
Fitzgerald, M. P., *896*
Fitzsimmons, J. A., 623, *631*
Fitzsimmons, M. J., 623, *631*
Flamm, J., *1955*
Flannery, B. P., *1693*
Fleig, J., 214, *224*
Fleischer, G. A., 2360, *2393,* 2394, *2405*
Fleischmann, M., 538, *541*
Fleishman, E. A., *944,* 985, *990, 992*
Fleiss, J. L., *2240*
Fletcher, J. D., 928, 929, *942*
Fletcher, R., 2551, 2552, 2562, *2566, 2567*
Flin, R., 2208, *2217*
Flores, F., 143, *154*
Floudas, C. A., 2564, *2567*
Flowers, P., *399,* 985, *990*
Flowers, S., 949, 951, 952, 961, *971*
Flynn, F. M., 1226, *1232*
Fogan, T., *944*
Fogarty, K., 83, *108*
Fogarty, R. B., *2465*
Fogel, L. J., 875, *895*
Fogelman, M., 1202, *1232*
Foley, J., 2507, *2520*
Foley, J. P., *2220*
Foley, R. D., 1524, *1525*
Föllinger, H., *588*
Fong, G. T., 2201, *2218*
Fong, P., 602, *618*
Foo, Y. S., 1778, *1788*
Forcier, L., 1082, *1104*
Ford, Henry, 1620
Ford, J. K., 860, *865,* 933, *941*
Ford, L. R., 2574, *2580*
Ford, R. N., 874, *895*
Fordyce, W. E., *1101*
Forge, S., *258*
Forger, G., 2056, 2058, *2068*
Forkel, M., 213, *224*
Fornell, C., 623, 624, 629, *631*
Fortes, J. A. B., *618*
Fortin, C., 1050, *1102*
Foss, M. L., *1109*
Foster, F., Jr., 834, *836*
Foster, G., *154*
Foster, I., 229, *257*
Foster, J. W., *1583*
Foster, R. W., *993*
Fouad, R. H., *561*
Foulke, J. A., *1101, 1230, 1232*
Fourer, R., 2535, 2536, *2538*
Fournies, F. F., 938, *942*
Fowler, A., 902, *918*
Fowler, Jim, 352
Fox, B. L., *2494,* 2645, *2647*
Fox, J. G., 1136, 1150, *1153*
Fox, M., 1776, *1788,* 2035, 2044, *2053*
Fox, M. L., 911, *919*
Fox, M. S., 2014, *2019*
Fox, R., *2052*
Fox, W. M., 915, *918*
Fraiman, N. M., *1739*
Francett, B., 83, *108*
Francis, R., 1380, *1390*
Francis, R. L., *2580*
Francis, T., 2067, *2069*
Franke, J., 432, 433, *445*
Frankenhaeuser, M., 874, *895*
Franklin, G. F., 1885, *1885*
Franzblau, A., *1230*
Fraser, N. M., 128, *150*
Fraser, S. L., 923, *942*
Frazelle, E., *2109*
Frazelle, E. H., 1524, *1525, 1526,* 2093, *2108*
Freed, L., 228, *257*
Freeman, J. A., 163, *175*
Freer, M., *1106*
Frei, B., 183, *224*
Freidenfelds, J., *2405*
Freimer, M., 2611, *2621*

Freivalds, A., 1050, *1102, 1103,* 1410, 1418, 1419, 1423, 1424, 1426, 1436–1439, 1450, 1456–1459, 1461, *1462*
French, J., 1179, *1188*
French, J. R. P., *895*
French, J. R. P., Jr., *989*
French, S., *1740*
Frepoli, C., 207, *224*
Freund, R. J., *2293*
Frey, S. L., *1547*
Friedman, J. W., 2173, 2209, *2218*
Friedman, T. L., 1888, *1917*
Fries, B., 746, *749*
Frigo, C., *1101*
Frings, S., 642, *650*
Frisch, D., 2178, *2218*
Fritz, R., 15, *24*
Fröhlich, B., 2507, *2520*
Frohman, A. L., 1710, 1714, *1715*
Frost, P. J., 849, 851, *865*
Fruchtbaum, J., 1504, *1525*
Fruin, W. M., *560*
Frymoyer, J. W., 1071, *1102,* 1119, *1128*
Fujigaki, Y., 1222, 1223, *1230*
Fujii, M., 523, *528*
Fujimoto, M., *973*
Fujimoto, R. M., 2464, *2465*
Fujimoto, T., 545, 556, *559, 560,* 1288, *1295*
Fulford, L. A., *2442*
Fulkerson, D. R., 2574, *2580*
Fuller, B., *864*
Fuller, R., 163, *175*
Fulmer, R. M., 858, *867*
Fulton, R. E., *175*
Fung, W., 759, *771*
Funge, J., *1128*
Furlani, Cita, 352
Furman, J. P., *941*
Furner, S., *1106*
Furtado, G. P., 171, 172, *176*
Futrell, D., *897*

Gabathuler, J. P., *588*
Gablenz-Kolakovic, S., *1155*
Gad, S. C., *2240*
Gadd, K. W., 1713, *1715*
Gaddie, P., 1080, *1103, 1104, 1109*
Gadh, R., 2498, *2519*
Gadiesh, Orit, 848
Gaebler, T., *992*
Gael, S., 869, *895*
Gage, J. R., *1128*
Gagné, E. D., 931, *942*
Gagnon, M., 1072, *1102*
Gaidosch, T., *258*
Gal, T., 2621, *2621, 2623*
Gale, B. T., 623, *631*
Galinsky, T. L., *1236*
Gallagher, C. C., 870, 889, *895*
Gallessich, J., 938, *942*
Gallo, G., *2581*
Gallupe, R. B., 144, *150*
Gallwey, T. J., 1896, *1917*
Galt, S., *399*
Galvin, Paul, 860
Gamberale, F., 1071, *1102*
Gambino, B., *2220*
Ganesham, R., *2053*
Ganesharajah, T., 1525, *1525*
Ganz, W., 636, *650,* 1293, *1295*
Gappmaier, M., 1705, *1715*
Garcia, B., *823*
Garcia-Diaz, A., *2581*
Gardell, B., 1186, *1188*
Gardner, J. T., 2135–2138, *2139, 2140*
Garey, M., 794, *823*
Garey, M. R., 2594, 2595, *2600*
Garg, A., 981, *990,* 1071, 1072, 1076, 1087, 1088, *1102, 1105, 1107, 1108,* 1119, *1128, 1129, 1155*
Gargano, J., *991*
Garnett, J., 2452, *2465*
Gärtner, R., *588*
Garvin, A. D., 1246, *1251*
Garvin, D., *1807*
Garvin, D. A., 625, 626, *631,* 638, 639, *650*
Gass, S., 2528, 2538, *2538*
Gass, S. I., 126, *150*
Gasser, L., 965, *972*
Gattiker, U., 952, *973*
Gattorna, J., 785, *786*
Gaucher, E., 747, *749*
Gauld, S., *1104*
Gautrat, J., *1153*
Gawron, V. J., 2412, *2442*
Gay, D. M., *2538*
Gearhart, W. B., 2614, *2621*
Gebhardt, A., 191, *224*
Gebhart, A., *588*
GeeLee, W., *1104*
Geiger, M., *588*
Geiger, R., *588*
Gelatt, J. R., *2600*
Gen, M., *1790,* 2620, *2622*
Genaidy, A., 1043, 1053, *1102, 1108, 1109*
Genaidy, A. M., *1100, 1104, 1105, 1108*
Gendreau, M., *824*
General Accounting Office (GAO), 1697, 1701, *1715*
Geneste, L., 1781, *1788*
Geng, X., 278, *279*
Gentner, D., 1213, *1232*
Geoffrion, A. M., 808, 810, *823*
Georgakopoulos, D., 490, *528*
George, A., 2534, *2538*
George, A. D., 2464, *2465*
George, J. F., *2220*
Georgenson, D. L., 931, *942*
Gerard, H. B., 933, *942*
Gerard, M. J., 1201, *1232*
Gerhard, M., 438, *445*
Gershwin, S. B., 1668, *1668,* 2168, *2171*
Gersick, C. J. G., 982, 983, *990*
Gerson, J., 1201, *1230, 1235*
Gerstein, M. S., *1010*
Gertman, D., 1941, *1954*
Gertman, D. I., 2189, 2191, 2192, *2218*
Gery, G. J., 940, *942*
Gescheider, D. A., *1108*
Geweke, J., *2393*

Ghaleb-Harter, T. E., 759, 760, 765, 766, 768, 770, *771*
Ghedira, K., *1788*
Ghiselli, R., 826, *836*
Ghosh, S., *1807*
Giannakourou, M., *1038*
Gibbons, T. C., 852, 854, *864*
Gibbs, W., 949, *971*
Gibson, A., *109*
Gibson, J. J., 1014, *1037*
Gibson, J. L., 846, *865*
Gibson, R. R., 2461, *2465*
Giddens, A., 952, *971*
Gieles, I., 1904, *1917*
Gigante, M., *2520*
Gilad, I., *1407, 1408*
Gilbert, R., *1102*
Gilbert, T. F., 925, *942*
Gilbreth, F. B., 739, *749,* 871, 874, *895*
Gilbreth, Frank, 20
Gilbreth, L. M., 739, *749*
Gilbreth, Lillian, 20
Gill, J. M., *1106*
Gill, P. E., 2530, *2538, 2567*
Gill, Philip, 2564, 2565
Gillet, B., 803, *823*
Gillies, R. R., *989*
Gimpel, R. J., *2392*
Ginnett, R. C., *865*
Ginsberg, B. C., *353*
Ginsberg, E., 623, *631*
Ginsburg, D., 1284, *1295*
Gissler, A., *226*
Githens, P. B., *1104*
Giunipero, L. C., 2113, 2120, 2133, 2134, *2139, 2140*
Givoni, B., 1139, *1153*
Gjessing, C. C., *989*
Gladstein, D. L., 877, *895,* 933, *942,* 984, 987, *990*
Glantschnig, W. J., *599*
Glanz, F. H., *1789*
Glaser, R. B., 926, *942*
Glasserman, P., 1669, 1690, 1692, *1693,* 2633, *2647*
Glassey, C. R., 538, *541*
Glatz, R., 193, *224*
Gleick, J., *57*
Glenn, F. A., *1127*
Globerson, S., 847, *865,* 886, *895, 1251, 1278*
Glover, F., 800, *823,* 1731, *1739,* 1752, *1766,* 2447, *2465,* 2573, *2580,* 2590, 2591, *2600*
Glynn, P. W., 2492, *2494,* 2634, 2635, *2647*
Goble, J., 2459, *2465*
Göbler, T., 213, *224*
Goddin, D., 745, *750*
Godfrey, A. B., 984, 986, 987, *990*
Gödicke, P., 213, *224*
Goel, S., 1360, *1389*
Goetschalckx, M., 2093, *2108*
Goetz, W. G., 1525, *1525*
Goff, D. R., *2003*
Goggin, J. E., *1102*
Gogoll, A., 640, *650*
Goldbaum, L., 95, *108*
Goldberg, A., 1328, *1330*
Goldberg, A. J., 740, *749*
Goldberg, D., 1780, 1781, *1788*
Goldberg, D. E., 2591, *2600*
Gold-Bernstein, B, 341, *352*
Golden, A. L., 1063, 1070, *1104*
Golden, B., 2061, *2068*
Goldfarb, D., 2552, *2566,* 2575, *2580*
Goldhar, J. D., 955, *973*
Goldman, G. L., 527, *528*
Goldman, R. F., 1139, *1153*
Goldman, S. L., 955, *971*
Goldman Sachs & Co., 759, *771*
Goldratt, E., 2035, *2052*
Goldratt, E. M., 557, *560*
Goldsman, D., 2482, 2488, *2494*
Goldsman, D. M., *2239*
Goldsmith, F., *57*
Goldstein, I. L., 926, 934, *942*
Goldstein, W. M., 2201, 2204, 2206, *2218*
Golembiewski, R. T., 981, *992*
Golhar, D. Y., 545, *560, 1806*
Golm, F., 205, *224*
Gomberg, W., 908, *918*
Gomez-Mejia, L. R., 862, *865,* 915, *918*
Goncalves, M., 229, *257*
Gonzales, F. G., 2464, *2465*
Gonzalez, M., *2216*
Goodall, S., *1787*
Goodman, P. S., 870, 877, 880, *895,* 933, *942*
Goodstein, L., *1039*
Goodstein, L. P., *152*
Goonetilleke, R. S., 1912, *1917*
Göpel, W., *399*
Gordon, C. C., 1113, *1128*
Gordon, J., 2187, *2218*
Gordon, J. R., *1583*
Gordon, R. D., 1922, *1954*
Gordon, W. J., 1656, *1668*
Gore, B., 2434, *2444*
Gore, B. F., 2440, *2442*
Gorman, W. M., 2605, *2621*
Gormley, J. T., *353*
Gotlieb, C. C., 1225, *1231*
Gottlieb, H. S., *1235*
Gottlieb, W., *2520*
Gou, L., 697, *707*
Gould, J. D., 1207, *1232, 1919*
Gould, Nick, 2564
Gould, R. W., *1918*
Govaert, G., *1108*
Govindarajan, V., *149*
Gowan, M. A., *919*
Gower, J., 774
Goyal, A. K., *2393*
Goyal, S., 499, *528*
Grabot, B., 1781, *1788*
Grabowski, H., 191–193, *224*
Grady, J. F., *992*
Graeber, R. C., 963, *971*
Graedel, T. E., 530, 533, 537, *541,* 598, 599, *600*
Graham, D., *257*
Graham, R., 1723, *1739*
Graham, R. J., *1262*

Graham, S., 914, *918*
Gramopadhye, A., *1152, 1917*
Gramopadhye, A. K., 1896, 1897, *1917*
Grams, R., 916, *918, 919*
Granda, R. E., 532, *541*
Grandjean, E., 871, 876, *896,* 1061, *1102, 1104,* 1120, *1128,* 1193–1195, 1197, 1204, *1232, 1234*
Grant, E. L., *1855,* 1861, 1876, *1876, 2003*
Grant, K. R., *152, 707*
Grant, R. M., *1807*
Grantham, C., *57*
Grassia, J., *2218*
Grätz, J.-F., 181, *224*
Grauer, M., *2623*
Graves, R., *617*
Graves, S. C., 1524, *1526,* 2044, *2052,* 2105, *2108*
Gray, P., 135, *150,* 2538, *2539*
Gray, W. D., 1209, *1232,* 2410, *2442*
Grayson, C., 950, *971*
Grayson, K. A., *632*
Greco, E. C., *1100*
Green, D. M., 1016, *1037*
Green, G. C., 559, *560*
Green, H., 671, *683*
Green, P. E., 702, 703, *707*
Greenberg, H. J., 537, *541*
Greenberg, Harvey J., 2563
Greene, R. T., 1798, *1806*
Greening, C. P., 1895, *1917*
Greenspan, A., 344, 345, 347, *352*
Greenwood, D. C., *1330*
Greenwood, E., 488, *528*
Greey, G., 1056, *1103*
Greif, I., 1222, *1232*
Greiner, B., *1153*
Grey, S., *1037,* 1378, *1389*
Greyser, S. A., 704, *707*
Grieco, A., *1101*
Grieve, D., 1043, 1047, *1102*
Griffin, R. W., 874, 893, *896,* 979, *990, 994*
Griffith, J., 747, *749*
Griffith, L. E., 534, *541*
Griffith, R. E., *2566*
Grigoriadis, M. D., 2575, *2580*
Grobelny, J., 1050, *1108*
Groenevelt, H., 545, *560*
Gronbaek, K., 964, *970*
Grönroos, C., 624, 626, 628, *631, 632*
Grossman, K., *1106*
Groth, K. M., 1145, *1153*
Groupement des Entreprises Sidérurgiques et Minières (GESIM), 1145, *1153*
Grove, D. M., *2240*
Grover, V., 1703, 1704, 1714, *1715, 1716, 1806*
Grudin, J., 133, 134, *151*
Gruenfeld, D. H., 2212, *2218*
Grundeman, R. W., *993*
Gryna, F. M., *2003,* 2228, 2229, *2239*
Guélaud, F., 1145, *1153*
Guengerich, S., 229, *257*
Guertin, F., *824*
Guest, R. H., *898*
Guggenbuhl, U., 1201, *1232*
Guha, S., 1702, 1705, 1706, *1715, 1716*
Guide, V. D. R., Jr., 538, *541*
Guimaraes, T., 1700, 1710, 1711, *1716*
Guion, R. M., 921, 922, 924, *942, 2216*
Gulati, R., 1561, *1582*
Gumbel, E. L., 1932, *1954*
Gunasekaran, A., *1716*
Gunatilake, P., *1919*
Gungor, A., 440, 443, *445*
Gunst, R. F., *2240*
Gunther, R., 955, *970*
Gupta, M. C., 1777, *1788*
Gupta, N., 911, *918*
Gupta, S., 538, *541,* 679, *683*
Gupta, S. M., 440, 443, *445, 446*
Gupta, S. S., 2488, *2494*
Gupta, V. K., 538, *541*
Gupta, Y. P., 499, *528, 1788*
Gustafson, D. H., *918,* 984–986, *990, 2217*
Güth, W., 2210, *2218*
Gutierrez, S., 536, *541*
Gutman, E. G., *631, 633*
Guttman, H., 2192, *2222*
Guttman, I., *2293*
Guzzo, R. A., 855, *865,* 870, 877, *896,* 984, 986, 987, *990*
Gyi, D. E., 1120, *1129*
Gyllenhammar, P. G., *1190*

Ha, Y. W., 2198, *2219*
Haaland, P. D., *2240*
Habes, D. J., 1061, *1102*
Hackathorn, R. D., 83, *108*
Hacker, W., 1145, *1153*
Hackett, E. J., *992*
Hackett, R. D., *864*
Hackman, J. R., 871, 874, 880, 887–889, *896, 897, 943,* 977, 979, 982, 987, *990, 991,* 1899, 1900, *1917*
Hackman, R. J., 1794, *1806*
Hackman, S., 2093, *2108*
Hackman, S. T., 2044, *2052,* 2093, *2108*
Hackos, J. T., 1206, 1207, 1210, 1212–1216, *1232*
Hadavi, K., 1733, *1739,* 2044, *2052*
Haderspeck, K., *1107*
Haeckel, S. H., *353*
Hagberg, M., 1061, *1100, 1102,* 1202, *1233*
Hagel, J., 147, *151*
Hagel, J., III, 327, *352*
Hägg, G., 1086, *1102*
Hagglund, G., *1188*
Hagopian, J. H., 1176, *1188*
Hahler, B., 2410, *2442*
Hahn-Woernle, C., 2103, *2109*
Haimes, Y. Y., 2620, 2621, *2621–2623*
Haims, M. C., 981, *991,* 1229, *1231, 1233*
Haines, H. M., 980, 981, *994,* 1229, *1236*
Haischer, M., 647, *650*
Håkansson, H., 2118, 2123, *2139*
Halal, W. E., 148, *151*
Hale, J., 745, *749*
Hale, T. S., *1526*

Halevi, G., *483*
Hall, A. D., 127, *151*
Hall, G., 954, *971*
Hall, J., *942, 947*
Hall, J. K., 934, *942*
Hall, M., 78, *108*
Hall, N. G., *1525*
Hall, R., 803, *823*
Hall, R. E., *1583*
Hall, R. W., *528,* 819, 821, *823,* 2062, *2069,* 2146, 2157, *2171*
Hall, W. A., 2620, *2621*
Hallberg, B., 228, *257*
Hamill, L. S., *990*
Hamilton, D. B., 2434, *2442*
Hamilton, V., 2208, *2218*
Hamm, R. M., *2218*
Hammer, E. G., 922, *942*
Hammer, M., 217, *224, 353, 528,* 747, *749,* 1696, 1700, 1706, 1709, 1714, *1716,* 2049, *2052,* 2126, 2132, *2139*
Hammer, W., 2189, *2218*
Hammond, J., *399*
Hammond, J. S., 129, *151*
Hammond, K. R., 129, *151,* 2195, 2196, 2200, 2205, 2209, 2214, *2218*
Hammond, R. W., 874, *896*
Han, M. H., 1524, *1526*
Han, S. P., 2562, *2566*
Hancock, W. M., 1889, *1917*
Handelsmann, R., *2109*
Handfield, R., *1807,* 2113, 2116, *2139*
Handley, T. R., *152*
Haney, L., 1909, *1917*
Hann, R. I., *2444*
Hannam, R., 229, *257*
Hannam, R. G., 501, 507, *529*
Hanne, T., 2621, *2621*
Hänninen, K., *1106*
Hannum, W. J., *941*
Hänsel, M., *588*
Hansen, K. S., *2003*
Hansen, P., *2623*
Hanson, M. L., *944*
Hansson, T., 1071, *1103*
Hansson, T. H., *1100, 1101*
Happ, A., *1235*
Harbeson, M. M., *943*
Hardin, G., 2210, *2218*
Harding, F. D., 915, *918*
Hardy, G. L., *1919*
Harel, D., 1774, *1788*
Harker, P. T., *707,* 2074, *2081*
Harkey, D., *736*
Harkins, S., *896*
Harkins, S. G., 877, 880, *896*
Harless, J. H., 925, *942*
Harmon, J., 2212, *2218*
Harmon, R. L., *1583*
Harrell, C. R., 2447, 2459, 2461, *2465*
Harrington, H., 2449, *2467*
Harrington, H. J., *57,* 1700, 1710, 1714, *1716*
Harrington, J., 485, *528*
Harris, C. M., 126, *150*
Harris, D. H., 1897, *1917*
Harrison, A. W., 988, *989*
Harrison, C. L., *543*
Harrison, E. L., *991*
Harrison, H. R., 134, *151*
Harrison, J. M., 1655, 1656, *1668*
Harrison, M., 133, *151*
Harrison, R. A., *1955*
Harry, M., 1368, *1390*
Hart, D. M., *1152*
Hart, J. C., *2519*
Hart, K., *2217*
Hart, P., 2212, *2218*
Hart, S. L., *895*
Harten, A. V., *541*
Hartley, R., 2615, *2621*
Hartline, M. D., 625, *632*
Hartman, Mark, 539
Hartmann, H. I., 916, *919*
Hartung, J., 206, *224*
Hasegawa, Hideyoshi, 352
Hasegawa, T., *707*
Haslam, D. R., *1109*
Hassan, S., *1129*
Hasselquist, R. J., 1131, *1153*
Hastie, R., 2205, 2207, *2221*
Hater, J. J., 849, 851, *865*
Hatvany, J., 697, *707*
Haug, E. J., 2498, 2515, *2520*
Haugen, E. B., 1940, *1954*
Haugh, L. D., *1101*
Haurie, A., 2645, *2647*
Hauschild, M., *542*
Hauser, J. R., 703, *709*
Hauser, N., 745, *749*
Hausman, W. H., 1524, *1526, 2108*
Hauss, I., *1295*
Hawker, J., 1772, *1788*
Hax, A. C., *823, 2081, 2538, 2581*
Hay, E. N., *918*
Hay Associates, 908, *918*
Hayes, R. H., 2318, *2329*
Hayes-Roth, B., 1024, *1038*
Hayes-Roth, F., 1328, *1330*
Hays, R. T., 932, 933, *942, 943*
Hazen, G. B., *2623*
He, Y., *1918*
Head, T. C., 978, 979, *991*
Headley, D., *2441, 2444*
Healy, A. F., *945*
Healy, M. J. R., 2266, 2273, *2293*
Heap, H. F., 1618, *1623*
Hearne, S., *541*
Heath, C., 2204, *2218*
Heath, L., 2199, 2203, 2212, *2218*
Hecht, J., *2003*
Heckmann, A., 201, *226*
Heflin, D. L., 2447, 2459, *2465*
Hegge, H. M. H., 694–696, *707*
Heinze, R., *588*
Heisig, P., 214, 215, 217, *223–225*
Heislitz, F., *588*
Helander, M., 705, *707,* 1016, *1038,* 1373, *1389*

Helander, M. G., 706, *707*, 1203, *1233*
Held, M., 1726, *1739*
Helgason, R. L., 2575, *2580*
Helmreich, R. L., 957, 960, 961, *971*
Helms, K., *618*
Hemyari, D., *588*
Henderson, W. B., 1752, *1766*
Henderson-King, E., *2218*
Hendrick, H. W., *1233*
Hendricks, K. B., *1807*
Hendriksen, G., 932, *943*
Hendry, L. C., 2035, *2052*
Henin, C., *2393*
Henkoff, R., 623, *632*
Henley, E. J., 1936, 1937, *1954*
Henry, R. A., 2193, *2222*
Henry Dreyfuss Associates, 1043, 1048, *1103*
Henz, J., *618*
Herek, G. M., 139, *151*
Herfurth, H.-J., *225*
Herfurth, K., *588*
Herko, R. G., 130, *150*
Herman, R., 532, 535, *541*
Hermann, F., *541*
Herrin, G. D., *1101, 1102, 1108, 1109, 1128*
Herrmann, C., 443, *445*
Hertz, D. B., *2393*
Hertzberg, H. T. H., 876, *896*
Herzberg, F., 871, 874, *896*, 926, *943*
Herzum, Peter, 352
Heskett, J. L., 623, *632*, 641, 643, *650*, 956, *972*, 2092, *2109*
Heslop, B., 229, *257*
Hess, J. L., *2240*
Hess, T., 218, *224*
Hesse, J., *399*
Hesse, S., *399*
Hesselbach, J., 443, *445*
Hesselbein, F., *57*
Hestenes, M. R., 2561, *2566*
Hester, K., *864*
Hewer, N. D., *399*
Hewitt, F., 2123, 2127, *2139*
Hext, G. R., *2567*
Heyman, D. P., 538, *541*, 2643, *2647*
Hiam, A., 1793, *1806*
Hicks, C. R., 2228, 2232, *2239*
Hicks, D., *1583*
Hicks, D. T., 2318, *2329*
Hickson, D. J., 960, *972*
Hidalgo, J., *1108, 1109*
Higgins, C. D., *1188*
Higgins, L. R., 1550, *1582*, 1610, 1612, *1622*
Higgs, A. C., *894, 989*
Higuchi, K., *400*
Hildebrandt, V. H., 1070, *1103*
Hilgard, E. R., *943*
Hill, T., 2072, *2081*
Hillier, F. S., 129, *151, 2393, 2567, 2580*
Hills, F. S., 912, *918*
Hills, M., 228, *257*
Hilton, R. W., 2319, *2329*
Hilweg, D., 1101, *1108*
Himsworth, F. R., *2567*
Hinderer, K., 2636, 2645, *2647*
Hindle, J., *1106*
Hines, W. M., 2258–2261
Hines, W. W., 2375, *2392*
Hipel, K. W., 128, *150*
Hiriart-Urruty, J. B., 2636, *2647*
Hirsch, B., 603, 604, *617*
Hirschvogel, M., 582, *587*
Hirtle, S. C., 1213, *1233*
Hitch, G. J., 2434, 2435, *2442*
Hitomi, K., *541*
Hix, D., 134, *151*
Hlupic, V., 2449, *2465, 2466*
Ho, J. L., 127, *151*
Ho, S. J. K., 1705, 1707, 1709, *1716*
Ho, Y. C., 2633, *2647*
Hoagland, D., 2413, *2442*
Hoban, C. F., 928, *943*
Hochberg, M., 1061, *1104*
Hochberg, Y., 2490, *2494*
Hochman, N., *771*
Hochwater, W. A., *989*
Hockey, G. R. J., 2208, 2209, *2220*
Hocking, R. R., 2276, *2293*
Hodges, R., 1775, *1788*
Hodgins, J. K., 1126, *1128*
Hoehn, R., *541*
Hoerl, A. E., 2290, *2293*
Hof, R. D., 672, *683*
Hoffer, J. A., *109*
Hoffman, J., *895*
Hofmann, H., 635, 642, *650*
Hofstede, G., 881, *896*, 957, 958, *971*
Hogarth, R. M., 1023, *1037*, 2196, 2197, 2199, 2201, 2204, 2206, *2217, 2218*
Hogg, R. V., 2243, 2248, 2254, *2263*
Hohnston, J. H., *1039*
Hokel, T. A., 303, *307*
Holder, G. W., 862, *865*
Holding, D. H., *1917*
Holdreith, J. M., *540*
Holford, R. R., *1104*
Hollan, J., *1038*
Holland, J. H., 2591, *2600*
Holland, M., 193, *224*
Hollander, M., 2256, *2263*
Hollenbeck, J. R., 870, 877, *896*, 922, *943*
Hollnagel, E., 1024, 1032, *1040*, 1894, 1909, *1917, 1918*, 2436, *2443*
Holsapple, C., 67, 107, *107–109*
Holsapple, C. W., 67, 70, 107, *108*, 214, *224*
Holtzblatt, K., 964, *970*
Holtzman, S., 2191, *2218*
Holyoak, K J., 2200, 2208, *2216*
Holzmann, P., 1061, *1103*
Homem-de-Mello, T., 2636, *2648*
Hommel, G., *618*
Hooke, R., 2549, *2566*
Hoover, D. J., 915, *919*
Hopcroft, J. E., *822*
Hopkins, B., *1190*
Hopkins, B. L., 1182, *1188*
Hopp, W., 2037, *2052*
Hopp, W. J., 2168, *2171*

Hoppock, R., 874, *896*
Horkeby, I., 531, *541*
Hornick, C. W., *991*
Hornick, M., *528*
Horton, R. L., *2293*
Hoshino, T., 538, *541*
Hou, T.-S., 1905, 1906, 1913, 1914, *1918*
Houck, E. C., *2494*
Hough, L. M., *944*
House, R. J., 842, 845, 846, 853–855, 858, *865–867*
Hout, T. M., *1010*
Houtzeel, A., *483*
Hovden, J., 2202, *2221*
Hovey, P. W., 1909, *1918*
Howard, A., 939, *943*
Howard, M. T., *707*
Howard, R. A., 2187, 2190, 2191, *2218, 2393*
Howell, J. M., 843, 849, 851, 854, *865, 867*
Hseih, D. A., 759, *771*
Hsiang, T., *529*
Hsiao, H., *1109*
Hsu, J. C., 2494, *2494*
Hsu, W., *2466*
Huang, C. Y., 155, 170, 171, 175, *176,* 603, 606, 612, *617, 618*
Huang, H., 1777, *1788*
Huang, S., *2466*
Hubbard, T. N., *353*
Huber, G. P., 135, *151*
Huber, J., 2194, *2218*
Huber, V. L., 913, *918*
Hübner, G., 2645, *2647*
Hudock, B., *1501*
Hudson, J. A., *1129*
Huemer, K. H., *1233*
Huffman, J. R., 2095, *2109*
Hughes, E. F., *989*
Hughes, E. F. X., *993*
Hughes, R. L., 844, *865*
Huitema, C., 229, *257*
Hukill, E., 984, 987, *993*
Hulin, C. L., 886, *896*
Hull, F., *994*
Hull, R., 1696, 1700, *1715*
Hullinger, D. G., 2460, *2466*
Humphrey, W. S., 648, *650*
Humphreys, K., *540*
Hundt, A. S., *990*
Hung, R., 1763, *1766*
Hung, Y. Y., *1918*
Hung, Y-F., 2044, *2052*
Hunsaker, P. L., *970*
Hunsicker, P. A., 1056, *1103*
Hunt, R., *540*
Hunt, R. G., *543*
Hunter, B., 745, *749*
Hunter, J., *945*
Hunter, J. E., 921, *943*
Hunter, J. S., 2225, 2226, 2229, *2239*
Hunter, W. G., *2239,* 2273, *2293*
Hunting, W., *1104*
Hurley, J. R., *109*
Hurlock, R. E., 931, *943*
Hurrell, J. J., Jr., *991, 992*
Hurth, P., *151*
Husband, T. M., 910, *919*
Hüser, M., *323*
Huson, A., *1102*
Hussey, J. R., 2491, *2494*
Huston, R., *1108, 1109*
Hutcheson, T. D., *946*
Hutchins, C., *2443*
Hutchins, E., 1018, 1026, *1038*
Huy, Q. H., 846, *867*
Hwang, C., 2074, *2081*
Hwang, C. L., *1109,* 2606, 2608, 2614, *2622*
Hwang, H., 1524, 1525, *1526*
Hwang, S., 131, 145, *151*
Hyer, N. L., 889, *896*
Hyman, J., 981, *991*
Hyvärinen, T., *1918*

Iacobucci, D., 623, 628, 629, *632, 633*
Iacocca, Lee, 848
Iansiti, M., 952–954, 963, *971*
Ibbotson, R., 752, *771*
Ida, K., 2620, *2622*
Iglehart, D. L., 2488, *2494*
Ignall, E., 1728, *1739*
Ignizio, J. P., 2614, *2622*
Ilg, R., 1202, *1233*
Ilgen, D. R., *896,* 933, *943*
Illuminating Engineering Society, *1153*
Imada, A., 980, *991*
Imai, M., 551, 553, *560*
Iman, R. L., *540*
Imrhan, S. N., 1055, 1056, *1103*
Inderfurth, K., 538, *541*
Indik, B. P., 933, *943*
Industrial Modeling Corporation, 2460, *2466*
Informationscentrum des Deutschen Schraubenverbandes (ICS), 410, *445*
Information Week, 950, *971*
Ingle, S., *991*
Inman, R. A., 559, *560*
INPO, 1030, *1038*
Institute for Interconnecting and Packaging Electronic Circuits (IPC), 423, *445*
Institute of Industrial Engineers, 2446, 2461, *2466*
Instrument Society of America (ISA), *1885*
IntelliQuest, 702, 703, *707*
International Atomic Energy Agency (IAEA), 960, *971*
International Civil Aviation Organization (ICAO), 1135, *1153*
International Labour Office (ILO), 1394–1400, *1408,* 1426, *1462*
International Labour Organization (ILO), 1157, 1165, *1188*
International Organization for Standardization (ISO), *1153, 1974*
Inuiguchi, M., *2622*
Iovine, J., 2503, 2508, *2520*
Irani, R. K., 494, *529*
Irion, A. L., 930, *944*
Irish, B., *1789*

Irvine, C. H., *1107*
Isaacs, J. A., 538, *541*
Isen, A. M., 2206, *2219*
Isenberg, D. J., 881, *896*, 2212, *2219*
Ishii, K., 689, *707*
Ishikawa, K., 552, *560*, *1827*, 1857, 1859, *1876*
ISO 10303-1, 204, *224*
Isobe, T., *400*
IT logistiek, Logistiek Krant, and Berenschot, 2103, *2109*
Ito, K., *400*
Ivancevich, J. M., *865*
Iverson, G., 2204, *2219*
Iverson, R. D., 1160, 1161, *1188*
Iwanowa, A., *1153*
Iwasaki, R., *528*
Iwata, H., *2520*
Iyer, A. V., *2032*
Izui, C., *989*

Jablin, F., 145, *151*
Jablin, F. M., 174, *176*
Jablonski, S., 204, *224*
Jackson, J. M., 956, *971*
Jackson, J. R., 1650, *1668*, 1723, *1739*, 2164, *2171*
Jackson, Ric, 352
Jackson, S. E., 855, *865*
Jacobs, J., 602, *618*
Jacobs, J. W., 929, *943*
Jacobson, A., *151*
Jacobson, I., 124, *151*, 490, 507, *529*
Jacobson, S. H., 2492, *2494*
Jacoby, J., 2200, *2219*
Jacoby, R., *2520*
Jaffe, D. T., 2126, *2140*
Jäger, M., 1072, 1076, *1103*, *1109*
Jaikumar, R., 796, 803, *823*
Jaillet, P., 801, 802, *822*, *823*
Jain, S., *2466*
Jakimcius, A., *1920*
Jamaldin, B., *1109*
Jambekar, A. B., 953, 954, *971*
James, C., *991*, *1236*
James, D. S., *2220*
James, J. M., *1039*
James, L. R., *991*
Jamieson, G. H., 1899, *1918*
Jang, R., *1104*, *1109*
Jani, Arpan, 352
Janik, A., *154*
Janis, I. J., *151*
Janis, I. L., 140, *151*, 877, 881, *896*, 983, *991*, 2176, 2184, 2193, 2206, 2212, *2219*
Jansen, H., *225*
Janssen, T. J., 139, *151*
Japanese Industrial Standards Committee (JIS), *1806*
Japan Society for the Promotion of Machine Industry, *483*
Jaques, E., 910, *918*
Jarrar, Y. F., 1697, *1716*
Jarrard, S. W., 931, *943*
Jarrell, S. L., *1806*
Jarvenpaa, S. L., *224*
Jaschinski, C., *650*
Jaschinski, C. M., 645, *650*
Jayaram, J., 2112, *2138*
Jayaraman, V., 538, *541*
Jeanneret, P. R., *1154*
Jeeves, T. A., 2549, *2566*
Jeffrey, T., *2240*
Jeffries, R., 1209, 1210, *1233*
Jenkins, G. D., *918*
Jennings, K. R., 979, *991*
Jensen, A. R., 921, *943*
Jentsch, F., 963, *971*
Jetly, N., 88, 92, 94, *108*
Jha, N. K., 490, *529*
Jiang, B. C., *1109*
Jiao, J., 685, 687, *707*, *708*
Jobe, J. M., 1889, 1891, *1920*
Jochem, R., 218, *225*, *226*
Johan, S., *2519*
Johannes, J. D., *2316*
Johannessen, T. B., *1807*
Johansen, R., 134, 142, 143, *151*, 2214, *2219*
Johansson, G., 874, *896*, 1221, 1222, *1230*, *1233*
John, B. E., *1232*, *2442*
John, W. M., 2225, *2239*
Johnson, B., 952, 962, *971*
Johnson, D., 794, *823*
Johnson, D. S., 2594, 2595, *2600*
Johnson, E. J., *2219*, *2221*, *2443*
Johnson, E. L., *2600*
Johnson, E. M., 1023, *1038*
Johnson, G., 134, *153*
Johnson, G. R., *1106*
Johnson, H. T., 953, *971*, 2318, *2329*
Johnson, J. K., 532, *541*
Johnson, L. A., *2053*
Johnson, M., 538, *541*
Johnson, M. C., 857, *865*
Johnson, N. E., 985, 986, *991*
Johnson, P. W., *1230*
Johnson, S., 1353–1359, 1366, 1367, 1369, 1372, 1375–1386, *1389*, 1391, 1395, 1401–1405, 1407, *1408*
Johnson, W. B., 1895, *1918*
Johnson, W. L., 1112, 1127, *1128*
Johnson, W. R., *991*
Johnston, A. N., 957, *971*
Johnston, J. C., *2443*, *2444*
Johnston, R., *1668*
Johnston, W. A., 934, *941*
Joint Commission Accreditation for Healthcare Organizations (JCAHO), *991*
Jones, A., 1769, 1778, *1788*, *1789*
Jones, A. P., 855, *867*, *991*
Jones, A. T., 1779, *1790*
Jones, Al, 352
Jones, B., *540*
Jones, B. W., *2405*
Jones, C., *353*
Jones, D., *543*
Jones, D. M., 1133, *1153*, *2442*
Jones, D. T., 5, 8, *24*, *561*

Jones, H., *2520*
Jones, J. C., 1912, *1918*
Jones, J. P., 680, *683*
Jones, L. V., 2182, *2216*
Jones, M. B., *943*
Jones, R., *399*
Jones, R. V., *1886*
Jones, S. W., *483*
Jones, W. W., 914, *919*
Jonsson, B., 1061, *1100, 1103, 1109*
Jordan, C. S., *2442*
Jorgensen, C. C., *2442*
Joseph, B. S., 981, *991*
Joshi, K. D., *108*
Journal of the Society for Health Systems, 745, 747, 748, *749*
Joyce, C. R. B., 2200, *2216*
Judd, C. H., 932, *943*
Judt, C., *1788*
Juel, C., *1129*
Juengel, C., *707*
Jung, E. S., 1050, *1103*
Juran, D., 985–987, *991*
Juran, J. M., 626, *632,* 1794, *1806, 1807, 2003,* 2228, 2229, *2239*
Juran, Joseph, 747

Kaas, H.-W., *650*
Kachhal, S., 745, 747, *749*
Kachhal, S. K., 745, 746, *749*
Kacmar, K. M., *864*
Kacprzyk, J., *1955, 2217*
Kafka-Lutzow, A., *1233*
Kahai, S. S., *866*
Kahmeyer, M., *398*
Kahn, Bob, 238
Kahn, H., 1226, *1233*
Kahn, H. D., 1455, 1456, *1462*
Kahn, J. A., *895*
Kahn, J. F., 1057–1059, *1103*
Kahn, K., *154*
Kahneman, D., 953, *971,* 1023, *1039,* 2173, 2196–2199, 2202, 2203, 2214, *2219, 2222, 2444*
Kalakota, R., 229, *257*
Kalantari, B., *918*
Kalawasky, R., *2520*
Kalb, B., 949, *971*
Kalimo, R., 1170, *1188*
Kall, P., 2630, *2647*
Kalpakjian, S., 450, 467, *483*
Kälviäinen, H., *1918*
Kamada, A., *588*
Kamal Zafar, S., 1572, *1582*
Kamauff, J., *2140*
Kaminsky, P., *109,* 2019, *2019*
Kamwendo, K., 1224, *1234*
Kanet, J., 1735, 1736, *1739*
Kanin-Lovers, J., 901, *918*
Kanje, M., *1104*
Kano, N., 978, *991*
Kansi, P., *1103, 1128*
Kanter, E. M., 1888, *1918*
Kanter, R. M., 981–983, 988, *991, 1010*
Kanungo, R. N., 846–849, 854, 855, 859, *864, 865*
Kanz, J., 953, 955, *971*
Kaplan, P., 752, *771*
Kaplan, R., 997, 998, *1010*
Kaplan, R. P., *1189*
Kaplan, R. S., 21, *24,* 321, *322,* 645, *650,* 953, *971,* 2318, 2319, *2329, 2330*
Kapur, K. C., 1922, 1925, 1932, 1937, 1940, 1945, *1954, 1955*
Kar, K., *541*
Karasek, R., 980, *991*
Karasek, R. A., 874, *896*
Karat, J., 1206, 1216, 1219, *1233*
Karetta, B., *1189, 1233*
Karhu, O., 1061, *1103,* 1117, *1128*
Karlqvist, L., 1202, *1233*
Karlsson, J., *399*
Karlsson, R., *542*
Karltun, J., 1185, *1188*
Karmarkar, N., *2539,* 2543, *2566*
Karmarkar, U., 551, *560*
Karmarkar, U. S., 2037, 2044, *2052*
Karp, R. M., 1726, *1739*
Karsh, B., 1226, *1231, 1236*
Karsh, B.-T., *1190*
Karush, W., 2543, 2553, 2554, 2556, *2566*
Karwan, M., *1916*
Karwan, M. H., *1919*
Karwowski, W., 1042, 1043, 1050, 1053, 1064, 1068, 1070, 1071, 1080, 1082, 1085, 1086, *1100–1105, 1108, 1109,* 1229, *1233,* 1898, *1918*
Kashyap, B. R. K., *1693*
Katircioglu, K., *1693*
Kato, K., *2622*
Katsigris, C., 833, *836*
Katz, H. C., 877, *896, 992*
Katz, R., 146, *151*
Katz, Z., *399*
Katzel, J., 1566, *1582, 1583*
Katzenbach, J. R., 975, 978, *991*
Katzenbach, R. J., 1250, *1251*
Kaufman, L., *2442*
Kaufman, S., *943*
Kaun, R., *399*
Kawaguchi, T., 1723, *1739*
Kawalik, J. S., *2581*
Kay, A., 1213, *1233*
Kazarian, E. D., 826, 827, 829, 830, 833, *836*
Kazazi, A., 545, *560*
Kearsley, G., 1226, *1233*
Kececioglu, D., 1925, 1932, 1940, 1945, 1946, *1955*
Kee, D., 1064, *1104, 1109*
Keefe, J. H., *896*
Keen, P. G. W., 136, *151*
Keeney, R. L., 129, *151,* 2173, 2183, 2187, 2189, 2194, 2195, 2209, 2214, *2219,* 2605, *2622*
Kees, J. R., 2105, *2109*
Keiningham, T. L., *632*
Keithly, D., *1583*
Kelleher, Herb, 849

Keller, A. Z., 545, *560, 561*
Keller, B., *750*
Keller, E., *1107*
Keller, G., 490, *529*
Keller, L. R., 127, *151*
Keller, P., 2464, *2466*
Keller, R. T., 851, *865, 897, 993*
Kellerman, S. A., 229, *257*
Kelley, A., *1583*
Kelley, M. R., 952, 961, *971*
Kelley, R., *1920*
Kelley, R. E., 854, *865*
Kelley, T., *2441, 2442, 2444*
Kelling, C., 607, *618*
Kelloway, K. E., *864*
Kelly, F. P., 2156, 2166, *2171*
Kelly, J. P., 2447, *2465*
Kelly, K., *57*
Kelly, M. L., 1891, *1918*
Kelsey, J. L., 1061, 1063, 1070, 1071, *1104*
Kelton, W. D., 2389, *2392,* 2449, 2455, 2456, *2466,* 2471, 2478, 2485, 2488, 2494, *2494*
Kemeny, John, 74
Kemp, S., *2222*
Kemper, M., 523, 525, 526, *529*
Kempis, R.-D., *650*
Kemppainen, M., *1107*
Kendrick, D., *2538*
Keningham, T., *664*
Kennard, R. W., 2290, *2293*
Kennedy, M., *2405*
Kennedy, R. S., 929, *943*
Kennington, J., *2580*
Kennington, J. L., 2575, *2580*
Kensing, F., 964, *972*
Kent, W., 121, *151*
Kenyon, R., *2519*
Keoleian, G., 536, *541*
Keoleian, G. A., 536, *541*
Keppel, G., 931, *947*
Keren, G., 2201, *2219*
Kerlinger, F. N., 1134, *1153*
Kern, P., *1231*
Kernighan, B., 800, *823*
Kernighan, B. W., *2538*
Kerr, L. N., 2212, *2219*
Kerschberg, L., 122, *151*
Kervahut, T., *823*
Kerzner, H., *1262*
Keserla, A., 1524, *1526*
Keshav, S., 228, *257*
Kesselman, K., 229, *257*
Ketchel, J., 1216, *1235*
Ketscher, N., *588*
Kettinger, W. J., 1697, 1714, *1715, 1716*
Keyserling, M., 1366, *1389*
Keyserling, W. M., 1061–1063, 1071, 1087, *1101, 1104, 1109,* 1131, 1134, 1143, 1146, *1153*
Khalid, H. M., 706, *707*
Khalil, T. M., *1100*
Khanna, N., 603, 604, 606, 612, *618*
Khoshafian, S., *152*
Khumawala, B. M., 2074, *2081*
Khuri, A. I., 2239, *2239*
Kiefer, J., 2548, *2566*
Kieras, D. E., 931, *943, 2219*
Kierswelter, E., *1154*
Kiesewetter, T., 195, *224*
Kiesler, S., 1217, 1220, 1225, *1233, 1236*
Kihara, N., *618*
Kilbom, A., 1061, *1100, 1101, 1103, 1109*
Kilböm, A., 1071, *1102*
Kilduff, P. A., *2467*
Kilgour, D. M., *150*
Kilmann, R. H., 956, *972*
Kim, C., 1780, *1788, 1789*
Kim, C. O., 604, *618*
Kim, C. W., 1525, *1526*
Kim, G. H., 1780, *1788*
Kim, H., 1213, *1233,* 1780, *1789*
Kim, N., *2567*
Kim, S. H., 1525, *1526*
Kimura, F., *707*
Kindle, K. W., *2648*
Kindrick, J., *1789*
King, D., *2520*
King, D. R., 145, *150*
King, F. J., *941*
King, J. L., 134, 141, *151, 152, 1231*
King, N., *972*
Kingman-Brundage, J., 641–643, *650*
Kingsman, B. G., 2035, *2052*
Kinkade, R. G., 875, *898,* 932, *943*
Kinkel, S., 316, *323*
Kinlaw, D. C., *991*
Kinoshita, I., *400*
Kirk, D., 826, *836*
Kirkpatrick, D., 87, 88, *108*
Kirkpatrick, D. L., 935–937, *943*
Kirkpatrick, S., 2591, *2600*
Kirkpatrick, S. A., 851, 852, 855, *866*
Kirsch, W., *588*
Kirwan, B., 1025, 1028, *1038,* 1145, *1153,* 1209, 1211, *1233,* 1941, *1955*
Kitamura, T., *1918*
Kivi, P., 1145, 1146, *1154*
Kiviat, P. J., 2455, *2466*
Kjeldgaard, E. A., *543*
Kjellberg, T., *707*
Klamorth, K., 2621, *2621*
Klaucke, D. N., 1095, *1104*
Klauer, P., *152*
Klausen, T., *1038*
Klayman, J., 2198, *2219*
Kleijnen, J. P. C., 2491, *2494*
Kleiman, L. A., 922, *942*
Kleiman, L. S., 923, *943*
Klein, D. A., 147, 148, *151*
Klein, G., 1023, 1028, 1031, *1038*
Klein, G. A., 129, 137, *152, 1038,* 2173, 2177, 2198, 2205, 2208, 2209, 2214, *2219*
Klein, J., 988, *991*
Klein, K. J., 855, *866*
Klein, L., *650*
Klein, M. M., 1702, 1712, *1716*
Klein, W. M., 2201, *2223*
Kleinbaum, D. G., *2293*

Kleiner, A., *1010*
Kleiner, B. M., *1152,* 1894, 1897, 1900, 1912, *1917*
Klein Haneveld, W. K., 2630, *2647*
Kleinmuntz, B., 2200, *2219*
Klein Wassink, R. J., 423, *446*
Klemperer, P., 277, *279*
Kleywegt, A. J., 2635, 2645, *2647*
Klimer, F., 1146, *1153*
Klingman, D., 2573, *2580*
Klion, J., 1928, *1955*
Klutke, G., *749*
Knapp, B., 2423, *2443*
Knappenberger, W., *529*
Knauth, P., 1367, *1389*
Knight, J. L., 1895, *1918*
Knight, W., *1330*
Knight, W. A., 538, *542,* 870, 889, *895*
Knödler, R., 581–584, *587*
Knörr, M., *587*
Knothe, K., 201, *225*
Knott, K., 1460, 1461, *1462*
Knutilla, Amy, 352
Kobacker, E., *1739*
Kobayashy, M., *588*
Koch, F., 747, *749*
Kochan, A., 497, *529*
Kochan, T. A., *896,* 981, *992*
Kochar, B., *1330*
Koehler, J. J., 2197–2199, *2219*
Koehler, M. D., *542*
Koelsch, J. R., 539, *541*
Koenig, D. T., 493, *529*
Koenig, L. W., 2471, 2488, *2494*
Koenig, W., *588*
Kogi, K., 1144, *1153*
Kohdate, A., 689, *708*
Kohlberg, L., 853, *866*
Kohlhoff, S., 200, *225*
Koli, S. T., 1132, 1134, 1141, *1153*
Koller, S., *399*
Kolli, S., *152*
Komatsu, K., 1892, *1918*
Komorowski, J. P., 1897, *1918*
Kompier, M. A., *993*
König, W., 208, *225, 588*
Konold, P., 418, *446*
Konolige, K., *2217*
Konsynski, B. R., *149*
Kontogiannis, T., 1028, 1031, *1038*
Kontoravdis, G., 796, 800, *823*
Konz, S., 1353–1360, 1366, 1367, 1369, 1372, 1375–1385, *1389,* 1391, 1395, 1401–1405, 1407, *1408,* 1445, 1447, *1462*
Koo, P. H., *1526*
Koon, J. F., *540*
Koop, G. J., 1751, 1760, *1766*
Koriat, A., 2201, *2219*
Körner, E., 581–584, *587*
Kornhauser, A., 874, *896*
Korukonda, A. R., 914, *918*
Korunka, C., 1179, *1189,* 1226, 1228, *1233*
Kosiba, E., *1918*
Kosiur, D., 228, *258*
Kossek, E. E., *992*
Kotchevar, L. H., *836*
Kotler, P., 327, *352*
Kotter, J., 939, *943*
Kotter, J. P., 956, *972,* 996, 997, 1008, *1010*
Kouvelis, P., 1524, *1526*
Kouzes, J. M., 857, *866*
Kovacs, V., *175*
Kozaczynski, Voitek, 352
KPMG Global Consulting, *1350*
KPMG U. S., *1350*
Kraemer, K. L., 134, *152, 1231*
Krafcik, J. F., 555, *560*
Kraft, L. G., III, *1789*
Kralovec, O., 747, *749*
Kralovec, P., *989*
Kramer, G. P., *2219*
Kramlinger, T., 926, *947*
Kramme, K., *2444*
Krantz, D. H., *2218*
Krarup, J., 2076, *2082*
Kraus, N. N., 2192, *2219*
Krause, F.-L., 182, 191, 193, 197, 207, 209, 212, *225, 226,* 691, *707*
Kraut, A. I., 860, *866*
Kreiger, A. M., *707*
Kreinin, A. Y., *1668*
Krepchin, I. P., 2446, *2466*
Krieger, A. M., 702, *707*
Krikke, H. R., 538, *541*
Krimi, S., 435, *445*
Krishnamoorthy, M., *2465*
Krishnan, R., *1807*
Krist, R., 1117, 1120, *1128*
Kriwet, A., 207, *225*
Kroeck, G. K., 862, *866*
Kroeck, K. G., *866,* 923, *942*
Kroemer, H., *1109*
Kroemer, H. J., *1104*
Kroemer, K., *1109*
Kroemer, K. H. E., 1043, 1044, 1049, 1080, *1104, 1106, 1109,* 1201, 1202, *1233, 1234*
Kroemer-Elbert, K., *1109*
Kroemer-Elbert, K. E., *1104*
Kroenke, D. M., 117, *152*
Krogoll, T., *1155*
Krol, E., 229, *257*
Kroll, D. E., 2568, *2581*
Krucky, J., 1781, *1788*
Krueger, H., 1201, *1232*
Krueger, K. W., 2106, *2109*
Krüger, W., 2507, *2520*
Kruth, J. P., 209, 210, *225*
Kuglin, F., 775, *786*
Kuh, E., *2292*
Kuhberger, A., 2203, *2219*
Kuhl, J. G., *2520*
Kuhlmann, T., *617*
Kuhn, H. W., 2543, 2553, 2554, 2556, *2566*
Kuhnert, W., 312, *323*
Kukkonen, S., *1918*
Kulik, C., 928, *943*
Kulik, J. A., 928, *942, 943*
Kulwiec, R., 1519, 1520, *1526*

Kumagai, C., *560*
Kumamota, H., *1954*
Kumar, K., 348, *352*
Kumar, K. R., 2568, *2581*
Kumar, S., 1072, *1104, 1105*
Kundel, H. S., 1895, 1896, *1918*
Kunhert, K. W., 857, *866*
Kuntsevich, Alexei V., 2565
Kunze, H.-D., *588*
Kunzinger, C., *258*
Kuo, B. C., 160, *176*
Kuo, T., 538, *543*
Kuorina, I., *1128*
Kuorinka, I., 1082, *1100, 1103, 1104,* 1144, *1153*
Kupferschmid, M., 2563, *2567*
Kupper, L. L., *2293*
Kuroiwa, S., 550, 551, *560*
Kurta, Thomas, 74
Kurzhanski, A., *2623*
Kushner, H. J., 2635, 2645, *2647*
Kusiak, A., 462, *483*
Kusunoki, K., *561*
Kuswanti, Christiana, 539
Küttner, K.-H., *587*
Kyan, S., 1723, *1739*
Kylian, H., *1153*
Kyllonen, P. C., 930, *943*
Kyral, E., *2520*

Laabs, J. J., 901, *918*
Laandauer, J., *2519*
LaBant, Robert, 654
Lach, A., 1561, *1582*
Ladd, A., 86–88, 90, 91, *108*
LaFasto, F. M. J., 982, *992*
Laffel, G., *993*
Laflamme, L., 1160, *1189*
Lafollette, B. S., 1895, 1896, *1918*
Lagrange, J. L., 2530, 2531, 2533, *2539*
Laguna, M., 2447, *2466,* 2590, 2591, *2600*
Lai, Y. J., *2622*
Laibson, L., *540*
Laios, L., 1034, *1038*
La Londe, B. J., 2112, *2139*
Lam, D., 955, *971*
Lamberson, L. R., 1922, 1932, 1937, 1940, 1945, *1954, 1955*
Lambert, A. J. D., 538, *542*
Lambert, D. M., 2110–2112, 2114, 2116, 2118–2120, 2123–2127, 2132–2137, *2139, 2140*
Lamm, R. M., 759–762, 765, 766, 768, 770, *771*
Lammers, C. J., 960, *972*
Lancraft, R., *2442*
Landau, A., *1104*
Landau, K., 1050, *1104, 1106,* 1132, 1138, 1144–1146, *1153, 1154*
Landau, O., 849, *866*
Landers, T. L., *2109*
Landis, D., 953, 957, 961, *972*
Landrigan, P. J., *1189*
Landy, F. J., 987, *992*
Lang, R. G., *1583*
Lang, U., 2512, *2520*
Lange, J. E., *2220*
Lange, K., *588*
Langeard, E., 624, 625, *632*
Langley, G., *1827*
Langley, J., *1827*
Langston, E. A., *2220*
Lanning, S., *154*
Lapide, L., *353, 2019,* 2055, *2069*
Lapierre, J., *632*
Lappe, W., 412, *446*
Lara, M. A., 171, *176*
Larkin, J., 930, *943*
Larrow, R., 559, *560*
Larson, C. E., 982, *992*
Larson, P. D., *1807*
Larson, R. C., 2146, *2171*
Lascelles, D., *1807*
Lasdon, L., 2562, 2564, *2566, 2567*
Lasdon, L. S., *2565, 2567*
Lasserre, J.-B., 2044, *2052*
Lassister, D. L., 984, 985, *992*
Lassiter, D. L., 933, *944*
Latane, B., 882, *896*
Latham, G. P., *992*
Latham, H. C., *543*
Lathrop, R. G., 2197, *2219*
Latorella, K., 1909, *1918*
Lau, J. H., 424, 431, *446*
Laubacher, R. J., 1001, *1010*
Laufenberg, L., *224*
Laufer, J., *175*
Laughery, K. R., 1177, *1189, 2442–2444,* 2458, *2465*
Laughery, R., *2442,* 2458, *2466*
Launsby, R. G., *2240*
Laurent, F., *1102*
Laurig, W., 1050, *1106,* 1117, 1119, *1128*
Lavealle, S., *400*
Lavender, S. A., *1105*
LaVine, N., *2442*
LaVine, N. D., 2427, 2429, *2443*
Law, A. M., 2389, *2392,* 2449, *2466,* 2471, 2478, 2485, 2488, *2494*
Lawler, E., 794, *823,* 2060, *2069*
Lawler, E. E., 874, *896, 897,* 975, 979, *992, 993,* 1179, 1186, *1189*
Lawler, E. E., III, 17, *24, 918,* 926, *943,* 976, 977, 979, 987, 988, *992, 1807*
Lawler, E. L., *919,* 1735, *1739,* 2582, *2600*
Lawless, M. L., 2414, *2443*
Lawrence, Craig, 2563
Lawrence, J. S., *1109*
Lawrence, P. R., 874, *898,* 954, *972*
Lawshe, C. H., Jr., 910, 915, *919*
Layman, M., *2220*
Le, L. V., *1717*
Leachman, R. C., 2044, *2052, 2053*
Leamon, T. B., *1109*
Leavenworth, R. S., *1855,* 1861, 1876, *1876, 2003*
Leblanc, L. J., *750*
L'Ecuyer, P., 2635, 2645, *2647*

Ledford, G. E., *1807*
Ledford, G. E., Jr., *992*
Lee, B. H., 689, *707*
Lee, C. R., 1704–1706, 1708, *1716*
Lee, C. S. G., 1780, *1788*
Lee, E., *152*
Lee, E. S., 2620, *2622*
Lee, G., 163, *176*
Lee, H., 785, *785,* 1669, 1692, *1693*
Lee, H. F., 1524, *1526*
Lee, H. L., *707, 2032,* 2112, 2113, 2127, *2140*
Lee, I. H., *1236*
Lee, K. H., 1777, *1789*
Lee, M. K., 1524, *1525*
Lee, P. M., 2401, *2405*
Lee, R. G., 1696, 1697, *1716*
Lee, S. M., 2614, *2622*
Lee, W. G., *1109*
Lee, W. H., *1789*
Lee, Y.-H., 1725, *1739*
Lee, Y. R., 2620, *2622*
Legarth, J. B., 537, *540*
Legeard, D., *1918*
Legg, S. J., 1071, *1104, 1106, 1109*
Legrand, R., *1330*
LeGrande, D., *1235*
Lehmann, E., 2256, *2263*
Lehnerd, A. P., 688, *708*
Lehto, M., *2220*
Lehto, M. R., 1177, *1189,* 2181, 2186, 2196–2198, 2203, 2205, 2206, *2219, 2220*
Lei, M., *708*
Leigh, J. P., 1157, *1189*
Leino, P., 1084, *1104*
Leino, T., 1223, *1234*
Leitmann, G., 2621, *2623*
Leitner, K., 1145, *1153*
Leksan, M. P., 2320, *2330*
Lemarechal, C., 2636, *2647*
Lemke, C. E., 2556, *2566*
Lemmink, J., 628, 629, *631, 632*
Lemmo, T., *399*
Lemon, K., *664*
Lenat, D., *1330*
Lennartz, J., *588*
Lenstra, J., *823, 2069*
Lenstra, J. K., *1739, 1740,* 2590, 2591, *2600*
Lenz, J., 506, *529*
Leonard, D., 964, *972*
Leonard, M., *1313, 1330*
Leonard-Barton, D., 955, 963, *972*
Leplat, J., 1025, *1038*
Leppard, J., 663, *664*
Leredde, C., *1918*
Lesso, W. G., *2393*
Lester, J. C., *1128*
Lester, R., *970*
Leston, J., 2498, *2520*
Leszinski, R., 669, *683*
Leube, H., *588*
Leurgans, S. E., *1105*
Levenberg, K., 2551, *2566*
Leventhal, G. S., 880, *896*
Levi, Itzhak, 842
Levi, L., 1222, *1234*
Levin, I. M., 984–987, *992*
Levin, L. P., 2197, *2220*
Levine, E. L., 934, *944*
Levy, J., 745, *749*
Lew, A., 2645, *2646*
Lewellen, M., 2459, *2466*
Lewis, B. T., 1558, *1582*
Lewis, C. T., 916, *919*
Lewis, E. E., *1955*
Lewis, I., 2056, *2069*
Lewis, J., 781, *786*
Lewis, P., 857, *866*
Lewis, P. M., *1740*
Lewis, T. R., 263, *279*
Li, D., 2621, *2622*
Li, H., *1716*
Li, L., 1669, *1693*
Li, R. J., 2620, *2622*
Li, X., *1918*
Li, Y., *2567*
Liang, B. T., 130, 145, *152*
Liang, T., *1790*
Liang, T. J., 2577, *2581*
Liang, T. P., *152*
Liberson, J., 826, *836*
Lichtenstein, S., 2192, 2196, 2197, *2217, 2219–2221*
Lieberman, G. J., 129, *151,* 2243, 2248, 2255, *2263, 2567, 2580*
Liebowitz, J., 147, 148, *152*
Lieu, J., 759, *771*
Lieu, K. C., *353*
Lifshitz, Y., *1100,* 1365, *1389*
Lifshitz, Y. R., *992*
Light, D., *2109*
Liker, J. K., 556, 557, *560,* 980, *992*
Likert, J. G., 2211, *2220*
Likert, R., 874, *896,* 2211, *2220*
Lilegdon, W. R., 2454, 2458, *2466, 2467*
Liles, D., *1100*
Liles, D. H., *1104*
Lille, F., *1106*
Lim, C., *2466*
Lim, J. M., 1524, *1526*
Lim, L., 602, *618*
Lim, S Y., 1224, *1231, 1234, 1235*
Lin, C., 608, *618,* 1780, *1789*
Lin, C.-T., 163, *176*
Lin, Chih-Jen, 2565
Lin, G. Y., 697, *707, 1693*
Lin, J., *823*
Lin, J. T., *1525*
Lin, L., *1918*
Lin, L.-C., 2095, *2109*
Lin, S., 800, *823*
Lindblom, C. E., 1024, 1036, *1038*
Lindsay, R. W., 1021, *1038*
Lindsay, W., 1889, 1900, *1917*
Lindstrom, B. O., *896*
Lindstrom, K., *1188,* 1223, *1234*
Ling, S. H., *708*
Lingle, R., 1781, *1788*
LINGO Systems Inc., 2076, *2082*

Link, C. H., 473, 477, *483*
Linn, R. J., 1524, *1526*
Lintern, G., 932, *944*
Linton, J., 536, *542*
Linton, S. J., 1224, *1234*
Lioukas, S., *1038*
Lippitt, R., 939, *944*
Lissner, H. R., 1070, *1108*
List, G. F., *543*
Litschauer, B., *1233*
Little, R., 2423, *2443*
Littman, I. D., 980–982, *989*
Liu, B., *1693*
Liu, C., *2053*
Liu, C. L., 1724, *1739*
Liu, C. R., 1313, 1314, 1328, 1329, *1330*
Liu, J. W.-H., 2534, *2538*
Liu, L., 1674, 1675, 1691, *1693*
Liu, L. Richard, 1311
Liu, X.-G., *1668*
Liu, Y., *1918*
Livernash, E. R., 900, *919*
Llaneras, R. E., 928, 930, 931, 940, *946*
Lloyd, C. J., 1893, *1918*
Lloyd, M. H., 1071, *1104*
Lloyd, R. F., *993*
Lock, M. W. B., 1909, *1918*
Locke, E. A., 851, 852, 855, *866,* 983, *992*
Lockett, J., *2441, 2442*
Lockhart, R. S., 929, *942*
Lofgren, J., *1153*
Loftus, G. R., 930, *944*
Logan, M., *2443*
Loher, B. T., 887, *896*
Lombardi, Vince, 13
Lombardo, M. M., *944*
Long, R. J., 949, 950, *972*
Loomba, A., *1807*
Lootseen, P., *1919*
Lord, J., *1102*
Lorenson, W., *1789*
Lorenz, B., 574, 575, *587*
Lorenz, D., 1050, *1101*
Lorie, J. H., 2335, *2351*
Lorsch, J. W., 954, *972*
Lorussin, E., *993*
Los Angeles Times, 949, *972*
Lotter, B., *399,* 416, 418, *446*
Lounsbury, J. W., *917*
Louveaux, F., 2630, *2646*
Love, P. E. D., 1696, 1699, 1701, 1705, 1709, 1710, *1716*
Lovelock, C., *632*
Lovelock, C. H., 642, *650*
Low, Y., 2464, *2466*
Lowe, K. B., 842, 852, *866*
Lowerre, J. M., *1766*
Lowery, J., 745, *749*
Lowery, P. E., 856, 857, *866*
Lozito, S., *2444*
Lozo, P., *1789*
Lu, Q., *542*
Lu, Qin, 539
Lu, S. C.-Y., *707*
Lubin, J., 2434, *2443*
Luce, R. D., 2173, 2178, 2204, *2219, 2220*
Luchs, R., *445*
Lucier, C. E., 1699, 1701, *1716*
Luczak, H., 1138
Luenberger, D. G., *2567*
Lueng, Y. T., *1717*
Luftig, P. D., *2240*
Luh, C. W., 930, *944*
Luh, P. B., *707*
Luhn, G., 214, *225*
Luisi, T., *1955*
Luk, B. L., *399*
Lumai, R., 1904, 1905, 1912, *1918*
Lummer, S. L., 761, *771*
Lummus, R. R., *560*
Lundborg, G., 1086, *1104*
Luo, H., 508, *529*
Lusted, L. B., 1023, *1038*
Lustig, I., *2539*
Lustig, I. J., 2055, *2069*
Luthans, F., 911, *919*
Luttmann, A., 1072, 1076, *1103, 1109*
Lutz, S., 616, *618*
Lynch, C. L., 147, *153*
Lynch, D. C., *258*
Lynch, R. P., *57*
Lyons, L., *1885*

McAdams, J. L., *990*
McAtamney, L., 1117, *1128*
McBride, R. D., 2573, *2581*
McCall, M. W., 939, *944*
McCarl, B., *2580*
McCarthy, J. C., 1892, 1912, *1918*
McClagan, P. A., *919*
McClain, J. O., *2052*
McClelland, C. L., 870, 871, 877, 887, 892, 893, *894*
McClelland, G. H., *151*
McClelland, I., 1131, *1154*
McClelland, J., *1789*
McConville, J. T., 1049, *1106, 1128*
McCormick, E., *1109,* 1177, *1189*
McCormick, E. J., 871, 875, *897,* 1132, 1134, 1135, 1138, *1154*
McCormick, E. S., 1016, *1038*
McCormick, G. P., 2530–2532, *2538,* 2540, 2551, 2555, 2560, *2565–2567*
McCormick, S. T., 1732, *1740*
McCormick, W. T., 1137, *1154*
MacCoun, R. J., *2219*
McCracken, J. H., 2410, 2423, *2443*
McCullagh, P., 702, *708*
McDaniel, J. W., 1050, *1105, 1107*
McDaniel, S., *1789*
McDermott, R., 148, *152,* 963, *972*
Macdonald, J., 1711, *1716*
McDonald, S., *257*
MacDuffie, J. P., 951, *972, 973*
McElroy, F. E., 1568, *1582*
McFadden, F. R., 102, *109*
McGeoch, J. A., 930, *944*
McGill, S. M., 1076, *1105, 1106*

McGinnis, L., *1390*
McGinnis, L. F., *1526*
McGlennon, D., 1101
McGlothlin, J. D., *989,* 1084, 1085, *1107*
McGrath, J. E., 127, *152,* 877, 880, 881, *896,* 984, *992*
McGrath, M. E., *353*
MacGregor, D., *1038,* 2201, *2217*
McGuire, F., 745, *750*
McGuire, W. J., 2201, *2220*
McGwire, T. W., *1233*
Mach, R. S., *1108*
McHenry, J. J., 921, *944*
Machner, B., 193, *224*
McIntyre, R. M., *990*
Mack, J. D., *1110*
Mack, R. L., 1213, 1214, *1231*
McKay, A., 193, *225, 707*
McKay, K. N., 1738, *1740*
McKenna, D. D., *866*
McKenzie, R. M., 1899, *1918*
Mackenzie, S. B., *866*
Mackey, J., *560*
McKnight, A. J., 2206, *2220*
McKnight, A. S., *2220*
McLaney, M. A., *991, 992*
McLean, R. A., 2225, 2229, *2239*
Mcleary, K. J., *631*
McLeod, R. J., 79, 80, *109*
McLeod, R. W., 1209, *1234*
McMahan, G. C., 856, *867*
McMahon, C., 229, *258*
McMillan, C., 1752, *1766*
McMillan, G. R., 2410, *2443*
MacNair, E., *2467*
Macneil, I. R., 2126, *2140*
McNichol, D., 1897, 1898, *1919*
MacPherson, B., *1102*
McShane, S. L., 916, *919*
McSweeney, M., 2256, *2263*
Madden, J. M., *918*
Madeley, S. J., *1101*
Madigan, K., 671, *683*
Madigan, R. M., 914, 915, *919*
Maeda, K., 1071, *1104*
Maffioli, F., *2581*
Magazine, M., *2053*
Mager, R. F., 925, 926, *944*
Magnanti, T., *2068*
Magnanti, T. L., *822, 823, 2081, 2538, 2580, 2581*
Magora, A., 1071, *1105*
Magretta, J., *57*
Mahachek, A. R., 745, *749*
Mahajan, P., *1104*
Mahalik, N. P., 165, *176*
Maher, J., 1900, *1918*
Mahmood, M. A., 914, *919*
Mahon, J. J., 738, *750*
Maida, J., *1129*
Main, M., 72, *108*
Maister, D. H., *1350*
Majchrzak, A., 870, 889, *896,* 953, 955, 961–965, *972, 1232*
Major, D. A., *943*
Makens, P. K., *750*
Malakooti, B., 2621, *2622*
Malarkey, R., 74, *108*
Malcolm, J. A., *1462*
Malde, B., 1131, *1154*
Malhotra, A., *972*
Malhotra, M., 1793, *1806*
Malhotra, Y., 1696, 1715, *1716*
Mallet, L., 2208, *2220*
Mallows, C. L., 2284, *2293*
Malone, D. M., *1916*
Malone, M. S., 142, 147, 148, *150*
Malone, T., 1287, *1295*
Malone, T. B., 1145, *1154*
Malone, T. W., 144, *152,* 697, *707,* 1001, *1010*
Malorny, C., 648, *650*
Maltz, A., 2070, *2082*
Maluso, N., 1704, *1716*
Mambretti, J., 229, *258*
Manenica, I., 1063, *1101*
Manenica, J., *1101*
Manganelli, R. L., 1702, 1712, *1716*
Mangano, J. M., *1583*
Mangasarian, O. L., *2567*
Manion, M. M., *541*
Mann, L., 140, *151, 1526,* 2176, 2184, 2206, 2212, *2219*
Mann, N. R., 1946, *1955*
Mann, W. S., 463, *483*
Mannix, E. A., *2218*
Mantel, R. J., *2109*
Manuaba, A., *993*
Manufacturing Studies Board (MSB), 950, *972*
Marakas, G. M., *152*
Marascuilo, L., 2256, *2263*
Marathe, V., 746, *749*
March, J., 140, *152*
March, J. G., *152*
Marchal, P., *1918*
Marchant, H., 928, *946*
Marciniak, Z. K., *617*
Marcus, A., 1215, *1234*
Marcus, J. S., 228, *258*
Margolin, B. H., 2491, *2495*
Marion, L., 89, 93, *108*
Markle, S. M., 940, *944*
Markowitz, H., 2455, *2466*
Markowitz, H. M., 752, *771*
Markowitz, S. B., *1189*
Marks, M. L., 978, 979, *992*
Markus, A., 697, *707*
Markwell, J. S., *2465*
Marmaras, N., 1023–1025, 1032–1035, *1038*
Marn, M. V., 669, 677, *683*
Marqardt, D. W., 2551, *2566*
Marquardt, M., *57*
Marras, W. S., 1042, 1080, 1082, 1085, 1092, *1102, 1103, 1105, 1107, 1109*
Marron, J. P., 1558, *1582*
Marrs, F., *57*
Marrs, J. A., *542*
Marshall, J., *1190*
Marsten, R., 2530, 2534, 2535, *2539*

Martel, A., 1212, 1213, *1234*
Martel, S., *750*
Martello, S., 811, *823*
Martin, A. J., 1675, *1693*
Martin, B. J., *1101, 1128, 1129, 1230, 1232*
Martin, C. H., 2586, *2600*
Martin, C. L., 922, *944*
Martin, E., *2442*
Martin, H., 214, *225,* 1186, *1189*
Martin, J., 745, *749, 1262*
Martin, J. B., *1109*
Martin, L., *2217*
Martin, M. V., 689, *707*
Martin, N. A., *1100, 1107*
Martin, R. K., 2587, *2600*
Martin, S., 350, *352*
Martino, J. P., 952, *972*
Martinson, F. K., 2620, *2622*
Martocchio, J. J., 2200, *2220*
Marty-Mahe, P., *1918*
Maslow, A. H., 853, *866*
Mason, J. S., 1504, *1526*
Mason, R., 981, *991*
Mason, R. L., *2240*
Massie, C., 2071, *2082*
Massow, C., *617*
Masters, J. M., 2071, *2081*
Mastrangelo, C. M., 1863, *1876*
Masud, A., 2074, *2081*
Masud, A. S. M., 2606, 2608, *2622*
Masuoka, T., *400*
Matheson, D., 129, 148, *152*
Matheson, J., 129, 148, *152*
Mathias, K., *1789*
Mathis, R. H., 555, *560*
Matsui, M., 607, 608, *617, 618*
Matsuo, H., 1731, *1739*
Mattes, W., *225*
Mattheus, R., *2520*
Matthews, M., 1900, *1920*
Mattila, M., 1062, *1105,* 1145, 1146, *1154*
Mattison, R., 83, *108*
Mattson, J., *945*
Mattsson, J., *631*
Maule, A. J., 2173, 2208, 2209, *2220, 2222*
Maurer, D., *225*
Maxey, J., *1236*
Maxwell, W. L., *1740,* 2576, *2581*
May, S. F., *1129*
Mayer, R. J., 490, *529*
Mayhew, D. J., 1206–1208, 1210, 1212, 1215, 1216, *1234*
Maynard, Harold B., 1429
Mayo, E., 874, 892, *896*
Meacham, A., 538, *542*
Mead, R., 2550, *2566*
Meccham, R. C., *1154*
Medsker, G. J., *894, 895,* 976, 977, 987, *989, 992*
Meedt, O., 440, *445, 446*
Meeraus, A., *2538*
Megalino, B. M., 855, *866*
Megaw, E. D., 1897, 1914, *1919*
Mehle, T., 1023, *1038*
Mehra, S., 559, *560*
Meidan, A., 2074, *2082*
Meier, R., *588*
Meindl, R. S., *1129*
Meiren, T., 635, 637, 639, *650*
Meister, D., 875, *897,* 1941, *1955*
Meller, R. D., 1524, *1526*
Mellers, B. A., 2197, 2198, *2216*
Meltzer, A. L., *946*
Melum, M. M., *992*
Menasce, D. A., *736*
Menckel, E., 1160, *1189*
Mendelssohn, R., 2645, *2647*
Menerey, D., 536, *541*
Menges, B., *2520*
Menges, R., *1106*
Menoni, D., *1101*
Mento, A. J., *993*
Mentzer, J. T., 674, 675, *683*
Mercier, D., 2573, *2580*
Merritt, A., 961, *971*
Mertins, K., 213, 214, 218, *225, 226*
MESA International, 1782, *1789*
Mesenbourge, T. L., *353*
Meshkati, N., 953, 957, 959, 963, *972*
Messick, D. M., 2210, *2220*
Messmer, M., 857–860, *866*
Metcut Research Associates, Inc., 458, 467, *483*
Metev, B. S., 2621, *2622*
Métivier, M., *2646*
Metz, C., 2572, *2581*
Metzel, E., *446*
Meyer, L. H., *353*
Meyer, M., 688, *708*
Meyer, M. H., 685, 686, *708*
Meyer-Nolkemper, H., *588*
Meyers, R. A., *2216*
Meyn, S. P., 2166, *2171*
Miars, M. D., *2465*
Michalowski, W., 2621, *2622*
Michel, R., 2055, 2064, *2069*
Michels, K. M., *2240*
Michman, R., 2115, *2140*
Micro Analysis and Design, 2418, *2443*
Miedema, M. C., 1064, *1105*
Miettinen, Kaisa, 2564
MIL-HDBK 1974, 1928, *1955*
MIL-HDBK 1979, *1955*
Milkovich, G., *919*
Milkovich, G. T., 880, *897,* 903, 904–910, 914, 915, *919*
Miller, A., *57*
Miller, C. E., 2558, *2566*
Miller, D., *25*
Miller, G., 1015, *1038*
Miller, J. M., 1177, *1189*
Miller, L., 803, *823*
Miller, L. A., *1919*
Miller, M., *257*
Miller, R., *2218*
Miller, R. B., 932, *944*
Miller, R. W., *1740*
Miller, S., *2467*

Miller, W. L., 148, *152*
Miller, W. T., 1780, *1789*
Mills, J., *617*
Mills, P. K., 623, *632*
Milner, N. P., 1063, *1105*
MIL-STD-1969, 1940, *1955*
Min, H., 1780, *1788, 1789,* 2080, *2082*
Minahan, T., 2057, *2069*
Minieka, E., 809, *823*
Minoli, D., 228, *258*
Minoux, M., *2567*
Minton, P. D., *2293*
Mintzberg, H., 113, *152,* 311, *323,* 880, *897*
Mir, A. H., 1136, 1139, 1142, 1144, *1154*
Mirchandani, P., 2067, *2069*
Mirvis, P., *993*
Mirvis, P. H., *992*
Misra, K. B., *1955*
Misterek, S. D., 985, 986, *992*
Mital, A., 1052, 1053, 1071, 1072, *1100, 1105, 1108, 1152*
Mitchell, D., *2441, 2444*
Mitchell, Mary, 352
Mitchell, T. R., 874, *897,* 2207, *2215*
The MIT Commission on Industrial Productivity, *970*
Mitroff, I. I., 956, *972*
Mittelmann, H. D., 2563
Miya, T., *618*
Miyake, D. I., 555, *560*
Mo, Y. W., 524, *529*
Mockler, R. L., 2214, *2220*
Moder, J. J., 1455, 1456, *1462*
Modigliani, V., 931, *945*
Modl, A., 532, *542*
Moeller, B., 1701, *1716*
Moeller, N. L., *896*
Moen, R. D., *1827, 2240*
Mohr, R., *749*
Mohrman, S. A., 975, 979, *992, 1807*
Molay, Ken, 2387
Molleston, J. L., *991*
Molteni, G., *1101*
Molyneux, L., 663, *664*
Monden, Y., 545, 546, *560,* 1890, 1895, *1919,* 2035, *2052*
Mondy, R. W., *1583*
Monge, P., 145, 147, *150*
Monma, C., *2068*
Monod, H., 1056–1059, *1103, 1105, 1106*
Monod, H. A., 1056, *1105*
Monostori, L., 697, *707*
Monro, S., 2634, *2648*
Monroe, K. B., 670, 671, 674, 675, *683*
Montague, W. E., 931, *943*
Montevecchi, M., *2467*
Montgomery, D. C., 1863, 1868, 1872, 1876, *1876, 2003, 2053,* 2225, 2228, 2229, 2232, 2239, *2239, 2240,* 2258–2261, *2293,* 2375, *2392*
Montgomery, H., 2207, *2220*
Moodie, C. L., *1787*
Moody, J., 1780, *1789*
Moor, W. C., *2392*
Moore, D. M., *1583*
Moore, F. G., 915, *919*
Moore, J. S., 981, *990,* 1087–1089, *1105*
Moore, P. R., 165, *176*
Moorman, R. H., *866*
Moraleda, J., 1902, *1919*
Moran, P., *1037*
Moran, T. P., 134, *150, 153, 1231, 1310, 2442*
Morawski, T. B., 1895, 1913, *1919*
Moray, N., 1894, 1895, *1919*
Moré, J. J., 2536, *2539,* 2563, *2566,* 2575, *2581*
Moré, Jorge, 2565
Morgan, B. B., Jr., 933, *944, 947,* 984, 985, *992*
Morgan, C. P., *989*
Morgan, J., *1236*
Morgan, K., *2520*
Morgan, R. L., 929, *946*
Morgenstern, O., 2173, 2178, 2180, 2182, *2222*
Morgenthal, J. P., *258*
Morin, T., 2645, *2647*
Morin, T. L., *2623*
Moro, F. B., *1190*
Morris, D., 1703, 1704, *1716*
Morris, J. R., 1705, *1716*
Morris, L., 148, *152*
Morris, N. M., 928, *944*
Morris, P. W. G., 1242, *1251*
Morris, W. T., *25*
Morrisey, S. J., *1109*
Morrison, A. M., *944*
Morriss, S. B., 160, *176*
Morrow, R. L., 1423, 1448, *1462*
Morse, J. N., *2623*
Morton, T., 1724, *1740*
Moscovici, S., 2212, *2220*
Mosel, D., 984–987, *992*
Moser, G., *588*
Mosgaller, T., *993*
Mosges, R., *400*
Mosier, J. N., 132, 133, *153*
Moskowitz, A. D., 1453, 1454, *1462*
Moskowitz, H., *1790*
Mosteller, F., 2283, 2286, *2293*
Motazed, R., *2393*
Mote, C. D., *1235*
Motwani, J., 559, *560*
Mountford, S. J., 1212, 1213, *1234*
Mourier, P., 926, *944*
Mowday, R. T., 870, 874, *897*
Mower, N. R., *992*
Moyer, L. K., 440, *446*
Mozrall, J. R., *1152*
Muchinsky, P. M., 915, *918*
Mueller, W. S., *895*
Muench, D., 1242, *1251*
Mukherjee, A., 1329, *1330*
Mulford, M., 2198, *2216*
Müller, E., *399*
Müller, F., *588*
Muller, J., *1788*
Müller, K. W., *1107*
Muller-Schwenn, H. B., *1154*

Mulvey, J. M., 2080, *2081*
Mumford, M. D., 856, *866*
Mumpower, J., *151*
Mundel, M. E., 874, *897,* 1424, 1425, *1462*
Mungwattana, A., 1524, *1526*
Munk-Madsen, A., 964, *972*
Muñoz, J., *2316*
Muralidharan, R., *2442*
Murgatroyd, R. A., 1909, *1919*
Murhammer, M. W., 229, *258*
Murphy, A. H., 2193, 2196, 2197, 2201, *2220, 2223*
Murphy, F. H., 131, *152*
Murphy, K. R., 923, *944*
Murray, M., 938, *944*
Murray, W., *2538*
Murray, Walter, 2564, 2565
Murrell, K., *1109*
Murtagh, Bruce A., 2564
Murty, K. G., 2526, 2528, *2539*
Musselman, K. J., 2461, *2466*
Myers, J., 2202, *2218*
Myers, J. L., *2220*
Myers, P. S., 147, *152*
Myers, R. H., *2240, 2293, 2494*
Myles, W. S., 1071, *1104*
Myopoulos, J., 122, *152*

Nachemson, A., *1107*
Nachemson, A. L., *1100, 1101*
Nadler, D. A., 1000, *1010*
Nadoli, G., 171, *176*
Nagamachi, M., *992*
Nagao, D. H., 922, *944*
Nagel, R. N., *528, 971*
Nagle, B., *1583*
Nahmias, S., 1669, *1693, 2032,* 2039, 2040, *2052*
Naik, N., 759, *770*
Naim, M. M., *2140*
Nakajima, S., 553, *560*
Nakajima, Y., *1918*
Nakamura, K., *588*
Nakane, T., *588*
Nalebuff, B., 2212, *2217*
Nance, R. E., 2446, *2466*
Napolitano, M., 2079, *2082*
Narahari, Y., 1668, *1668*
Narasimhan, R., 1763, 1764, *1766, 1767*
Narayanan, S., 2464, *2466*
Nardi, B. A., 1193, 1206, 1209, *1234*
Naruo, N., *2220*
Narus, J. A., 617, *618*
Nash, A. N., 907, *919*
Nash, D. B., 985, 986, *991*
Nash, S. G., *2567*
Nathemson, A., *1107*
National Academy of Sciences (NAS), 1070, *1105,* 1195, 1197, 1198, 1199, 1204, *1234*
National Health and Nutrition Examination Survey (NHANES), 1113, *1128*
National Institute for Occupational Safety and Health (NIOSH), 1048, 1052, 1070, 1082, 1084, 1086, *1105,* 1119, 1121, *1128,* 1139, *1154,* 1162, 1163, 1167, 1168, 1170, 1171, *1189,* 1195, 1197, 1198, 1200–1202, 1205, 1224, *1234*
National Institute of Standards and Technology (NIST), 8, *24, 1806, 1807*
National Research Council (NRC), 1888, *1919*
National Safety Council, 1070, *1105,* 1158, 1168, 1172, 1183, *1189*
National Software Testing Laboratory (NTSL), *1262*
National Transportation Safety Board (NTSB), 960, *972*
Navathe, S. B., 80, *108*
Navon, D., 2198, *2220*
Nawaz, S., *1919*
Nayak, P. R., *57*
Nayar, N., 1050, *1106*
Naylor, J. C., *2222*
Neale, D. C., 1213, 1214, *1234*
Neale, M. A., 2212, *2215, 2218*
Nebel, D. C., III, 825, *836*
Negroponte, N., *58*
Nelder, J. A., 702, *708,* 2550, *2566*
Nelson, B. L., *2171, 2465,* 2488, 2492, *2494*
Nelson, G. D., *989*
Nelson, L. S., 1836, *1855*
Nelson, P. A., 953, 954, *971*
Nelson, R. A., *1547*
Nelson, S. R., *750*
Nelson, W., 1945, 1946, *1955*
Nemeth, C. J., 983, *992*
Nemhauser, G., *2068*
Nemhauser, G. L., 805, 811, 813, 814, *823,* 2582, *2600,* 2636, *2647*
Neslin, S., 679, *683*
Ness, J. A., 672, *683*
Netanyahu, Benjamin, 842
Neter, J., *2293*
Netherton, D., *1583*
Neubert, K. H. P., *1885*
Neugebauer, J., *399, 400*
Neukom, C., *2443*
Nevins, J. L., 1313, *1330*
Newbrough, E. T., *1584*
Newcomb, P. J., 688, *708*
Newell, A., *150,* 1026, *1037, 1039, 1310,* 2208, *2220,* 2410, 2412, *2442, 2443*
Newell, A. F., 1216, *1234*
Newell, A. L., *1231*
Newell, G. F., 1656, *1668*
Newhouse, R., 1615, *1622*
Newman, D. G., 835, *836*
Newman, J. M., 880, *897,* 903, 904–910, 914, 915, *919*
Newton, D., *542*
Newton, I., 2530–2532, 2533, 2534, *2539*
Nguyen, V., 1655, 1656, *1668*
Nichols, E. L., Jr., 2113, *2139*
Nicholson, A. S., *1105*
Nicholson, L. M., 1071, *1106*
Nicol, D. M., *2494*
Niebel, B., 1410, 1414–1419, 1423–1426, 1436–1439, 1450, 1456–1459, 1461, *1462*

Niebel, B. W., 473, *483,* 871, 874, *897,* 1311, *1330*
Nielsen, J., 1193, 1206–1210, 1212, 1215–1217, *1234*
Nielson, C. P., *542*
Nielson, J., 143, *152*
Nieva, V. F., 933, *944,* 984, 985, *992*
Nikkan Kogyo Shimbun, 548, *560*
Nikoukaran, J., 2449, *2466*
Nisanci, A., 2461, *2466*
Nisanci, H. I., 2446, 2454, *2467*
Nisbett, R., 2200, 2201, *2220*
Nisbett, R. E., *2218*
Niu, A., *1919*
Nixon, R. M., 590, *600*
Nocedal, J., *2567*
Nock, A. J., 739, *750*
Noe, R. A., *896*
Noesen, S., *540*
Nof, S. Y., 155, 157, 159, 161, 162, 167, 168, 170–172, 175, *176, 399,* 603–609, 612, *617–619,* 1525, *1526*
Nolan, K., *1827*
Nolan, T., *1827*
Nolan, T. W., *1827, 1855, 2240*
Nolen, J., *483*
Nonaka, I., 148, *152,* 214, 215, *225,* 1293, *1295*
Norberg-Bohm, V., 531, *542*
Nord, W. R., *990*
Nordgren, W. B., 2460, *2466*
Nordhaus, W. D., 344, *352*
Nordin, M., *1107*
Noreen, E., 557, *560*
Nori, V. S., *2647*
Norkin, V. I., 2630, *2647*
Norlan, S., *1233*
Norman, C., *1827, 1855*
Norman, D., 1014, 1017–1019, 1026, *1038, 1039*
Norman, D. A., 930, *945,* 1193, 1209, 1213, *1235,* 1297, *1310*
Norman, R. W., 1076, *1105, 1106*
Norman, V., 2461, *2466*
Norris, G., 94, *109, 540*
North, R. A., 2423, *2443*
Norton, D., 997, 998, *1010*
Norton, D. P., 21, *24,* 321, *322,* 645, *650*
Nott, H., *2465*
Novelli, L., 979, *992*
Nowacki, H., 193, *225*
Nowlan, F. S., 1618, *1623*
Nozari, A., 2491, *2495*
Nuclear Regulatory Commission (NRC), 875, *897*
Nunamaker, J. F., *149,* 2214, *2220*
Nunnally, J. C., *992*
Nurani, R. K., 1890, *1919*
Nurminen, M., *1107*
Nussbaum, M. A., 1119, *1129*

Oak Associates, Inc., 1254, *1262*
Oakford, R. V., 2400, *2405*
Oakland, J. S., 1713, *1715*
Oark, S. J., *992*
Obata, H., *2622*
Oberstone, J., 2575, *2581*
Object Management Group (OMG), 1773, 1783, *1789*
Oblak, J. M., *707*
O'Brien, C., *1105*
O'Brien, J. L., *989, 993*
O'Brien, T. G., 1889, *1919*
Occhipinti, E., *1101*
Occupational Safety and Health Administration (OSHA), 593, *600,* 980, *992,* 1070, 1082, 1098, *1106,* 1133, 1145, *1154*
Occupational Safety and Health Authority (Australia), *1154*
O'Connor, E., 222, *223*
O'Connor, E. J., *993*
O'Connor, J., *866*
O'Connor, K., *749*
O'Connor, R. O., Jr., *2215*
O'Connor, T., *1104*
Odegard, O., *2520*
Odoni, A., 801, 802, *822, 823*
Odoni, A. R., 2146, *2171*
Oesterreich, R., *1155*
Office of Industrial Technologies, 534, *542*
Office of Management and Budget, 329, *352*
Office of Technology Assessment (OTA), 950, *972,* 1217, 1220, 1222, *1235*
Ofrali, R., *736*
Ogawa, K., 1145, *1155*
O'Grady, P., 1777, *1789*
O'Hara, D., 1031, *1039*
O'Hara, J. M., 931, *944,* 2410, *2443*
Ohno, T., 545, *560*
Ohta, H., *1787*
Ojanen, K., *1107*
Okamura, K., *973*
Oldham, G., 979, *990*
Oldham, G. R., 871, 874, 887–889, *896*
Olfman, L., 135, *150*
Oliver, R. K., 2112, *2140*
Oliver, R. L., 623, 628, 629, *632*
Olle, T. W., 300, 301, *307*
Ollero, A., *1919*
Olsen, R. F., 2125, *2140*
Olson, C., 559, *561*
Olson, D. E., *1790*
Olson, J. R., 1210, *1231*
Omori, T., *1918*
O'Neill, D., *993*
Ones, D., *945*
Onisawa, T., *1955*
Open Application Group, 343, *352*
Opitz, H., 462, *483*
Optimal Planning Techniques, *2069*
Orasanu, J., 2208, 2209, *2219–2221*
Orbell, J. M., *2216*
Ordonez, L. D., *2216*
O'Reilly, C., 893, *897*
O'Reilly, C. A., 956, *973*
O'Reilly, J. J., 2449, 2454, 2458, 2461, *2466, 2467*

Organization for Industrial Research, Inc. (OIR), 462, 477, *483*
Orlicky, J., 2034, *2052*
Orlikowski, W. J., 952, *973*
Orlin, J. B., *822, 2580*
Orr, G. B., 1097, *1100*
Orta-Anes, L., 981, *992*
Orte, M., *1919*
Ortengren, R., *1107*
Ortiz, A., 745, *750*
Ortony, A., 1213, *1235*
Osada, T., 553, *561*
Osborn, F., 2213, *2221*
Osborne, D., *992*
Osburn, H. G., 1134, *1154*
Osman, I., *824*
Østerle, H., *223*
Osterman, P., 952, 954, *973*
Ostrom, A. L., *632*
Ostwald, P. F., 2303, 2307, *2316*
Otte, R., *529*
Ouederni, B. N., *2393*
Ould, M. A., 1708, *1717*
Ounpuu, S., *1128*
Ovalle, N. K., *993*
Overbach, W., *1918*
Overbeck, J., *399*
Overington, I., 1895, *1919*
Owen, D. B., *1876*
Owen, J., *225*
Owen, V., *749*
Owens, J., *540*
Owens, J. W., 537, *542*
Ozaki, S., *836*
Ozan, T. M., *2581*

Pacific Northwest Pollution Prevention Resource Center, 534, *542*
Packman, E., 229, *257*
Padberg, M., *2600*
Paese, P. W., 2203, 2212, *2221*
Page, A. L., 689, *708*
Page, R. C., *918*
Pagh, J. D., 2072, *2082,* 2112, 2114, 2116, 2119, 2123, 2124, 2126, *2139, 2140*
Pahl, G., 207, *225*
Pajak, J., *1919*
Palacios-Gomez, F., 2562, *2566*
Palko, E., 1560, *1582*
Pallantino, S., *2581*
Palmer, G., *1918*
Palmer, J. D., 114, *153*
Palmer, M. S., *1127*
Pan, S. C., *1236*
Panchapakesan, S., 2488, *2494*
Pandy, M. G., 1126, *1129*
Pandya, A., 1118, *1129*
Pandzic, I., 2499, *2520*
Panico, J. A., *1462*
Panier, Eliane R., 2563
Panter, W., 1145, *1154*
Papadimitriou, C. H., 2582, 2594, *2600*
Papanicolaou, V., 1524, *1526*
Papastavrou, J., 607, *618,* 2181, 2186, *2220*
Papastavrou, J. D., 1177, *1189*
Pape, E. S., 1450, 1455, *1462*
Papp, D. S., *149*
Papper, E. M., *895*
Paraboschi, S., *150*
Paramore, B., *1037*
Parasuraman, A., 623, 625–630, *632, 633,* 640, 641, *650, 1965*
Parent, R. E., *1129*
Pareto, V., *2624*
Pareto, Vilfredo, 1859
Park, B. G., *1918*
Park, C. S., *2351,* 2370, 2372, 2375, *2392,* 2394, 2403, *2405*
Park, H. S., *446*
Park, K. S., *1108, 1109*
Park, S., *1789, 1790*
Park, S. J., 981, *992*
Parker, M., 988, *993*
Parker, R. G., 2582, 2588, 2589, 2594, *2600*
Parkinnen, J., *1918*
Parlette, G., *897*
Parrish, R. G., *1104*
Parsaei, H. R., *152*
Parsaye, K., 122, *152*
Parsons, H. M., 932, *944*
Parsons, K. C., 1132, 1134, 1145, *1154*
Parsons, P. J., 930, *944*
Partsch, W., *2140*
Partyka, J. G., 819, 821, *823,* 2062, *2069*
Parunak, H., 1777, *1789*
Parunak, H. V., 1213, *1235*
Pascoletti, A., 2614, *2622*
Passino, E. M., *917*
Patel, C., *2466*
Patel, S., 1135, *1154*
Patel, S. C., 1899, *1917*
Patrick, P., *529*
Patterson, E., *864*
Patterson, T. T., 910, *919*
Patty, B., *2580*
Pätzold, B., 212, *225*
Paul, R. J., 2449, *2466*
Paulin, M., 628–630, *632*
Paulus, P. B., 877, 881, *897*
Pausch, R., *2520*
Pavard, B., 1025, 1032, *1038*
Pavett, C., 1794, *1807*
Pawllek, G., 207, *225*
Payne, J. W., 1023, *1039,* 2173, 2176, 2200, 2205, 2207, 2209, *2219, 2221*
Payne, P., *1232*
Payne, W., 2436, *2443*
Peace, G. S., 2237, *2239*
Pearcy, M. J., 1061, *1106*
Pearlman, A., 1569, *1582*
Pearlstein, G. B., 938, *944*
Pearlstein, R. B., 938, 939, *944*
Pearson, J. N., *561*
Pearson, R. G., 875, *897*
Pearson, T., *2218*
Peck, E. A., *2293*
Peck, H., 2116, *2139*
Pedigo, P. R., *866*

Pedotti, A., *1101*
Peele, T. T., 1556, *1582*
Peeters, M., *993*
Pegden, C., 505, *529*
Pegden, C. D., 2455, *2467, 2495*
Pejtersen, A. M., *152*
Pejtersen, M., *1039*
Peltu, M., 1706, 1707, *1716*
Penev, K., 538, *542*
Penev, K. D., 538, *542*
Penkuhn, T., *542*
Pennington, N., 2205, 2207, *2221*
Pennypacker, B., *989*
Peppers, D., 662, *664,* 704, *708*
Perakath, B., *529*
Perez, R. S., *946*
Permenter, K. E., *1154*
Perrakis, S., *2393*
Perrewe, P. L., *864*
Perrien, J., 628–630, *632*
Persensky, J. J., *2443*
Persson, J., *1103, 1109*
PERT Coordination Group, 1272, *1277*
Perusse, M., *1233*
Peter, J. P., 630, *631, 632*
Peters, B. A., 1524, *1526*
Peters, D. S., *750*
Peters, G. A., 1177, *1189*
Peters, R. W., 1588, 1589, 1593, 1597, 1598, 1600, 1604, 1606, 1607, 1609, *1623*
Peters, S. D., *2443*
Peters, T., 653, *664*
Peters, T. J., *1807*
Peterson, E. L., *2565, 2567*
Peterson, R., *1693*
Peterson, R. A., 630, *632*
Peterson, R. O., 926, *945*
Peterson, W. J., 953, 965, *970*
Petrozzo, D. P., 1704, *1716*
Peuranemi, A., *1107*
Pfeffer, J., 148, *152,* 861, *866*
Pflug, G. C., *2647*
Pflug, T. K., 191, *224*
Phadke, M. S., 2237, *2240*
Pheasant, S., 1043, 1045, 1047–1049, *1102, 1106*
Phelps, C., 1556, *1582*
Phelps, R. A., *2467*
Phillippe, D., *1767*
Phillips, C., *2442*
Phillips, C. B., *1127*
Phillips, C. L., 603, *618*
Phillips, D. T., *1583, 2539,* 2559, *2565, 2567, 2581*
Phillips, L. D., *2220*
Phillips, T., 2456, *2467*
Phizacklea, A., 1220, *1235*
Phuon, M., *1106*
Piersol, D., *1918*
Pierson, D., 901, *919*
Pil, F., 951, *973*
Pine, B. J., 685, 686, 701, *708,* 955, *973*
Pinedo, M., *1740,* 2045, *2052*
Pinedo, M. L., 1719, 1722, 1732, *1739, 1740*
Pinneau, S. R., *895*
Pinneau, S. R., Jr., *989*
Pinnell, J., *278*
Pipe, P., 925, *944*
Piper, J., 1564, 1566, 1568, 1569, *1582*
Piper Jaffray Equity Research, 2115, *2140*
Pirsig, R., 1794, *1806*
Pisanich, G., *2444*
Pisanich, G. M., 2437, *2442, 2444*
Pittenger, D. J., 937, *945*
Pitz, G. F., 2197, *2221*
Plante, R. D., *2648*
Plato, 213
Platzman, L., 802, *822*
Ploenzke, A. G., 195–198, *225*
Plotnicoff, J. C., *1739*
Plott, B., 2426, *2443,* 2458, *2467*
Plsek, P. E., 986, *993*
Pluym, B. V. D., 1777, *1789*
Podnar, G., *708, 1919*
Podsakoff, P. M., 851, *866*
Poe, V., 122, *152*
Pöhlau, F., 433, *445, 446*
Poirer, D. F., 11, 17, *24*
Poirier, David, 17
Polito, J., *2442*
Pollard, N. S., 1126, *1128*
Pollard, P., 2197, *2217*
Polterauer, A., *1295*
Pond, K., *352, 353*
Pongpatanasuegsa, N., 1053, *1103*
Pool, R., 146, *152*
Poole, M., 952, *970*
Poole, M. S., *154,* 952, *973*
Pope, M. H., *1102, 1128*
Pople, H. E., Jr., *1039*
Popovic, Z., 1126, *1129*
Popper, M., *866*
Porras, J., 656, *664*
Porras, J. I., 7, 8, 10, *24*
Porter, J. M., 1050, *1106*
Porter, L. W., 880, *897*
Porter, M., 702, *708*
Porter, M. E., 10, *24,* 33, *58,* 955, *973,* 2117, *2140*
Porter, M. J., 1120, *1129*
Porter, Michael, 33
Portillo Sosa, J., 1145, *1154*
Posavac, E. J., *2218*
Pottier, M., 1056, *1106*
Potting, J., 537, *542*
Potvin, J., 795, 796, 800, *823, 824,* 1076, *1106,* 1779, *1789*
Powell, J. D., 1885, *1885*
Powell, M., 2551, 2552, *2566*
Powell, M. J. D., 2561, 2562, *2566*
Powell, T. C., *1807*
Power, F. P., 1184, *1189*
Powers, J., *2443*
Prabhu, P., *1152, 1154,* 1909, *1917–1919*
Prabhu, P. V., *1917*
Praemer, A., 1082, *1106*
Prakash, A., 171, *176*
Prasad, B., 490, *529*

Prassana, V., 608, *618*
Pratt, D. B., *2351, 2405*
Prediger, D. J., 921, *945*
Preiss, K., *323,* 527, *528, 971*
Premeaux, S. R., *1583*
Premerlani, W., *1789*
Prenninger, J., *1295*
Prescott, M. B., *109*
President's Council on Sustainable Development, 533, *542*
Press, W. H., 1683, *1693*
Price, B., *2293*
Price, S. M., *1668*
Priel, V. C., *1109*
Priest, J. W., *1330*
Prieto, J., 1292, *1295*
Prince, C., *943*
Priouret, P., *2646*
Pritchard, J. P., 1697, 1710, *1716*
Pritsker, A., 505, *529*
Pritsker, A. A. B., 1080, *1106,* 2036, *2052,* 2449, 2454, 2455, 2458, 2461, *2467*
Pritsker, Alan B., 2494
Pritsker Corporation, 2461, *2467*
Probst, G., 213–215, 218, *225,* 1292, *1295*
Project Management Institute (PMI), 1242, 1243, 1248, *1251,* 1254, *1262,* 1266, 1272, *1277, 1350*
Promisel, D., *2441, 2444*
ProModel Corporation, 2459, *2467*
Prophet, W. W., 931, *945*
Proschan, F., 1932, 1935, *1954*
Prosser, P., 697, *706*
Provenmire, H. K., 932, *945*
Provost, L., *1827, 1855*
Provost, L. P., *1855, 2240*
Pruitt, D. G., 881, *897*
Prusak, L., 147, 148, *150, 152,* 213, 215, *224*
Pruzan, P. M., 2076, *2082*
Ptak, C. A., 2039, *2052*
Ptak, R. L., 229, *258*
Puchert, H., *542*
Pulat, B. M., 1131, 1132, 1145, *1152, 1154*
Punnett, L., *1104,* 1202, *1235*
Purba, S., 117, *152*
Purcell, D., 759, *771*
Purcell, J. A., *1039*
Purkiss, M., *1501*
Puterman, M. L., 2636, 2640, *2647, 2648*
Putnam, L., 145, *151*
Putz-Anderson, V., 1061, 1082, 1086, 1087, 1092, *1102, 1106–1108, 1129, 1154, 1155,* 1361, *1389*
Pyke, D. F., *1693*
Pyykonen, M., *1107*

Qiao, Jianhong, 539
Quaid, M., 902, *919*
Quercia, V., 228, *258*
Quesenberry, C. P., 1863, *1876*
Quick, J. H., *1462*
Quinlan, J., 1776, *1789*
Quinn, F., 785, *786*
Quinn, F. J., 2054, *2069*
Quinn, L. B., 1284, *1295*
Quinn, R., *993*
Quiroz, M., *2108*

Raar, D. J., *2053*
Raban, A., 979, *989*
Rabelo, L., 1778–1780, *1789*
Rabideau, G. F., 875, *897*
Rada, R., 1216, *1235*
Rademacher, L., *258*
Rafaeli, A., 979, *993*
Rafuse, M., 2116, *2140*
Rahman, S., 557, *561*
Rahn, A., 429, *446*
Raiffa, H., 129, *151, 152,* 2173, 2177, 2178, 2181, 2183, 2187–2189, 2191, 2193–2195, 2203, 2209, 2211, 2212, 2214, *2219–2221,* 2605, *2622*
Rainer, G., 192, *225*
Rajulu, S. L., *1105*
Ramadge, P. J., 2635, *2646*
Raman, N., *1789*
Ramaswamy, R., 635, *650*
Randall, R. M., *57*
Randhawa, S., 744, *750*
Randolph Hood, L., *770*
Rangaswami, M., 171, *176*
Ranney, T. A., *2220*
Rantzau, D., 1289, *1295,* 2512, *2520*
Rao, C. R., 2284, *2293*
Rapport, D., 146, *153*
Rardin, R. L., 2582, 2587–2589, 2594, *2600*
Raschke, U., 1119, 1126, *1129*
Rasmussen, J., 145, *152,* 930, *945,* 1014, 1019, 1020, 1023–1026, 1028, *1039,* 1145, *1154,* 1894, *1919,* 2176, 2205, 2214, *2221,* 2435, *2443*
Rathnow, P. J., 206, *225*
Ratliff, H. D., 2093, 2105, *2108, 2109*
Ratnam, C. S. V., 856, 862, 863, *866*
Raub, S., *225, 1295*
Ravden, S., 134, *153*
Ravindran, A., 2524, 2526, 2528, *2539,* 2621, *2622*
Ravlin, E. C., *866, 895*
Ray, S., 1783, *1789*
Ray, Steve, 352
Raymond, L., 1561, *1582*
Rayner, G. T., *941*
Rayport, J., 964, *972*
Raz, T., *1278*
Razouk, R. R., 174, *176*
RDG-376, 1925, *1955*
Rea, C. P., 931, *945*
Realff, M. J., 538, *542*
Reason, J., 1020, 1026, *1039,* 1145, *1154,* 1894, 1909, *1919,* 2206, *2221*
Reason, J. T., *2444*
Rebholz, M., *588*
Rebiffé, R., 1117, 1120, *1129*
Reboh, R., *2217*
Redding, R. E., 1028, *1039*
Redish, J. C., 1206, 1207, 1210, 1212–1216, *1232*

Redmond, W. H., 675, *683*
Ree, M. J., *947*
Reed, M. P., 1122, *1129*
Reeves, C. A., 625, 626, *632*
Reeves, C. M., 2552, *2566*
Reeves, C. R., 2590, *2601*
Reger, H., 418, *446*
Régie Nationale des Usines Renault (RNUR), 1145, *1154*
Reiche, H., *1583*
Reicher, G. M., *942*
Reichheld, F., *664*
Reichheld, F. F., 623, *632,* 641, *650*
Reick, A. M., *992*
Reid, J. K., *2538*
Reilly, Charles, 2494
Reiman, M. I., 2167, *2171*
Reinertsen, D., 1285, *1295*
Reinke, D. P., *540*
Reinwald, B., 203, *226*
Remick, H., 917, *919*
Remington, R. W., 2434, *2443, 2444*
Rempel, D., 1201, 1224, *1230, 1232, 1235*
Rempel, D. M., *1230*
Ren, D., *1918*
Renaud, P. E., *736*
Rentz, O., *542*
Resch, M., *1155*
Research Institute for Human Engineering for Quality of Life (HQL), *1130*
Rethans, A. J., 2197, *2221*
Reuter, B., *2109*
Reynolds, A., *57*
Reynolds, D. H., *944*
Reynolds, P. D., 957, *973*
Rezendes, C., 685, *708*
Rheaume, R. A., 914, *919*
Rice, D. P., *1106*
Rice, J. O., 1550, *1582*
Rice, J. R., *2293*
Rice, R., 952, 962, *973*
Rice, R. E., 952, *971, 972,* 1217, *1235*
Richardson, H. L., 2115, *2140*
Richer, C., *2441, 2444*
Richter, A. S., *918*
Richter, P., *1153*
Rickel, J. W., *1128*
Ridenhower, G. J., 2120, 2133, 2134, *2140*
Rieck, A., *944*
Rieke, R., *154*
Riepe, M. W., *771*
Rifkin, J., *1919*
Riggs, D., 775, 776, *786*
Riihimäki, H., 1070, 1071, *1106*
Rijckaert, M. J., 2559, *2566*
Riley, V., 2423, *2442, 2443*
Rim, K., 1126, *1128*
Ring, K., *2520*
Ringdahl-Harms, K., *1236*
Rinnooy Kan, A., *823, 2069*
Rinnooy Kan, A. H. G., *1739, 1740*
Riviere, J. W. M. L., 537, *542*
Rivlin, A. M., *353*
Rob, P., 80–82, *109,* 117, *153*
Robbins, H., 2634, *2648*
Robbins, S., 775, 776, *786*
Robert, H. M., 2213, *2221*
Roberts, C., *1010*
Roberts, C. A., *176,* 2464, *2467*
Roberts, E. B., 685, 686, *708*
Roberts, H. K., *974*
Roberts, L., 1702, 1703, *1716*
Roberts, S. D., 745, *750*
Roberts, T. L., 134, *153*
Robertson, G., 1206, *1236*
Robey, D., 94, *107,* 952, *973,* 1226, *1235*
Robinette, K. M. M., 1049, *1106*
Robinson, S. M., 2635, *2648*
Rochat, Y., 800, *823*
Rockafellar, R. T., 2561, *2566*
Rodahl, K., 871, 876, *894*
Roddick, Anita, 843, 846, 849
Rodgers, S. A., 1364, *1389*
Rodrigues, J. M., 2446, 2461, *2467*
Rodriguez-Clare, A., 602, *618*
Roebuck, J. A., 1043, *1106,* 1113, *1129*
Rogers, A., 1900, *1919*
Rogers, D. F., 2645, *2648*
Rogers, E., 962, *973*
Rogers, K. J. S., *1234*
Rogers, M., 662, *664,* 704, *708*
Rogers, P., 487, *529*
Rogers, T. R., *1583*
Rohmert, W., 1050, 1056–1058, *1106, 1109,* 1117, 1119, *1129,* 1132, 1138, 1144–1146, *1153, 1154*
Rohrbaugh, J., 2212, *2218*
Röhrbor, D., *1295*
Rohrer, D., *1920*
Rohrer, K. N., *1235*
Rohrer, M., 2461, *2467*
Rohrer, M. W., *946*
Rolfes, J. D., 1698, *1715*
Roll, Y., *1526*
Rollier, D. A., *176, 2392*
Rollins, D., 984–987, *993*
Romanin Jacur, G., *748*
Rombach, V., 1050, *1106*
Romhardt, K., 213, *225, 1295*
Rommel, G., 648, *650*
Ron, A. D., 538, *542*
Ron, A. J. D., 538, *542*
Roos, D., *561*
Roos, G., *153*
Roos, J., 147, 148, *153*
Rosaler, R. C., 1550, *1582*
Rösch, A., *2519*
Roscoe, S. N., 932, *945*
Rose, D., *399*
Rose, M. T., 228, *258*
Rosemau, R. D., 1902, *1919*
Rosen, B., *1806*
Rosen, D. W., *708*
Rosen, J. B., 2560, *2566*
Rosen, J. C., *1102*
Rosenau, M. D., Jr., *58*
Rosenbaum, H. F., 689, *708*
Rosenblatt, M. J., 1524, *1526,* 2093, *2108*

Rosenbloom, R. S., 955, *970*
Rosenbrock, H. H., 2550, *2566*
Rosenkrantz, D. J., 1724, *1740*
Rosenstein, R., 986, 987, *993*
Rosenthal, A. S., 2105, *2109*
Rosenthal, J., *971*
RosettaNet, 343, *352*
Rosiello, R. L., 677, *683*
Rospond, K. M., 1574, *1582, 1583*
Ross, D. T., 508, *529*
Ross, J. R., 1567, 1571, *1583, 1584*
Ross, L., 2200, 2201, *2220*
Ross, R., *1010*
Ross, S. M., 1679, 1691, *1693,* 2146, 2162, 2166, *2171,* 2636, 2640, *2648*
Rosselot, K. S., 531, 533, 534, *540*
Rossett, A., 926, *945*
Rössler, A., 2516, *2520*
Roth, E. M., 1024, 1028, 1032, *1039, 1040*
Roth, J. T., 2427, *2443*
Roth, N., *445*
Rothenburg, U., *225*
Rottbauer, H., 402, 406, *445*
Rouse, W. B., 111, 114, 116, 126, 136, 146, 147, 149, *153,* 928, *944,* 1206, *1235,* 1297, 1298, *1310,* 1909, *1919*
Rousseau, D. M., 874, *897*
Rousseau, J., 795, 796, *823, 1789*
Roustang, G., *1153*
Rouwenhorst, B., 2084, *2109*
Roy, B., 2604, 2614, 2621, *2622*
Roy, K.-P., *650*
Roy, M., *529*
Rubinstein, R. Y., 2634–2636, *2648*
Ruckelshaus, W. D., 590, *600*
Rudas, I. J., *399*
Ruffner, J., 931, *945*
Ruggeberg, B. J., *990*
Ruggles, R. L., 220, *226*
Ruggles, R. L., III, 147, 148, *153*
Rule, J., 1225, *1230*
Rumbaugh, J., 1774, 1785, *1789*
Rumelhart, D., 1778, 1779, *1789*
Rumelhart, D. E., 930, *945*
Rummel, W. D., 1909, *1919*
Rummler, G. A., 925, *945*
Rupp, M., *588*
Ruppert, D., 2635, *2646, 2648*
Russek, H., 1179, *1189*
Russell, C. J., 856, 857, *866,* 923, *945*
Russell, C. S., 538, *542*
Russell, E. C., 2455, *2467*
Russell, R., 795, *823*
Russell, R. G., 1284, *1295*
Rust, R., 652, 654, *664*
Rust, R. T., 623, 628, 629, *632*
Ruszczynski, A., *2647*
Rutenfranz, J., *1153, 1154*
Ryan, G. A., *1109*
Ryan, N., 685, *708*
Ryan, P. W., *1106,* 1122, *1129*
Rybko, A. N., 2167, *2171*
Ryder, J. M., *1039*
Ryding, S. O., 531, *542*
Rynes, S., 916, *919*
Rzehak, H., 168, *175*

Saaty, T., 2214, *2215*
Saaty, T. L., 1899, *1919,* 2074, *2082,* 2173, 2183, 2192, 2195, 2214, *2221,* 2605–2608, *2622*
Sackett, G. C., 2572, *2581*
Sackett, P. R., 921, *945*
Sadiq, M., 2093, *2109*
Sadowski, D., 2456, *2467*
Sadowski, D. A., *2466*
Sadowski, R. P., *2466, 2467*
SAE International RMS Committee, *1955*
Safayeni, F., *1740*
Sagar, A., *542*
Sage, A., 2184, 2214, *2221*
Sage, A. P., 111, 112, 114, 116, 122, 126, 128, 139, 141, 146, 147, 149, *151, 153*
Saha, J., 2538, *2539*
Sahinoglu, M., *1789*
Sahlman, W. A., 602, *618*
Sahney, V., 745, 747, 748, *749, 750*
Sahney, V. K., 738, *750*
Sainfort, F., *991, 993*
Sainfort, P. C., 1159, 1170, *1189,* 1194, 1217, 1224, 1228, *1234, 1235*
Saipe, A. L., *2108*
Sakawa, K., 2620, *2622*
Sakawa, M., *2622*
Salas, E., *897,* 933, 934, *941–943, 945–947, 971,* 984, *993,* 2208, 2209, *2217, 2221, 2443*
Salazar, A., 2400, *2405*
Salkin, H. M., 2538, *2539*
Salmoni, A. W., 929, *945*
Salomon, M., 538, *543*
Salon, R., *970*
Saltzman, M., *2539*
Salus, P. H., 228, *258*
Salvendy, G., 871, 874, 875, *897,* 983, *990, 1038,* 1042, 1043, *1103,* 1193, 1217, 1228, 1229, *1231–1233, 1236,* 1405, *1408, 1584,* 1895, *1918, 1919, 2220*
Salvi, L., *2441, 2444*
Samet, M. G., *1038*
Saminathan, M., *600*
Sampson, S. E., 636, *650*
Samuelides, M., *1787*
Samuelson, B., 1061, *1107*
Samuelson, P. A., 344, *352*
Sanchez, P., *749*
Sanders, K., *1236*
Sanders, K. J., *1235*
Sanders, M., *1109*
Sanders, M. S., 871, 875, *897,* 1016, 1027, *1038*
Sanderson, P. M., *1039*
Sanderson, S., 689, *708*
Sandin, D. J., *2519*
Sandry, D. L., *2444*
Santi, J., 532, *542*
Santner, T. J., *2239, 2494*
Sappington, D. E., 263, *279*

Sarhadi, M., 602, *619*
Sarin, S., 1222, *1232*
Sarker, B. R., 1524, *1526*
Sarter, N., 963, *973*
Sasser, W., *664*
Sasser, W. E., *650*
Sasser, W. E., Jr., 623, *632*
Satava, R. M., *2520*
Sathe, P., *1917*
Saunders, C. E., 745, *750*
Saunders, M. A., *2538*
Saunders, Michael A., 2564, 2565
Saunders, P. B., *2580*
Sauser, W. I., 914, *918*
Sauter, S. L., 1222, *1235, 1236*
Savage, D., 599, *600*
Savage, L. J., 2173, 2178, 2180, 2182, 2184, 2202, *2221*
Savard, G., *1129*
Savelsbergh, M., 794, 800, *823, 824*
Savelsbergh, M. W. P., *2647*
Savitch, W., 72, *108*
Sawaragi, Y., *2624*
Saxena, M., 494, *529*
Scarf, H., 1678, *1693*
Schaefer, S. K., 1524, *1526*
Schaffer, B., *1501*
Schaffer, J., *824*
Schaffer, R., *1807*
Schallock, B., *225*
Scharlacken, J. W., 2112, *2140*
Schaub, K., 1050, *1106*
Scheck, D. E., 473, *483*
Schecter, M., 826, *836*
Scheder, H., 193, *226*
Scheepers, F., 1126, *1129*
Scheer, A.-W., *307,* 507, 512, 513, *529*
Scheidt, L.-G., *541*
Schein, E. H., 15, *24,* 938, *945,* 956, 957, *973*
Scheipers, P., *588*
Scheller, H., *445*
Schelling, T., 2210, 2211, *2221*
Schenbach, T. E., *2351*
Scherkenbach, W. W., 938, *945*
Scherrer, J., 1056, *1105*
Schiettekatte, J., *1102*
Schillling, W., 416, *446*
Schilperoort, B. A., *483*
Schittkowski, K., 2564
Schlecht, L., *2519*
Schleifer, L. M., 1222, *1235*
Schlesinger, L. A., *650*
Schlie, T. W., 955, *973*
Schlopp, W., *542*
Schmeiser, B., 2494, *2495*
Schmeltzer, J., 671, *683*
Schmenner, R. W., 638, *650, 1668*
Schmidt, A., 228, 229, *258*
Schmidt, F., 923, *945*
Schmidt, K.-H., *1153, 1154*
Schmidt, R. A., 929, *945*
Schmidt, S. R., *2240*
Schmierer, G., *399, 400*
Schminke, M., *895*
Schmittberger, R., *2218*
Schmoeckel, D., *588*
Schneeweis, T., 759, *771*
Schneidemans, J., 559, *559*
Schneider, B., 625, *633*
Schneider, L., 951, *970*
Schneider, N. L., *2466*
Schneider, R., 214, *224*
Schneider, V. I., 929, *945*
Schneiderman, B., 132, 133, *153*
Schoenfeldt, L. F., *865*
Schoenmarklin, R. W., 1092, *1107, 1109*
Scholpp, C., *399*
Scholtes, P. R., 982, 986, *993*
Scholz-Reiter, B., 193, *226*
Schon, D., 1024, *1039*
Schönbach, T., 207, *226*
Schonberger, R. J., *1807*
Schoonhard, J. W., 1896, *1919*
Schor, J. B., 1888, *1919*
Schorn, E. C., *2108*
Schraft, R. D., *399, 400*
Schrage, L. E., 1728, *1739, 2494*
Schragenheim, E., 2039, *2052*
Schramm, W., 745, 746, *750,* 928, *945*
Schramm, W. S., *749*
Schreckenghost, D. L., *1039*
Schreiber, G., 213, 214, 218, *226*
Schreyer, M., 699, *708*
Schriber, T. J., 2455, *2467*
Schrijver, A., 2582, *2601*
Schroeder, R., 1368, *1390*
Schroeder, R. G., *894*
Schroeder, W. H., 932, *943*
Schroyer, D., 745, *750*
Schruben, L., 2478, 2482, 2494, *2494, 2495*
Schruben, L. W., 2480, 2482, 2491, 2492, *2494, 2495*
Schuh, G., *224*
Schuler, R. S., 855, *865*
Schulman, M. A., 229, *258*
Schulte, H., 207, *225*
Schulte, J., *400*
Schultetus, W., 1117, *1129*
Schultz, A., 1061, *1107*
Schultz, R., *58,* 2630, *2648*
Schum, D. A., 137, *153*
Schurman, D., 1909, *1920*
Schustack, M. W., 1023, *1039*
Schutte, L., *1129*
Schuur, P. C., *541*
Schvaneveldt, S. J., 559, *559, 561*
Schwab, D., 916, *918, 919*
Schwab, D. P., 914, *919*
Schwab, John L., 1429
Schwab, R. E., 2446, 2447, 2454, *2467*
Schwartz, K. D., 1703, *1716*
Schwarz, F., *2109*
Schwarz, L. B., *1526, 2108*
Schwarze, B., *2218*
Schweiger, C., 2564
Schweiger, D. M., *992*
Schweitzer, P. J., 2645, 2646, *2648*
Scott, B., *353*

Scott, C. A., 680, *683*
Scott, C. D., 2126, *2140*
Scott Morton, M. S., 136, *151*
Seaborne, A. E. M., 1899, *1920*
Seabright, M. A., *2215*
Seamster, T. L., 1028, *1039*
Seashore, S. E., *993*
Seeber, A., 1138, *1154*
Seering, W. P., 689, *708*
Segal, M., 1752, *1767*
Sego, D. J., *896, 943*
Seidl, A., 1122, *1129*
Seidl, J., 2103, *2109*
Seidler, K. S., *1039*
Seidman, A., 2646, *2648*
Seidmann, A., 1524, *1526*
Seifert, D. J., *2444*
Seiford, L., 2618, 2619, 2621, *2622, 2623*
Seilstorfer, H., *588*
Seip, H., *107*
Sekutowski, J. C., *600*
Selen, W., 518, *528*
Seliger, G., 441, 442, *446*
Selin, K., *1233*
Selke, S., *540*
Selkirk, C. G., 1750, 1751, *1767*
Sell, S., *1716*
Sellers, J. D., *543*
Sellers, R. L., *540*
Sellie, C. N., 1428, 1433, 1446, *1462*
Sells, S. B., *991*
Seltzer, J., 843, 845, 849, 851, *866*
Selye, H., *2221*
Semiconductor Equipment and Materials International (SEMI), 165, *176*
Sena, M., 107, *108*
Senge, P., 999, *1010*
Senge, P. M., 22, 24, *24*
Sennott, L. I., 2636, 2640, 2645, *2648*
Serafini, P., 2614, *2622*
Serfaty, D., 1016, *1039*
Serfozo, R. F., *2108,* 2643, *2648*
Seshadri, R., 1777, *1789*
Seubert, Michael, 352
Seymour, J. W. D., 1895, *1919*
Seymour, W., 1405, *1408*
Shadbolt, N., *226*
Shafer, G., 2173, 2186, *2221*
Shafer, R. E., *1955*
Shahraray, M. S., *2052*
Shamanna, S. K., *2109*
Shamir, B., 842, 845–848, 853–855, 858, 862, *865, 866*
Shamlou, P. A., 1504, *1526*
Shamp, M. J., 984–987, *992*
Shani, R., *1807*
Shankar, R., *2442, 2444*
Shanmugham, S. G., 165, *176*
Shan-No, D., *2539*
Shanno, D. F., 2552, *2566*
Shanno, David, 2535
Shannon, R. E., *2467*
Shanthikumar, J. G., 1640, 1652, 1655, 1656, 1660, 1668, *1668,* 1669, 1691, 1692, *1693,* 2146, 2150, 2168, *2171*
Shapiro, A., 2634–2636, *2646–2648*
Shapiro, C., *58,* 147, *153,* 350, *352*
Shapiro, J., *943*
Shapiro, J. F., *353,* 2043, *2052, 2567*
Sharit, J., 1223, *1231, 1917, 1920*
Sharp, G. G., 2088, *2109*
Sharp, G. P., 2084, 2089, 2092, 2093, 2107, *2108, 2109*
Sharp-Bette, G. P., *2351,* 2370, 2372, 2375, *2392,* 2394, 2403, *2405*
Sharpe, W., 768, *771*
Sharplin, A., *1583*
Shaw, J. B., *865*
Shaw, M., 1776, *1789*
Shaw, M. E., 880, *897,* 933, *946*
Shaw, M. J., 697, *708*
Shaw, M. J. P., 697, *708*
Shaw, R. B., *1010*
Shawver, D. M., 2499, *2520*
Shea, G. P., *865,* 870, 877, *896*
Shea, J. F., 929, *946*
Sheehan, J. J., 1898, *1919*
Sheina, M., 222, *226*
Shen, Y., *1789*
Sheng, P., 538, *542*
Shenk, D., 2013, *2019*
Shepard, J., *2019*
Shepard, J. M., 874, *897*
Shepherd, A., 1028, *1039*
Sherali, H. D., *2565, 2567*
Sheridan, J. J., 1899, *1919*
Sheridan, T., 1023, *1039*
Sheridan, T. B., 112, *153,* 1208, 1211, *1235*
Sherif, M., *1100*
Sherman, H., *58*
Sherman, J. D., 956, 957, *973*
Sherwood-Jones, B. M., 1209, *1234*
Sheth, A., *528*
Shetty, B., *2580*
Shetty, C. M., 2530, *2538, 2565, 2567*
Shewhart, W. A., 1828, 1830, 1834, 1835, 1837, *1855,* 1861, *1876*
Shi, Y., 2620, *2622*
Shibano, T., 2620, *2622*
Shieh, J. S., *1526*
Shih, H. S., 2620, *2622*
Shin, C., *1189*
Shin, W. S., 2621, *2623*
Shina, S. G., 1312, 1313, *1330*
Shindo, W., 1890, *1919*
Shingo, S., 545, 547, 548, *561*
Shirose, K., 553, *561*
Shively, J., *2443*
Shively, R. J., 2431, 2436, *2442, 2443*
Shmoys, D., *823, 2069*
Shmoys, D. B., *1739*
Shneiderman, B., 1018, *1039,* 1212, *1235*
Shoaf, C., *1109*
Shodhan, R., 1777, *1790*
Shoemaker, C. A., 2646, *2646*
Shor, N. Z., 2562, *2566*
Short, J., 1696, *1715*
Shortell, S. M., *989, 993*
Shortliffe, E. H., 2187, *2218*
Shostack, G. L., 624, *633,* 641, *650*

Shreve, S. E., 2636, *2646*
Shtub, A., 1242, 1245, *1251, 1278*
Shuell, T. J., 931, *941*
Shutter, J., *278*
Sibik, L. K., *540*
Siebold, D. R., 2213, *2221*
Sieburg, H. B., *2520*
Siegel, A. I., 2423, *2443*
Siegel, J., *1233*
Siegel, L. B., *771*
Siegel, M., 1902, *1919*
Siemens AG, 425, *446*
Siemieniuch, C. E., 1888, *1920*
Sier, D., *2465*
Siha, S., 559, *561*
Siha, S. M., *560*
Sihn, W., 318, *323*
Sikora, R., 697, *708*
Silbert, J., 2507, *2520*
Silver, E. A., 1669, *1693*
Silverstein, B., *1233,* 1366, *1389*
Silverstein, B. A., 1088, *1100, 1104, 1107, 1153*
Silvestro, R., *1668*
Sim, S. K., 1779, *1789*
Simchi-Levi, D., 94, 95, *109,* 2010, 2012, 2014, 2019, *2019*
Simchi-Levi, E., *109, 2019*
Simeone, B., 2575, *2581*
Simmons, L., *2220*
Simon, A., *2519*
Simon, H., 676, *683,* 1024, 1026, 1027, *1037, 1039*
Simon, H. A., 135, *153,* 1209, *1232,* 2180, 2206, 2208, 2214, *2217, 2220, 2221,* 2410, *2443*
Simon, R., *918*
Simpson, D. L., *1918*
Simpson, G. C., 1145, 1150, *1154*
Sims, E. R., Jr., *1501*
Sims, H. P., 889, *897, 993*
Sinclair, M. A., 1136, 1145, *1155,* 1888, 1891, 1893, 1898, 1912, 1916, *1917, 1920*
Singer, B. D., *770*
Singer, M., 147, *151,* 327, *352*
Singer, M. J., 932, 933, *942*
Singer, T., 1622, *1623*
Singh, H., 83, *109, 2495*
Singh, M., 107, *108*
Singh, N., 488, *529*
Singhal, S., 229, *258*
Singhal, V. R., *1807*
Singleton, W. T., 1892, *1920,* 2202, *2221*
Singpurwalla, N. D., *1955*
Sinha, A., *1807*
Sinha, K. K., *894*
Sinioris, M. E., *992*
Sink, D. S., 11, 15, 22, *24, 25*
Siprelle, A. J., 2460, *2467*
SISCO, Inc., 165, *176*
Sistrunk, F., 931, *947*
Sitompul, D., 744, *750*
Sivasubramaniam, N., 855, 856, 862, 863, *866*
Sjogaard, G., *1100,* 1119, *1129*
Skapura, D. M., 163, *175*
Skogan, W. K., *994*
Skyrme, D. J., 148, *153,* 213, *226*
Slany, W., 1781, *1789*
Slaughter, J., 988, *993*
Sloan, F., 738, *750*
Slovic, P., *971,* 2192, 2196, 2199, 2200, *2217, 2219–2221*
Sly, D. P., 171, *176*
Smalley, H. E., 739, *750*
Smith, A., 870, *897*
Smith, B., *1010,* 2431, *2442, 2444*
Smith, D., *560*
Smith, D. E., 143, *153*
Smith, D. K., 975, 978, 988, *991,* 1001, 1006, 1007, 1009, 1010, *1010*
Smith, G. B., 984, 987, *993*
Smith, G. L., 22, *24*
Smith, G. W., 2350, *2351*
Smith, H., *2293*
Smith, J. D., *1547, 1623*
Smith, J. L., 1072, *1105, 1107, 1109*
Smith, J. M., 2446, *2467*
Smith, J. S., *1526*
Smith, K. D., 1250, *1251*
Smith, K. U., *1189*
Smith, K. V., 2318, 2320, *2330*
Smith, L. D., *1766*
Smith, L. M., *895*
Smith, M., 1182, *1190*
Smith, M. J., 874, *897, 991, 993,* 1159, 1161, 1170, 1171, 1178, 1183, 1184, 1186, *1188–1190,* 1193–1195, 1197, 1198, 1200–1202, 1204, 1217, 1224–1226, 1228, *1231, 1233–1236*
Smith, P., *483*
Smith, P. G., *1188*
Smith, R. A., 2044, *2052*
Smith, R. E., 229, *258*
Smith, R. L., *2646*
Smith, R. P., 1056, *1101*
Smith, S. F., 2044, *2053*
Smith, S. L., 132, 133, *153*
Smith, T. J., 1184, *1190*
Smith, W. E., 1723, *1740*
Smulders, P. G., *993*
Smyth, G., 1072, *1102*
Smythe, C., 228, *258*
Snedecor, G. W., 2255, *2263*
Snee, R. R., 2275, *2293*
Snehota, I., 2118, 2123, *2139*
Snelgar, R. J., 914, *919*
Snell, S. A., 856, *867*
Sniezek, J. A., 2193, 2212, *2221, 2222*
Snijders, C. J., 1061, *1107*
Snook, S. H., 1055–1058, 1071, 1072, 1089, 1091, *1101, 1107, 1110,* 1117, 1118, *1128*
Snowdon, J. L., 2456, *2467*
Snyder, K., 2036, *2052*
So, R. H. Y., *708*
Soares, C. G., *1955*
Sobel, M. J., 2643, *2647*
Socolow, R., 533, *542*
Socrates, 213
Sodhi, M., 538, *542*
Sofer, A., *2567*

Sohal, A. S., 545, *561*
Sohn, J. E., *600*
Solberg, J. J., 697, *707, 2539*
Solid Waste Technology Staff, *1584*
Solomon, I., *57*
Solomon, M., *540,* 795, 800, *824*
Solomon, M. M., *823*
Solomon, M. R., 624, *631, 633*
Somerville, M. A., 146, *153*
Sommerville, R. M., 534–537, *543*
Song, J. S., 1692, *1693*
Song, Y., *225*
Sorenson, P. F., *991*
Sosik, J. J., 851, 855, *866*
Soumis, F., *823*
Sounder, W. E., 956, *973*
Soutar, C. A., *1104*
Southwick, W. O., *1104*
Sparks, D., 671, 672, *683*
Sparks, L., *785*
Spear, N. E., *944*
Spear, S., 551, *561*
Spearman, M. L., 2037, 2042, *2052, 2053,* 2168, *2171, 2646*
Spears, S., 16, *24*
Speh, T. W., 2071, 2079, *2082*
Spellucci, P., 2563
Spencer, F., 1909, *1920*
Spencer, F. W., 1909, *1917*
Spencer, M. S., *353, 541,* 557, *561*
Spendley, W., 2550, *2567*
Spengler, D. M., *1100, 1101*
Spengler, D. M. J., 1070, *1107*
Spengler, T., 538, *542*
Sperry, B., *993*
Spingler, J., *399*
Spitzer, D. R., 938, *946*
Spooner, R. L., *1038*
Sprague, R. H., 2079, *2082*
Sprague, R. H., Jr., 113, 114, 125, *153*
Sprengel, A., *107*
Springer, W. E., 1122, *1129*
Sprott, D., *352*
Sproull, L., 1217, 1220, 1225, *1236*
Spur, G., 182, 218, *226,* 407, 413, *446, 588*
Spurgin, R., 759, *771*
Srikanth, M., 557, *561*
Srinivasan, M. M., 1525, *1525, 1526*
Srinivasan, S., 277, *279*
Sriskandarajah, C., *1525*
Srivastava, R., *541*
Stacey, R. D., 147, *154*
Stadler, W., 2602, 2621, *2621, 2623*
Staffon, J. D., 1021, *1038*
Stahl, D. O., *279*
Staines, G. L., *993*
Stalick, S. K., 2126, *2138*
Stalk, G., 1000, *1010*
Stallaert, J., *279*
Stambough, J., *1108, 1109*
Stamm, C. L., 545, *560*
Stammerjohn, L. W., *1235*
Stammwitz, G., *226*
Standard & Poor’s Statistical Service, 2395, *2405*
Standring, P., *588*
Stanek, J., 192, *226*
Stanke, A., 636, *650*
Stanney, K. M., 1193, *1236*
Stanoulov, N., 2184, 2195, *2222*
Stanton, N., 1209, *1236*
Stanton, S., *353*
Star, S. L., 141, *151*
Starbuck, W. E., 140, *154*
Starkweather, T., 1781, *1789*
Stasser, G., 2212, *2222*
Steber, M., 410, *446*
Stedman, C., 86, 88, 92, 93, *109*
Steel, R. P., *993*
Steele, D., *1806*
Steele, W., 1562, *1583*
Steen, B., *542*
Steenkamp, J. B. E. M., 625, *633*
Steers, R. M., 870, 874, *897*
Stefanacci, E. F., *599*
Steffens, K., *588*
Stefik, M., 142, 144, *154*
Stegemerten, Gustave J., 1429
Steiglitz, K., 2582, *2600*
Stein, T., 91, *108*
Steiner, I. D., 877, *897,* 933, *946*
Steiner, P. O., 626, *631*
Steiner, V. M., 1565, *1583*
Steinman, J. S., *2468*
Steinmann, D. O., *2218*
Steinmetz, V., 1900, 1902, 1906, *1920*
Steinwelds, B., 738, *750*
Stenger, A. J., 1675, *1693*
Stephan, M., 198, *225*
Stephan, P. E., 602, *618*
Stephanidis, C., 1217, *1236*
Stepper, J. C., 1704, *1716*
Stern, A. A., 677, *683*
Stern, E. K., *2218*
Stern, L. W., 2117, 2118, *2140*
Sternberg, R. J., 1023, *1039*
Sterns, R. E., *1740*
Stetson, D. S., *1104, 1153*
Steudel, H. J., 478, *483*
Steuer, R., 2074, *2082*
Steuer, R. E., *2624*
Stevens, C. K., 916, *919*
Stevens, G. C., 2112, *2140*
Stevens, M. J., 884, *894*
Stevens, S. S., 1071, *1107, 1110*
Stevenson, M. K., 2195, 2196, 2200, 2202, *2222*
Stewart, D. M., 559, *559*
Stewart, J. G., *2003*
Stewart, L. C., 229, *258*
Stewart, R. A., 2562, *2566*
Stewart, R. D., *2316*
Stewart, T. A., *58,* 147, *154*
Stewart, T. F. M., *1152*
Stewart, T. R., *2218*
Stigler, G. E., 2123, *2140*
Stillwell, H. R., 1315, *1330*
Stock, J. R., 2110
Stockram, V., *2109*
Stodgill, R. M., 938, *946*

Stoer, J., 2562, *2567*
Stöferle, T., 407, 413, *446*
Stogdill, R. M., 882, *897*
Stohr, E. A., 131, *152*
Stok, T., 1900, *1920*
Stolyar, A. L., 2167, *2171*
Stone, M., 2205, *2222*
Stork, K., 1703, *1716*
Stoughton, S., 349, *352*
Stougie, L., *2648*
Stout, R., 228, *258*
Stranden, E., *1108*
Strasser, H., 1058, 1060, *1107*
Strassman, P. A., 1710, *1716*
Strat, T. M., 2186, *2222*
Stravinskas, J., *993*
Streufert, S., 922, *946*
Stringer, D. Y., *864*
Stroebe, W., *895*
Strojwas, A. J., *1919*
Stroud, J., 87, 90, 95, *109*
Stroustrup, B., 72, *109*
Strub, M., *2217*
Strub, M. H., *2443*
Strutt, J. E., 1909, *1918*
Stuart, J. A., 531, 532, 534–539, *542, 543*
Stubler, W. F., *2443*
Stukey, E., 2178, *2222*
Stutts, A. S., 835, *836*
Su, C. J., *708*
Suarez-Balcazar, Y., *2218*
Subrahmanian, E., 687, *708*
Subramanian, R., *2539*
Suchman, L., *154*
Sugimori, Y., 545, *561*
Suh, C. J., *1739*
Suh, N. P., 686, 695, *708*
Sullivan, A. C., 2318, *2330*
Sullivan, D. W., 2488, *2495*
Sullivan, R. S., *1739*
Sullivan, W. G., *2351, 2393,* 2401, *2405*
Sullo, G. C., 124, *154*
Summitt, Pat, 843, 844
Sun, J., *1918, 2567*
Sun, T., *824*
Sun, Y., 1779, *1790*
Sundad, H., *2622*
Sundararajan, K., 1056, *1103*
Sunday, C., *2520*
Sundelius, B., *2218*
Sundstrom, E., 877, 882, *897*
Suokas, J., 1173, *1190*
Supel, T. M., 914, *919*
Supply Chain Council, *353*
Suri, R., 2168, *2171*
Surprenant, C. F., *631, 633*
Sutherland, I. E., *2520*
Sutjana, D. P., 981, *993*
Sutton, R., 1780, *1790, 2443*
Sutton, R. I., 148, *152*
Sutton, R. S., 2646, *2648*
Suurnakki, T., *1107*
Suwa, K., *1918*
Suydam, M. M., *2220*
Suzue, T., 689, *708*
Suzuki, K., *400*
Suzuki, T., 553, *561*
Sveiby, K. E., 148, *154*
Svenson, O., 2173, 2205, 2208, 2209, *2217, 2222*
Svensson, H.-O., 1071, *1107*
Svetkoff, D., 1904, *1920*
Swain, A. D., 2192, *2222*
Swain, J. J., *2494*
Swan, J. E., *631*
Swanson, N. G., 1202, *1236*
Swartz, T. A., 623, *633*
Sweat, J., 955, *973*
Swenson, D. F., 759, *771*
Swerdlow, A. J., *1188*
Swets, J. A., 1016, *1037,* 2185, *2222*
Swezey, L. L., *946*
Swezey, R. W., 922–924, 926, 928, 930–933, 936, 940, *946*
Swift, K. G., *483*
Swigger, K. M., 1734, *1739*
Sy, J. N., *2465*
Sydenham, P. H., *1886*
Sydow, M., *1189*
Syed, J. R., 220, 221, *226*
Sykes, R., 95, *109*
Symix Systems/Pritsker Division, 2454, 2458, *2467*
Syslo, M. M., *2581*
Szabo, B. A., *226*
Szabo, S., *2442*
Szapiro, T., 2621, *2622*
Sze, Cedric, 536, *543*
Szigeti, F., *1582*
Szilagyi, A. D., *897, 993*

Taboun, S. M., *1110*
Tabrizi, J. L., 2196, 2197, *2217*
Taghaboni, Fataneh, 2494
Taguchi, G., 498, *529,* 1889, *1920,* 2237, *2240*
Taillard, E., 800, *823, 824*
Takahashi, D., 1193, *1236*
Takeda, E., 2605, 2608, *2623*
Takeda, H., *400*
Takefuji, Y., 1778, *1788*
Takeuchi, H., 148, *152,* 214, 215, *225,* 1293, *1295*
Takeuchi, Y., *400*
Talalayevsky, A., 2056, *2069*
Talavage, J., 501, 506, 507, *529,* 1777, *1787*
Talavage, J. J., 1777, *1787, 1790*
Tamais, G., *2442*
Tamesh, L., *771*
Tamhane, A. C., 2490, *2494*
Tamura, S., *707*
Tanaka, S., 1084, 1085, *1107*
Tanchoco, J., *2109*
Tanchoco, J. M. A., *1526*
Tanchoco, Jose M., 2360
Tang, T. L., 981, *993*
Tanino, T, *2624*
Tannenbaum, S. I., 870, 877, *897,* 929, *945, 946, 993*
Tanner, W. P., 2185, *2222*
Tardif, V., 2042, *2053*

Tarquin, A. J., *2351*
TASS Investment Research, 759, *771*
Tasto, D., *1189*
Tattershall, A. J., *2442*
Tatum, R., 1550, 1566, *1583*
Tausz, A., 2058, 2068, *2069*
Taveira, A., *991*
Taveira, A. D., 982, *993*
Taylor, D., 350, *352*
Taylor, D. A., *2140*
Taylor, E., 228, *258*
Taylor, F. W., 871, 874, *897*
Taylor, Frederick, 20
Taylor, G. D., *2109*
Taylor, J. C., 869, 870, 884, *895*, *897*, 964, *973*, 1889, 1899, *1920*
Taylor, J. R., *1885*
Taylor, R. H., *400*
Taylor, S. A., 628, 630, *631*
Tayur, S., 2034, *2053*
Tchernev, N., 2464, *2466*
Teather, D. C. B., 928, *946*
Tebbets, I. O., *1128*
Tecnomatix Technologies, 171, *176*
Teel, K. S., 1896, *1916*
Temple, Barker, and Sloane, Inc., 2057, *2069*
Templeton, J. G. C., *1693*
Teng, J. T. C., *1715*, *1716*
Teresko, J., *708*
Terhune, A., 347, *352*
Terrell, M. E., *836*
Terzopoulos, D., *1128*
Tesauro, G., 1780, *1790*
Tetreault, M. S., *649*
Tetzlaff, U. A. W., 501, *529*
Teukolsky, S. A., *1693*
Thacker, S. B., *1104*
Thalemann, J., *587*
Thalmann, D., *2520*
Thalmann, N. M., *2520*
Thangiah, S., 800, *824*
Thayer, P. W., 870, 871, 875, 883, 892, *894*, 1217, *1231*
Thein, B., *2442*
Theodore, L., 537, *543*
Theodore, M. K., 537, *543*
Theorell, T., 1224, *1236*
Thiagarajan, S., 938, *946*
Thimbleby, H., 133, *151*
Thomas, C., 833, *836*
Thomas, H., *57*, *2393*
Thomas, J. C., 1213, *1231*
Thomas, J. P., *2442*
Thomas, L. F., 1899, *1920*
Thomas, L. J., *2052*
Thomas, P., 1289, *1295*
Thomas, R. J., *58*
Thomasson, C., *993*
Thomke, S., 1285, 1288, *1295*
Thompson, J. D., 4, *24*, 880, *897*
Thompson, K. R., *989*
Thompson, T. J., 2577, *2581*
Thomson, D. M., 930, *946*
Thomson, H., *1807*
Thomson, W. G., *1106*
Thorndike, E. L., 932, *946*
Thorndike, R. L., 921, *946*
Thornton, D. C., 1900, *1920*
Thornton, G. C., *944*
Thorpe, J. A., 929, *946*
Thorsrud, E., 1186, *1188*
Thuesen, G. J., *2351*, 2395, 2399, *2405*
Thusen, G. J., *2393*
Tiberwala, R., 1746, *1767*
Tichauer, E. R., 871, 876, *897*, 1061, *1107*
Tien, J. M., *2393*
Tierney, L., *2495*
Tijms, H., 1678, *1693*
Tijms, H. C., 608, *618*
Tillman, B., *2443*
Tillman, J., 536, *543*
Tindale, R. S., *2218*
Tippett, L. H. D., 1448, *1462*
Tirre, W. C., *946*
Tits, Andre, 2563
Titus, W., 2212, *2222*
Tiwana, A., 214, 220, *226*
Tobin, D. R., 148, *154*
Todd, J., *540*
Toffler, A., 685, *708*
Toint, Philippe, 2564
Tola, S., *1106*
Tolga, K., *2520*
Tolle, D. A., *543*
Tollers, G. V., 478, *483*
Tollison, P. S., *993*
Tomlin, J. A., *2538*
Tomlingson, P. D., 1552, *1583*
Tomlinson, W., *1919*
Tompkins, J., 5, *24*, *1501*, 1588, *1623*, 2089, 2092, *2109*
Tompkins, J. A., *1501*, 1502, 1503, 1508, 1525, *1526*, *1547*, 1611, 1618, 1621, *1623*
Tönshoff, H. K., 402, *446*
Topkis, D.M., 2643, *2648*
Toquam, J. L., *944*
Torione, R., *150*
Tornow, W. W., *918*
Torrance, E. P., 2208, *2222*
Torsilieri, J. D., 1699, 1701, *1716*
Toth, P., 811, *823*, *2581*
Toulmin, S., 137, 138, *154*
Towers, S., 2127, *2140*
Towill, D. R., 2112, *2140*
Townsend, J. T., 2205, *2216*
Toyota, T., *560*
Tracy, M. F., 1052, *1107*
Traffton, L. L., *2138*
Trankle, U., 1202, *1236*
Trappey, J. C., 1313, 1314, 1328, *1330*
Trautner, S., *445*
Travers, R. M., 928, *946*
Treese, G. W., 229, *258*
Treiman, D. J., 916, *919*
Tremont Partners, Inc., 759, *771*
Trengove, C. D., *1668*
Trevino, J., *1526*, *2109*
Triadi, M. N., 2464, *2467*

Triandis, H. C., *972*
Trick, Michael, 2575
Trist, E., 976, *993*
Trist, E. L., 874, *895*
Tritsch, C., 439, *446*
Trivedi, V., 746, *750*
Troup, J. D., 1069, *1107*
Trowbridge, R. L., *1104*
Troy, K., *1807*
Trucks, H. E., *1331*
Trussell, H. J., 2285, *2292*
Tsai, F. F., *2520*
Tsai, J., *1789*
Tseng, M. M., 685, 687, 697, 699, 701, *708*
Tsitsiklis, J. N., 2645, 2646, *2646, 2648*
Tsuda, M., *400*
Tsujimura, Y., 1781, *1790*
Tu, X., *1128*
Tubbs, M. E., *2221*
Tucker, A. W., 2543, 2553, 2554, 2556, *2566*
Tucker, J. S., *1716*
Tuckman, B. W., 2210, *2222*
Tukey, J. W., 2283, 2286, *2293*
Tukiainen, M., 2401, *2405*
Tulkoff, C., 536, *541*
Tulkoff, J., *483*
Tullar, W. L., 922, *946*
Tulving, E., 930, *946, 947*
Tumay, K., 2449, 2459, 2461, *2465–2468*
Tung, K., 691, *708*
Turbak, F. A., *152*
Turban, D. B., 855, *867*
Turbini, L. J., *543*
Turk, M., 1206, *1236*
Turnage, J. J., 889, *898*
Turnbull, P. W., 2125, *2140*
Turner, A. N., 874, *898*
Turner, S. J., *2466*
Turner, W. C., 355, *398, 2405*
Turnquist, M. A., 538, *543*
Tustin, A., *2444*
Tuttle, D. B., *896*
Tuttle, T. C., *25*
Tversky, A., *971,* 1023, *1039,* 2179, 2196–2198, 2202–2204, 2214, *2219, 2222, 2444*
Tweedie, R. L., 2166, *2171*
Twiss, B., 952, *973*
Tyburski, D., *1128*
Tyler, S., 2431, *2443*
Tyndall, G., 2112, *2140*
Tyre, M. J., 952, *973*

Uchida, T., *400*
Uchikawa, S., *561*
Ukelson, J., *1232*
Ulbrich, A., *225*
Ulgen, O., *749*
Ullman, J. D., *822*
Ulrich, D., 147, *154*
Ulrich, K., 688, 689, 691, *708*
Ulrich, K. T., 688, 689, *708*
Umbers, I. G., 1894, *1920*
Umble, M., 557, *561*
Underwood, B. J., 930, 931, *946, 947*
U.S. Army Material Command (AMC), 1331, 1923, *1955*
U.S. Bureau of the Census, 826, *836*
U.S. Department of Defense, 1133, *1152,* 1243, *1251,* 1266, *1278*
U.S. Department of Defense MIL-I-45208A, 1967, *1974*
U.S. Department of Defense MIL-Q-9858A, 1967, *1974*
U.S. Department of Energy, 539, *540*
U.S. Department of Health and Human Services (HHS), 1157, 1158, *1190*
U.S. Department of Labor, 869, *898*
U.S. Environmental Protection Agency (EPA), *543*
University of Michigan, 1054, *1107*
Unterweger, P., 950, *973*
Upton, D. M., *529,* 697, *709*
Urban, G. L., 703, *709*
Urban, V., *400*
Ushakov, I., *1955*
Utterback, J. M., *154,* 686, *708*
Uzsoy, R., 538, *540, 542*
Uzumeri, M., 689, *708*

Vail, R. G., 917, *917*
Vaill, P., *25*
Vaill, P. B., 938, *947*
Vaillancourt, D. R., *1107*
Valacich, J. S., *2220*
Valenzi, E., 2195, *2222*
Valfer, E. S., 869, 874, *895*
Vallespir, B., *528*
Valverde, H. H., 932, *947*
Van Breda, L., *2443*
Van Cott, H., *1037*
Van Cott, H. P., 875, *898*
van de Kragt, A. J. C., *2216*
Van den Besselaar, P. A., 964, *970*
Vanderheiden, G. C., 1217, *1236*
van der Laan, E., 538, *541, 543*
Van der Vlerk, M. H., 2630, *2647, 2648*
Van de Ven, A. H., *918,* 984–986, *993, 2217*
Van de Welde, W., *226*
Van Dieën, J. H., 1119, *1129*
Van Dijk, N., 2645, *2648*
Van Fleet, D. D., 882, *894*
Van Harrison, R., *895, 989*
Van Hillegersberg, J., 348, *352*
van Houtum, G. J., *2109*
van Nunen, J. A. E. E., *541*
Van Ormer, E. B., 928, *943*
Van Rensselaer, G., *993*
Van Roy, B., 2646, *2648*
Van Slyke, R., 2630, *2648*
van Veen, E. A., 694–696, *709*
van Wassenhove, L. N., *541*
Vardeman, S. V., 1889, 1891, *1920*
Vargas, L. G., 2074, *2082*
Varian, H. R., *58,* 147, *153,* 350, *352*
Varney, G. H., *993*
Vassilakopoulou, P., *1038*
Vaughan, W. J., 538, *542*
Vaughn, L. T., *736*

Vaught, C., *2220*
Vayrynen, S., 1061, *1107*
VDI Gesellschaft Entwicklung Konstruktion Vertrieb (VDI-EKV), 191, 206, *226*
Vecchi, M. P., *2600*
Velleman, P. F., 2286, *2293*
Venable, T., 2080, *2082*
Venema, W. J., *1917*
Venkatadri, U., *542*
Venkataraman, R., *2053*
Ventura, J. A., 1904, *1920*
Venzin, M., 213, *226*
Vepsalainen, A., 1724, *1740*
Verein Deutscher Ingenieure, 407, *446*
Verguld, M. M. F., 423, *446*
Verhoef, C. W. M., 1209, *1236*
Verma, D., *2393*
Verma, S., 2436, *2443*
Vermeeren, A. P. O. S., 1206, *1236*
Vernadat, F. B., 301, *307,* 508, 510, *529*
Vertelney, L., 1216, *1236*
Verter, V., 402, *446*
Vessey, I., 88, *108*
Vettering, W. T., *1693*
Vicente, K., 1020, 1024, *1039*
Vicente, K. J., 1193, 1205, *1236*
Vicere, A. A., 858, *867*
Vick, A. L., *990*
Victor, B., 214, *226*
Videman, T., 1070, *1106, 1107*
Vidulich, M., *2444*
Vigon, B., 536, *540*
Vigon, B. W., *540, 543*
Viikari-Juntura, E., *1100*
Vikki, M., *1105*
Vilhjalmsson, H., 1127, *1128*
Villanueva, R., 2455, *2466*
Vincke, P. H., 2604, *2622*
Vink, P., 980, 981, *993*
Vinogradov, O., *1955*
Viscusi, W. K., 2212, *2222*
Visual Thinking International Inc., 2460, *2468*
Viswanadham, N., 1668, *1668*
Vogel, D. R., *2220*
Vogelsang, K., 559, *560*
Voigt, K., 1733, *1739, 2052*
Vojta, G., 623, *631*
Volberg, O., *1129*
Volger, P., *223*
Volkema, R. J., 127, *154*
Vollbach, A., *225*
Vollman, T. E., 2039, 2040, *2053*
Vollmann, T. E., 546, *561*
Volpe, C., *942*
Volpe, C. E., *942*
Volpert, W., 1145, *1155*
Von Alven, W. H., 1937, *1955*
von Briel, R., 318, *323*
von Krogh, G., 213, *226*
von Neumann, J., 2173, 2178, 2180, 2182, *2222*
Vorbeck, J., 215, 217, *224*
Voss, C., *1668*
Voss, D. T., 2235, *2239*
Voss, P. A., *2393*
Vrielink, H. H. E., 1119, *1129*
Vroom, V. H., 874, *898*

Waag, W. L., *941*
Wachtel, J. A., *2443*
Wacker, G., 1025, *1037*
Wacker, G. L., 869, 884–886, *895*
Waddell, H. L., 1448, *1462*
Wade, J., *971*
Wadsworth, H., 2227, *2240*
Wageman, R., *991,* 1794, *1806*
Wagenaar, W. A., 2206, *2222*
Wagner, Gerald, 134
Wagner, J. A., 847, *867*
Wagner, M., 193, *226,* 441, 442, *446*
Wagner, R., 1138, 1146, *1155*
Waguespack, B. G., 984, 986, *994*
Waites, C., *1919*
Waldman, D. A., 858, 859, *864, 867*
Waldner, J. B., 492, 493, *529*
Walker, C. R., *898*
Walker, R. A., *1128*
Wall, T., 870, *898*
Wall, T. D., *989*
Wallace, E., 1782, *1790*
Wallace, M., 1061, *1107*
Wallace, S., 1783, *1789*
Wallace, S. W., 2630, *2647*
Wallenius, J., 2620, *2623*
Waller, M. A., *1806*
Wallsten, T. S., 2198, 2201, 2205, *2222*
Wall Street Journal, 949–951, 965, *973, 974*
Walraven, E. J. C., 2559, *2566*
Walter, C. B., *945*
Walter, S. D., *1104*
Walters, B., *2443*
Walton, Sam, 782
Wang, B., *1918*
Wang, E. H., *1919*
Wang, H. P., *483*
Wang, H.-P., *483*
Wang, J. Y., 1524, *1526*
Wang, L. C., 166, *176*
Wang, L.-C., 166, 175, *176*
Wang, M., 538, *541*
Wang, M. J., 1893, *1920*
Wang, M-J., 1898, *1920*
Wang, P. Z., *2624*
Wang, Q., 963, *972*
Wang, R., 277, *279*
Wang, S. P., *919*
Wang, Y., 1669, 1692, *1693*
Wangenheim, M., 1061, *1107*
Wapler, M., *400*
Warburton, D. M., 2208, *2218*
Ward, A. F., *258*
Warden, D. M., 746
Warden, G., 747, *750*
Warnecke, G., 218, *226*
Warnecke, H.-J., *323, 399, 400,* 404, *446*
Warner, D. M., 744, *750*
Warner, M., 746, *750*
Warr, P., 870, *898*
Warschat, J., 207, *223,* 1285, 1293, *1294, 1295*
Wasil, E. A., 2575, *2581*

Wassenhove, L. N. V., *540*
Wasserman, W., *2293*
Watabe, K., 107, *109*
Waterman, D., *1330*
Waterman, R., 653, *664*
Waterman, R. H., *1807*
Waters, T., *1109*
Waters, T. R., 1076–1079, *1102, 1107, 1108,* 1117, 1119, *1129,* 1133, *1155*
Watford, B., *749*
Watkins, C., 1780, *1790*
Watkins, M. J., 930, *947*
Watson, H. J., *153*
Watson, J., *944*
Watson, R. T., 143, *154*
Watson, S. R., 1898, *1920*
Watson, T. W., *947*
Waurzyniak, P., 536, *543*
Wayne, S. J., 978, *994*
Weaver, J. L., 922, 929, *947*
Webb Associates, *1110*
Webber, B. L., *1100, 1127*
Webber, M. D., 2112, *2140*
Weber, B., *2442*
Weber, C., *919*
Weber, E., 2196, 2197, 2203, *2222, 2223*
Weber, T., *864*
Webster, B. S., *1107*
Webster, J., *2220*
Wechsler, D., 976, *994*
Weck, M., 201, *226*
Wee, W. G., *1919*
Weggeman, M., 215, 218, *226*
Wegner, D., *2223*
Wei, L., 607, *618*
Weibull, W., 1945, *1955*
Weicher, M., 1696, 1707, 1710, *1717*
Weick, K., 174, *176*
Weick, K. E., 140, *154,* 963, *974,* 980, *994,* 1024, *1039*
Weidman, C. H., *1152*
Weigand, R. E., 2115, *2140*
Weigl, A., 443, *446*
Weil, M., 2058, 2065, *2069*
Weil, U., *1104*
Weill, R. D., *483*
Weinfurter, A., *258*
Weinstein, N. D., 2197, 2201, *2223*
Weisbecker, A., 642, *650*
Weisener, T., *400*
Weisman, G., *1101*
Weiss, A., *1189, 1233*
Weiss, G., 2167, *2171*
Weiss, H., *943*
Weiss, H. M., 858, *867*
Weiss, J. J., *1920*
Weiss, R., *2216*
Weiss, W., 2202, *2216*
Weissberg, R. W., 745, *750*
Weitz, K., *540*
Welch, D. A., 139, *154*
Welch, D. D., 2213, *2223*
Welch, P. D., *2495*
Welford, A. T., 875, *898,* 933, *947,* 2175, *2223*
Wells, R., 1202, *1233, 1236*
Wells, S., *529*
Welsch, R. E., 2286, *2292, 2293*
Welsh, J. R., 921, *947*
Welti, N., 90, *109*
Wen, U. P., *1525*
Wenblad, A., *542*
Wengel, J., 316, *323*
Wenner, C., 1896, 1908, *1920*
Wenner, C. L., 1895, *1920*
Wenner, F., *1920*
Werbos, P., 1778, *1790*
Werners, B., 2620, *2623*
Werning, H., *588*
Wesel, E. K., 229, *258*
Wesolowsky, G. O., *2293*
Wessels, H., 201, *225*
Wessinger-Baylon, T., 140, *152*
Westerberg, A., *708*
Westercamp, C., *587*
Westerlund, J., *588*
Western Electric Company, *2003*
Westgaard, R. H., 1086, *1108*
Westin, B., *1236*
Westkämper, E., 315, *323*
Westley, B., *944*
Weston, R., 1713, *1717*
Weston, R. H., 520, 522, 523, *528*
Wets, R. J. B., 2630, *2648*
Wettberg, W., 1185, *1190*
Wetzels, M. G. M., 628, *631, 632*
Wexelblat, A., *2520*
Wheaton, G. R., 932, *943*
Wheelwright, S. C., *2329*
Wherry, R. J., *1127*
Whinston, A., 107, *107–109*
Whinston, A. B., 67, *108,* 214, *224,* 229, *257,* 278, *278, 279*
Whiston, T. G., 953, *974*
White, A., *541,* 599, *600*
White, A. A., *1104*
White, D. J., *2624,* 2645, *2648, 2649*
White, J., *1390, 2109*
White, J. A., *1501,* 1509, *1525, 1526, 1547,* 2350, *2351, 2405*
White, R. E., 553, *561*
White, T. S., *399*
Whitehouse, D. J., *1886*
Whiteside, H. D., *993*
Whitley, D., *1789*
Whitman, M. V., 1888, *1920*
Whitney, D. E., 1313, *1330*
Whitney, G., 1794, *1807*
Whitt, W., 1655, *1668,* 2163, 2168, 2169, *2171,* 2645, *2649*
Whybark, D. C., *561, 2053*
Wichern, D. W., 2277, *2293*
Wick, C., *483*
Wick, J., 1131, 1146, *1152*
Wick, W., *945*
Wickens, C., 1014, 1015, *1040, 2443, 2444*
Wickens, C. D., 1146, *1155,* 1895, 1898, *1920,* 2185, *2223,* 2421, 2423, *2443, 2444*
Wickens, C. W., 2434
Wicks, E. M., *2405*
Wielinga, B., *226*

Wiendahl, H. P., *618*
Wiener, E., *974*
Wiener, E. L., 1137, *1152*
Wierzbicki, A. P., *2623*
Wiget, K., *749*
Wiggins, M., *1039*
Wightman, D. C., 931, *947*
Wiig, K. M., 213, 218, *226*
Wikner, J., *2140*
Wilbur, K., 5, *24*
Wilcox, L. C., 147, *152*
Wildberger, A. M., 2464, *2468*
Wilde, D. J., *2567*
Wilder, D. G., *1102*
Wilhelm, W. E., *399*
Wilhelmy, R. A., *942*
Wilkins, C. L., 925, *945*
Wilkinson, P. L., *1101*
Will, H. J., 125, *154*
Willemse, M. A., *400*
Willets, L. G., 1710, *1717*
Willham, C. F., 2197, 2198, *2216*
Williams, A., *1039*
Williams, D. J., *529*
Williams, G., *600*, 2449, *2465*
Williams, K., *896*
Williams, K. Y., *2218*
Williams, L., *1127*
Williams, M., 1070, *1108*
Williams, N. P., 606, *619*
Williams, R. J., 2167, *2171*
Williams, R. L., 148, *150*
Williams, T., 1769, *1790*
Williams, T. J., 157, 159, 161, 162, 168, 175, *176*, 507, *529*
Williamson, O. E., 2126, *2140*
Willke, H., 213, *226*
Willy, A., *398*
Wilson, C., 1207, 1208, 1212, *1236*
Wilson, J., *1101*, 1378, *1389*
Wilson, J. R., *561*, 980, 981, *994*, 1151, *1155*, 1229, *1236*, 2488, *2495*
Wilson, R. B., 2562, *2567*
Wilson, R. C., 2576, *2581*
Wilson, W. R., 630, *632*
Wilt, A., 745, *750*
Wilt, C. A., *540*
Winer, B. J., 2227, *2240*
Winkler, R. L., 2193, 2196, 2197, 2201, *2220*, *2223*
Winograd, T., 143, *154*
Winstanley, N., 916, *919*
Winter, D. A., 1069, *1108*, *1130*
Winterfeldt, D. V., 2173, 2178, 2182, 2183, 2185, 2187, 2191–2195, 2201, *2223*
Winters, J., 1112, *1129*
Wise, L., 745, *749*
Wise, R., 602, 603, *619*
Wisner, P., *1584*
Witkins, A., 1126, *1129*
Wittink, D. R., *2218*
Witzerman, J. P., 167, 171, *176*, 603, *619*
Wixon, D., 1207, 1208, 1212, *1236*
Wixted, J. T., 930, *947*
Wogalter, M., *1103*
Wogalter, M. S., 931, *943*, 1177, *1189*
Wojciechowski, J. Q., *2467*
Wolf, A., *400*
Wolf, G., *991*
Wolf, J. A., 2423, *2443*
Wolf, K. U., 422, *445*, *446*
Wolf, L. D., *990*
Wolf, S., 2206, *2219*
Wolfe, D., 2256, *2263*
Wolfe, P., 2074, *2081*, 2558, 2559, *2567*
Wolff, R., 1672, 1686, 1689, *1694*
Wolff, R. W., 2146, *2171*
Wolkowitz, C., 1220, *1235*
Wolsey, L. A., 805, 811, 813, 814, *823*, 2582, 2587, *2600*, *2601*
Womack, J. P., 555, 556, *561*
Womack, J. T., 5, 8, *24*
Wong, C. S., 886, 887, *898*
Wong, H., *152*
Wong, P., *2580*
Wong, P. J., 2645, *2649*
Wong, R. T., *2648*
Wong, W., *1039*
Woo, S. L. Y., 1112, *1129*
Wood, B., 2459, *2465*, *2468*
Wood, R., 979, *994*
Woodcock, C. R., 1504, *1526*
Wooden, John, 843
Woodring, S. D., *353*
Woods, D., 1022–1024, 1032, *1040*
Woods, D. D., 963, *973*, 1030, 1032, *1037*, *1039*, *1040*
Woods, J. A., 148, *150*
Woods, M. S., 222, *226*
Woodson, W. E., 875, *898*
Woodworth, R. S., 932, *946*
Woolford, B., *1129*
Woolley, C., *1128*
Worhach, P., 538, *542*
Workflow Management Coalition, 507, *529*
Works, M., 949, *974*
World Health Organization (WHO), 1083, 1084, *1108*
Wörner, K., 1288
Worrall, G. M., *1919*
Wortel, E., *993*
Wortley, M. D., *1100*, *1101*
Wortman, D. B., 2410, *2444*
Wortmann, J. C., 695, 696, *707*
Wössner, J. F., *399*
Wright, D. T., 1703, *1717*
Wright, I., *109*
Wright, M. A., *2538*
Wright, Margaret H., 2564
Wright, P., 1135, *1155*
Wright, P. M., 856, *867*
Wright, S. J., 2536, *2539*, 2563, *2566*, *2567*, 2575, *2581*
Wright, T., 1400, 1402, *1408*
Wright, W. O., 492, *529*
Wu, C., 490, 496, *528*
Wu, C. T., 1524, *1525*
Wu, M., 104, *109*

Wu, S., 104, *109*
Wu, S. D., 1777, *1790*
Wu, S.-Y., *176*
Wu, T. P., *1739*
Wu, Y., *1788,* 1889, *1920*
Wu, Y. I., 2237, *2240*
Wubker, G., *683*
Wurster, T. S., *57*
Wycoff, M. A., *994*
Wysk, R. A., *483, 1330,* 1524, *1526,* 1777, 1779, *1788, 1790*
Wyskida, R. M., *2316*

Xiao, D. Y., 524, *529*
Xiaoming, X., *618*
Xie, B., *528*
Xu, C., *400*
Xu, S., *1693*

Yager, R. R., *2217*
Yago, G., 759, *771*
Yalch, R. F., 680, *683*
Yamamoto, A., *1787*
Yamashina, H., *560*
Yammarino, F. J., 844, 851, 858, 859, *867*
Yan, P., *1790*
Yang, C. L., 1221, *1231, 1236*
Yang, C.-L., 1223, *1236*
Yang, G., *618*
Yang, S., 146, *150*
Yang, W., 2488, *2494*
Yaniv, I., 2212, *2216*
Yano, C., 538, *540*
Yano, C. A., *540,* 1734, *1740*
Yantis, S., *2443, 2444*
Yanushefski, A., *990*
Yao, D. D., 1669, 1690, 1692, *1693, 1694*
Yarberry, L. S., 2485, *2494*
Yates, F., 1421, *1462*
Yates, J., 952, *973*
Yates, J. F., 2173, 2202, *2223*
Ye, S., *1918*
Yeadon, Fred, 352
Yeh, C., *400*
Yeh, M. S., 1525, *1526*
Yeh, W. C., 1525, *1526*
Yen, C. K., 1525, *1525*
Yeo, K. T., *1789*
Yeong, M. Y., *399*
Yeremeyev, A., *2109*
Yih, Y., 1524, *1526,* 1776, 1778, 1779, 1781, *1788–1790*
Yin, G. G., 1668, *1668*
Yoon, C. S., 2084, *2109*
Yoon, K., *1109,* 2606, 2608, 2614, *2622*
Yordanova-Markova, I. T., 2621, *2622*
Yoshida, H., 1145, *1155*
Yoshitake, A., *528*
Yost, P. R., *865*
Young, C. E., *1716*
Young, D., *2351, 2393*
Young, J., *540*
Young, M., 2440, *2444*
Young, M. L., 228, *258*
Youngdahl, W. E., 559, *559*
Younger, M. S., *2293*
Yourdon, E., 123, *150*
Yu, B., 1703, *1717*
Yu, D., *1717*
Yu, L. C., 2464, *2468*
Yu, P. L., 2605, 2606, 2608–2612, 2614–2619, 2621, *2621–2623*
Yuan, Mingjian, 2494
Yuguchi, R., *400*
Yukl, G., 929, *946*
Yukl, G. A., 842, 845, 854, *867*
Yura, K., *541*
Yusuf, Y. Y., 602, *619*

Zaccaro, S. J., *866,* 985, *990*
Zacharias, G., *2442*
Zachman, A. C., 302
Zack, M. H., 148, *154*
Zadeh, L. A., 1898, *1920,* 2620, *2621*
Zahorik, A., *664*
Zahorik, A. J., *632*
Zahorjan, J., *1152*
Zajac, F. E., 1126, *1129*
Zajonc, R. B., 877, 880, 882, *898*
Zakai, E., *866*
Zakay, E., 849, *866*
Zander, A., 2210, *2223*
Zandin, K. B., 1439, *1462*
Zanettin, M., *528*
Zangwill, W. I., 2560, *2567*
Zauchner, S., *1233*
Zawack, D., *1739*
Zeckhauser, R., 2178, *2222*
Zeh, J., *1107*
Zehel, D., *1236*
Zehner, G. F., 1115, *1129*
Zeigler, B. P., 124, *154*
Zeithaml, V., *664,* 1964, *1965*
Zeithaml, V. A., 623–625, 629, *632, 633, 650*
Zeleny, M., 2614, 2618, *2623, 2624*
Zemel, J. N., *399*
Zemke, R., 926, *947*
Zener, C., *2565, 2567*
Zerega, B., 95, *109*
Zey, M., *2223*
Zhand, D., *2621*
Zhang, C., 1777, 1778, *1790*
Zhang, H., 538, *543*
Zhang, J., 2562, *2567*
Zhang, M., 1777, *1790*
Zhang, P., *617*
Zhang, Q., 1668, *1668*
Zhang, W., *528,* 1780, *1790*
Zhang, X., *1128*
Zheng, S., 1692, *1693, 1694*
Zhongjun, Z., 607, *618*
Zhou, D. N., 1778, *1790*
Zhou, Jian, 2563
Zhou, M. C., 490, 504, 505, *529*
Ziegenfuss, J. T., *994*
Ziembra, W., 756, *770*
Zijm, W. H. M., *2109*
Zimmer, A., 2201, *2223*

Zimmermann, H. J., 2620, *2623*
Zimolong, B., 1183, 1186, *1190*
Zink, K., 1184, *1190*
Zinn, C. D., *2393*
Zionts, S., 2620, *2623*
Zipkin, P., 1669, *1694*
Zipkin, P. H., *2032*
Zohman, B., 1179, *1189*
Zoran, P., *1129*
Zoutendijk, G., 2560, *2567*
Zowe, J., 2562, *2567*
Zsambok, E., *2219*
Zweben, M., 2044, *2053*
Zwick, R., *2222*
Zyda, M., 229, *258*
Zyser, V., *1526*

SUBJECT INDEX

ABC, *see* Activity-based costing
ABCD Model, 10, 16
ABC FlowCharter, 304
Ability tests, 921–922
ABM, *see* Activity-based management
Absolute value functions, 2527
Absorption cost accounting, 1272
Abstraction:
  in enterprise models/modeling, 281–283
  levels of, 281
  in OOP, 1328
Acceptability, as measurement issue for successful design, 1299, 1301
Acceptable day's work, 1392, 1405, 1406
Acceptance region (hypothesis testing), 2243, 2244
Access control (security technology), 734
Accessibility:
  as enterprise issue, 662
  as new marketing paradigm, 660–662
Accessories, computer, 1202
Accidents, *see* Occupational safety and health
Accountability (work packages), 1268
Accounting and finance. *See also* Financial asset management
  cost estimating, accounting data for, 2309, 2310
  and ERP function, 336
  ERP tools for, 91
  performance, financial, 49, 50
  and performance management, 1002
  and plant engineering, 1562–1564
  and transportation management software, 2065
ACD (activity cycle diagram), 506
ACE team benchmarking system, 1598–1601
ACID properties (of transactions), 721–723
Acid Rain Program, 593
Acoustical control, 1200
Acquisition tools (data), 83
Action-cycle model (cognitive tasks), 1017–1019
Action-decision diagrams, 1376, 1378
Actions (human–computer interface design), 1213
Action language, 131–132
Active redundancy, 1933
Active Server Pages (ASPs), 79
Active X controls, 76
Activities:
  in business processes, 44
  definition of, 40
Activity-based costing (ABC), 1272, 2317–2319
  conventional costing vs., 2319
  definition of, 2319
  for energy costs, 1576–1577
  for process design and reengineering (PDR), 1704
Activity-based inventory assessment, 531
Activity-based management (ABM), 2317–2329
  case study of, 2319–2329
  definition of, 2317
Activity cycle diagram (ACD), 506
Activity databases, 1260
Actual dollar analysis:
  of after-tax cash flow, 2403–2404
  of before-tax cash flow, 2402, 2403
  with differing inflation rates for component cash flows, 2400
  with differing inflation rates per time period, 2401
  of economic equivalence calculations with inflation, 2399–2400
Actual dollar cash flows, constant dollar vs., 2397–2398
ADA (Americans with Disabilities Act), 1592
Adaptive directed synthesis control, 161
Additivity, in linear models, 2525–2526
Addresses:
  Internet, 237, 241–243
  10 net addresses, 238
  URLs, 244, 245
Adept Technologies, 167
Adherent grippers, 414
Adhesives, 413, 431
Ad hoc teams, 976, 982
Administrative solutions:
  for management of work-related musculoskeletal disorders, 1092–1093
  for reduction of musculoskeletal disorders, 1363, 1365
Advanced planning and scheduling (APS), 2034–2036, 2045–2052
  components of, 2045–2046
  implementation of, 2046–2052
    access considerations in, 2047
    and accuracy of manufacturing data, 2049–2050
    and business process reengineering, 2049
    and ERP integration, 2047–2048
    quality considerations in, 2049
    strategies for, 2051
    timing considerations in, 2047
  software for, 338

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy Copyright © 2001 John Wiley & Sons, Inc.

Advance shipping notice (ASN), 2087, 2097
Advertising:
  and electronic commerce, 272, 273
  online classified, 275
ADW, 304
AEM, *see* Assemblability Evaluation Method
Aerospace industry, human modeling in, 1121
Aesthetics, as performance measure of quality, 1247
AET, *see* Arbeitswissenschaftliches Erhebungsverfahren zur Tätikgkeitsanalyse
Affinity diagrams, 1813, 1815
A fortiori (strength of the argument) principle, 2367
After-ANOVA range tests, 2261
After-tax cash flow analysis, inflation in, 2403–2405
Agendas (for groups), 2213
Agents:
  in business model, 32
  in scheduling problems, 1777
Agent-based control systems, 174–175
Agent-based manufacturing, 697
Agent technology, 221
Aggregates (data), 84
Aggregations, in object-oriented enterprise modeling, 293
Agile manufacturing, 486
  and collaborative manufacturing, 602, 603
  and computer integrated manufacturing (CIM), 527
Agriculture, 346
AGV systems, *see* Automated guided vehicle systems
AHA (American Hospital Association), 739
AHP (analytic hierarchy process), 2195
AI, *see* Artificial intelligence
AIM, 2458
Airbus door manufacture, 571–572
Air conditioning systems, 1581
Airline industry, *see* Aviation industry
Air pollution control. *See also* Clean Air Acts
  air permits, 595–596
  emissions, estimation of, 596–598
    factors, emission, 597–598
    mass balance approach to, 596–597
  energy-improvement possibilities for, 1581
  and total enclosure concept, 598
Aisin Seiki, 551
Aisle allowances (storage), 1535–1537
Alignment, achievement of (performance management), 1005–1010
  integrated system, creation of, 1009–1010
  leadership of both formal and informal organization, 1008–1009
  working arenas, identification/alignment of, 1006–1007
Allaire, 78, 79
Allergic dermatitis, 1167
Alliances, 30, 34
  architecture of, 39
  in business model, 46, 48–49
  as external forces, 39
  strategic, 48
AlliedSignal Turbocharging Systems, 86
Allocation of function (inspection and test systems), 1892
Allocation of reliability requirements, 1937, 1938
Allowances:
  delay, 1398, 1400
  fatigue, 1394–1400
  personal, 1394
  in time studies, 1426–1427
Alternative design, 1049
Alternative hypothesis, 2245
Amazon.com, 262, 263, 265, 266, 272, 273, 773, 783, 2071
Ambiguity, organizational, 140
Ambulatory surgery centers, 742
American Customer Satisfaction Index, 624
American Hospital Association (AHA), 739
American Institute of Industrial Engineers, 1714
American Manufacturing Excellence, 553
American National Standards Institute (ANSI), 72, 1093–1095, 1164–1165, 1967–1968. *See also* ANSI standards
American Production and Inventory Control Society, 949
American Society for Quality (ASQ), 1967, 1968
American Society for Testing and Materials (ASTM), 1165, 1967
American Society of Mechanical Engineers (ASME), 1967
Americans with Disabilities Act (ADA), 1592
Ameritech, 654
AMICE, *see* European Computer Integrated Manufacturing Architecture
AMPL (A Mathematical Programming Language), 2536
Amputations, 1169
AMRL, 1112
Analysis:
  in EPEM model, 1798
  of experimental design, 2232–2234
  in ISO 9001:2000 QMS standard, 1971
Analysis of variance (ANOVA), 2233–2236
Analytical models, 1630
  assumptions of, 1634–1635
  for client/server (C/S) system evaluation, 728, 729
Analytic hierarchy process (AHP), 2195
Anchoring, 1023
Anchoring and adjustment heuristic, 2199
Andersen Consulting, 950
Andersen Windows, 783
Andon (visual control system), 549
Angular transducers, 1902
Anheuser-Busch, 2129
Animation (digital human modeling), 1116, 1120, 1125–1127
ANN, *see* Artificial neural networks
Annual worth method (cost estimating), 2347–2348
Annuities, 764, 765

Anonymity:
  of customer information, 267–268
  types of, 268
Anonymity services, 268–269
Anonymizer.com, 268
ANOVA, *see* Analysis of variance
ANSI, *see* American National Standards Institute
ANSI standards:
  for informing employees about hazards, 1176–1177
  Q9004, 1968
  Q9004–2000, 1968
  Z1.8–1971, 1968
  Z1.11 standard, 1973
  Z1.15–1979, 1968
Anthropometric Survey (ANSUR), 1113, 1114
Anthropometry, 1043–1050
  alternative design, 1049
  body position, description of, 1043
  computer-aided models of man, 1050
  databases, 1113, 1114
  definition of, 1043
  design criteria, 1048, 1049
  digital human modeling, 1113–1115
    databases, 1113, 1114
    methods, 1113–1115
  method of limits, 1048
  physical vs. functional anthropometry, 1043
  range-of-joint mobility, 1043, 1046
  statistical descriptions, 1043, 1048
  test population, 1120–1121, 1122, 1123
Anxiety, as obstacle to performance management, 1002
Anxiety allowances, 1397
APICS, 348
APIs, *see* Application program interfaces
Apparent tardiness cost (ATC), 1724–1725
Appeals, handling, 913
Apple Computer, 86
Applets, 78
Application class level of abstraction, 281–283
Application hosting, 343
Application of knowledge, 215
Application program interfaces (APIs), 78, 340
Application proxies, 735
Applications consistency (user interfaces), 133
Application service providers (ASPs), 107, 343
Approach directions, 452, 453
APS, *see* Advanced planning and scheduling
Aptitude tests, 921–922
AR, *see* Augmented reality
Arbeitswissenschaftliches Erhebungsverfahren zur Tätikgkeitsanalyse (AET), 1138–1140
Architects, selection of, 1496–1499
Architectures (CAD), 210–213
Architecture of integrated information systems (ARIS), 293–300, 507, 512, 513
  and CIMOSA, 302
  house of business engineering (HOBE), 294, 299–300
  and IFIP ISM, 301
  methods for conceptual modeling in, 295–297
  phases in, 295–298
  SAP R/3 integration with, 304–306
  views of, 294–295
  and Zachman framework, 302
Architecture structure model (CIMS), 520–521
Archiving (of project documentation), 1349
Area level of functioning, 1771
Arena, 2446, 2456
ARIS, *see* Architecture of integrated information systems
Arithmetic gradient conversion factor (interest), 2342–2343
Arm(s):
  holding time for, 1066–1067
  static efforts of, 1058–1062
  and work posture, 1359
  workstation guidelines related to, 1359, 1361–1362
Armed Services Vocational Aptitude Battery (ASVAB), 921
Armour, 654
Armstrong Laboratory, 1050
ARPANet, 235, 238
Articulated robots, 374, 375
Artificial intelligence (AI), 107, 121, 160–164
  in agent-based control systems, 174–175
  in capacity modeling, 2044
  and control, 1775–1782
    commercial software, 1782
    fuzzy set theory, 1781–1782
    genetic algorithms (GA), 1780–1781
    knowledge-based systems, 1775–1776
    neural networks, 1777–1780
  as decision support tool, 2014
  in expert database model, 122
  fuzzy logic, 163, 164
  genetic algorithms, 164
  hybrid intelligent control models, 164
  knowledge-based systems, 160, 162
  knowledge management in, 213
  in model base management systems, 131
  neural networks, artificial, 162–163
  and shop floor scheduling, 1775–1782
    commercial software, 1782
    fuzzy set theory, 1781–1782
    genetic algorithms (GA), 1780–1781
    knowledge-based systems, 1775–1776
    neural networks, 1777–1780
Artificial neural networks (ANN), 162–163
Art network applications, 250
ASAP, *see* AutoSchedAP
Asbestosis, 1169
Ascending bid auctions, *see* English auctions
Ascension MotionStar, 1125
"As-if" bias, 1023
ASME (American Society of Mechanical Engineers), 1967
ASN, *see* Advance shipping notice
AS 9100 standard, 1973
Aspiration level principle (decision theory), 2378

ASPs, *see* Active Server Pages; Application service providers
ASQ, *see* American Society for Quality
AS/R (automated storage/retrieval) systems, 1524
Assemblability Evaluation Method (AEM), 368–369
Assembler (programming tool), 71
Assemble to order (ATO), 330, 1685–1689, 1692
Assembly, 355–398
  and assemblability evaluation, 368–369
  automated assembly systems, 358–362, 418–419
  automotive, 388–392
    gearboxes, unpacking of, 389, 391–392
    steering components, 389, 391
  Boothroyd-Dewhurst method for analysis of, 369, 370
  categories of, 356–358
  cells, assembly, 408–409
  computer-aided methods for, 386–388
    layout planning/optimization, 386–388
    simulation of material flow, 388
  current developments in, 402–407
  definition of, 355, 407
  design for assembly (DFA), 367–370, 1328
  diagnosis of assembly processes, 422–423
  diagrams, assembly, 1376, 1377
  disassembly, 439–445
    applications of, 443–444
    ecological factors in, 440
    goals of, 439
    integrated approach to, 444–445
    processes/tools for, 440–443
  electronic, 392–396
    fiberoptic connectors, 395, 396
    luminaire wiring, 394–395
    measuring instruments, 392–394
    overload protector, 392
  of electronic devices/systems, 423–439
    feeding, PCB, 426–428
    interconnection materials, application of, 424–425
    interconnection technology for, 429–431
    and miniaturization, 423, 424
    molded interconnect devices (MIDs), 432–439
    placement of components, 425–429
    process chain in, 423
    quality assurance in, 431–432
    substrates, 424
  feeding for, 381–383, 415, 426–428
  fixturing of workpieces for, 384
  flexible assembly systems, 403, 419–422, 1633
    CAD-CAM process chain, 420–422
    for changing amounts of different versions of a product, 419–420
    handling equipment, 420, 421
  flexibly varying assembly system, 366, 367
  in food industry, 396–398
  functions in assembly systems, 407
  global, 402–404
  impact of electronics on, 404–407
  joining technologies for, 371–373, 409–413
    classification of, 409–410
    clinching, 373, 411, 412
    press-fitting, 372
    riveting, 372, 411, 412
    screwing/bolting, 371, 410, 411
    self-pierce riveting, 372
    sticking, 412–413
    welding, 413
  lines, assembly, *see* Assembly lines
  and linkage, 415–416
  magazines for storage of workpieces in, 383, 384
  manual assembly systems, 356, 358, 359, 416–418
  microassembly, 395–397
  in pharmaceutical/biotechnological industries, 398
  rationalization of, 364–367, 402, 403–404
  scope of, 355
  selection of assembly system, 362–364
  sensors for use in, 384–386
    force/torque sensors, 385
    tactile sensors, 385
    ultrasound sensors, 385, 386
    video-optical sensors, 385–386
  simultaneous engineering for efficiency in, 369, 371, 372
  structures of systems for, 407–409
  technological alternatives in, 403, 404
Assembly language, 71
Assembly lines, 330, 331
  balancing, 1382–1385
  non-progressive vs. progressive, 1355
  and process engineering, 334
  software-based design of, 386, 387
Assembly modeling (computer aided design), 185–187
Assessment centers, for leader selection, 856–857
Assessments, energy, 1578–1579
Asset classes, 758–761
  currencies, 761
  enhanced index products, 761
  hedge funds, 759, 760
  insurance-linked products, 761
  private equity and venture capital, 759–761
  Treasury inflation-protected securities (TIPs), 761
Asset management, *see* Financial asset management
Assignment problem (network flow models), 2572
Assistance robots, 381, 382
Associations:
  and electronic commerce, 273
  in object-oriented enterprise modeling, 293
ASTA, 2163
Asthma, 1167
ASTM, *see* American Society for Testing and Materials
ASVAB (Armed Services Vocational Aptitude Battery), 921

Asynchronous transfer mode (ATM), 250
ATC, *see* Apparent tardiness cost
ATM (Asynchronous transfer mode), 250
ATO, *see* Assemble to order
ATP (available to promise), 2046
Attention, limited-resource model of, 1016
AT&T Laboratories, 268, 913
Attribute control charts, 1844–1851
Attribute data, 1856–1857
Attribute modeling, 2279–2280
AT&T runs rules, 1863–1868
Attributes:
  in object-oriented enterprise modeling, 291, 292
  relationship database model, 80
  on statistical process control (SPC) charts, 1871–1875
Attributes data, 1836
Auctions, online, 271, 273–277
  B2B trading markets, 275
  double auctions, 277
  Dutch auctions, 274
  English auctions, 273–274
  first price auctions, 274
  reverse auctions, 275–276
  second price auctions, 274
Audi, 212
Audio or video conferences, computer supported, 142
Audit(s):
  of customer service, 662, 663
  human factors, *see* Human factors audits
  for maintaining standards, 1407
  of warehousing operations, 1544–1547
    methodology for, 1546–1547
    performance categories, 1544–1546
Auditory environment, and human–computer interaction, 1200
Augmented Langrangian methods, 2561–2562
Augmented reality (AR), 235, 251, 2501
Aurum Software, 95
Australia, quality standards in, 1968
Authentication protocol, 733, 734
Authority:
  for effective teamwork, 982
  team, 985
  and work packages, 1268
Auto-by-tel, 266
Autodesk, 783
Automated assembly systems, 358–362, 418–419. *See also* Assembly; Robots
Automated drafting (CAD), 494
Automated guided vehicle (AGV) systems, 1524–1525
Automated manufacturing devices, 500
Automated material-handling systems, 500
Automated storage/retrieval (AS/R) systems, 1524
Automated systems (for material handling), 1524–1525
Automated Visual Inspection Systems (AVIS), 1904–1907
Automatic assembly systems, *see* Automated assembly systems
Automatic screw machine, 1320, 1321
Automatic transfer lines, 1632–1633
Automation:
  building systems for, 1566
  economic climate for investment in, 363
  human-centered, 962
  in JIT, 548–549
  in parts production vs. assembly, 364
  of project management core processes, 1256–1260
    change management, 1259
    communications management, 1259, 1260
    risk management, 1257–1259
    scope/time/cost/resource management, 1256, 1257
  of project management support processes, 1260
  in test and inspection, 1900–1907
    image processing, 1904–1907
    materials handling, 1902
    sensing, 1902–1904
    setup, 1901–1902
    signal processing, 1904
Automation technology, 155–175
  applications of, 155–156
  artificial intelligence, 160–164
    fuzzy logic, 163, 164
    genetic algorithms, 164
    hybrid intelligent control models, 164
    knowledge-based systems, 160, 162
    neural networks, artificial, 162–163
  control systems, automatic, 156–161
    definition of, 157
    instrumentation of, 158
    models for, 159–161
  integration of, 164–167
    and distributed vs. central control, 166, 167
    networking, 165–166
    object orientation, 166
    Petri net, 166
    in robot simulator/emulator, 166, 167
  physical, 156
  trends in, 167–175
    agent-based control systems, 174–175
    concurrent flexible specifications, 172–174
    Facility Description Language, 171–173
    tool perspective, 169–174
    virtual machines, 168–170
AutoMod, 2456, 2457
Automotive industry:
  activity-based costing case study in, 2319–2329
  advanced technology failure example in, 951
  assembly systems in, 388–392
    gearboxes, unpacking of, 389, 391–392
    steering components assembly, 389, 391
  automated assembly in, 364
  automated test and inspection in, 1907
  best practices study of, 555
  component suppliers, 365
  conditions for global assembly in, 403
  high-involvement work practices in, 951–952
  Japanese/U.S. comparison, 1313

Automotive industry (*Continued*)
  JIT in, 544–545
  kanban in, 550
  multiskilled workforce in, 547, 548
  robots in, 420
  scheduling in, 1734
Autonomous agents, 174
Autonomous maintenance, 553
Autonomy, in quality-related teamwork, 979
AutoSchedAP (ASAP), 2457–2458
Availability:
  in client/server (C/S) systems, 727
  definition of, 1924
  and maintainability/reliability, 1949–1951
  measures of, 1949–1950
  product, 2131
Availability bias, 1023
Availability heuristic, 2199
Available to promise (ATP), 2046
Average inflation, 2395–2396
Avianca Airlines, 960–961
Aviation industry:
  advanced technology failure example in, 951
  civil aviation:
    automated test and inspection in, 1907
    nonproduction test and inspection in, 1908–1912
  lean production principles in, 559
  MIDAS case studies, 2436–2440
    air traffic control, extension of model to, 2439–2440
    flight crew performance, prediction of, 2436–2439
  online services, 266, 267, 275, 276
  and "safety culture" concept, 960–961
  scheduling in, 1734–1735
AVIS, *see* Automated Visual Inspection Systems
Avis (company), 662
AweSim, 2446, 2454–2455
Axial die rolling, 570, 584
Axioms of rational choice, 2178–2179

Baan, 87, 88, 95, 492, 1738
BaanERP, 95, 304, 492
Backlogged demands, 1636–1637
Back propagation (BP), 163
Back-propagation neural nets, 1778–1779
Balanced scorecard, 321–322, 997–998
Balance (with introduction/use of computer technologies), 1228, 1229
Balance model of occupational safety and health, 1159–1162
Baldrige criteria, 1956–1957, 1964
  for customer and market knowledge, 1962
  for customer satisfaction and relationships, 1963
  for employee education, training, and development, 1960
  for employee well-being and satisfaction, 1961
  for organizational leadership, 1958
  for performance excellence, 1957
  for work systems, 1960
Bandwidth(s), 213–232
  and managed bandwidth services, 250
  for networked collaboration, 234
  and object caches, 245
  and speed of information transfer, 236
Banking industry:
  client/server systems in, 735–736
  personnel scheduling in, 1743
Barlett's test, 2255–2256
Barnes & Noble, 262
Barrier control (workplace hazards), 1175–1176
Barrier function method, 2560, 2561
Base classes (computer programming), 72
Base-stock control:
  production-inventory systems, 1672–1675
    demand over lead time, 1674–1675
    normal approximations, 1673–1674
  queueing models for coordination of production, 1663–1664
BASIC (computer language), 74
Bata International, 609
Batch(es):
  definition of, 2087
  sizes of, 2037
Batch facilities, 334
Batching, 2051
Batch picking (warehouse operations), 2093–2095, 2098
Baxter, 654
Bayesian inference, 138–139, 2184–2185
Bayes' rule, 2184–2187
BE analysis, *see* Break-even analysis
Before-tax cash flow analysis, inflation in, 2401–2403
Behavior:
  and hazard control, 1181–1182
  skill-, rule-, and knowledge-based (SRK) model of, 1019–1021
Behavioral decision theory, *see* Decision theory (behavioral)
Behavioral models, 1014
Behavioral system (TQL), 1796
Behavior-driven change, 1008, 1009
Behr, 314–315
Belief form, subjective (decision making), 2191
Bell Telephone Laboratories, 1936
Belt conveyors, 1513, 1514, 1516
BEMs (boundary element methods), 199
Benchmarking, 1703, 1811, 1814
  for client/server (C/S) system evaluation, 719
  in job classification, 903–905
  of maintenance operation, 1593–1597
    assessment for, 1594–1597
    external benchmarking, 1593–1594
    internal benchmarking, 1593, 1597
  for plant/facilities engineering, 1561
  as TQL success factor, 1805
Benefit–cost method (cost estimating), 2349–2350
Best practices:
  automotive industry, 555
  change process and inclusion of, 965
  maintenance, 1610–1620

Between-operations analysis (methods engineering), 1374–1385
BFGS formula, 2552
Bias(es):
  in diagnosis, 1023
  in evaluation, 893
  gender, 916
  in group decision making, 2212
  and heuristics, 2198–2199
  in human judgments, 2201
  in statistical estimation/inference, 2198–2199, 2201
  in work sampling, 1449, 1456
Big bang approach to APS implementation, 2051
Bill of materials (BOM), 85, 2039, 2050, 2314–2316
  generic (GBOM), 695–697
  for variant handling, 694–695
Binding constraints, 2541
Biodata, 922–923
Biological job design, 873, 876, 877, 884, 888
Biomechanical design, 1072, 1076
Biomechanics, occupational, 1068–1070
Biotechnology, 38, 398
Birth-and-death models, 2156
Birth rates, market influence of, 37
Bisection method (decision making), 2191
"Black box" diagram, 100
Black & Decker, 784
Black Forest Group, 350
Blank sheet concept of change, 1700
Blind rivets, 372, 411
Block diagram, reliability, 1933–1936
Blocking (in experimental design), 2228
Blocks (in experimental design), 2225–2226
Block stacking, 1520–1521
Blow molding, 1325, 1327
BLS, *see* Bureau of Labor Statistics
Blue-collar workers, and cognitive tasks, 1013
Blueprinting, service, 641, 642
Body dimensions (chart), 1044
Body discomfort map, 1363
Body force allowances, 1396
Body movements, 1047
Body position, description of, 1043
Boeing, 7, 956
Boeing Aircraft, 1112
Bolting, 410, 411
BOM, *see* Bill of materials
Books, online retailing of, 266
Boolean operations, 183
Boothroyd-Dewhurst DFA method, 369, 370
Borg-Warner, 913
Boring, 1322
  cost of machinery for, 467
  geometric capabilities of, 464
  technological capabilities of, 469
Borland, 72, 304
Bottleneck queues, 2162
Bottom-up networking, 254
Boundary element methods (BEMs), 199
Boundary manikins, 1115, 1123, 1124
Boundary representation models (B-reps), 182
Bounded rationality, 139, 140, 1020
Bowl feeders, 415
BP (back propagation), 163
BPI (business process improvement), 304
BPM (business process management), 1697
BPR, *see* Business process reengineering
Brainstorming, 127, 2213
Branch and bound procedures, 1728–1729, 2592–2593
Brand image, managing, 39
Break-even (BE) analysis, 99, 2361
Breaks, work, *see* Work breaks
B-reps (boundary representation models), 182
Brightness, 2506
British Airways, 951
British Coal, 1145
British Standards Institution (BSI), 1185
Broaching, 1322
  cost of machinery for, 467
  geometric capabilities of, 464
  technological capabilities of, 469
Broadcast addressing, 242
Broad-range tasks (service systems), 1633–1634
Browsers, *see* Web browsers
BSI (British Standards Institution), 1185
B2B electronic commerce, *see* Business-to-business electronic commerce
B2C (business-to-consumer) electronic commerce, 70
Budgeting:
  and cost accounts, 1273
  by plant engineers, 1562
  for professional services projects, 1343–1346
    compiling/reconciling, 1346
    personnel costs, 1343–1344
    support/overhead/contingency factors, 1344
    time-phased budget, 1345
  project, 1347
Building codes, 1565, 1566
Building model:
  alliances/relationships in, 34, 46, 48–49
  business processes in, 34, 40–48
    activities included in, 44
    and controls, 45–48
    core business processes, 43
    definition, 40
    identification of, 40
    inputs, 44
    key business processes, 40
    objectives of, 43, 44
    outputs from, 45
    resource management processes, 43
    risks related to, 45–48
    strategic management process, 41–43
    subprocesses, 40
    and supporting systems, 45
  core products/services in, 34, 49
    categories of, 49
    measurement of, 49–50
  customers in, 34, 35
  external forces/agents in, 32–33, 35–40
    alliances, 39

Building model (*Continued*)
    capital markets, 39–40
    and changing playing field, 35–36
    community, 39
    competitors, 39
    customers, 38
    and data access vs. traditional reporting, 37
    demographic trends, 37
    economy, 40
    and globalization of, 36
    and information technology, 36
    and knowledge work, 36–37
    owners, 39
    political trends, 37–38
    regulators, 39
    social trends, 37–38
    stakeholders, 39
    suppliers, 39
  markets in, 34, 40
  performance management in, 48–49
Buildings and grounds department, 1566–1567
Buildtime, 297
Built to Last (Collins and Porras), 7
Bulk metal forming techniques, 567–570
Bulk recycling, 538
Bullwhip effect, 546, 2010
Bundle trading, 277
Bundling, price, 676
Bureau of Labor Statistics (BLS), 1082, 1164, 1174, 2395
Business intelligence tools, 84
Business model, 27–57
  alliances/relationships in, 34, 46, 48–49
  application of, 51–57
    business performance measurement for, 54–56
    business process analysis for, 52–54
    by communicating the nature of the business, 51
    and risk assessment, 56–57
    strategic analysis for, 51–52
  building of, 31
  business processes in, 34, 40–48, 58–60
    activities included in, 44
    and controls, 45–48
    core business processes, 43
    definition, 40
    identification of, 40
    inputs to, 44
    key business process of, 40
    objectives of, 43, 44
    outputs from, 45
    resource management processes, 43
    risks related to, 45–48
    strategic management process, 41–43
    subprocess of, 40
    and supporting systems, 45
  comprehensive framework for, 29–31
  content of, 31–35
  as context for IE, 28–29
  core products/services in, 34, 49–50
    categories of, 49
    measurement of, 49–50
  creating consensus for, 31
  customers in, 34, 35, 50
    categories of, 50
    markets and, 50–51
  elements of, 29–30
  external forces/agents in, 32, 35–40
    alliances, 39
    capital markets, 39–40
    and changing playing field, 35–36
    community, 39
    competitors, 39
    customers, 38
    and data access vs. traditional reporting, 37
    demographic trends, 37
    economy, 40
    and globalization, 36
    and information technology, 36
    and knowledge work, 36–37
    owners, 39
    political trends, 38
    regulators, 39
    social trends, 37–38
    stakeholders, 39
    suppliers, 39
  and globalization, 28
  IE's use of, 30–31
  information requirements for, 31
  and key questions about the enterprise, 29
  level of detail in, 28
  markets in, 34, 40
  meeting consensus, 31
  performance management in, 48–49
  purpose of, 28, 31
  scope of, 31
  strategic alliances in, 48
  team for development of, 31
Business network applications, 250
Business partner relationship management, 23
Business process(es), 30, 34, 40–48
  activities included in, 44
  in business model, 34
  categories of, 40–43
  core, 43
  core business processes, 58–59
  definitions of, 40, 41, 286, 1696
  energy as, 1574
  identification of, 40
  inputs to, 44
  key, 40
  as knowledge management application areas, 215
  knowledge management (KM) with, 218–220
  as knowledge processing processes, 223
  models for improvement of, 284–285, 318, 319
  objectives of, 43, 44
  outputs from, 45
  process-oriented enterprise modeling of, 286–291
    data views, 288–290
    function views, 287, 288
    organization views, 286–287
    output views, 287–289

      process view, 290–291
    resource management, 43, 59–60
    risks related to, 45–48
    strategic management, 41–43, 58
    subprocess of, 40
    and supply chain management, 2118–2125
      customer order-fulfillment process, 2121–2122
      customer relationship management process, 2121
      customer service management process, 2121
      demand management process, 2121
      information flow, 2124
      links, business process, 2118–2120, 2123–2124
      manufacturing flow management process, 2122
      procurement process, 2122
      product development/commercialization, 2122
      returns process, 2122
    and supporting systems, 45
Business process analysis:
    in business model, 52–54
    and interrelatedness of processes, 29
Business process improvement (BPI), 304
Business process management (BPM), 1697
Business process-oriented knowledge management, 218–220
Business process redesign, 1696–1697
Business process reengineering (BPR), 18–20, 217, 218, 306, 1696–1697
    and advanced planning and scheduling, 2049
    and ERP tools/systems, 88
    and knowledge management (KM), 217, 218
Business purpose, and new technology implementation, 955
Business results (in EPEM model), 1800
Business-to-business (B2B) electronic commerce, 70, 262–265
    emerging models of, 349
    and logistics management, 264–265
    manufacturing, contract, 263–264
    modeling for, 306
    procurement, Web-based, 262–263
    supply chain operations-ERP interfaces, 343
    trading markets, 275
Business-to-consumer (B2C) electronic commerce, 70
Business Week, 266
Buyer behavior, influence of pricing on, 666, 668–671
Buzz group analysis, 2213
Byssinosis, 1169

C++ (programming language), 72–73
CA, *see* Cost of assembly
CAA, *see* Clean Air Acts
Caching, 232, 233
CACI Products Company, 2455, 2459
CAD, *see* Computer-aided design
CAD-CAM, *see* Computer-aided design/computer-aided manufacturing
Cadre family, 1121
CAESAR project, *see* Civilian American and European Surface Anthropometric Resource project
CAFM (computer-assisted facility management), 1566
Calendar management software, 142
Calendering, 1325, 1326
Calibration, 1881–1882, 2193
California, 949
California Ergonomic Standard, 1166
Call Center MAESTRO, 2461
Call centers (customer service), 658, 659
CAM, *see* Computer-aided manufacturing
CAM-I automated process planning system (CAPP), 474–475
Canada, quality standards in, 1968
Cancers, occupational, 1169
CAN Financial, 654
Cantilever racks, 1523
Capability analysis:
    in design and process platform characterization methodology, 1995–1996
    process, 1869–1871
Capable-to-promise (CTP), 2046
Capacitated MRP (MRP-C), 2042–2043
Capacity, 2037–2038
    algorithms for, 2038–2045
      finite capacity, 2042–2045
      infinite capacity, 2039–2042
    condition, capacity, 2158
    network, 213–232
    planning of, with client/server (C/S) systems, 723–728
    queueing models for determining, 1631
Capacity requirement planning (CRP), 2042
Cap Gemini, 95
Capital:
    expenditures, capital, 2332
    weighted cost of, 2334
Capital costs, in hospitality industry, 834–835
Capitalized costs, 2350–2351
Capital markets, 39–40
Capital recovery factor (interest), 2340–2341
CAPM, *see* Computer-aided project management
CAPP, *see* CAM-I automated process planning system; Computer-aided process planning
CAPS Logistics Inc., 2058
Cardiopulmonary capacity reducers, 1170
Cardiovascular disease, 1170
Career planning, 938
Carousel principle (electronic components), 425
Carousels (storage retrieval), 1524
Carpal tunnel syndrome (CTS), 1084, 1085, 1092
Carrying allowances, 1396
Carrying costs, 2021
Cartesian placement systems, 436, 438
Cartesian robots, 374, 375
Carton, 2087
Cart-on-track conveyors, 1518

Cascade control, 161
CASE, *see* Computer-aided software engineering
Cash flow(s). *See also* Inflation
  actual vs. constant dollar, 2397–2398
  differing rates of inflation for component, 2400
  uncertain, risk analysis with, 2371
  uncertainties in, 2361
  with uncertain timing, 2369–2371
Cash flow diagram (table), 2332–2333
Cash flow profiles, 2332–2333
  computer spreadsheet, 2333
  engineering economy, 2332–2333
Casting, 453–456, 566–568, 571–573
  design for, 1316–1318
  machining vs., 453–455
  obtainable accuracy values, 565
  of plastics, 1324, 1326
  thixocasting, 568
Catalog retailers, warehouses for, 2086
Catastrophe bonds, 761
Categorization (hypotheses for), 137
Cathode ray tube (CRT) screens, 1195, 1197
Cause-and-effect (C&E) diagrams, 1816, 1819, 1859, 1860
CBs, *see* Common bases
CBS (cost breakdown structure), 1273
CBT, *see* Computer-based training
CCB (change control board), 1276
CCDs (charge-coupled devices), 1904
C charts, 1844, 1847–1851, 1874–1875
CCSO, 966
CDC, *see* Centers for Disease Control and Prevention
CDF, *see* Cumulative distribution function
CDM, *see* Critical Decision Method
CDNow, 266
CDs (compact discs), online retailing of, 266
CDSS (control-decision support system), 1777
CDSSs, *see* Clinical decision support systems
C&E diagrams, *see* Cause-and-effect diagrams
Ceilings, and acoustical control, 1200
Cell *j* queueing model, 1663–1664, 1666–1667
Center for Creative Leadership, 857
Center for Customer Driven Quality (Purdue), 660
Center for Health Statistics, 1164
Centers for Disease Control and Prevention (CDC), 1163, 1164, 1168, 1195
Central control, distributed vs., 166, 167
Centralization, 657, 1471
Centralization warehousing strategy, 2071
Centralized mainframe systems, 711, 721
CERs, *see* Cost estimating relationships
Ceramics, automated test and inspection of, 1907
Ceramic industry, 518
CERCLA, *see* Comprehensive Environmental Response, Compensation and Liability Act
CERN (European Laboratory of Particle Physics), 244
Certainty analysis, 2361–2362
Certainty equivalent method (utility function assessment), 2193, 2194
CFR, *see* Code of Federal Regulations
CFS, *see* Concurrent flexible specifications
CGI, *see* Common Gateway Interface
Chain conveyors, 1516, 1517
Chairs:
  adjustable, 1359
  costs of, 1359
  design of, 1204
  ergonomic recommendations for, 1196
Chakko-hiki, 551
Change:
  agreement on process for, with new technology, 963–965
  decision- vs. behavior-driven, 1008–1009
  frequency of, with computer technologies, 1228, 1229
  perceived, 958
  and performance management, 996–997
  resistance to:
    and new technologies, 955
    and organizational culture, 956
    overcoming, 889
  speed of, 311–313
  work breakdown structure (WBS) and control of, 1274, 1276
Change control board (CCB), 1276
Change leadership, 14–15
Change management, automation of, 1259
Channels, *see* Marketing channels
Character design, computer, 1196–1197
Charge-coupled devices (CCDs), 1904
Charismatic leadership, *see* Transformational leadership
Charts. *See also* Control charts
  Gantt, 103–104
  relationship, 826–828
  tolerance, 472–473
Chase Manhattan, 7
Chebyshev's inequality, 2372, 2373
Checklists, 1385, 1387
Chemical industry:
  automated test and inspection in, 1907
  as process industry, 518
Chemical machining, 1323
Chemicals (as cause of disease/injury), 1170
Chief maintenance officer (CMO), 1621
China, user differences in design requirements for, 1228
Choice:
  in behavioral decision theory, 2201–2205
  in classical decision theory, 2178–2184
    and axioms of rational choice, 2178–2179
    and dominance, 2179
    and elimination by aspects (EBA) rule, 2179–2180
    and expected utility theory, 2182–2183
    and holistic comparison, 2184
    and lexicographic ordering principle, 2179
    and maximization of expected value, 2181
    and minimax cost/regret, 2180–2181

    - minimum aspiration level, 2180
    - and multiattribute utility theory, 2183
- Chronic obstructive pulmonary disease, 1167
- Chrysler Corporation, 2134
- Chrysler Financial, 951
- Chute conveyors, 1513
- CIE (Council of Industrial Engineering), 23
- CIM, *see* Computer integrated manufacturing
- CIMOSA, *see* Computer Integrated Manufacturing Open System Architecture
- CIMS Application Integration Platform for Manufacturing Enterprises (MACIP), 517–518
- Cisco, 662, 783
- Citicorp, 7
- Civil aviation:
  - automated test and inspection in, 1907
  - nonproduction test and inspection in, 1908–1912
- Civilian American and European Surface Anthropometric Resource (CAESAR) project, 1113, 1114
- Clarify, 95
- Clarity of models, 284
- Classes (object-oriented programming), 70–71, 291–293
- Classical decision theory, *see* Decision theory (classical)
- Classification (AVIS), 1906
- Classification (group technology), 461
- Classification data, 1837, 1844–1847
- Classification method (job evaluation), 903
- Clean Air Act Amendments, 591–593
- Clean Air Acts (CAA), 590–593, 1164
- “Clean” facilities, 1489
- Cleaning robots, 380
- Clean manufacturing, 530–539
  - energy audits in, 534
  - and environmental management systems, 539
  - focus of, 530
  - and legal requirements/regulations, 531, 532
  - and life-cycle assessment, 536–538
  - and life-cycle design, 534–536
    - process design, 536
    - product design, 534–536
  - metrics related to, 531
  - and production planning, 538
  - and responsibility of manufacturer, 532
  - scope of, 530
  - terminology related to, 533
  - waste audits in, 533, 534
- Clean Water Act (CWA), 595, 1164
- Clearance, 1048–1049
- Clickshare, 272
- Click-through advertisements, 273
- Client/server (C/S) systems, 711–736
  - advantages of, 714
  - banking industry case example, 735–736
  - capacity planning with, 723–728
  - centralized mainframe systems vs., 711, 721
  - communication methods in, 718–722
    - CORBA, 719–722
    - DCOM, 721
    - Java RMI, 721
    - remote procedure call (RPC), 719
    - socket interface, 718–719
  - and computer technology trends, 712–714
  - disadvantages of, 714–715
  - distributed data management in, 723, 724
  - distributed transaction management in, 721–723
  - features of, 711
  - functional elements of, 715
  - interaction functions of, 715
  - logical functions of, 715
  - network-management protocols for, 730–732
  - and open system technologies, 714
  - performance evaluation with, 728–729
  - performance objectives/criteria for, 726–727
  - roles of client and server in, 711, 712
  - security management with, 732–735
    - services, security, 732
    - technologies for, 733–735
    - threats, 732
  - system management with, 729–730
  - three-tier architecture for, 716–718
  - two-tier architecture for, 715–717
  - and Web technologies, 712–713
  - workload modeling with, 727–728
- Client–server scheme, 240–241
- Clients, network, 240
- Climate allowances, 1398
- Clinching, 373, 411, 412
- Clinical decision support systems (CDSSs), 747–748
- CLM, *see* Council on Logistics Management
- Closed-loop control system, 22
- Close phase (professional services projects), 1348–1349
- Closest-open-location (COL) rule, 1509–1510
- Cluster sampling, 1136
- CM, *see* Configuration management; Construction management
- CMAC networks, 1780
- CMIP, *see* Common management information protocol
- CMMS, *see* Computerized maintenance management systems
- CMO (chief maintenance officer), 1621
- CMs (configuration mechanisms), 691
- CMS Research Inc., 2458
- CNC lathe programming, 1032–1034
- CNMA (Communications Network for Manufacturing Applications), 165
- Coaching, 938
- Coalitions (group), 2211
- Coal miners’ pneumoconiosis, 1169
- Coca-Cola, 2135
- Cochran’s test, 2255
- Code of Federal Regulations (CFR), 591, 1097–1098, 1162
- Code reuse (computer programming), 71–72
- Codesign, 604, 606
- Coding (group technology), 461–462
- Cogeneration (energy), 1577

Cognitive aids, 1027, 1032–1037
  CNC lathe programming example, 1032–1034
  computer technologies as, 1223–1224
  managerial planning, 1034–1037
Cognitive continuum theory, 2200
Cognitive design/engineering, 1014, 1205–1217
  and contextual task analysis, 1206–1211
  and requirements definition, 1206
  and usability evaluation, 1216–1220
  and user interfaces, 1212–1216
Cognitive ergonomics, 1014
Cognitive mapping (for decision structuring), 2190
Cognitive task(s), 1013–1039
  and action-cycle model, 1017–1019
  analysis of, *see* Cognitive task analysis (CTA)
  and blue-collar workers, 1013
  decision making as, 1023–1024
  definition of, 1013
  design of cognitive aids for, 1032–1037
    CNC lathe programming example, 1032–1034
    managerial planning, 1034–1037
  diagnosis as, 1022–1024
  ergonomic interventions in area of cognitive tasks, 1014
  and human information-processing model, 1014–1017
  and skill-, rule-, and knowledge-based (SRK) model, 1019–1021
Cognitive task analysis (CTA), 1024–1031
  active participation in, 1027
  examples of, 1032–1033
  functional model derived from, 1027
  scope of, 1025
  stages in, 1025
  techniques for, 1025, 1028–1031
    Critical Decision Method (CDM), 1028, 1030–1031
    hierarchical task analysis, 1028, 1029
  and theoretical models, 1026
Cognitive tunnel vision, 1023
Cognos, 83
Cold extrusion, 565
Cold-formed components, 575–580
  extrusion, 575–577
  orbital pressing, 579–580
  swaging, 577–579
Cold forming, 568
ColdFusion (programming language), 78–79
Cold heading, 1319, 1321
Cold molding, 1325, 1326
Cold standby components, 1933
Collaboration(s):
  networked, 234
  strategic, 34
  types of, 604, 605
  World Wide Web, supported by, 246, 256
Collaborative customer/demand planning, 968
Collaborative forecasting and replenishment, 779–780, 785
Collaborative manufacturing, 601–617
  and agile manufacturing, 602, 603
  coordination/control requirements in, 603–604
  and coordination cost, 607–609
  and distributed environment, 604, 607–609
  distributed manufacturing case example, 609–616
  and e-work, 606
  framework for, 604
  future of, 617
  implementation of, 604–606
  and knowledge-based economy, 602
  variable production networks for, 616–617
Collaborative technology, implementation of, 961–962
Collect-and-place principle (electronic components), 425
Collective inquiry methods, 127
Collective intelligence, 976
Collectivism (in national cultures), 957
Colliery (human factors audit), 1150–1151
Color coding, 548
Color saturation, 2506
COL rule, *see* Closest-open-location rule
COMBIMAN (computer-aided model), 1050, 1112
Combination-location storage, 1534, 1535
Comfort, digital human modeling of, 1120
Command and control systems, *see* Group decision support systems
CommerceNet, 269
Commercialization, and supply chain management, 2122
Commercial network applications, 250
Commission of European Communities, 1144
Commitment, and effective teamwork, 982
Commonality (product families), 688–689
Common bases (CBs), 690–691
Common Gateway Interface (CGI), 77–78
Common management information protocol (CMIP), 731–732
Common object request broker architecture (CORBA), 714, 719–722, 732
Common random numbers (CRN) technique, 2492–2493
Commonwealth Edison, 654
Communication(s):
  among professional groups, 23
  in client/server (C/S) systems, 718–722
    CORBA, 719–722
    DCOM, 721
    Java RMI, 721
    remote procedure call (RPC), 719
    socket interface, 718–719
  electronic, 232, 233
  interpersonal/interagent, 174
  and knowledge engineering, 1291–1293
  for knowledge management, 220
  of nature of the business, 51
  network, *see* Networks/networking
  percent of GDP in, 346
  and relationship management, 49
  in supply chain design, 2131

  in supply chain management, 2125
  three types of languages for, 132
  through WBS as project dictionary, 1277
  of total quality leadership (TQL) philosophy, 1801, 1802
  and work team, 880, 881
  World Wide Web as tool for, 246
Communications management:
  automation of, 1259, 1260
  project, 1248
Communications Network for Manufacturing Applications (CNMA), 165
Communications planning, 1248
Communication system, 17
Communication technologies:
  and integration of automation technologies, *see* Integration technology
  and rapid product development, 1284
Community, as external force, 39
COMNET, 2447
Compact discs (CDs), online retailing of, 266
Compaq, 265, 662
Comparability (of models), 284
CompareNet, 671
Comparison method (of cost estimating), 2301–2302
Compensable factors (in jobs), 907–909
Compensating non-linear elements, 1883
Compensation:
  incentive pay, time/job determination for, 1392
  and job evaluation and job evaluation systems, 910–911
  and leadership, 861–862
Compensatory decision rules, 2178
Competence, as company asset, 1888
Competencies, 937–938
Competition-based neural networks, 1779–1780
Competition/competitiveness, 39
  analysis, competitive, 1212
  and business environment, 32–33
  collaboration, competitive, 605
  customer service as key success factor in, 634–635
  encouragement of, by Toyota, 556
  global, 1888
  major issues in, 1313
  and pricing, 668, 682
  promotion via double auctions, 277
  and retail supply chains, 781
Competitive advantage:
  and customer service, 651
  as driver of new technology implementation, 954
  and just-in-time (JIT), 544
  speed as source of, 2116
Compiler (computer tool), 71
Complaints, customer, 641, 658
Complementary slackness, 2554
Complete anonymity, 268
Completely randomized experimental designs, 2230
Complexity:
  of megaprojects/megaprograms, 1002–1003
  in modeling, simplicity vs., 1631
Component decomposition analysis (enterprise resource planning), 349
Componentware, 285
Composite component concept (group technology), 462–463
Compound amount factor (interest), 2337–2339, 2345
Compound interest, 2336–2346
  continuous compounding, 2352, 2353
  continuous uniform cash flows compounded continuously, 2345–2346
  discrete cash flow compounded continuously, 2343–2345
    compound amount factor, 2345
    geometric conversion factor, 2345
  discrete cash flows compounded discretely, 2337–2343
    arithmetic gradient conversion factor, 2342–2343
    capital recovery factor, 2340–2341
    compound amount factor, 2337–2339
    present worth factor, 2338–2339, 2341
    sinking fund factor, 2339–2340
  discrete compound interest factors, 2354–2357
  geometric series factors, 2358–2359
Comprehensive Environmental Response, Compensation and Liability Act (CERCLA), 594, 1164
Comprehensive services business model, 603
Compressed air systems, 1581
Compression molding, 1324, 1326
COMPU-RATE stopwatch, 1412, 1414
Computational complexity theory, 2594–2595
Computer-aided analysis of working postures, 1061
Computer-aided assembly methods, 386–388
  layout planning/optimization, 386–388
  simulation of material flow, 388
Computer-aided design (CAD), 178–195, 494, 2498, 2499, 2509–2511, 2515, 2517–2519
  assembly technology in, 185–187
  with ECAD systems, 188, 190
  and engineering solution center, 1290, 1291
  feature technology in, 185
  graphic 3D simulation systems, 378, 379
  integration of process planning with, 191
  interfaces for, 191–195
    classification of, 191
    definition, 191
    IGES, 192–193
    product data exchange, 192
    standardization of, 191–195
    STEP, 193–195
  parametrical modeling with, 183–185
  with pipe design, 187, 189
  solid modeling with, 182–184
  surface modeling with, 180–182
  3D, 180–183
    solid modeling with, 182–183
    surface modeling with, 180–182
  2D, 178–180, 190

Computer-aided design (CAD) (*Continued*)
    dimensional-oriented modeling, 180
    geometrical elements available in, 178–179
    strategies for using, 179
    3D vs., 178
  underlying architecture for, 210–211
  with weld design, 188, 190
Computer-aided design/computer-aided manufacturing (CAD-CAM):
  and computer integrated manufacturing, 494–496
  integration of, 495, 496
  for manufacturing design, 1328–1329
  process chains, CAD/CAM, 420–422
Computer-aided diagnosis, of assembly systems, 422–423
Computer-aided manufacturing (CAM), 495, 949
Computer-Aided Manufacturing International, 474
Computer-aided models, of man, 1050
Computer-aided process planning (CAPP), 460–482, 494–495
  advantages of, 473
  capability analysis, 465, 468–471
  CAPP, 474–475
  cost model, 465–467, 472
  development of, 473–474
  generative approach to, 477–478
  geometry analysis in, 452
  group technology, 461–463
  mapping, 463–466
  selection criteria for, 478–482
  tolerance charting, 472–473
  variant approach to, 475–477
Computer-aided project management (CAPM), 1252–1262
  automation of, 1256–1260
  future of, 1261–1262
  history of, 1253
  implementation of, 1260–1261
  and project concentric circle model, 1253–1255
Computer-aided quality-management system, 497–498
Computer-aided software engineering (CASE):
  for IS system development, 105
  as modeling tools, 304
  Teamwork software package, 173
Computer-assisted facility management (CAFM), 1566
Computer-based instruction, 928
Computer-based scheduling, 1735–1738, 1765
Computer-based training (CBT), 222, 940
Computer conferencing systems, 142
Computer control systems, 500
Computer industry, channel marketers in, 264
Computer integrated manufacturing (CIM), 85, 485–528, 491–499
  and agile manufacturing, 527
  CAD/CAPP/CAM system, 494–496
  components of, 491–499
    CAD/CAPP/CAM system, 494–496
    computer-aided quality-management system, 497–498
    computer networks, 498–499
    database management system, 499
    hardware, 491
    management information system, 491–494
    manufacturing automation system, 496, 497
    software, 491
  computer networks, 498–499
  for control, 1772–1775
    component architecture of, 1773–1774
    component-specification methodology of, 1774
    shop-floor application modules in, 1774–1775
  database management system, 499
  definitions of, 487–488
  and enterprise model, 507–514
    ARIS, 512, 513
    CIMOSA method, 510–512
    function view, 508, 509
    GIM, 512–514
    information view, 509, 510
    organization view, 510
    process view, 507–508
    resource view, 510
  Flexible Manufacturing Systems, 499–507
    benefits of, 506, 507
    definition of, 499–500
    design of, 500–501
    limitations of, 507
    modeling/simulation in, 503–506
    planning/scheduling/control in, 501–503
  future of, 527–528
  and green manufacturing, 527
  hardware, 491
  and human factors, 488
  and human resource quality, 526–527
  implementation of, 514–518
    detailed system design phase, 515
    feasibility study phase, 514, 515
    implementation phase, 516
    and integration platform technology, 516–518
    overall system design phase, 514, 515
  integration as core of, 489–491
  management benefits of, 526
  management information system, 491–494
  manufacturing automation system, 496, 497
  and manufacturing environment, 485–486
  and manufacturing systems, 486–487
  in process industry, 518–526
    and architecture structure model, 520–521
    definitions, related, 518–519
    and hierarchical structure model, 521–522
    and information integration, 522–523
    key technologies, development of, 519
    refinery enterprise example, 523–526
  software, 491
  system vs. information view of, 488–489
  technical benefits of, 525–527
  and virtual manufacturing, 527–528

Computer Integrated Manufacturing Open System Architecture (CIMOSA), 301–302, 489–490, 507–509, 510–512
  cube, CIMOSA, 511, 512
  in process industry, 519
Computerized enterprise modeling, 303
Computerized maintenance management systems (CMMS), 1591–1592, 1605–1610
Computerized work sampling, 1458
Computer languages, 2446, 2449–2450, 2454–2456. *See also specific languages*
Computer-mediated group decision making, 2214
Computer networking, *see* Networks/networking
Computer records, security for, 1568
Computers, 36
  for cost estimating, 2316
  human interaction with, *see* Human–computer interaction
  human modeling using, *see* Digital human modeling
Computer Sciences Corporation, 775
Computer simulation, 2445–2465
  advances in, 2463–2465
  applications of, 2461
  for cost estimating, 2306
  evaluation of tools for, 2449–2454
    hardware characteristics, 2451
    input characteristics, 2450
    language characteristics, 2449–2450
    methodology for, 2452–2454
    operational characteristics, 2450
    output characteristics, 2450–2451
    vendor characteristics, 2451–2452
  human performance modeling, *see* Human performance modeling
  for human strength design, 1054
  languages, simulation, 2446, 2449–2450, 2454–2456
  packages, simulation, 2446, 2456–2461
  pitfalls in use of tools for, 2461–2463
  of revised NIOSH lifting equation, 1079, 1080
  selection of tools for, 2447–2448
  support software, 2446–2447
Computer software, 68–69. *See also specific programs*
  for artificial intelligence approaches to control, 1782
  componentware, 285
  for information systems, 285–286
  for linear programming, 2534–2536
  for network flow models, 2572
  for nonlinear programming, 2563–2565
  scheduling and development/implementation of, 1737–1738
  surveys of, 1260
Computer-supported collaborative work (CSCW), 603–604, 606, 1284, 2503
Concentration allowances, 1397
Conceptual data modeling:
  in decision support systems, 119, 120
  for information systems, 102–103
Conceptual design document (human-centered product planning/design), 1307
Concurrent collaboration, 605
Concurrent engineering, 206–207, 485, 486, 556
  collaboration required by, 603
  and design by customers, 701–702
  plant engineer's involvement in, 1551
Concurrent flexible specifications (CFS), 172–174
Concurrent truck travel, sequential vs., 1510–1511
Concurrent validity, 1134
Conditional statements (Structured English), 101
Conditioning arguments (decision making), 2192
Conditions for success, 15–18
  communication system, 17
  culture system, 15–16
  infrastructure, 16
  ISE's role in, 6
  learning system, 17
Conference Board, 965
Conference estimating, 2300–2301
Conferences, computer-supported, 142
Confidence intervals, 2253–2254, 2485–2487
Configuration management (CM), 1259, 1274, 1276
Configuration mechanisms (CMs), 691
Confirmation bias, 1023
Conflict:
  affective vs. substantive forms of, 2211
  in decision making, 2174–2175
  and group decision making, 2210–2212
  management of, and team effectiveness, 982–983
  sources of, 2175–2178
Conflict analysis, 128
Conflict-driven decision making, 2176
Conflict resolution:
  in decision making, 2175–2178
  in group decision making, 2211–2212
Conformance, as performance measure of quality, 1246
Confounding, 2276, 2277
Confusion, 997
Congestion, 2037, 2044
Congressional Office of Technology Assessment, 950
Conjoint analysis, 702, 703
Conjoint measurement theory (decision making), 2195
Conjugate direction method, 2552, 2553
Conjugate gradient methods, 2552–2553
Connectionist processing model, *see* Neural networks
Connections, *see* Joining technologies
Connectivity, Internet, 254–255
CONOPT, 2563

Consensus:
  meeting, in business model, 31
  as preferred team decision-making strategy, 982
Consequentialism (decision making), 2178
Constant cone domination structures, 2615–2616
Constant dollar analysis:
  actual dollar vs., 2397–2398
  of after-tax cash flow, 2404
  of before-tax cash flow, 2402
  with differing inflation rates for component cash flows, 2400
  with differing inflation rates per time period, 2401
  of economic equivalence calculations with inflation, 2398–2400
Constant tasks, 740
Constrained optimization (nonlinear programming), 2553–2562
  feasible directions, methods of, 2559–2560
  geometric programming problems, 2558–2559
  Karush–Kuhn–Tucker conditions for, 2554–2555
  Lagrange multipliers for, 2553–2554
  and nonsmooth optimization, 2562
  quadratic programming problems, 2555, 2562
  separable programming problems, 2556–2558
  sequential unconstrained minimization techniques for, 2560–2562
  successive linear programming, 2562
  successive quadratic programming, 2562
Constraint(s):
  binding/nonbinding, 2541
  control, constraint, 161
  definition of, 557
  in mathematical programs, 2540, 2541
  process, 466
  selection, 691
  Theory of
    and JIT, 557–558
    and just-in-time (JIT), 557–558
  in transportation management, 2055
  work constraints, 1024–1025
Constraint-based computational models, 348
Construction. *See also* Site selection and construction
  facilities for, 330, 331
  percent of GDP in, 346
  and plant engineering, 1565–1566
Construction management (CM), 1493–1494
Constructive Solid Models (CSGs), 182
Construct validity, 1134
Consultants, for site selection, 1476
Consumer goods, online auctions of, 275–276
Consumer Goods Manufacturer Magazine, 775
Consumer Price Index (CPI), 2395
Consumer surplus, 669, 676
Containerization, 1503
Containment, hazard, 1175
Content, network:
  classification of, 247–248
  generation/provision of, 246–249, 251–252
  rating/filtering of, 248–249
Content industry, 251
Content validity, 1134
Context diagrams, 100
Contextual task analysis (human–computer interaction), 1206–1211
  background information for, 1206
  data collection/analysis, 1208–1210
  task allocation, 1210–1211
  user profiles, 1207–1208
  work practices models, 1210
Contingency factors:
  budgeting for, 1344
  warehousing, 1530
Contingent decision making, 2207
Continual improvement, 1972
Continuity index, 1734
Continuous change (term), 1228, 1229
Continuous data, 1837
Continuous improvement:
  in just-in-time (JIT), 548
  motivation for, 963
  and performance management, 1000
  and process design and reengineering, 1712
  in total quality leadership (TQL) process, 1802
  as TQL success factor, 1805
Continuous operation/process facilities, 329–331
  major process engineering task in, 334
  personnel scheduling for, 1743–1744, 1755
Continuous quality improvement (CQI), 747
Continuous reliability improvement (CRI), 1610–1611
Continuous sensors, 1903–1904
Continuous-time dynamic-simulation, 128
Continuous timing, 1420
Continuous uniform cash flows compounded continuously (compound interest), 2345–2346
Contract engineering, 330–331
Contract maintenance, 1622
Contract manufacturing, 263–264, 330–331
Contractors, selection of, 1499
Contracts, 49
  and enterprise resource planning, 336, 337
  management of, through project life cycle, 1250
  personnel scheduling and changes in, 1755, 1757
  personnel scheduling and negotiation of, 1765
  transportation service, tracking of, 335
Contribution ratio, *see* Profit-volume ratio (PV)
Control(s), 1768–1787. *See also* Monitoring
  AI approaches to, 1775–1782
    commercial software, 1782
    fuzzy set theory, 1781–1782
    genetic algorithms (GA), 1780–1781
    knowledge-based systems, 1775–1776
    neural networks, 1777–1780
  in business processes, 45–48
  CIM Framework for, 1772–1775
    component architecture of, 1773–1774

    component-specification methodology of, 1774
    shop-floor application modules in, 1774–1775
  cost, *see* Cost control
  distributed vs. central, 166, 167
  engineering, 1175–1176
  feedback, 158
  in Flexible Manufacturing Systems, 501–503
  human factors controls, 1176–1179
  manufacturing execution systems (MESs) for, 1782–1787
  panels, control, 1016
  project, 1347–1348
  Purdue Enterprise Reference Architecture (PERA) for, 1769–1772
    control hierarchy in, 1769–1771
    equipment organization in, 1771–1772
  safety features on, 1178
  span of, 1264
Control charts, 1818, 1821, 1825, 1826, 1832–1834
  C charts, 1847–1851
  for determining performance of processes, 1830–1831
  in health care systems, 745–746
  P charts, 1844–1847
  R control charts, 1850–1855
  Shewhart, *see* Shewhart control charts
  for statistical process control (SPC), 1861–1875
    and AT&T runs rules, 1863–1864
    data patterns on, 1863
    variables, charts for, 1864–1871
  U charts, 1847–1849, 1851
  with work sampling, 1457
  X-bar charts, 1850–1855
Control-decision support system (CDSS), 1777
Control limits, 1840
  standards known, 1864–1866
  standards not known, 1866–1868
Control message standard, 168
Control systems, 156–161. *See also* Artificial intelligence (AI); Integration technology
  definition of, 157
  instrumentation of, 158
  kanban, 549–551
    alternatives, 550–551
    appropriate environments for, 545
    case study, 551
    control parameters, 550
    limitations, 550
  models for, 159–161
  robotic, 376–378
  shop floor, 699–701
Conversational structuring (in software), 142–143
Convex combinations, 2543
Convex functions, 2543, 2544
Convexity (in linear programming), 2543–2546
Convex sets, 2543
Conveyors (material handling), 1504, 1513–1520
  belt conveyors, 1513, 1514, 1516
  cart-on-track conveyors, 1518
  chain conveyor, 1516, 1517
  chute conveyors, 1513
  power-and-free conveyor, 1518, 1519
  roller conveyor, 1514, 1516
  skate wheel conveyor, 1515–1517
  slat conveyor, 1515, 1517
  sortation conveyor, 1518–1520
  tow-line conveyor, 1517
  trolley conveyor, 1517–1518
Cooperation, knowledge-intensive, 1291, 1292
Cooperative collaboration, 605
Cooperative processing/database management, 125
Coordination cost, 607–609
CORBA, *see* Common object request broker architecture
Core business processes, 30, 41, 43, 58–59
  analysis of, 52
  ERP tools for use with, 89–92
    accounting and finance, 91
    human resources, 91–92
    manufacturing and procurement, 90–91
    sales and distribution, 90
  knowledge management (KM), 215, 216
  project management, 1254
  risks for, 45
Core competencies, 42
Core products and services, 30, 49–50
  in business model, 34
  categories of, 49
  definition, 49
  measurement of, 49–50
Corning Asahi Video, 1712
Corporate culture, *see* Organizational culture
Corporate portals, 271
Correctness of models, 284
Correlation, 2271
Cost(s):
  capitalized, 2350–2351
  CIM implementation and reduction of, 526
  classification of, 672–673
  of digital products, 270
  dynamic decision problems, 2638–2639
  of electricity, 1575–1576
  of energy, 1576
  of engineering changes at different life cycle stages, 1312, 1313
  estimating, *see* Cost estimating
  of failed technology implementation, 949–950
  in Internet economy, 267
  of machinery, 467
  manufacturing, 455
  minimax, 2177, 2180–2181
  and outsourcing, 263
  as performance management metric, 1005
  personnel, determination of, 1343–1344
  and postponement, 2115–2116
  and pricing, 663, 667, 672–674
  in retail supply chains, 774, 775
  and speculation, 2116
  of work-related injuries/deaths, 1157
Cost accounts, 1272–1273

Cost allocation:
  learning and, 1405
  time/job determination for, 1392
Cost-benefit analysis:
  activity-based costing (ABC) for, 1704
  IS systems, 98–99
Cost breakdown structure (CBS), 1273
Cost control, 207
  in material handling, 1355, 1356
  by plant engineers, 1563–1564
  for projects, 1246
Cost drivers, 2319
Cost estimating, 2298–2316
  accounting data, use of, 2309, 2310
  alternatives, comparison of, 2346–2350
    annual worth method, 2347–2348
    benefit–cost method, 2349–2350
    future worth method, 2348
    payback period method, 2349
    present worth method, 2346–2347
    rate of return method, 2348–2349
  analytical methods of, 2302–2304
  comparison method of, 2301–2302
  conference estimating, 2300–2301
  and economic want, 2299
  factor method of, 2302
  forecasting techniques for, 2310
  historical data, use of, 2307
  indexes, cost-estimating, 2310–2311
  and labor analysis, 2307–2308
  and material analysis, 2308–2309
  and operations estimating, 2311–2314
  power law technique for, 2303, 2304
  probability and statistical techniques for, 2304–2306
    computer simulation, 2306
    expected value, 2304
    percentile method, 2305
    PERT, 2305–2306
  and product estimating, 2314–2316
  request for estimate, 2299–2300
  standard data, use of, 2306, 2307
  types of, 2298–2299
  unit method of, 2301
Cost estimating relationships (CERs), 2302–2304
Costing:
  activity-based, 2317–2319, *see* Activity-based costing (ABC)
  conventional systems of, 2317–2318
  by plant engineers, 1562–1563
Cost management, 2317–2318
  automation of, 1256, 1257
  improvements in, *see* Activity-based management (ABM)
  project, 1245–1246
Cost matrix, 2376, 2377
Cost model (for process planning), 465–467, 472
Cost of assembly (CA), 363–364
Cost-performance ratio, 727
Cost pools, 2319
Council of Industrial Engineering (CIE), 23
Council on Logistics Management (CLM), 348, 2113
Count data, C and U charts for, 1847–1851
Counterbalanced lift trucks, 1506, 1508, 1509
Coupled joints, 1115
Coupled process chains, 204
Coupling classification, 1078
Coupling of work, 880
Coupons, 678
Courier robots, 379–380
Covariates, 2280
COVISE, 2512
CPI (Consumer Price Index), 2395
CPLEX, 2535, 2575
CP rule, *see* Critical path rule
CPSC Children (anthropometric database), 1114
CQI (continuous quality improvement), 747
Creation of knowledge, 215
"Creative destruction," 1888
Creativity models, 1812, 1814
CrewChief, 1050, 1112, 1118
Crew scheduling, 1743–1744, 1755–1757
Crew workload, evaluation of, 2420–2427
  future command and control process, modeling workload of, 2421–2425
  other environments, extension to, 2424–2427
CRI, *see* Continuous reliability improvement
Criteria for Performance Excellence, 1956
Critical Decision Method (CDM), 1028, 1030–1031
Critical path, for professional services projects, 1341, 1342
Critical path (CP) rule, 1722, 1724
CRM, *see* Customer relationship management
CRN technique, *see* Common random numbers technique
Cross-docking, 778
Cross-training, 934
Crowds (privacy service), 268–269
CRP (capacity requirement planning), 2042
CRT screens, *see* Cathode ray tube screens
Cryptographic systems (cryptosystems), 733
CSCW, *see* Computer-supported collaborative work
CSGs (Constructive Solid Models), 182
C/S systems, *see* Client/server systems
CTA, *see* Cognitive task analysis
CTDs, *see* Cumulative trauma disorders
CTP (capable-to-promise), 2046
CTS, *see* Carpal tunnel syndrome
Cuban Missile Crisis, 139
Culture. *See also* National culture; Organizational culture
  and alignment of technology/organizational structure, 956–961
  safety, 959–961
Culture shift, 14, 16
Culture systems, 15–16, 1798
Cumulative distribution function (CDF), 2385–2386
Cumulative trauma disorders (CTDs), 1082, 1083
Currencies, 761

Customer(s), 30, 34, 35, 50–51
  categories of, 50
  commitments, customer, 1962
  decision making, customer, 703–704
  definition of, 50
  design by, 701–703
  determining requirements of, 1708
  as external force, 38
  focus on, in EPEM model, 1798
  internal, 14, 23
  Internet and security/privacy of, 267–269
  lifetime value of, 651, 652, 654
  and markets, 50–51
  nature of, as industry categorizer, 329
  needs/wants/demands of, 327
  relative importance of, 1381, 1382
Customer-based approach to service quality, 639–640
Customer-driven organizations, 1797
Customer-driven quality results, 1805
Customer knowledge, 1962
Customer management, 637
Customer orders, 329–331
Customer relationships, 1962–1963
  as aspect of lean production, 557
  Baldrige criteria for, 1963
  globalization and changes in, 1888
Customer relationship management (CRM), 14, 34, 35
  as dimension of competitive advantage, 326–327
  and ERP, 95–96, 337
  software, 90, 95
  and supply chain management, 2121
  systems, CRM, 69
Customer satisfaction, 651–654, 657, 1962–1963
  Baldrige criteria for, 1963
  and iCollaboration tools, 968
  measurement system for, 657
  and price of product, 668
  and queueing models, 1629
  and service encounter, 624–625
  and service quality, 640
  service quality vs., 628–629
  and site selection, 1468–1469
  wheel of success, 652
Customer service, 651–663
  applications of, 654, 655
  audit of, 662, 663
  and complaint management, 658
  and customer satisfaction, 651–654
  department, customer service, 657–660
    call centers in, 658
    and centralization, 657
    complaint management by, 658
    hiring/training/retaining employees for, 659
    and internal customers, 659–660
    organization of, 657
    and service-quality standards, 657–658
  ensuring focus on, 654, 656–657
  future of, 660–662
  importance of, 653–654
  and internal customers, 659–660
  as key success factor in competition, 634–635
  in retailing, 779
  revenues from improved, 652–653
  and service-quality standards, 657–658
  and supply chain design, 2130–2131
  and transportation management software, 2065
Customer service employees, training of, 659
Customer Service Group, 657
Customer service management, 2121
Custom information systems applications, 285–286
Customization of products and services, 261–262. *See also* Mass customization
Custom-made software, 68
Cutting tools, 457, 459
CWA, *see* Clean Water Act
Cyberglove, 1125
CyberGold, 273
Cybermediaries, 271
Cybertec, 1738
Cycle time, variation in, 1829
Cyclical fluctuations (in space planning), 2088–2089
Cyclic coordinate search method, 2549
Cyclic scheduling, 1746–1747
Cylindrical robots, 375

DAMES (design acronym), 1387–1389
Dark fiber optical cables, 250
Data:
  collection of, *see* Data collection
  in context of knowledge management, 214
  flow of, in process-oriented enterprise modeling, 288–290
  independence, data, 115, 116
  integration, data, 89
  models, data, *see* Data models
  objects, data, 118
  redundancy, data, 115–117
  sharing, data, 94–95
  standard, for work measurement, 1443–1445, 1447–1448
  stores, data, 99
  types of, 1836–1838
Data access, traditional reporting vs., 37
Data analysis, 221
  in contextual task analysis, 1208–1210
  in human factors audits, 1145–1146
Data backups (warehousing operations), 2103
Database management systems (DBMSs), 113–125, 114–125
  for computer integrated manufacturing (CIM), 499
  cooperative approach to, 125
  data models for, 117–124
    expert database model, 122–124
    external, 119–124
    generic types of, 119–120
    levels of, 117–120

Database management systems (DBMSs) (*Continued*)
        object-oriented database models, 122–124
        record-based models, 120
        single-level data model, 118–119
        structural models, 120–122
    definition of, 80
    design of, 116–117
    distributed, 124–125
    integration of, with decision support system, 117
    model base management systems (MBMSs) vs., 125
    objectives for, 115, 116, 124, 125
    organization of data in, 80
    primary objectives of, 125
    selection of, 117
    tools used in
        data warehouses, 83–85
        for information systems, 79–85
        object-oriented databases, 82–83
        and relational database model, 80–81
Databases:
    anthropometric, 1113, 1114
    gateway, database, 84
    as human-centered product planning/design tool, 1302, 1303
    object-oriented, 82–83
    project activity/historical, 1260
    servers, database, 240
    for warehouse operations, 2095–2103
        backups, data, 2103
        equipment masters, 2097, 2099–2102
        flow control, 2097, 2098
        hardware controllers, links to, 2103
        products and orders, 2096–2097
        protocols, 2102–2103
    World Wide Web, 245
Database systems, in client/server (C/S) systems, 723
Data clouds, 1123, 1124
Data collection:
    in contextual task analysis, 1208–1210
    forms for, 1810, 1811, 1813
    in human factors audits:
        AET, 1138–1140
        checklists, 1137–1145
        ERGO/EEAM/ERNAP, 1141–1143
        Ergonomic Checkpoints, 1144
        Ergonomics Audit Program, 1139, 1141
        IEA Checklist, 1137, 1138
        Position Analysis Questionnaire (PAQ), 1137–1139
        Upper-Extremity Checklist, 1143–1144
    major activities of, 1770
Data/control flow diagrams (DFD/CFDs), 173
Data definition languages (DDLs), 119
Data dictionaries (DDs), 102–103, 119
Data Encryption Standard (DES), 733
Data flow diagrams (DFDs), 99–101
Data gloves, 1125
Data-improvement process (TQL), 1805
Data manipulation language (DML), 118, 119
Data Mark, 84
Data mining, 83, 84, 2013
Data models, 117–124
    component sets in, 119
    conceptual, 119, 120
    expert database model, 122–124
    external, 119–124
    internal, 119, 120
    levels of, 117–120
    object-oriented database models, 122–124
    record-based models, 120
    single-level model, 118–119
    structural models, 120–122
DataMyte 1010 stopwatch, 1412, 1455
Data-presentation elements (measurement systems), 1878
Data query languages (DQLs), 119
Data requirements (for modeling), 1631
Data resource control, 115, 116
Data warehouses/warehousing, 83–85, 221
    terms related to, 84–85
    tools used in, 83–85
Day sleeping, 1367
Day's work:
    acceptable, 1392, 1405, 1406
    fair, 1411
DBMSs, *see* Database management systems
DBR (drum-buffer-rope) scheduling, 558
DCOM (distributed component object model), 721
DDLs (data definition languages), 119
DDoS (distributed denial of service) attacks, 278
DDs, *see* Data dictionaries
DEs (differentiation enablers), 691
Deadheading, 1513
Death rates (from occupational injuries/diseases), 1157
Debt management (ERP), 336
DEC, *see* Digital Equipment Corporation
Decentralization warehousing strategy, 2071–2072
Decentralized business (human factors audit), 1146–1150
Decentralized decision making, 698
Decentralized planning (rapid product development), 1288
Decision aids, computerized, 965–968
    iCollaboration, 966–968
    TOP Modeler, 965–966
Decision analysis, 129, 2187–2195
    and decision trees, 2187–2188
    preference assessment in, 2194–2195
    probability assessment in, 2191–2193
    structuring of decisions in, 2187–2191
        cognitive mapping for, 2190
        decision matrices for, 2187–2188
        event trees (networks) for, 2189–2190
        influence diagrams for, 2190–2191
        value trees for, 2188–2189
    utility function assessment in, 2193–2194
Decision banding, 910
Decision-driven change, 1008–1009
Decision making, 2173–2178. *See also* Optimization; Risk analysis; Sensitivity analysis
    cognitive probes for, 1026

  as cognitive task, 1023–1024
  conflict-driven, 2176
  consensus as preferred team strategy for, 982
  contingent, 2207
  customer, 703–704
  decentralized, 698
  decision rules for, 2177–2178
  dynamic, 2176, 2205
  elements of, 2174–2175
  employee involvement in, for interactive system design, 122, 1221
  explanation-based, 2207–2208
  forming, storming, norming, performing model of, 2210
  group, 2176, 2209–2214
    and biases, 2212
    computer-mediated, 2214
    and conflict, 2210–2212
    prescriptive approaches for improving, 2212–2214
    and social norms/ethics, 2209–2210
  individual vs. group, 141
  information technology and options in, 1889
  integrative model of, 2175–2178
  knowledge-based, in inspection systems, 1898–1899
  logistical, *see* Logistics management
  material-handling, 1504
  model development for, 1630
  multicriteria, *see* Multicriteria optimization
  and national culture, 958
  naturalistic, 2205–2209
    contingent decision making, 2207
    and dominance structuring, 2207
    explanation-based decision making, 2207–2208
    and image theory, 2207
    recognition-primed decision making, 2205
    and shared mental models, 2208
    and team leadership, 2208
  organizational
    and health/safety performance, 1179
  for plant/facilities engineering, 1561
  pricing, 674–677
  problem solving vs., 2173
  recognition-primed, 2205
  routine, 2176
  rule-based, in inspection systems, 1896–1898
  stochastic models for, *see* Stochastic models
  and test and inspection, 1890
Decision making hierarchy (manufacturing), 487
Decision matrices, 2187–2188
Decision process (inspection systems), 1896–1899
Decision rooms, 134
Decision rules, 2177–2178
Decisions. *See also* Decision making
  framing of, 2202–2203
  related to team design, 977–978
  under risk, 2377–2378
  structuring of, 2187–2191
    cognitive mapping for, 2190
    decision matrices for, 2187–2188
    event trees (networks) for, 2189–2190
    influence diagrams for, 2190–2191
    value trees for, 2188–2189
  under uncertainty, 2377–2381
Decision structure tables, 1385, 1387
Decisions under uncertainty, 2378–2381
Decision support systems (DSSs), 67, 84, 110–149, 2011–2019
  abilities supported by, 111, 113
  analytical tools for, 2013–2015
  components of, 110
  control in, 130
  database management systems, *see* Database management systems
  definition of, 110
  dialog generation and management systems, 113, 115, 131–134
  distributed group, 145
  and DSS generator software, 114
  enterprise resource planning (ERP)
    algorithm development, 348
    applications, 339, 340
  feedback in, 130
  flexibility in, 130
  frameworks for engineering of, 113–114
  group, 134–145
    distributed group decision support systems, 145
    engineering of, 141–145
    information needs for, 135–141
  for health care delivery systems, 747–748
  increased consistency in, 130
  input data for, 2012–2013
  interface in, 130
  knowledge management for, 145–149
  and logistics systems, 2011–2019
    analytical tools, 2013–2015
    input data, 2012–2013
    presentation tools, 2015–2018
  manufacturing, 348
  MIS/PMIS vs., 112–113
  model base management systems (MBMSs), 113, 115, 125–131
    and issue analysis, 127–129
    and issue formulation, 126
    and issue interpretation, 129
    and model base management, 129–131
    objectives for, 125
  presentation tools for, 2015–2018
    algorithms and GIS, integration of, 2018
    geographic information systems, 2016–2018
  redundancy reduction in, 130
  in supply chain planning, 2010
  in transportation planning, 2011
  and types of decisions, 111, 112
  for warehousing, 2079–2081
Decision theory (generally), 2376–2382
  aspiration level principle in, 2378
  dominance principle in, 2377
  expectation principle in, 2377–2378
  Hurwicz principle in, 2379–2380
  Laplace principle in, 2380–2381
  maximax principle in, 2379
  maximin principle in, 2379
  minimax principle in, 2378–2379

Decision theory (generally) (*Continued*)
  minimin principle in, 2379
  most probable future principle in, 2378
  naturalistic, 2177
  Savage principle in, 2381
Decision theory (behavioral), 2195–2205
  preference/choice in, 2201–2205
    and framing of decisions, 2202–2203
    labile preferences, 2204–2205
    and prospect theory, 2203–2204
    and subjective expected utility, 2202
  statistical estimation and inference in, 2196–2201
    biases, 2198–2199, 2201
    and human judgment models, 2200–2201
    and human limitations, 2196–2198
    selective processing, 2199–2200
Decision theory (classical), 2178–2187. *See also* Decision analysis
  choice procedures in, 2178–2184
    and axioms of rational choice, 2178–2179
    and dominance, 2179
    and elimination by aspects (EBA) rule, 2179–2180
    and expected utility theory, 2182–2183
    and holistic comparison, 2184
    and lexicographic ordering principle, 2179
    and maximization of expected value, 2181
    and minimax cost/regret, 2180–2181
    minimum aspiration level, 2180
    and multiattribute utility theory, 2183
  rationality in, 2178
  statistical inference in, 2184–2187
    Bayesian inference, 2184–2187
    Dempster–Schafer method, 2186–2187
    and signal-detection theory, 2185–2186
Decision times, 2636–2637
Decision trees, 1776, 2187–2188, 2382–2385
  deterministic, 2382–2384
  influence diagrams vs., 2190–2190
  stochastic, 2384, 2385
Decision under risk, 2377
Decision under uncertainty, 2377–2378
Decision variables (mathematical programs), 2540, 2541
Declarative knowledge, 1775
Decomposition, 2167–2170
Decomposition-based approach (production-inventory systems), 1692
Decomposition heuristics (scheduling), 1729–1731
Deconstructing value/supply chains, 43
DE (design efficiency), 369
Dedicated material-handling systems, 1660
Dedicated storage, 2092
DEDS, *see* Discrete event dynamic system
Deep-lane warehousing systems, 2089
Defect concentration diagram (SPC), 1860, 1861
Defect databases, 432
Deflation, 2394–2395
Degrees of freedom (DOF), 413, 414
Delay allowances, 1398, 1400
Delays, 1459
Deliverables (professional services projects), 1336, 1339–1341
Delivery, 2058, 2059. *See also* Pickup and delivery operations
Dell Computers, 90, 264, 265, 266, 272, 660, 662, 706, 782, 783, 969
Delphi technique, 127, 2213
Demand(s), 2020–2032
  average (constant), 2020–2023
    and economic order quantity (EOQ) model, 2022, 2023
    and inventory costs, 2021–2022
    training of flight attendants case example, 2022–2023
  customer, 327
  electric, 1575, 1576
  exponential smoothing model for forecasting of, 2029–2032
  forecasting/management of, 781–782
  and lead time, 2025–2027
  managing, 1742, 2121
  as mixture of distributions, 2027–2029
  over a single period, 2023–2025
  over multiple periods, 2025–2027
  price elasticity of, 668–669
  and pricing, 667, 668–672, 682
  in transportation, 789, 794
Demand chains, *see* Supply chain(s)
Demand over lead time, 1674–1675
Dematerialization, 535, 536
Deming, W. Edwards, 1831–1833
Deming Prize, 555
Demographic trends, 37
Demonstration (as measurement issue for successful design), 1299, 1301
Demonstration-based team training, 934
Demotivation, 997
Dempster–Schafer method (for decision making), 2186–2187
Deneb, 1050
Deneb Robotics, Inc., 2460
Deregulation, 38, 39
Derivative products, 49
Derived classes (computer programming), 72
Dermatitis, allergic/irritant, 1167
DES (Data Encryption Standard), 733
Descending bid auctions, 274
Descriptive knowledge, 67
Design. *See also* Design and process platform characterization methodology; Process planning
  action-cycle model in, 1018–1019
  alternative, 1049
  chair, 1204
  of cognitive aids, 1032–1037
    CNC lathe programming example, 1032–1034
    managerial planning, 1034–1037
  of computer–human interfaces, 1212–1216
  concurrent (simultaneous) engineering, 556
  by customers vs. for customers, 701–703
  DAMES steps for, 1387–1389
  and environmental regulations, 589–590
  ergonomic, *see* Ergonomic design

  evaluation of, 450
  of experiments, *see* Experimental design
  hierarchy of, 1313–1314
  of human–computer interaction, 1193
  of human factors audits, *see* Human factors audits
  integrated approaches to, 604
  in ISO 9001:2000 product realization clause, 1971
  job, *see* Job design
  for manufacturing. *see under* Manufacturing
  for mass customization, *see* Design for Mass Customization (DFMC)
  physical task criteria, 1048, 1049
  plant engineers involved in, 1565–1566
  probabilistic approach to, 1940
  process, *see* Process design and reengineering
  product, *see* Product design
  quality of, 1797
  for reliability, 1922, 1937, 1939–1940
  reliability program applications during, 1953
  review of, 1908, 1939
  rules/guidelines, 207
  for safety, 1177–1178
  supply chain. *see under* Supply chain(s)
  support/verification/evaluation/acceptance tests of, 1942–1943
  team, *see* Team design
  for test and inspection systems, 1914–1916
  of training, 926–927
  of work breakdown structure (WBS), 1268–1272
    geography-based, 1269, 1271
    logistics-based, 1271
    project-life-cycle-based, 1269, 1270
    technology-based, 1269
  workplace, *see* Workplace analysis/design
  workstation, 1202–1205
Design-bid-build, 1492–1493
Design-build, 1494–1495
Design characterization, 1978–1980
Design by consensus (hospitality industry), 826–830
  relationship charts for, 826–828
  relationship diagrams for, 829
  supervision, designing for, 829, 830
  utility use, design for, 830
Design efficiency (DE), 369
Designers, 1301
Design for assembly (DFA), 367–370, 384, 1328
Design for environment, 527
Design for manufacture and assembly (DFMA), 403
Design for Mass Customization (DFMC), 687–694
  commonality, 688–689
  common bases, 690–691
  customers, design by, 701–703
  derivation processes, 692–694
  modularity, 688–689
  multiple views, synchronization of, 691–692
  and product family concept, 688
  variety, product, 689–690
Design for the extreme, 1048
Design limits, 1056, 1057
Design manuals, 1998–1999
Design models, 448–449, 1979–1980
Design and process platform characterization methodology, 1976–2003
  capability analysis steps in, 1995–1996
  deployment of, 1999–2003
  and design characterization, 1978–1980
  linkage of product design and process platforms in, 1996–1999
  management commitment to, 2000
  measurement system characterization steps in, 1984–1987
  model development steps in, 1987–1993
  performance measures for, 2002–2003
  process definition steps in, 1982–1984
  and process platform development, 1980–1982
  and product development, 1977–1978
  statistical process control steps in, 1993–1995
Design review committee, 1939
Design-to-cost methods of product development, 207
Deskilling, 962
Desktop manufacturing, 586
Detailed design document, 1307
Detailing, process, 457–459
  optimization, process, 458
  parameters, determination of, 458
  and tool selection, 457–459
Detectability, error, 1371
Detection, 1030
Deterministic decision trees, 2382–2384
Deterministic multiperiod model (production-inventory systems), 1671
Development, 937–939
  Baldrige criteria for, 1960
  future of, 940
  and individual performance enhancement, 937–938
  and leadership, 859–861
  and organizational performance enhancement, 938–939
  of team design, 877
  of training, 926–927
  training vs., 937
DFA, *see* Design for assembly
DFD/CFDs (data/control flow diagrams), 173
DFDs, *see* Data flow diagrams
DFMA (design for manufacture and assembly), 403
DFMC, *see* Design for Mass Customization
DGMSs, *see* Dialog generation and management systems
Diagnose phase (process design and reengineering), 1697, 1708, 1709
Diagnosis:
  of assembly processes (electronic products), 432
  as cognitive task, 1022–1024
  specialized, 1634

Diagnosis systems (assembly), 422–423
Diagnostic related group (DRG) payment system, 738–739
Diagnostics, 2282–2288
  example of, 2286–2288
  internal validation, 2284
  notation for, 2283
  partial plots, 2286
  questions, diagnostic, 2282
  residuals, 2284–2285
  row deletion, 2284
Diagrams:
  activity cycle, 506
  "black box," 100
  context, 100
  data flow, 99–101
  entity relationship, 102, 103
  PERT, 104
  relationship (food service kitchen design), 829, 830
Dialog design, primary objectives for, 132–133
Dialog generation and management systems (DGMSs), 113, 115, 131–134
  and DBMS design, 117
  primary purpose of, 131
Die casting, 565
Die forging, 565, 569
Differential pricing, 677
Differentiation enablers (DEs), 691
Differentiators, service, 1957
Diffuse hypotheses, 137
Digital computer simulation, *see* Monte Carlo simulation
Digital economy, 107
  as basis for electronic commerce, 261
  flexibility in, 262
  industrial vs., 261
  outsourcing in, 263–264
  pricing in, 270–271
Digital Equipment Corporation (DEC), 487, 489, 2127
Digital human modeling, 1112–1127
  and anthropometry, 1113–1115
    databases, 1113
    methods, 1113–1115
  of comfort, 1120
  of fatigue, 1188–1119
  and immersive virtual reality, 1124
  kinematic representation in, 1112–1113
  of low-back injury, 1119–1120
  motion/animation in, 1116, 1120, 1125–1127
  and performance models, 1126
  posturing, human figure, 1115–1116
  and product design, 1121–1124
    accommodation, 1122–1124
    usability, 1123
  of strength, 1116, 1118
  tools for, 1116, 1117
  and workplace analysis, 1120–1121
Digital mock-up (DMU), 209, 210, 1289, 1290, 2501
Digital products:
  online retailing of, 266, 267, 270–271
  pricing of, 270–271
Digital prototyping, 1288–1290
Digitization (signal processing), 1904
DII (dynamic invocation interface), 720
Dillard's, 263
Direct-assessment methods (decision analysis), 2195
Directed systnesis control, 161
Direction of enterprise:
  and comprehensive business model, 30–31
  strategic analysis to learn, 51–52
Direct labor, 2299
Direct materials, 2300, 2308
Direct numerical assessment (decision making), 2191
Direct optimization, 2541
Direct product replenishment, 780
Direct radiant lighting, 1198
Direct release, 533
Direct work, 1459
Dirt allowances, 1399
Disassembly, 439–445
  applications of, 443–444
  diagrams, disassembly, 1376, 1378
  ecological factors in, 440
  goals of, 439
  integrated assembly/disassembly approach, 444–445
  manual vs. automated, 440–441
  planning models for, 538
  processes/tools for, 440–443
Discrete batch process, 330, 331
Discrete cash flows compounded continuously (compound interest), 2343–2345
  compound amount factor, 2345
  geometric conversion factor, 2345
Discrete cash flows compounded discretely (compound interest), 2337–2343
  arithmetic gradient conversion factor, 2342–2343
  capital recovery factor, 2340–2341
  compound amount factor, 2337–2339
  present worth factor, 2338–2339, 2341
  sinking fund factor, 2339–2340
Discrete compound interest factors, 2354–2357
Discrete event dynamic system (DEDS):
  modeling, 503–504
  simulation, 506
Discrete-event sensors, 1903
Discrete event simulation models, 128
Discrete optimization, 2582–2600
  backtracking in, 2591–2592
  branch and bound procedures in, 2592–2593
  and computational complexity theory, 2594–2595
  heuristic search in, 2589–2591
  modeling in, 2582–2583
  problem-solving strategy, choice of, 2595–2596
  relaxation in, 2584–2589
    Lagrangean relaxations, 2587–2589
    linear programming relaxations, 2585–2587
  solutions in, 2583–2584

  standard models for, 2596–2600
  total enumeration in, 2584
Discretion, 147
Discrimination, price, 681–682
Diseases, occupational. *See also* Occupational safety and health
  definition of, 1168–1170
  descriptions of, 1167, 1169–1170
  statistics related to, 1157, 1173–1174
Dispatching, 1723–1725
  basic rules, 1723–1724
  composite rules, 1724–1725
  first-come-first-served (FCFS), 1511–1513
  major activities of, 1770
  notation used in modeling of, 1719–1722
  rules, dispatching, 1511, 1513
Dispensing (solder paste), 425, 426
Display systems, virtual environment, 2502
Distance, computer screen viewing, 1197
Distance learning, 940
Distributed commerce model, 271–272
Distributed component object model (DCOM), 721
Distributed control, 166, 167
Distributed database management systems, 124–125, 723, 724
Distributed denial of service (DDoS) attacks, 278
Distributed environment (collaborative manufacturing), 604, 607–616
Distributed group decision support systems, 145
Distributed Operator Model Architecture (DOMAR), 2440–2441
Distributed/parallel processing model, *see* Neural networks
Distributed problem solving (DPS), 174
Distributed processing, 233
Distributed Systems Project, 173
Distributed transaction management, 721–723
Distribution, 2147
  contract manufacturing in, 264
  ERP tools for, 90
  flexible, 1471
  and information systems, 1472
  of knowledge, 215
  linear programming applications for, 2056
  modular, 1471
  retail supply chains, 777
  reverse, 1470
  of test and inspection effort, 1889–1890
  types of, 2129
Distribution centers, management of, 334
Distribution control business model, 603
Distribution management, *see* Logistics management
Distribution network planning, 1472–1475
Distributor relationships, 557
Diversity, as TQL success factor, 1805
Division of labor, 1264, 1266–1267
DML, *see* Data manipulation language
DMU, *see* Digital mock-up
DNS, *see* Domain name system
Documentation:
  archiving project, 1349
  of RPD projects, 1286–1287
  of time standards, 1406
Document control, 1770
Document distribution networks, 1473–1474
Document holders (at computer workstations), 1204–1205
DOF, *see* Degrees of freedom
Dollars, constant vs. actual, 2397–2398
Domain name system (DNS), 237, 240, 242–243
Domains (Internet), 242–243
DOMAR, *see* Distributed Operator Model Architecture
Domestic appliances, conditions for global assembly of, 403
Dominance, 2179
Dominance decision rule, 2177
Dominance principle (decision theory), 2377
Dominance rules, 1727
Dominance structuring, 2207
Domination structures:
  constant cone, 2615–2616
  variable cone, 2616–2617
DONLP2, 2563
Double auctions, 277
Downsizing, 1888
Downstream business models, 602, 603
DP, *see* Dynamic programming
DPS (distributed problem solving), 174
DQLs (data query languages), 119
Drawing(s), 304, 1314–1315
DRG payment system, *see* Diagnostic related group payment system
Drill down (term), 84
Drilling, 1322
  calculation of time needed for, 460
  cost of machinery for, 467
  geometric capabilities of, 464
  technological capabilities of, 468
Driven magazines, 384
Drivers. *See also under* Transportation
  cost, 2319
  performance, 55
Drug testing, 923
Drum-buffer-rope (DBR) scheduling, 558
DSSs, *see* Decision support systems
D2D, Ltd., 1713
Dual feasibility, 2554
Dummy variables, 2265
Duplex assembly cells, 409
Durability, 1246
Duration estimates, 1341
Dust allowances, 1399
Dutch auctions, 274
Dynamic characteristics, 1884–1885
Dynamic decision making, 2176, 2205
Dynamic fault management, 1022–1023
Dynamic invocation interface (DII) (CORBA), 720
Dynamic job shops, queueing models for, 1650–1656
  general service times, 1654–1656
  multiple-job-class open Jackson queueing network model, 1652–1654

Dynamic job shops, queueing models for (*Continued*)
  single-job-class open Jackson queueing network model, 1650–1652
Dynamic programming (DP), 2636–2646
  decisions in, 2637–2638
  decision times in, 2636–2637
  finite horizon, 2641–2643
  infinite horizon, 2643–2645
  policies in, 2639–2640
  rewards/costs in, 2638–2639
  states in, 2637
  transition probabilities in, 2638
Dynamic response, 157
Dynamic scheduling, 497, 503
Dynamic standing forces, 1055
Dynamic vs. static strengths, 1052, 1053

EAI tools, *see* Enterprise application integration tools
EAM, *see* Enterprise asset management
Earliest due date rule (EDD), 1722–1724
EA3, 2563
EBA rule, *see* Elimination by aspects rule
eBay, 273–275
E-business suites, 95
ECAD systems, 188, 190
Echelon, 166
Eco-labels, 532
Ecological interfaces, 1020–1021, 1024
E-commerce, *see* Electronic commerce
Econometrics models, 128
Economic analysis, *see also* Risk analysis; Sensitivity analysis
  effects of inflation in, 2401–2405
  role of, 2396
Economic equivalence calculations, 2398–2400
Economic feasibility (IS systems), 98
Economic growth, science and, 602
Economic life, 2332
Economic order quantity (EOQ) model, 545, 1670, 2022, 2023
Economic service life, 2332
Economic tax life, 2332
Economic want, 2299
The Economist, 36
Economy:
  changes in, 146
  digital, *see* Digital economy
  and enterprise resource planning (ERP), 344–347
  as external force, 40
  industrial vs. digital, 261
  Internet, 261
    layers of, 260
    pricing in, 267–268
  knowledge, 107
  network, 107
  networked, 262
  service-based, transition to, 623
eDC, *see* Electronic design and commerce
EDD, *see* Earliest due date rule
EDI, *see* Electronic data interchange
EDM, *see* Engineering data management
Edosomwan Performance Excellence Model (EPEM), 1798–1801
EDS, 966
Education. *See also* Learning; Training
  Baldrige criteria for, 1960
  for creating service-driven workforce, 1959–1961
  for product design/process platforms methodology, 1999–2000
  training vs., 924, 925
EEAM human factors checklist, 1141, 1142
Effective interest rates, 2337
Effectiveness:
  employee involvement and organizational, 976
  reliability and system, 1922
  team, 983–987, 987
    outcome variables affecting, 987
    process variables affecting, 985–987
    structure variables affecting, 983–985
Efficiency. *See also* Job design/redesign
  in modeling, 1631
  production, 526
  team, 987
Efficient frontiers, 753, 754, 759
  after-tax, 765
  in different macroeconomic environments, 763
  value-at-risk approach to, 767, 768
Effort estimates, 1341
Eight pillars of quality, 1796–1798
EIS (executive information systems), 84
"E-Lance economy," 1001
Elasticity of demand, 668, 669
Elastic linkage, 416
Elderly people:
  anthropometric estimates for, 1045
  performance models for, 1126
Electrical discharge machinery, 467
Electrical systems, energy-improvement possibilities for, 1580
Electrical technology, 365
Electric discharge machining, 1322
Electricity costs, 1575–1576
Electric power industry, 518
Electrochemical machining, 1323
Electroforming, 1320, 1321
Electromagnetic/electrohydraulic machinery, 467
Electromagnetic grippers, 414
Electron beam machining, 1323
Electronic assembly, 392–396. *See also* Electronic devices/systems assembly
  fiberoptic connectors, 395, 396
  luminaire wiring, 394–395
  measuring instruments, 392–394
  overload protector, 392
Electronic auctions, 271
Electronic boardroom, 134
Electronic commerce (e-commerce), 259–278
  and advertising, 272, 273
  and associations/referrals, 273

  auctions, online, 273–277
    B2B trading markets, 275
    consumer markets, 275
    double auctions, 277
    Dutch auctions, 274
    English auctions, 273–274
    first price auctions, 274
    reverse auctions, 275–276
    second price auctions, 274
  B2B, 262–265
    and logistics management, 264–265
    manufacturing, contract, 263–264
    procurement, Web-based, 262–263
    trading markets, 275
  bundle trading via, 277
  CRM software and, 95
  and customization of products/services, 261–262
  and design by customers, 701
  digital economy as basis for, 261
  and distributed commerce model, 271–272
  and enterprise models, 306
  and enterprise resource planning (ERP), 95–96, 306, 347–348
  ERP and, 95–96
  future of, 277–278
  intermediation models for, 271
  and Internet, 260
  and mass customization, 705–706
  modeling for, 306
  and organizational flexibility, 262
  pricing, 267, 269–271
    digital product, 270–271
    real-time, 269
  and pricing, 671–672
  privacy-security issues with, 267–269
  rationale for, 260
  and retailing, 265–267
    digital products, 266, 267, 270–271
    physical products, 266
    services, 266, 267
    storefronts, Web, 265–266
    supply chains, retail, 782–784
  transorganizational ISs as element of, 69–70
  and warehousing changes, 2070–2071
Electronic communication, 232, 233
Electronic data interchange (EDI), 85
  and e-commerce, 262
  for kanban systems, 551
  in retail supply chains, 776
  in supply chain management, 2124
Electronic design and commerce (eDC), 705, 706
Electronic devices/systems assembly, 423–439
  feeding, PCB, 426–428
  interconnection materials, application of, 424–425
  interconnection technology for, 429–431
  and miniaturization, 423, 424
  molded interconnect devices (MIDs), 432–439
  placement of components, 425–429
  process chain in, 423
  quality assurance in, 431–432
  substrates, 424
Electronic document management, 221. *See also* Engineering data management
Electronic mail (e-mail), 235, 243
Electronic performance support systems (EPSS), 940
Electronics:
  impact of, on assembly, 404–407
  scope for rationalization in assembly, 365
Electronics industry:
  case study (transportation management), 2059–2060
  conditions for global assembly in, 403
Electrostatic grippers, 414
Elements, job, 1418–1419
Elimination by aspects (EBA) rule, 2177, 2179–2180
E-mail, *see* Electronic mail
Embedded services business model, 603
Emergency Planning and Community Right-to-Know Act (EPCRA), 594
Emergency systems, personnel scheduling for, 1744
Emissions, estimation of, 596–598
  factors, emission, 597–598
  mass balance approach to, 596–597
Employee assessment systems, 938
Employees. *See also* Staffing
  and achievement of safety culture, 960
  characteristics of, and occupational safety and health, 1159–1160
  development of, as outcome of leadership, 852–855
    attitudes toward leader, 854–855
    group development, 855
    personal development, 853–854
  education/training/development of, 1960
  hazard information for, 1176–1177
  impact of teams, 98–99
  involvement of
    in employee/management ergonomics committee, 1187
    in occupational safety and health, 1186–1187
    and teamwork, 976
  participation of, in compensation setting, 913
  perception of hazards in workplace by, 1158
  in rapid product development, 1286
  well-being and satisfaction of, 883, 1961
Employment:
  in manufacturing jobs, 486
  unemployment rate as metric for, 344
Empowerment, and leadership, 853–854, 860
EMS, *see* Environmental management systems
Encapsulation (OOP), 70, 292, 349, 1328
Endurance data, 1119
Energy audits, 534
Energy costs, and site selection, 1471
Energy management, 1572–1582
  assessment, energy, 1578–1579
  demand and power factor charges, 1757
  environmental issues, 1577

Energy management (*Continued*)
  financial issues, 1576–1577
  process, energy, 1574
  productivity, energy, 1573
  programs, energy-management, 1578
  strategies and tactics, 1577–1582
  system, energy, 1574–1575
Engineering:
  cognitive, 1014
  costs of changes in, at different life cycle stages, 1312, 1313
  service, 635–636
  set-based, 556
  simultaneous, *see* Simultaneous engineering (SE)
Engineering controls:
  for management of work-related musculoskeletal disorders (WRMDs), 1092–1093
  for workplace hazards, 1175–1176
Engineering data management (EDM), 195–198, 1290, 1291
  architecture/components of, 196–198
  functions of, 195–196
  and multiple views of product families, 691, 692
  product data management vs., 195
Engineering design, 494, 1362, 1363, 1387–1389
Engineering phase (human-centered product planning and design), 1300, 1306–1308
Engineering solution center (ESC), 1290
Engineering time estimates, 1393
Engineer-to-order production, 330–331, 338
English auctions, 273–274
Enhanced index products, 761
Enterprise(s):
  areas of responsibility in, 1771
  as complex living system, 28
  definitions of, 27–28, 280
  key questions about, 29
Enterprise application integration (EAI) tools, 341, 342
Enterprise asset management (EAM), 1591–1592, 1605–1610
  and plant engineering, 1550
  software for, 69
Enterprise business model, *see* Business model
Enterprise-Control System Integration Part I: Models and Terminology (ANSI), 1769
Enterprise data management, *see* Engineering data management
Enterprise excellence models, 8
Enterprise information systems, 69, 107
Enterprise models/modeling, 280–306, 293–303. *See also* Business model
  abstraction in, 281–283
  architectures for, 293–303
    ARIS, 293–300
    CIMOSA, 301–302
    IFIP ISM, 300–301
    Zachmap framework, 302–303
  ARIS, 293–300
  benefits of, 284–286
    computerized enterprise modeling, 303
    for information system design, 285–286
    organizational processes, improvement of, 284–285
  CIMOSA, 301–302
  clarity of, 284
  comparability of, 284
  and computer integrated manufacturing (CIM), 507–514
    ARIS, 512, 513
    CIMOSA method, 510–512
    function view, 508, 509
    GIM, 512–514
    information view, 509, 510
    organization view, 510
    process view, 507–508
    resource view, 510
  computerized, 303
  correctness of, 284
  costs vs. benefits of, 284
  goal of, 280
  IFIP ISM, 300–301
  object-oriented enterprise modeling, 291–293
  outlook for, 306
  principles of, 283–284
  process-oriented enterprise modeling, 286–291
    data views, 288–290
    function views, 287, 288
    organization views, 286–287
    output views, 287–289
    process views, 290–291
  relevance of, 284
  systematic structure of, 284
  tools for, 303–306
  Zachman framework, 302–303
Enterprise resource planning (ERP), 83, 85–96, 325–351, 1738
  and achievement of interoperability, 348–351
  applications of, 89–92
    accounting and finance, 91
    and choice of ERP package, 92
    human resources, 91–92
    and implementation of ERP system, 92–94
    manufacturing and procurement, 90–91
    sales and distributions, 90
  architectures for, 341–343
  boundaries of/interfaces with, 336–339
    B2B supply chain operations, 343
    contracts management, 336, 337
    customer relationship management, 337
    external user-to-ERP interfaces, 343
    finance, 339
    human resource management, 339
    internal user-to-ERP interfaces, 343
    joint supply planning interfaces, 344
    manufacturing execution, 338–339
    product configuration management, 338
    product data management, 338
    standards development, 349–350
    supplier relationship management, 337
    supply chain execution, 338
    supply chain planning, 338
  and component decomposition analysis, 349

  configuration tools for, 340–341
  current market state of, 351
  and customer relationship management, 95–96
  data integration, 89
  and decision support algorithm development, 348
  decision support applications for, 339, 340
  definition of, 85, 325
  and e-business, 306
  and the economy, 344–347
  and electronic commerce, 95–96, 347–348
  enterprise application integration (EAI) tools, 341, 342
  evolution of, 85–88
  extended applications, 340
  external vs. internal views of, 326
  features of, 325
  financial asset management in, 336
  future of, 94
  global implementation of, 953
  implementation of, 339–341
  integration of advanced planning and scheduling (APS) with, 2047–2048
  and inter-ERP data sharing, 94–95
  and the Internet, 342–344
  and major business functions in manufacturing enterprises, 326–327
  for management information systems, 492
  market for, 87–88
  and open systems architecture, 88–89
  and operations planning (manufacturing), 327–329
  and partitioning of domain of manufacturing, 329–331
    customer, nature of, 329
    customer orders, nature of business in terms of, 329–331
    process, nature of, 329–331
    product, nature of, 329
  and "process view," 88
  scope of, in manufacturing enterprises, 331–332
  software for, 69, 304
  standards development for, 349–350
  and supply chain management, 94–95, 348
  teams for choice/implementation of, 92–94
  terminology related to, 88
  and transaction management, 332–336
    accounting, 336
    finance/management, 336
    human resource management, 335–336
    maintenance management, 334
    manufacturing management, 333
    materials acquisition, 332–333
    materials inventory, 332
    order entry/tracking, 333
    process specification management, 333–334
    transportation, 335
    warehousing, 334–335
  in transportation, 335
  trends in, 107
Enterprise resources, 1573
Enterprise-wide integration, 491
Entertainment robots, 381, 382
Entertainment sector network applications, 251
Entity relationship diagrams (ERDs), 102, 103
Entity-relationship (E-R) model, *see* Relational database model
Entity relationship method (ERM), 304
Entropy, 1776
Environment(s):
  business, 32–33, 35–36
  social, and human–computer interaction, 1217, 1220–1222
  virtual, *see* Virtual environments
  work, *see* Work environment
Environmental engineering, 596–599. *See also* Clean manufacturing; Waste management
  emissions, estimation of, 596–598
    factors, emission, 597–598
    and mass balance, 596–597
  and green engineering, 598–599
  industrial engineering vs., 530
  total-enclosure concept, 598
Environmental factors:
  in disassembly, 440
  in measurement systems, 1879
  and plant engineering, 1577
  in site selection, 1489
  types of, 531
Environmental law(s), 589–596
  CERCLA, 594
  Clean Air Acts, 590–593
  Clean Water Act, 595
  compliance with, 595–596
    air permits, 595–596
    water permits, 596
  Environmental Protection Act, 590
  Hazardous Materials Transportation Act, 594
  Resource Conservation and Recovery Act, 593–594
  Superfund Amendment Reauthorization Act, 594–595
  Worker Right to Know laws, 593
Environmental management systems (EMS), 539, 1185
Environmental Protection Act, 590
Environmental Protection Agency (EPA), 590–592, 596–598, 1164, 1168, 1489, 1592
Environmental stimulation, task vs., 1357, 1358
Envision phase (process design and reengineering), 1697, 1705–1707
EOQ model, *see* Economic order quantity model
EPA, *see* Environmental Protection Agency
EPC, *see* Event-driven process chain
EPCRA (Emergency Planning and Community Right-to-Know Act), 594
EPEM, *see* Edosomwan Performance Excellence Model
EPSS (electronic performance support systems), 940
Equal Pay Act, 908
Equipment:
  arrangement of, 1379–1382

Equipment (*Continued*)
  effectiveness of, 553
  and error reduction, 1369–1370
  estimating costs for, 2298
  maintenance, 1590
  material-handling, 1504, 1505
  replacement of, 1274, 2578, 2579
  safety and design of, 1178
  time-recording, 1411, 1412, 1414
  warehouse, 1541–1544
    hardware controllers for, 2103
    master table for, 2099
Equities, 757, 758, 764, 765
Equity, weighted cost of, 2334
Equity theory, 861
Equivalence, 2336
Equivocality reduction (of information), 141
ERDs, *see* Entity relationship diagrams
ERGO (computer-aided model), 1050
ERGO human factors checklist, 1141, 1142
ERGOMAN (computer-aided model), 1050
ErgoMOST, 1440, 1441
Ergonomic Checkpoints, 1144
Ergonomic design, 1042
  of manual workstations, 417–418
  programs, ergonomics, 1097
  for reduction of work-related upper-extremity disorders (WUEDs), 1086–1091
    programs, ergonomics, 1097
    proposed OSHA regulations, 1097–1100
    and quantitative models, 1087
    wrist/hand disorders, 1087–1091
Ergonomic guidelines:
  for job design, 1354
  for management of work-related musculoskeletal disorders (WRMDs), 1091–1092
Ergonomics, 1042, 1061, 1194–1195
  California Ergonomic Standard, 1166
  cognitive ergonomics, 1014
  for control panels, 1016
  definition of, 1042
  and digital human modeling, *see* Digital human modeling
  general checklist to prioritize potential problems, 1364
  for hospitality industry:
    tables, 833
    workstations, 834
  of human–computer interaction, *see* Human–computer interaction
  kitchen, 833–834
  and OSHA Proposed Ergonomic Program Standard, 1166
  participatory, 980–981
  underlying philosophy of, 1042
  and working postures, 1061
Ergonomics Audit Program, 1139, 1141
Ergonomics committee, employee/management, 1187
Erlang loss system, 2158–2159
ERM (entity relationship method), 304
E-R model, *see* Relational database model
ERNAP human factors checklist, 1141–1143
Ernst & Young, 781, 963
ERP, *see* Enterprise resource planning
Error:
  job design for reduction of, 1368–1371
  of measurement systems, 1883–1884
ESC (engineering solution center), 1290
Estimating minutes, 2314
Estimation:
  confidence interval, 2253–2254
  cost, *see* Cost estimating
  hypotheses for, 137
  maximum-likelihood estimators, 2254–2255
  of reliability, 1944–1946
  statistical, *see* Statistical estimation and inference
  uncertainties in, 2361
Estimation theory models, 128–129
Ethernet technology, 253
Ethics:
  in decision making, 2210
  and group decision making, 2209–2210
ETH Zurich, 321
eToys, 262, 273
European Agency for Safety and Health at Work, 1165
European Computer Integrated Manufacturing Architecture (AMICE), 489, 511
European Laboratory of Particle Physics (CERN), 244
European Quality Award, 645
European quality management systems standards, 1968
European standards for working postures (machinery operations), 1068
European Union, 1165
EV, *see* Expected value
Evaluate phase (process design and reengineering), 1698, 1711, 1712
Evaluation:
  design, 450
  formative, 934–936
  gulf of, 1018
  of human factors audits, 1134–1135
  integrated evaluation tool, 321, 322
  job, *see* Job evaluation and job evaluation systems
  of job design/redesign, 889, 892–893
    biases, potential, 893
    and data sources, 892
    example of, 893
    long-term effects, 892–893
    need for, 882–884
    with questionnaires, 889, 892
  as measurement issue for successful design, 1299, 1301
  of new processes, 1711–1712
  of process plans, 458–460
    quality estimation, 460
    time/cost estimation, 459–460
  of service quality, 1963–1964
  summative, 934–936
  of team design, 899–884

    biases, potential, 893
    and data sources, 892
    example of, 893–894
    long-term effects, 892–893
    need for, 882–884
    with questionnaires, 889–892
  technical/operational tests for, 1943
  of training, 934–937
Event-driven process chain (EPC), 290–291
Event predictions, 137
Event trees, 2189–2190
Evidence (for hypotheses), 137
EVIS, 1780
Evolutionary computation, *see* Genetic algorithms
E-work, 606
Exact algorithms, 2014
Exchange rates, 2401
Excite, 266, 272
Exclusive distribution, 2129
Execution, gulf of, 1018
Executive decision support systems, *see* Group decision support systems
Executive information systems (EIS), 84
Executive sponsorship (of ISE), 22–23
Expansion flexibility, 499
Expectation, principle of (decision theory), 2377–2378
Expected present worth, 2367–2368
Expected project life, 2392
Expected value (EV), 2304
  decision rule, 2177
  maximization of, 2181
  and SEU, 2182–2183
Expense work, standards for, 1459, 1461
Experiential learning, 938
Experimental design, 2225–2239. *See also* Hypothesis testing
  analysis of, 2232–2234
  blocking in, 2228
  checklist for, 2226
  completely randomized designs, 2230
  factorial designs, 2230–2231
  fixed-effect vs. random-effect models, 2229–2230
  and hypothesis testing, 2260–2264
  Latin square designs, 2230, 2231
  orthogonal arrays, 2232
  parameter designs, 2237–2238
  precautions for, 2228–2229
  randomization in, 2228
  randomized complete block design, 2230
  and replication, 2228
  screening designs, 2235–2236
  and strategies of experiments, 2238–2239
  terminology related to, 2225–2226
Experimentation, for continuous improvement, 963
Experiments:
  factorial, 2262
  one-factor, 2260
  size of, 2227
  stages of, 2226–2227
  statistical, 2225
  strategies of, 2238–2239
Expert database model, 122–124
ExpertFit, 2446
Expertise, team, 984–985
Experts, as human-centered product planning/design tool, 1303, 1304
Expert systems, 1328–1329
  for job evaluation, 914
  in model base management systems, 131
Explanation-based decision making, 2207–2208
Explanations (hypotheses for), 137
Explanatory variables, 2265
Explicit knowledge, 214, 1291–1292
Exponential distribution:
  of reliability, 1930, 1932
  reliability estimation, 1944–1945
Exponential service time:
  make-to-stock manufacturing/service systems, 1636–1637
  two-stage flow lines, 1639–1640
Exponential smoothing model (demand forecasting), 2029–2032
Ex post manufacturing, 276
Extended Resource Planning (XRP), 94. *See also* Enterprise resource planning (ERP)
Extensible Markup Language (XML), 77, 252, 306, 2124
External consistency (user interfaces), 133
External data, 117
External data model, 119–120
External forces and agents, 32, 35–40
  alliances, 39
  capital markets, 39–40
  and changing playing field, 35–36
  community, 39
  competitors, 39
  customers, 38
  and data access vs. traditional reporting, 37
  demographic trends, 37
  economy, 40
  on enterprise, 29
  and globalization, 36
  and information technology, 36
  and knowledge work, 36–37
  owners, 39
  political trends, 38
  regulators, 39
  social trends, 38–38
  stakeholders, 39
  suppliers, 39
Externalized *t* ratio, 2284
External risks, 45
Extraction, data, 84
Extranets, 220, 256
  as connection of intranets, 238
  definition of, 237
Extrusion, 569, 1319, 1321
  cold-formed components, 575–577
  hot-formed components, 582–584
  obtainable accuracy values, 565
  of plastics, 1324, 1326
  semihot formed components, 580, 581

Eye loss, 1169
Eye strain allowances, 1400

Face-to-face meetings, 142
Facilities engineering, 1586–1588. *See also* Maintenance; Plant engineering
  definition of, 1550
  and plant engineering, 1551–1552
  plant engineering vs., 1550
  work measurement in, 1562
Facilities management:
  energy-improvement possibilities for, 1579
  as resource-utilization issue, 1553
Facility Description Language (FDL), 171–173
Facility location models, 2067–2068
Facility surveys, 1564–1565
FACTOR/AIM, 2458
Factor comparison method (job evaluation), 903–907
Factorial designs, 2230–2231
Factorial experiments, 2262
Factor method (of cost estimating), 2302
Factors (in experimental design), 2225
Factory warehouses, 2085
Failure mode and effects analysis, 1940
Failure rate of product. *See also* Reliability
  in "infant mortality" period, 1925–1972
  in useful life period, 1927
  in wear-out period, 1927
Fair day's work, determination of, 1411
FairMarket, 275
Families, changes in, 37
Family-based storage, 2093
Family formation (group technology), 462
FAMs, *see* Fuzzy associative memories
Fast-food stores, personnel scheduling for, 1745
FastParts, 263, 275
Fatalities, occupational, 1157
Fatigue:
  allowances for, 1394–1400
  digital human modeling of, 1188–1119
  job design and reduction of, 1365–1368
Fault diagnosis, 1022–1023
Fault trees, 1936–1937, 2189, 2190
FCFS, *see* First-come-first-served
FDL, *see* Facility Description Language
FDMs (finite different methods), 199
Feasibility:
  dual, 2554
  primal, 2554
Feasibility analysis:
  CIM implementation, 514, 515
  for information systems, 98–99, 106
Feasibility stage (project life cycle), 1242
Feasible region, 2528, 2541
Feasible solutions, 2528, 2583, 2584
Feature extraction (AVIS), 1905–1906
Features:
  in CAD, 185
  manufacturing, 452, 454
  as performance measure of quality, 1246
Features mapping, 463–466
Federal Aviation Administration, 1141, 1909, 1910
Federal Express (FedEx), 264, 266, 662
Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), 1164
Federated Department Stores, 781
Feedback:
  about cues vs. outcomes, 2200
  in automatic control systems, 157–158
  cognitive probes for, 1026
  control, 158
  end-of-project, 1348
  on errors, 1370–1371
  keyboard, 1201
  in management system model, 22
  in quality-related teamwork, 979
  and team motivation/performance, 933–934
  time horizons of, 954
Feedback control, 158
Feedback control models, 160
Feedforward control, 161
Feeding, assembly:
  electronic components/PCBs, 426–428
  principles for, 415
  systems for, 381–383
Feet, as hand substitute, 1359, 1360
FEM, *see* Finite element methods
Females:
  hand work areas for, 1360
  maximum forces of pull for, 1055–156
  maximum wrist extension for, 1091
  maximum wrist flexion for, 1091
  reach distances for, 1361, 1362
  recommended weight of lift for (industrial workers), 1073–1075
Femininity (in national cultures), 957
Ferrographic oil analysis, 1614
Fiacco and McCormick algorithm, 2531–2532
Fiberoptic connectors, assembly of, 395, 396
Field bus, 165
Field (in database), 80
Field of view (FOV), 2505
Field use, reliability program applications during, 1954
FIFO, *see* First-in-first-out
FIFRA (Federal Insecticide, Fungicide, and Rodenticide Act), 1164
Figure posturing, 1122, 1123
Filament winding, 1327
Files (in database), 80
File servers, 240
File transfer, 243
File transfer protocol programs, 240
Filimant lamps, 1198
Filtering, content, 248–249
Finance, *see* Accounting and finance
Finance industries, 346
Financial asset management, 751–770
  and asset-allocation problem, 752–753
  client-tailored solutions in, 763–764
  and efficient frontier, 753, 754
  in ERP, 336
  extrapolation fallacy in, 756, 767
  and forecasting problem, 761–763
  with hedge funds, 759, 760, 768–770
  and mean-variance (MV) analysis, 752–756, 761–769
  new asset classes, 758–761

    currencies, 761
    enhanced index products, 761
    hedge funds, 759, 760
    insurance-linked products, 761
    private equity and venture capital, 759–761
    Treasury inflation-protected securities (TIPs), 761
  and optimization problem, 755–756
  selection of assets, 757, 758
  taxation issues, 764–766
  time horizon in, 766–767
  traditional approach to, 752
Financial management software, 336, 339
Financial rewards, 1182
Finished-goods warehouses, 1528
Finite capacity algorithms:
  artificial intelligence approaches, 2044
  capacitated MRP (MRP-C), 2042–2043
  capacity requirement planning (CRP), 2042
  congestion models, 2044
  and detailed scheduling, 2044–2045
  optimization approaches, 2043–2044
  rough-cut capacity planning (RCCP), 2042
Finite different methods (FDMs), 199
Finite element methods (FEM), 199–203
  applications of, 199
  CAD system interfaces for, 187
  postprocessing, 201–203
  preprocessing, 200, 201
  solution process with, 201, 202
  steps in, 200, 202
Finite horizon dynamic programs, 2641–2643
Fire safety systems, 1568
Firewalls, 734–735
  and distributed denial of service attacks, 278
  in intranets, 255
First Bank of Chicago, 654
First-come-first-served (FCFS), 1511–1513, 1685
First-in-first-out (FIFO), 1521, 2157, 2167
“First Man” program, 1112
First price auctions, 274
First-principle models of human behavior, 2413–2414. *See also* Man–Machine Integrated Design and Analysis System (MIDAS)
Fish diagrams, 1385, 1386
5S approach (industrial housekeeping), 553, 559
Five forces model of competition, 33
Fixed asset management (ERP), 336
Fixed-effect models, random-effect vs., 2229–2230
Fixed-income assets, 757–859
Fixed-location storage systems, 1534, 1535
Fixed ordering costs, 2021
Fixtures:
  planning, 455, 457
  workpiece, 384
Flags, for network content filtering, 248
Flat panel display ergonomics, 1195
Flexibility, 147, 262
  of distribution, 1471
  of manufacturing systems, 499
  in modeling, 1631
Flexible assembly cells, 408–409
Flexible assembly systems, 403, 419–422, 1633
  CAD-CAM process chain, 420–422
  for changing amounts of different versions of a product, 419–420
  handling equipment, 420, 421
  with industrial robots, 360–362
  layout of, 366, 367
  modular, 359–360
Flexible circuit technology, 424
Flexible flow lines, 1633
Flexible Manufacturing Systems (FMS), 499–507, 1633
  benefits of, 506, 507
  control in, 501–503
  definition of, 499–500
  design of, 500–501
  failures in implementation of, 949
  limitations of, 507
  machinery for, 467
  modeling/simulation in, 503–506
  planning/scheduling/control in, 501–503
  queueing models in, 1656–1662
    dedicated material-handling systems, 1660
    general single-class closed queueing network model, 1660–1661
    multiple-class model, 1661–1662
    single-class closed Jackson queueing network model, 1656–1660
Flexible transfer lines, 1633
Flex-link (programmable) assembly systems, 35, 356, 362
Flicker, computer screen, 1197
Flight attendants training case example, 2022–2023
Float pool scheduling, 1745
Float transducers, 1903
Floor coverings:
  and acoustical control, 1200
  reflectance of, 1199–1200
Flow(s), 4, 1503, 1504. *See also* Information flow
  data, 99
  in process-oriented enterprise modeling, 286
  in supply chain management, 94
Flow diagrams, 1811
  in healthcare, 740
  for methods engineering information gathering/organizing, 1374–1376
  process, 1983–1984
Flow lines, 1632, 1633
  balancing, 1382–1385
  queueing models for, 1638–1645
    general service times, 1640, 1643–1645
    multiple-stage flow lines with exponential processing times, 1642–1643
    paced systems, 1638–1639
    three-stage flow lines, 1640–1642
    two-stage flow lines, 1639–1640
    unpaced lines, 1639
Flow shops, 1721
Flow soldering, *see* Wave soldering
Flow time, determination of, 1631
Fluid-flow switches, 1902

Fluid limit, 2167
Fluorescent lighting, 1198
FMS, *see* Flexible Manufacturing Systems
Focus (market), 34
FOCUS-PDCA, 747
Fonts, computer, 1196–1197
Food Consulting, 396
Food industry:
  assembly in, 396–398
  automated test and inspection in, 1907
Footrests, 1204
Force field analysis, 1815
Force sensors, 385
Force transducers, 1903
Ford, 212, 659
Forecasting:
  of consumer demand, 781–782
  for cost estimating, 2310
  demand, 781–782
    average (constant), 2020–2023
    and economic order quantity (EOQ) model, 2022, 2023
    and exponential smoothing model, 2029–2032
    flight attendants training case example, 2022–2023
    and inventory costs, 2021–2022
    and lead time, 2025–2027
    as mixture of distributions, 2027–2029
    over a single period, 2023–2025
    over multiple periods, 2025–2027
  in financial asset management, 756, 757, 761–763
  hypotheses for, 137
  and retail supply chains, 779–781
  in retail supply chains, 776
  subjective/objective models of, 793
  in transportation planning, 792–793
Forestry industries, 346
Forgetting, 931
Forging:
  design for, 1317, 1319
  obtainable accuracy values, 565
  powder, 574–576
  precision (hot-formed components), 581–583
  semihot formed components, 581
  thixoforging, 568
Formal organizations, 1005–1006, 1008–1009
Formative evaluation, 934–936
Forming, 456
Forming, storming, norming, performing model of decision making, 2210
Forrester Research, 781
Forward-reserve allocation (warehouse operation), 2093
Forward selection, 2289
Fourth generation R&D, 148
FOV (field of view), 2505
Fox Meyer, 949–950
Fractals, 404
Fractures, 1169
Framework Software, Inc., 302
France:
  industrial robots in, 373
  quality standards in, 1968
F ratio, 2278–2279
Fraunhofer Institute for Production and Automation (IPA), 315, 316, 318, 320, 364, 381
Fraunhofer Institute for Production Systems and Design Technology, 218
Freedom (privacy service), 269
Free-form surfaces, modeling, 182, 1881
Free trade zones (FTZs), 1489–1490
Freight vehicles, 2063
Frequency plots, 1819, 1821, 1823, 1832–1834
Frontiers, efficient, *see* Efficient frontiers
FSQP, 2563–2564
FTE (full time equivalent), 742
Ftp programs, 240
FTZs, *see* Free trade zones
Full potential, 4–24
  conditions for successful achievement of, 15–18
    communication system, 17
    culture system, 15–16
    infrastructure, 16
    learning system, 17
  and definition of ISE, 4–6
  and enterprise excellence models, 8
  examples of, 7–8
  and “flow,” 4
  and IIE/CIE/CIEADH relationship management, 23
  operations improvement and achievement of, 18–22
    business processes, 18–20
    measurement systems, 20–22
    performance measurement, 21–22
  organizing for, 22–23
  personal mastery issues related to, 23–24
  planning system for achievement of, 11–15
    change leadership, 14–15
    policy deployment, 13
    relationship management, 13–14
Full range leadership model, 848–850, 854, 862
Full time equivalent (FTE), 742
Fully immersive virtual environments, 2507
Fumes allowances, 1399
Function(s):
  allocation of, 1892, 1912–1916
  convex, 2543, 2544
  inspection and test, 1892, 1893
  in object-oriented enterprise modeling, 291, 292
  in process-oriented enterprise modeling, 287, 288
  warehouse, 2102
Functional anthropometry, 1043
Functional information systems, 68–69, 107
Functionality testing, 1908
Functional organizational structure, 1265
Functional predetermined time systems, 1429
Function flow map (warehouses), 2102
Function view, modeling method for, 508, 509
Future worth method (cost estimating), 2348
Fuzzy associative memories (FAMs), 163, 164

Fuzzy constraint relaxation, 1781
Fuzzy logic, 163, 164
Fuzzy multicriteria optimization, 2620
Fuzzy neural networks, 164
Fuzzy sets:
  for artificial intelligence approaches to control, 1781–1782
  for inspection system decision making, 1898
  in shop floor scheduling, 1781–1782

GAs, *see* Genetic algorithms
Game theory, 2204, 2209
Gaming methods, 128
GAMS (General Algebraic Modeling System), 2536
Gantt charts, 1257
  with computer-based scheduling, 1736, 1737
  IS use of, 103–104
Garch model, 761–762
Garden.com, 265, 266
GATB (General Aptitude Test Battery), 921
Gateway(s), 783
  database, 84
  as Internet component, 238–239
  proxy, 735
Gateway (company), 264
GBOM, *see* Generic bill of materials
Gbps, 232
GCs, *see* General contractors
GDP (gross domestic product), 344
GDSS, *see* Group decision support systems
Gearboxes (automotive), unpacking of, 389, 391–392
Gear generating, 1323
Gear-shaping machinery, 467
Gender bias, 916
General Algebraic Modeling System (GAMS), 2536
General and administrative costs, 2300
General Aptitude Test Battery (GATB), 921
General contractors (GCs), 1492–1497
General Electric (GE), 80, 672, 783, 1419
General Electric Information Services, 263
Generalized enterprise reference architecture (GERAM), 1772
General ledger, 336
General linear model, 2265–2266
General Motors, 212, 856, 861, 966, 976
General product structure (GPS), 694
General Purpose Simulation System (GPSS), 2455
General service times:
  in dynamic job shops, 1654–1656
  in flow lines and series systems, 1640, 1643–1645
General single-class closed queueing network model, 1660–1661
Generative process-planning systems, 477–478
Generic bill of materials (GBOM), 695–697
Generic Guidelines for Quality Systems (ANSI/ASQ Z1.15–1979), 1968
Generic management system standards, 1185
Genetic algorithms (GAs), 164, 2591
  for artificial intelligence approaches to control, 1780–1781
  in shop floor scheduling, 1780–1781
Geographic information systems (GISs), 2016–2018
  integrating algorithms and, 2018
  in vehicle routing, 2062
  in warehousing model, 2080
Geography-based work breakdown structure, 1269, 1271
Geometric capability (of process), 457, 463–465
Geometric conversion factor (interest), 2345
Geometric modeling (CAD), 494
Geometric programming problems (for constrained optimization), 2558–2559
Geometric series factors (discrete compound interest), 2358–2359
Geometry, designed, 449, 450
GERAM (generalized enterprise reference architecture), 1772
German Automotive Industry Association, 193
Germany:
  industrial robots in, 373
  quality standards in, 1968
Gesture interaction (virtual environments), 2508–2509
Gigabits per second, 232
GIM, *see* GRAI integrated methodology
Ginna nuclear power incident, 1030–1031
GISs, *see* Geographic information systems
Glare, 1199–1200
Glass industry, 518
Globalization, 28, 36
  and assembly developments, 402, 403
  and collaborative manufacturing, 603
  e-business and, 306
  effects of, 1888
  manufacturing in context of, 601–602
  and need for standardization, 1967
  and retail supply chains, 782
  and test and inspection, 1887–1889
Globally dominant alternatives, 2377
Global production networks, 406–407
Global public network, *see* Internet
Gloves, data, 1125
GMP (guaranteed maximum price), 1496
GNP (gross national product), 623
Goals:
  activity- vs. outcome-based, 1002
  environmental, 539
  of near-net-shape processing, 564
  performance, 1708
  SMART goals, 1005, 1009
  stages for accomplishing, 1210
  in total quality leadership process, 1803
Goals, operators, methods, and selection rules (GOMS), 1208, 1209
Goal-seeking behavior, 2608
Gold bullion auctions, 273
Golden section method (unconstrained optimization), 2547–2549
Golfweb, 266

GOMS, *see* Goals, operators, methods, and selection rules
Goodness of fit test, 2256, 2258, 2259
Goods:
  services vs., 624, 636–637
  transportation of, 791–793
Governmental network applications, 251
Governments, and electronic commerce, 278
GPS (general product structure), 694
GPSHR (gross product by industry as percentage of GDP), 344
GPSS (General Purpose Simulation System), 2455
GRADENA Dynamic, 1257
GRAI, *see* Graphs with results and activities interrelated
GRAI integrated methodology (GIM), 507, 512–514
Grainger, 783
Grand strategy, 11
Graphical user interfaces (GUIs), 1253
Graphics (as types of language), 132
Graphs with results and activities interrelated (GRAI), 512–514
Gravity, and work postures, 1360
Great Britain, *see* United Kingdom
Green design, 527
Green engineering, 598–599
Green manufacturing, 527
GRG2, 2564
Grids (networking), 251
Grinding, 1322
  cost of machinery for, 467
  geometric capabilities of, 464
  obtainable accuracy values, 565
  technological capabilities of, 470
Gripping systems, 377, 413–415
Grocery stores, personnel scheduling for, 1745
Gross domestic product (GDP), 344
Gross national product (GNP), 623
Gross product by industry as percentage of GDP (GPSHR), 344
Groundskeeping, 1567
Group authoring software, 142
Group decision making, 2176
  and biases, 2212
  computer-mediated, 2214
  and conflict, 2210–2212
  prescriptive approaches for improving, 2212–2214
  and social norms/ethics, 2209–2210
  structuring, 2213–2214
Group decision support systems (GDSS), 134–145, 2214
  computer support in, 142–143
  current issues related to, 144–145
  distributed group decision support systems, 145
  engineering of, 141–145
  information needs for, 135–141
  levels of support in, 144
  needs for/met by, 135
Group memory management, 143
Groups:
  assessment of, for decision making, 2193
  for inspection work, 1900
  leadership and development of, 855
  training of, 933–934
  work groups, *see* Teams
Group technology (GT), 494–495
  as tool for process planning, 461–463
  variant planning based on, 475
Groupthink, 881, 2212
Groupware, 142–143, 220
GT, *see* Group technology
Guaranteed maximum price (GMP), 1496
Guarantees, 656–657
GUIs (graphical user interfaces), 1253

Hamtramck plant (Detroit), 951
Hand(s):
  disorders of the, 1087–1091
  two-hand vs. one-hand motions of, 1361–1362
  work areas for, 1360
  and work posture, 1359
  workstation guidelines related to, 1359, 1361–1362
Handling:
  definition of, 407
  robots for, 413, 414
Hanover University, 616
HAPs, *see* Hazardous air pollutants
"Hard automation," 1633
HASTE database, 1165
Haworth Company, 659–660
Hay System, 908
Hazards, workplace, 1168, 1171–1187
  engineering controls, 1175–1176
  human factors controls, 1176–1179
  and illness/injury statistics, 1173–1174
  improved work practices for reducing, 1181–1183
  and incident reporting, 1174
  informing employees about, 1176–1177
  inspection programs, 1171–1173
  measurement of potential for, 1171–1174
  new technologies, hazard control for, 1184–1187
  and safety programming, 1183–1184
  safety training for reducing, 1180–1181
  workplace/job design for reducing, 1177–1179
Hazard Communication Standard (OSHA), 593
Hazardous air pollutants (HAPs), 593, 599
Hazardous Materials Transportation Act (HMTA), 594
Hazardous Materials Uniform Safety Act, 594
Hazardous waste:
  cleanup of, 594–595
  regulations concerning, 593–594
  streams, waste, 1570
Hazard survey, 1186–1187
H.B. Maynard and Company, Inc., 1439
HCDs (head-coupled displays), 2502
HCFA (Health Care Financing Administration), 738

HCI, *see* Human–computer interaction
Head, and work posture, 1358
Head-coupled displays (HCDs), 2502
Headers, 467
Head-mounted displays (HMDs), 2502
Health, occupational, *see* Occupational safety and health
Health care delivery systems, 737–748
  decision support tools for, 747–748
  emerging trends in, 738–739
  future trends in, 748
  history of, 737–738
  information systems, use of, 747
  methods improvement in, 739–741
  optimization models for, 746
  payment for, 738
  personnel scheduling in, 744
  process reengineering with, 747
  quality-improvement tools for, 747
  queueing models for, 744–745
  regulation of, 738
  scheduling:
    optimization models for, 746
    personnel scheduling, 744
    work scheduling, 742–744
  scheduling in, 742–744
    optimization models, 746
    personnel scheduling, 744
    work scheduling, 742–744
  simulation models for, 745
  staffing in, 740, 742
  statistical methods for improvement of, 745–746
  work simplification in, 740
Health Care Financing Administration (HCFA), 738
Health care industry:
  personnel scheduling for, 1744, 1745, 1760–1762, 1764–1766
  scheduling for, 1742
Healthcare Information and Management Systems Society (HIMSS), 739, 748
Health maintenance organizations (HMOs), 738, 739, 748
Hearing loss, 1167
Heating, ventilating, and air conditioning (HVAC):
  ergonomic recommendations for, 1196
  and human–computer interaction, 1200–1201
Heating systems, energy-improvement possibilities for, 1580–1581
Heavyweight project manager system, 556
Hedge funds, 757, 759, 760, 765, 766, 768–770
Heights:
  table, 1203
  work, setting, 1359
Help desks, 221
HelpMate (robot), 379, 380
Hershey's, 93, 950
Hessian matrix (nonlinear programming), 2546
Heterarchical systems vs, 697
Heuristics, 2014, 2198–2199
Heuristic search, 2589–2591
Hewlett-Packard, 95, 654, 966, 2127
Hierarchical data model, 120–121
Hierarchical organizations, 284–285
Hierarchical structures, heterarchical systems vs., 697
Hierarchical structure model (CIMS), 521–522
Hierarchical task analysis (HTA), 1028, 1029, 1909–1912
Hierarchical workforce scheduling, 1744–1745
Hierarchy(-es):
  manufacturing, 487
  use, 1215
High-complexity network applications, 244
High-level programming languages, 71
High-level time/cost estimates, 1337
High-performance organizations, 1000–1001
Hill-Burton law, 738
HIMSS, *see* Healthcare Information and Management Systems Society
HIP, *see* Hot isostatic pressing
Hiring, *see* Recruiting
Histograms, 1821, 1823, 1856–1857
Historical data, 2307
Hitachi Corporation, 368, 369
Hit rate (job evaluation), 915
HMDs (head-mounted displays), 2502
HMOs, *see* Health maintenance organizations
HMSS (Hospital Management Systems Society), 739
HMTA (Hazardous Materials Transportation Act), 594
Hobbing, 1323
HOBE, *see* House of business engineering
Holding costs, 2021
Holistic comparison, 2177, 2184
Holonic manufacturing, 697
Home delivery, 783
Homogeneity of variances, 2255
Homomorphy, 281
Honda, 212
Honeycombing allowances (storage), 1535–1537
"Honeymoon effect," 892
Hong Kong University of Science and Technology, 700–701
Honing:
  geometric capabilities of, 464
  obtainable accuracy values, 565
  technological capabilities of, 471
Hopfield networks, 1778
Hospitality industry, 825–836
  control of capital costs in, 834–835
  design by consensus in, 826–830
    relationship charts for, 826–828
    relationship diagrams for, 829, 830
    supervision, designing for, 829, 830
    utility use, design for, 830
  kitchen ergonomics for, 833–834
    storage, 833–834
    tables, 833
    workstations, 834
  layouts, evaluating efficiency of, 830–833
    with distance charts, 831
    with move charts, 831

Hospitality industry (*Continued*)
    with travel charts, 831, 832
  life-cycle costing in, 834–835
  value engineering in, 834
Hospital Management Systems Society (HMSS), 739
Hotels, *see* Hospitality industry
Hot extrusion, 565
Hot-formed components, 568, 581–585
  axial die rolling, 584
  extrusion, 582–584
  precision forging, 581–583
Hot isostatic pressing (HIP), 572–574
Hot standby, 1933
House of business engineering (HOBE), 294, 299–300
Household growth, 37
Housekeeping, industrial, 553, 559
HQL anthropometric database (Japan), 1114
HRA, *see* Human reliability analysis
HRM, *see* Human resource management
HTA, *see* Hierarchical task analysis
HTML, *see* HyperText Markup Language
HTTP, *see* Hypertext Transfer Protocol
Hub-and-spoke model, 265
Human abilities limitations, 2196–2198
Human-centered automation, 962
Human-centered information systems, 962
Human-centered product planning and design, 1297–1310
  conceptual design document for, 1307
  detailed design document for, 1307
  engineering phase of, 1300, 1306–1308
  marketing phase of, 1300, 1303–1306
  measurement issues in, 1298, 1301
  naturalist phase of, 1299, 1301–1304
  objectives document for, 1306–1307
  objectives of, 1297
  requirement document for, 1307
  sales and service phase of, 1300, 1308–1310
  technology in, 1300
Human–computer interaction (HCI), 1193–1230. *See also* User interface(s)
  and auditory environment, 1200
  cognitive design considerations with, 1205–1217
    and contextual task analysis, 1206–1211
    and requirements definition, 1206
    and usability evaluation, 1216–1220
    and user interfaces, 1212–1216
  ergonomics of, 1194–1205
  and glare, 1199–1200
  and HVAC, 1200–1201
  and illumination, 1198–1199
  interfaces, computer, 1195–1200, 1201–1202
    accessories, 1202
    and character design, 1196–1197
    design of, 1212–1216
    and flicker, 1197
    and image stability, 1197
    keyboard, 1201–1202
    mouse, 1202
    swivel/tilt, screen, 1197–1198
    and viewing distance, 1195, 1197
  international considerations in, 1228
  and iterative design, 1228–1230
  and lighting, 1198
  and luminance, 1199
  management factors affecting, 1225–1228
  and noise, 1200
  organizational factors affecting, 1222–1225
  social environment for, 1217, 1220–1222
  usability evaluation of, 1216–1220
  for virtual environments, 2504–2509
    combined interaction, 2509
    direct manipulative interaction, 2508
    formal language interaction, 2508
    gesture interaction, 2508–2509
    natural language interaction, 2508
    and visualization, 2504–2507
  visual aspects of, 1198–1200
  and work practices, 1205
  and workstation design, 1202–1205
Human factors:
  in computer integrated manufacturing (CIM), 488
  in job design, 875
  in reliability, 1941
  in test and inspection, 1894–1900
    decision, 1896–1899
    and job design, 1899–1900
    present, 1895
    respond, 1899
    search, 1895–1896
    setup, 1894–1895
Human factors audits, 1131–1152
  at colliery (example), 1150–1151
  for decentralized business (example), 1146–1150
  design of, 1132–1146
    broad issues in, 1132
    data analysis/presentation, 1145–1146
    data-collection instrument, selection of, 1136–1145
    sampling scheme, 1135–1136
    standards, use of, 1133–1134
  evaluation of, 1134–1135
  need for, 1131–1132
  reliability in, 1134–1135
Human factors controls (safety/health), 1176–1179
Human Factors and Ergonomics Society, 1195
Human information-processing model, 1014–1017
Human judgment models, 2200–2201
Human–machine interaction, 876, 1020–1021
Human modeling, digital, *see* Digital human modeling
Human performance modeling, 2410–2441
  Distributed Operator Model Architecture (DOMAR), 2440–2441
  evolution of, 2410
  first-principle models, 2413–2414
  Man–Machine Integrated Design and Analysis System (MIDAS), 2413, 2429–2440
    aviation case studies, 2436–2440
    development of, 2429–2430

    structural architecture of, 2434–2436
    system architecture of, 2431–2434
  questions addressed by, 2411–2412
  reductionist models, 2413
  task network models, 2413–2429
    crew workload, evaluation of, 2420–2427
    design issues, 2420
    elements of, 2414–2419
    new task environments, extension of findings to, 2427–2429
    process control operator example, 2419–2420
Human reliability analysis (HRA), 1909, 2189
Human resource management (HRM):
  in EPEM model, 1798
  and ERP function, 335–336, 339
  and health/safety, 1179–1180
  for knowledge-management skills, 217
  and leadership, 855–863
    compensation/reward systems, 861–862
    inducement/involvement strategy, 862–863
    performance appraisal, 858
    performance management, 858–859
    recruiting, 856–858
    training/development, 859–861
  project, 1247
  in services, 641–642
  strategic approach to, 856
Human resources:
  ERP tools for, 91–92
  job design/redesign influences on, 869–871
  quality of
    benefits of CIM system implementation for, 526–527
    and computer integrated manufacturing (CIM), 526–527
  structure of, 1710
Humanscale, 1043, 1048
Human strength, design for, 1050–1058
  computer-simulation, use of, 1054
  joint strengths, maximum voluntary, 1052
  occupational strength testing, 1052
  and push–pull force limits, 1054–1058
  and static vs. dynamic strengths, 1052, 1053
Humidity, work area, 1200
Hurwicz principle (decision theory), 2379–2380
HVAC, *see* Heating, ventilating, and air conditioning
HyperText Markup Language (HTML), 76–77, 244, 245–246
Hypertext Transfer Protocol (HTTP), 244–245
Hypotheses:
  conditions for describing, 137
  in experimental design, 2232–2233
  uses of, 137
Hypothesis testing, 1023, 2243–2262
  in analysis of designed experiments, 2260–2264
  and confidence interval estimation, 2253–2254
  for equality of means and variances for *k* populations, 2255–2256
  for equality of two population variances, 2251–2252
  experimental design for, 2260–2264
  goodness of fit test, 2256, 2258, 2259
  and maximum-likelihood estimators, 2254–2255
  mean value with σ^2 known, 2244–2248
  mean value with σ^2 unknown, 2248–2249
  nonparametric tests, 2256
  in regression analysis, 2262
  single variance, 2249
  two means:
    variances known, 2249–2250
    variances unknown and not equal, 2252
    variances unknown but assured equal, 2250–2251
Hysteresis (in measurement systems), 1879–1880

I^2 Technologies, 94
I2 Technologies Inc., 2058
IBM, 83, 86, 95, 302, 654, 783, 1250
ICOH (International Commission on Occupational Health), 1067
iCollaboration software suite, 966–968
Idea-generation techniques, 2213
IDEF0 modeling method, 507, 508, 509
IDEF3 modeling method, 507
IDL, *see* Interface Definition Language
IDS system, *see* Integrated Data Store system
ID3 (Iterative Dichotomister 3), 1776
IEA, *see* International Ergonomics Association
IEA Checklist, 1137, 1138
IEC (International Electrotechnical Commission), 1968
IEC/TC 56, 1973
IEDF1X modeling method, 510
IEEE (Institute of Electrical and Electronic Engineers), 1967
IEM, *see* Integrated enterprise modeling
IEs, *see* Industrial engineers
IFIP (International Federation for Information Processing), 300
IFR (International Federation of Robotics), 379
IGES, *see* Initial graphics exchange specification
IIE, *see* Institute of Industrial Engineers
IIE Solutions, 1260
Illumination, 1177
  ergonomic recommendations for, 1196
  and human–computer interaction, 1198–1199
ILO (International Labour Organization), 1165
ILOG, 1738
Image-processing systems, 385–386, 1904–1905
Image stability, computer screen, 1197
Image theory, 2207
Immersive projection technology (IPT), 2499, 2516
Immersive virtual reality, 1124
Immigration, market effects of, 37
Impact assessment/analysis, 127
Impermanent organizations, performance management in, 1001
Implicit knowledge, 1291–1292

Improvement, 1808
  cycle process for, 11–13
  health care delivery systems, 745–746
  model development for, 1630
  quantifying, 1400–1404
  through knowledge management (KM), 217
  tools for, 1808–1827
    case study, 1823–1827
    information, gathering of, 1810–1813
    information, organization of, 1813–1820
    integration of tools, 1822–1823
    relationships, 1821–1822
    systems/processes, 1809, 1810
    variation, 1821
Incentive pay, 1392
Incident reporting (safety/health), 1174
Include conditions, 691
Income mobility, 37–38
Incomplete anonymity, 268
Independent variables, 2265
Indexes:
  cost-estimating, 2310–2311
  price, 2395
Indexing machines, 418–419
Indiana Department of Environmental Management, 596
Indifference methods (decision analysis), 2194–2195
Indirect labor, 2300
Indirect lighting, 1198
Indirect materials, 2300
Indirect measurement techniques (decision analysis), 2195
Indirect optimization, 2541, 2553
Indirect work standards, 1459–1462
Individuals, shift scheduling for, 1744, 1757–1764
Individual differences:
  in job design/redesign, 886–888
  and team design, 886–888
Individualism (in national cultures), 957
Individual learning, 1250, 1400
Individual measurement, control charts for, 1841–1844
Individual record data model, 120
Industrial ecology, 533
Industrial economy, 344
Industrial engineering:
  environmental engineering vs., 530
  plant engineering and techniques of, 1560–1562
Industrial engineers (IEs):
  environmental engineers vs., 530
  as plant/facilities engineers, 1553–1557
  role of, in transportation, 790
Industrial Management, 17
Industrial metabolism, 530
Industrial network applications, 250
Industrial robots, 360–362, 366, 373
  classification/types of, 374–375
  control systems, 376, 377
  definition of, 373
  in flexible assembly systems, 360–362
  as flexible handling equipment, 420
  gripping systems, 377
  measuring equipment, 376
  power supply for, 374, 376
  programming of, 377–378
  simulations, 378, 379
Industrial and systems engineers (ISEs):
  as change masters, 14–15
  definition of, 4–6
  and enterprise excellence models, 8–10
  and extended enterprise system, 5
  full-potential contribution from, 22–23
  and infrastructure for improvement, 16
  and learning system, 17
  and operations improvement, 18–22
  personal mastery issues, 23–24
  and planning system, 13
  and relationship management, 13–14
  roles of, 4, 6–7, 17–18
Industrial trucks, 1505–1513
  and closest-open-location (COL) rule, 1509–1510
  concurrent vs. sequential travel, 1510–1511
  counterbalanced lift truck, 1506, 1508
  first-come-first-served (FCFS) dispatching, 1511–1513
  number of trucks, determining, 1508–1509
  order picker truck, 1508
  pallet jack, 1505
  shortest-travel-time-first (STTF) rule, 1511, 1513
  sideloader truck, 1507, 1508
  straddle truck, 1506, 1508
  throughput capacity, 1509, 1510
  turret trucks, 1507
  walkie stacker, 1505, 1508
Industry classification systems, 329
Infeasible solutions, 2583
Infectious disease, 1167
Inference:
  human, 2196–2198
  statistical, *see* Statistical inference
Inference trees, 2189
Infinite buffer systems, 1639
Infinite horizon dynamic programs, 2643–2645
Infinite inventory banks, queueing models for, 1646
Inflation, 344, 2394–2405
  and actual vs. constant dollars, 2397–2398
  in after-tax cash flow analysis, 2403–2405
  average, 2395–2396
  in before-tax cash flow analysis, 2401–2403
  definition of, 2394
  differing rates of, for component cash flows, 2400
  differing rates of, per time period, 2400–2401
  economic equivalence calculations with, 2398–2400
  and exchange rates, 2401
  and interest rates, 2396–2397
  measures of, 2395
  periodic, 2395
  and purchasing power, 2395
Inflation-free interest rate, 2396–2398

Influence diagrams:
  for decision structuring, 2190–2191
  decision trees vs., 2190–2190
Informal monitoring, 1347
Informal organizations, 1005–1006, 1008–1009
Information. *See also* Communication(s)
  amount/availability of, with computer technologies, 1225
  caching of, 232, 233
  in context of knowledge management, 214
  distribution of, 1248
  in EPEM model, 1798
  and error reduction, 1368, 1369
  gathering/organizing (methods engineering), 1371–1387
    and arrangement of equipment, 1379–1382
    and balancing flow lines, 1382–1385
    between-operations analysis, 1374–1385
    with flow diagrams, 1374–1376
    with multiactivity charts, 1376–1379
    and SEARCH method, 1373, 1374
    by videotaping, 1371–1373
    within-operation analysis, 1385–1387
  on global network, 232
  on Internet:
    relevance/reliability of, 232
    security/privacy of, 267–269
  in networks:
    access to, 230–232
    provision of, 232, 233
    public vs. private, 232
    transmission of, for collaboration, 234
    value of, 232
  in process-oriented enterprise modeling, 288–290
  selective processing of, 2199–2200
  tools for gathering, 1810–1813
  tools for organizing, 1810, 1813–1820
  training and acquisition of learned, 929–930
  training and retention of, 930–931
  and transportation, 819
  in transportation management, 2056–2057
  WBS templates for storage/retrieval of, 1277
  workplace hazard, requirements for, 1176–1177
  World Wide Web as access to/technology for, 246
Information Age, 107
Information-based team training, 934
Information Builders, 83
Information economy, 344
Information flow:
  and material handling, 1503
  in supply chains, 2124, 2125
Information integration, 490–491
  model-driven approach to, 522–524
  in production process, 522–525
Information integration theory, 2200–2201
Information processing, 1014–1017
Information requirements:
  for business model, 31
  for group/organizational decision making, 135–141
    activities, decision-making, 140
    assessment of probability, 138–139
    attributes affecting quality/usefulnes of, 136–137
    decision perspective, 139
    and equivocality reduction, 141
    hypotheses, 137
    and imperfections in decision making process, 139–140
    logical reasoning, 137–138
    and organizational ambiguity, 140
    and representational appropriateness, 140
    and task relevance, 140
Information-retrieval tools, 221
Information Society, 249
Information systems (IS), 66–107
  and artificial intelligence technologies, 107
  classification approaches for, 67–68
  computer aided software engineering tools, 105
  database management tools for use in, 79–85
    data warehouses, 83–85
    object-oriented databases, 82–83
    and relational database model, 80–81
  data dictionaries, 102–103
  data flow diagrams, 99–101
  DBMS tools used in, 79–85
  development of, 70
  and distribution, 1472
  enterprise, 69
  enterprise models/modeling for design of, 285–286
  enterprise tools, *see* Enterprise resource planning
  entity relationship diagrams, 102, 103
  feasibility analysis, 98–99, 106
  flexibility of, 67
  functional, 68–69
  future trends in, 105, 107
  Gantt charts, 103–104
  in health care delivery systems, 747
  human-centered, 962
  joint application deployment, 105
  kanban as, 549
  local, 68
  management (MIS), 491–494
  operational vs. informational systems in, 83
  in operation of CIM, 489
  and organizational/system improvements, 147
  programming languages for building, 70–79
    ASP, 79
    C++, 72–73
    CGI, 77–78
    ColdFusion, 78–79
    HTML, 76–77
    Java, 78
    Visual Basic, 73–76
    web-based programming, 76–79
  rapid application deployment, 104, 105
  reasons for failure in, 961
  record keeping as core function of, 66–67
  and reports, 67
  systems development life cycle for, 96–106
    computer aided software engineering, 105

Information systems (IS) (*Continued*)
    data dictionaries, 102–103
    data flow diagrams, 99–101
    entity relationship diagrams, 102, 103
    feasibility analysis, 98–99
    Gantt charts, 103–104
    joint application deployment, 105
    PERT diagrams, 104
    rapid application deployment, 104, 105
    Structured English, use of, 100–102
  transorganizational, 69–70
    as element of electronic commerce, 69–70
  and types of knowledge, 67
  value of, 67
Information systems architecture (ISA), 302, 1782
Information system design methodologies (ISDMs), 286
Information Systems Methodology (ISM), 300–301
Information technology (IT), 36, 146, 949, 1888. *See also specific headings, e.g.:* Decision support systems
  as base for knowledge management, 213
  as driver of core business processes, 43
  as driver of knowledge management, 215, 216
  and economic variables interactions, 344
  for global production networks, 406–407
  in job evaluation and job evaluation systems, 914
  in kanban systems, 551
  as key trend, 38
  and knowledge management, 148
  for lean product development, 556
  migration for continuous improvement in, 349
  and organizational/system improvements, 147
  in process design and reengineering, 1706
  and retail supply chains, 782
  strategies for, 41
  and test and inspection, 1889
  trends in, 223
Information view, modeling method for, 509, 510
Infrared thermography, 1614
Inheritance (OOP), 71–72, 291, 292, 1328
Initial-condition bias, 2477–2483
  direction of, 2479–2483
  remedial measures for, 2478–2479
Initial graphics exchange specification (IGES), 192–193
Initiate phase (process design and reengineering), 1697, 1706–1708
Injection molding, 467, 1324, 1326
Injuries, work-related, *see* Work-related injuries
Inland Steel, 654
Inner pack, 2087
Innovation:
  and performance management, 1000
  in technology-organization solutions, 961–963
  as TQL success factor, 1805
In-process recycling, 533
Inputs to processes, 44
Input-output analysis models, 128, 2525–2526
Input–process–output (IPO) model of team effectiveness, 877–880
  input factors in, 878–879
  output factors in, 878, 880
  process factors in, 878, 879
Inspection, *see* Test and inspection
Instability, subcritical, 2167
Instance level of abstraction, 281, 283
Institute for Design and Construction, 321
Institute for Hygiene and Applied Physiology, 321
Institute for Information Systems (IWi), 290–291
Institute for Manufacturing Automation and Production Systems, 422
Institute of Electrical and Electronic Engineers (IEEE), 1967
Institute of Industrial Engineers (IIE), 23, 739
Institute of Production Systems, 616
Instructional systems development (ISD), 926–927, 935
Instruction sets (programming), 71
Instrumentation (automatic control system), 158
Instrument Society of America (ISA), 1772
Insufficient reason principle, 2380
Insurance industry:
  percent of GDP in, 346
  process design and reengineering case study, 1713–1714
Insurance-linked asset products, 761
Intangible products, services as, 636–637
Integrated Data Store (IDS) system, 80
Integrated enterprise modeling (IEM), 218–220
Integrated evaluation tool, 321, 322
Integrated modeling, 205–206
Integrated process chains, 204
Integrated product and process designs, 1329–1330
Integrated quality system, *see* Computer-aided quality-management system
Integrated simulation, 320
Integrated solutions business model, 603
Integrated system, creation of, 1009–1010
Integration:
  as core of CIM, 489–491
  data, 89
  in distributed environment, 604
  knowledge, 1293
  for product design/process platforms methodology, 2000–2002
  promoting, 23
  supported by World Wide Web, 246
  taxonomy of, 604, 605
  trend toward, 107
  using World Wide Web systems, 256
Integration management, project, 1244
Integration platform technology, 516–518
  evolution of, 516–517
  MACIP system architecture, 517–518
  requirements for, 516
Integration technology, 164–167

  and distributed vs. central control, 166, 167
  networking, 165–166
  object orientation, 166
  Petri net, 166
  in robot simulator/emulator, 166, 167
Integrative model of decision making, 2175–2178
Intel, 662, 783, 952
Intellectual capital, 147–149
Intellectual property, 268
Intelligence:
  artificial, *see* Artificial intelligence (AI)
  collective, 976
  of computer networks, 229, 230, 234
  organizational, 146
Intelligent control models, hybrid, 164
Intelligent transportation systems (ITS), 819, 822
Intensive distribution, 2129
Interactive planning method, 321
Interactive systems, *see* Human–computer interaction
Interactive video-based instruction, 928
Interactive voice response technology (IVR), 658
Intercorrelation, 2275–2277
  ambiguity in assessment of contributions, 2276–2277
  detection of correlation, 2277
  estimates, intercorrelated, 2276
  interactions vs., 2279
  variances, potentially enlarged, 2275–2276
Interest:
  compound, *see* Compound interest
  simple, 2336
Interest rate(s), 2334
  and inflation, 2396–2397
  nominal vs. effective, 2337
  selection of, 2335–2336
  and weighted average cost of capital, 2334–2335
Interfaces. *See also* User interface(s)
  in client/server (C/S) systems, 718–722
    CORBA, 719–722
    remote procedure call (RPC), 719
    socket interface, 718–719
  for computer aided design (CAD), 191–195
    classification of, 191
    definition, 191
    IGES, 192–193
    product data exchange, 192
    standardization of, 191–195
    STEP, 193–195
  computer–human, 1195–1200, 1201–1202
    accessories, 1202
    and character design, 1196–1197
    design of, 1212–1216
    and flicker, 1197
    and image stability, 1197
    keyboard, 1201–1202
    mouse, 1202
    swivel/tilt, screen, 1197–1198
    and viewing distance, 1195, 1197
  with control functions, 1772
  ecological, 1020–1021, 1024
  enterprise resource planning (ERP), 336–339
    and B2B supply chain operations, 343
    and contracts management, 336, 337
    and customer relationship management, 337
    external user-to-ERP, 343
    and finance, 339
    and human resource management, 339
    internal user-to-ERP, 343
    and joint supply planning, 344
    and manufacturing execution, 338–339
    and product configuration management, 338
    and product data management, 338
    and standards development, 349–350
    and supplier relationship management, 337
    and supply chain execution, 338
    and supply chain planning, 338
  geographic information system (GIS), 2016, 2017
  man-machine:
    and reliability, 1941
    and SRK model, 1020–1021
  personnel scheduling, 1765
  worker-machine, 548
Interface Definition Language (IDL), 720, 721, 1774
Interior point method (linear programming), 2530–2534
  computational efficiency of, 2534
  Fiacco and McCormick algorithm, 2531–2532
  Lagrange multiplier method, 2531
  Newton's method, 2530–2531
  simplex method vs., 2534
Intermediation models (electronic commerce), 271, 272
Intermittent process, 330
Internal consistency (user interfaces), 133
Internal customers, 14, 23, 659–660
Internal data, 117
Internal data model, 119, 120
Internalized *t* ratio, 2284
Internal risks, 45
International Commission on Occupational Health (ICOH), 1067
International Computers Ltd., 1713
International Electrotechnical Commission (IEC), 1968
International Ergonomics Association (IEA), 1067, 1144, 1195
International Federation for Information Processing (IFIP), 300
International Federation of Robotics (IFR), 379
International Harvester, 654
International issues:
  in human–computer interaction, 1228
  in transfer of organizational culture, 957–959
International Labour Organization (ILO), 1165
International Manufacturing Company, 2376
International Monetary Fund, 273
International Motor Vehicle Program (MIT), 545

International Organization for Standardization (ISO), 497, 731, 1133, 1185–1186, 1772, 1782, 1968, 1974
International quality management systems standards, 1968–1969
Internet, 237–243. *See also* World Wide Web
  addressing/naming system for, 237, 241–243
  architecture of, 238–239
  business effects of, 705–706
  caching of information on, 232, 233
  and client–server mechanism, 240–241
  connectivity, Internet, 254–255
  data communication via, 2124
  and design by customers, 701
  and electronic commerce, *see* Electronic commerce
  electronic commerce on, 260
  and enterprise resource planning (ERP), 342–344
  and firewalls, 734–735
  history of, 235–236, 238
  LP software on, 2536
  and maintenance systems, 1621–1622
  and packet switching, 239
  protocols for, 239–240
  relevance/reliability of information on, 232
  retailers on, 778–779
  revenues from/jobs on, 1988, 260
  security on, 278
  services available on, 243
  traffic volume on, 232
  use of, in selection, 939
  as world's largest network, 237–238
Internet banking, 735–736
Internet-based ERP, 342
Internet economy, 261, 267, 269–271
Internet portals, 271
Internet Protocol (IP), 240, 250
Internet service providers (ISPs), 237, 250
Internet Society, 238, 257
Interoperability (enterprise resource planning), 348–351
Interpretation, cognitive probes for, 1026
Interval of uncertainty, 2548
Interviews:
  for gathering task information, 1209
  as human-centered product planning/design tool, 1302–1304, 1309, 1310
  selection, 922
Intranets, 220, 255–256
  definition of, 237
  as private networks, 238
Intrinsic availability, 1924
Inventory. *See also* Production-inventory systems
  choosing policy for, 2021–2022
  control, inventory, 1392, 2104
  and demand forecasting, 2021–2022
  and direct product replenishment, 780–781
  increased turns in, 2071
  level of, 549
  management, inventory, 779
  queueing models for determining, 1631
  reduction of:
    from CIM implementation, 525
    and JIT, 545
  square-root rule of consolidation, 2071
  status, inventory, 2040
  storage analysis chart (SAC), 1532–1534
Inventory queues, 1672–1673, 1690–1692
Inverse kinematics, 1115–1116
Investigations, health hazards (NIOSH), 1163–1164
Investment(s):
  in automation, economic climate for, 363
  casting, investment, 565
  risk classes of, 2391–2392
Investors:
  life cycle of, 755
  risk preferences of, 753–755
Invoicing, 2065
Involvement:
  and leadership, 862–863
  and teamwork, 976
I/O point, 2087
IP, *see* Internet Protocol
IPA, *see* Fraunhofer Institute for Production and Automation
IP addressing, 241–242
IPO model, *see* Input–process–output model of team effectiveness
IPT, *see* Immersive projection technology
Irritant dermatitis, 1167
IS, *see* Information systems
ISA, *see* Information systems architecture; Instrument Society of America
ISD, *see* Instructional systems development
ISDMs (information system design methodologies), 286
ISEs, *see* Industrial and systems engineers
ISIS, 1776
ISM, *see* Information Systems Methodology
ISO, *see* International Organization for Standardization
ISO 9000–3:1997, 1972
ISO 9000–4:1993, 1972–1973
ISO 9000:2000, 1968–1969
ISO 9000 family standards, 1973
ISO 9000 QMS standards, 1972–1973
ISO 9001, 1973
ISO 9001:1994, 1972
ISO 9001:2000, 1968, 1969, 1972
ISO 9001:2000 QMS standard, 1969–1972
  continual improvement clause in, 1972
  management responsibility clause in, 1970
  measurement/analysis clauses in, 1971
  product realization clauses in, 1971
  resource management clause in, 1971
  scope of, 1969
ISO 9002, 1969
ISO 9003, 1969
ISO 9004, 1968, 1969
ISO 9004:2000, 1968, 1969
ISO 9004–2000 QMS standard, 1972
ISO 10006, 1969, 1972
ISO 10007, 1969, 1972
ISO 10011, 1969
ISO 10012, 1969, 1972

ISO 10013, 1969, 1972
ISO 10015, 1969, 1972
ISO 10017, 1969
ISO 19011, 1969, 1972
ISO/IEC JTC/SC7 standard, 1972
ISO Industrial Automation Technical Committee Number 184, 165
Isolation, for reducing environmental inputs effects, 1883
ISO/TR 10014, 1972
ISO/TR 10017, 1972
ISO/TS 16949, 1973
ISPs, *see* Internet service providers
Israel Defense Forces, 849, 850
Issue analysis models (DSS), 127–129
Issue formulation models (DSS), 126–127
Issue interpretation models (DSS), 129
I-STEPS Environmental Software, 596
IT, *see* Information technology
Italy, industrial robots in, 373
Items, 2087, 2525
Iterations (Structured English), 101–102
Iterative design, 1228–1230
Iterative Dichotomister 3 (ID3), 1776
ITS, *see* Intelligent transportation systems
i2, 1738
IVR (interactive voice response technology), 658
IWi, *see* Institute for Information Systems

JACK (computer-aided model), 1050, 1115
Jackson networks, 2164–2165
  closed, 1656–1660
  generalized, 2168
  open:
    multiple-job-class, 1652–1654
    single-job-class, 1650–1652
JAD (joint application deployment), 105
JAI, *see* Joint angles of isocomfort
Janitorial workforce, 1567
Japan:
  automobile industry in, 1313
  failures in FMS implementation, 949
  HQL anthropometric database, 1114
  industrial robots in, 373
  multiskilled workers in, 547
Japan Ergonomics Society (JES), 1195
Japanese Union of Scientists and Engineers (JUSE), 555
Japan Institute of Plant Maintenance (JIPM), 555
Java (programming language), 78, 306, 714
Java RMI (remote method invocation), 721
JBuilder, 304
J.D. Edwards Inc., 87, 88, 1738
JES (Japan Ergonomics Society), 1195
JIPM (Japan Institute of Plant Maintanance), 555
Jishu jozen, 553
JIT, *see* Just-in-time
Jobs:
  breakdown for time study of, 1418–1419
  from Internet commerce, 260
  tasks vs., 869
Job-comparison scale, 914–915
Job design/redesign, 869–877, 900, 904, 1353–1371. *See also* Team design
  biological, 876
  combining tasks in, 885–886
  and computer technologies, 1222
  criteria for, 1354
  and error reduction, 1368–1371
  evaluating need for, 882–884
  evaluation of, 889, 892–893
    biases, potential, 893
    and data sources, 892
    example of, 893
    long-term effects, 892–893
    with questionnaires, 889, 892
  and fatigue reduction, 1365–1368
  HR outcomes of, 869–871
  for human test and inspection systems, 1899–1900
  individual differences in, 886–888
  job analysis for, 926
  and job–task distinction, 869
  mechanistic, 870, 874
  motivational, 874–875
  Multimethod Job Design Questionnaire (MJDQ), 872–873
  and musculoskeletal disorders, 1362–1365
  organizational influences on, 869
  perceptual/motor, 875–876
  as prerequisite to training, 926
  for reduction of work-related musculoskeletal disorders, 1093–1097
    procedures, 1095
    risk factors, 1093–1094
    surveillance, 1095–1097
  for safe/healthful workplaces, 1178–1179
  strategies for, 884–885
    and basic decisions, 888–889
    existing jobs, redesign of, 885
    initial design, 884–885
  trade-offs among approaches to, 876–877
  workstation organization
    groups of workstations, 1354–1358
    individual workstations, 1357–1362
Job elements, 1418–1419
Job evaluation and job evaluation systems, 900–917
  acceptability of, 917
  appeals/reviews, handling, 913
  classification method, 903
  evaluation of, 914–917
    acceptability, 917
    legal defensibility, 916–917
    reliability, 914–915
    utility, 916
    validity, 915–916
  factor comparison method, 903–907
  future trends in, 914
  implementation of, 911–913
    macro objectives of, 911–912
    major decisions for, 911
    method, selection of, 912
    micro objectives of, 912
    participants, selection of, 912–913

Job evaluation and job evaluation systems (*Continued*)
  and knowledge-based pay systems, 911
  legal defensibility of, 916–917
  maintenance/administration of, 913–914
  and market-based pay systems, 910–911
  point method, 907–910
  ranking method, 902–903
  and relative worth of jobs, 900–901
  reliability of, 914–915
  role of individual in, 901
  single-factor systems, 910
  and skill-based pay systems, 911
  and societal values, 901
  traditional, 902–910
    classification method, 903
    factor comparison method, 903–907
    point method, 907–910
    ranking method, 902–903
    single-factor systems, 910
  using information technology in, 914
  utility of, 916
  validity of, 915–916
Job evaluators, 913
Job families, 912
Job rotation, 1363
Job satisfaction:
  and QC involvement, 979
  and queueing models, 1629
  and use of computer technologies, 1224
Job scheduling, 497
Job severity index (JSI), 1080, 1081
Job shops, 330, 331, 334, 608, 1632
Job standards, 1449
Job surveys, 1096
Joining, 407
Joining processes, 456
Joining technologies, 371–373, 409–413
  classification of, 409–410
  clinching, 373, 411, 412
  press-fitting, 372
  riveting, 372, 411, 412
  screwing/bolting, 371, 410, 411
  self-pierce riveting, 372
  soldering, 423–425, 429–431
  sticking, 412–413
  welding, 413
Joint angles of isocomfort (JAI), 1064, 1067
Joint application deployment (JAD), 105
Joints, coupled, 1115
Joint strengths, maximum voluntary, 1052
Joint supply planning (using Internet), 344, 350
JSI, *see* Job severity index
"J-standards," 1121, 1122
Junjo-hiki, 551
JUSE (Japanese Union of Scientists and Engineers), 555
Just-in-time (JIT), 492–494, 544–559
  autonomation in, 548–549
  continuous improvement in, 548
  elements of, 545
  and joint implementation of 3Ts, 553, 555
  kanban as decentralized control system for, 545, 549–551
    alternatives, 550–551
    case study, 551
    control parameters, 550
    limitations, 550
  and lean production, 555–557
  in service industries, 559
  smoothing of volume/variety in, 545–547
  and Theory of Constraints (TOC), 557–558
  and Toyota Production System, 544–545
  and TPM, 553
  and TQM, 552–555
  workforce for, 547–548
"Just wage" doctrine, 901

Kaizen, 547, 557
Kanban control, 549–551, 1692
  alternatives, 550–551
  appropriate environments for, 545
  case study, 551
  control parameters, 550
  drum-buffer-rope in, 558
  electronic, 551
  limitations, 550
  for production-inventory systems, 1689–1690
  as pull system of production, 545
  queueing models for coordination of production, 1664–1667
Karush–Kuhn–Tucker (KKT) conditions, 2554–2555
Kbps, 232
KBS, *see* Knowledge-based systems
K&E, 653, 654
Kerberos, 733, 734
Keyboards, computer, 1196, 1201–1202
Key business processes, 40, 44
Key competencies, 146
Key performance indicators (KPIs), 55–56
Kilobits per second, 232
Kinematics:
  definition of, 374
  in digital human modeling, 1112–1113
  in electronic production, 425
  inverse, 1115–1116
Kits, tracking, 492
Kitchens, food service, 833–834
  designing:
    by consensus, 826–830
    for supervision, 829, 830
    for utility use, 830
  ergonomic principles, 833–834
  evaluating layout efficiency of, 830–833
  productivity in, 826
  storage in, 833–834
  tables, 833
  workstations, 834
Kitting, 1379
KKT conditions, *see* Karush–Kuhn–Tucker conditions
KM, *see* Knowledge management
Knowledge:
  categories of, 213–214
  as company asset, 1888
  compensable factor scheme for, 908, 909
  customer/market, 1962

  declarative, 1775
  definition of, 213–214
  explicit, 214
  implicit, 1291–1292
  in industrial vs. digital economies, 261
  metaknowledge, 1775
  procedural, 1775
  types of, in information systems, 67
Knowledge, skills, and abilities (KSAs), 880
Knowledge assets, 148
Knowledge base, 131–132
Knowledge-based approach to schedule generation, 1736, 1737
Knowledge-based behavior, 1020
Knowledge-based pay systems, 911
Knowledge-based performance, 2206
Knowledge-based software systems, 1328–1329
Knowledge-based systems (KBS), 160, 162
  for artificial intelligence approaches to control, 1775–1776
  in shop floor scheduling, 1775–1776
Knowledge capital, 146, 147
Knowledge economy, 107, 602
Knowledge engineering, 1291–1293
Knowledge generation, 148
Knowledge integration, 1293
Knowledge in the world (action-cycle model), 1019
Knowledge management (KM), 213–223
  architecture for, 222
  business process-oriented, 218–220
  and business process reengineering, 217, 218
  core process of, 215, 216
  and definition of knowledge, 213–214
  design fields of, 215–217
  for DSSs, 145–149
  improvement through, 217
  origins of, 213
  technologies for, 220–222
  tools for, 220
  trends in, 222, 223
Knowledge networks, 42
Knowledge repositories, 963
Knowledge work/workers, 36–37, 147
Kolmogorov's backward and forward equations, 2155
Korea, 1114
KPIs, *see* Key performance indicators
KRISS anthropometric database, 1114
KSAs (knowledge, skills, and abilities), 880

Labor analysis, 2307–2308
Labor management, 1770–1771
Labor operations, indirect, 1458–1462
Labor productivity, 344
Lacerations, 1170
Lagrangean relaxations, 2587–2589
Lagrange multipliers, 2531, 2533, 2553–2554, 2561–2562
LANs, *see* Local area networks
LANCELOT, 2564
Land depth analysis, 2089, 2090
Land's End, 262, 266, 656
Language(s):
  business, 49
  computer, *see* Computer languages
  types of, 132
Laplace decision rule, 2177
Laplace principle (decision theory), 2380–2381
Lapping:
  geometric capabilities of, 464
  obtainable accuracy values, 565
  technological capabilities of, 471
Laser machining, 1323
Last-in-first-out (LIFO), 1521, 2157
Last-mile line speeds, 237, 249
Latin square designs, 2230, 2231
Law of requisite variety, 958
Layer manufacturing, *see* Rapid prototyping
Layout:
  of equipment, 1379–1382
  of keyboards, 1201–1202
Layout planning (warehousing), 1538–1541
  generating alternative layouts, 1538–1539
  objectives of, 1538
  philosophies of, 1539–1541
LBDs, *see* Low-back disorders
LBOs (leveraged buyouts), 760
LBP, *see* Low-back pain
LCDs, *see* Liquid crystal displays
Leadership, 841–863
  as critical success factor of knowledge management, 216
  and development, 859–861
  employee development as outcome of, 852–855
    attitudes toward leader, 854–855
    group development, 855
    personal development, 853–854
  in EPEM model, 1798
  of formal/informal organization, 1008–1009
  full range model of, 848–850
  and human resource management, 855–863
    compensation/reward systems, 861–862
    inducement/involvement strategy, 862–863
    performance appraisal, 858
    performance management, 858–859
    recruiting, 856–858
    training/development, 859–861
  and motivation, 841
  and naturalistic decision making, 2208
  organizational, 1958
  performance as outcome of, 851–852
  and performance management, 858–859, 1008–1009
  senior management, 1255
  service-driven, 1958–1959
  in supply chain management, 2126
  and team performance, 987
  and team success, 982
  total quality leadership (TQL), *see* Total quality leadership
  as TQL success factor, 1804
  and training, 859–861
  transactional approach to, 841–844
    calculative–rational basis of, 845
    extrinsic motivation in, 846

Leadership (*Continued*)
        individualistic orientation in, 846–847
    transformational approach to, 843–845
        collectivistic orientation in, 846–847
        emotional–expressive basis of, 845–846
        intrinsic motivation in, 847–848
    variation in, 1832
Leading principal minor test, 2545
Lead time(s):
    and demand forecasting, 2025–2027
    purchasing, 2050
    queueing models for determining, 1631
    utilization and mean/variance of, 2037
Lean Enterprise Institute, 8
Lean Enterprise Model (LEM), 8
Lean production, 545
    and just-in-time (JIT), 545, 555–557
    material handling in, 1502
    in service industries, 559
Learned information, training and acquisition of, 929–930
Learning. *See also* Education; Neural networks; Training
    individual, 1250
    machine, 222
    model development for, 1630
    team, 999
    time for, 1400–1406
        individual learning, 1400
        organization learning, 1400–1406
    from variation, 1832–1834
Learning organizations:
    project management in, 1250–1251
    and work breakdown structure, 1276–1277
Learning system, 17
Least-squares method, 2268–2270
Legal issues:
    in pricing, 680–682
    related to human factors, 1133
Legal requirements/regulations:
    for clean manufacturing, 531, 532
    for occupational safety/health, 1162, 1164–1166
Legs, posture checklist for, 1366
Lekin system, 1737, 1738
LEM (Lean Enterprise Model), 8
Less-than-truckload (LTL) industry, 2063, 2064
Level of abstraction, 281
Level coding, 2040
Levels (in experimental design), 2225
Levene's test, 2256
Leveraged buyouts (LBOs), 760
Lever systems, 1069–1070
Levi's, 783, 784
Lexicographic ordering principle, 2177, 2179
Liberty Mutual Insurance Co., 1118
Lie-detector tests, 921
Life characteristics curve, 1925–1927
Life cycle:
    project, 1242–1243
    system:
        reliability activities during, 1923–1925
Life-cycle assessment, environmental information from, 532, 536–538
Life-cycle costing:
    in hospitality industry, 834–835
    for justifying energy projects, 1577
Life-cycle design, environmental considerations in, 534–536
    process design, 536
    product design, 534–536
LIFO, *see* Last-in-first-out
Lift, recommended weight of, 1073–1075
Lifting allowances, 1396
Lifting equation, 1076–1080
Lighting:
    energy-improvement possibilities for, 1580
    and human–computer interaction, 1198
Likelihood ratio method, 2633–2634
Limited-resource model of human attention, 1016
Limited span of control, 1264
Limit switches, mechanical, 1903
LINDO, 2575
Linearity, in measurement systems, 1879
Linear models, 2524–2525
Linear optimization, *see* Linear programming
Linear position transducers, 1902
Linear programming (LP), 1725–1726, 2055–2056, 2524–2538
    and additivity, 2525–2526
    applications of, 2538
    computer software for, 2534–2536
    discrete optimization, *see* Discrete optimization
    and formulation of linear models, 2524–2525
    handling nonlinearities by, 2526–2527
        absolute value functions, 2527
        max-min problems, 2527
        piecewise linear functions, 2526–2527
    interior point method for, 2530–2534
        computational efficiency of, 2534
        Fiacco and McCormick algorithm, 2531–2532
        Lagrange multiplier method, 2531
        Newton's method, 2530–2531
    network flow models, *see* Network flow models
    and proportionality, 2525
    relaxations in, 2585–2587
    sensitivity analysis in, 2536–2538
        practical uses of, 2536–2537
        simultaneous variations in parameters, 2537–2538
    simplex algorithm for, 2527–2530
Linear programs, 2087, 2541
Line search techniques, 2547–2549
LINGO modeling language, 2076, 2080
Linkage:
    and assembly, 415–416
    assembly systems, 362, 415–416
    of business processes in SCM, 2118–2120
    of process platforms and product design, 1996–1999
    of supply chain members, 2123–2124
Linkage of processes (LOP), 1810, 1812
Linked process chains, 204
Liquid crystal displays (LCDs), 1195, 1197

Liquid state metal, designing for, 1316–1318
Little's law, 2037, 2162
Living companies, 7–8
The Living Company (A. DeGeus), 7
LNCs:
  and keyboards, 1201
  and viewing distance, 1197
  and workstation design, 1202
Local area networks (LANs), 165, 231, 238, 255, 256
Local information systems, 68, 107
Locality, axiom of, 1734
Local optima, 2590
Local search techniques (scheduling), 1731
Location problems (transportation management), 2067–2068
Lockheed Aircraft, 602
Logic, fuzzy, *see* Fuzzy logic
Logical data independence, 116
Logical data model, 119–120
Logical function allocation, 1912–1916
Logical reasoning, 137–138
Logistics, supply chain management vs., 2111–2115
Logistics-based work breakdown structure, 1271
Logistics management, 2007–2019. *See also specific topics, e.g.:* Warehousing
  and B2B electronic commerce, 264–265
  and decision support systems, 2011–2019
    analytical tools, 2013–2015
    input data, 2012–2013
    presentation tools, 2015–2018
  modeling of, 2008–2011
    network design/configuration, 2008–2010
    supply chain planning, 2009, 2010
    transportation planning, 2011
Lognormal distribution, 1931, 1932
London Ambulance Service, 949, 961
London Ambulance system, 951
Longest processing time first (LPT) rule, 1722, 1723
Longitudinal transfer systems, 358
Long-run behavior, determination of, 2161–2162
Long-term debt, weighted cost of, 2334
Long-term memory (LTM), 1015
LonWorks, 165–167
Loop principle (work sampling), 1456
Loose linkage, 415, 416
LOP, *see* Linkage of processes
Los Angeles, 949
Loss, probability of, 2367
Loss control, 1568
Lost sales, 1637–1638
Lot-for-lot policy, 1676–1678
Lotteries, reference (decision making), 2192
Lotus Notes, 143
Loughborough University, 1112
Low-back disorders (LBDs), 1167
  and manual materials-handling (MMH) tasks, 1070–1071, 1080–1082
  prevention of, 1080–1082
Low-back injury, digital human modeling of, 1119–1120
Low-back pain (LBP), 1070–171
Low-level network applications, 244
LP, *see* Linear programming
LPT rule, *see* Longest processing time first rule
LTL industry, *see* Less-than-truckload industry
LTM (long-term memory), 1015
Lucent, 269
Luminaire wiring, assembly of, 394–395
Luminance, 1199, 2506
Lung cancer, 1169
Lung disease, occupational, 1169

McDonald's, 93, 654
McDonnell Douglas, 7
Machine flexibility, 499
Machine learning, 222
Machinery:
  cost of, 467
  European standards for working postures during operation of, 1068
  hazards related to, 1160
  layout/use/design of, for safety, 1177
The Machine That Changed the World (Womack, Jones, and Roos), 555
Machine tools, conditions for global assembly in, 403
Machining, 453–455, 1322, 1323
Machining Data Handbook, 458, 467
Machining time, calculation of, 459–460
McHugh Software International, 2058
MACIP, *see* CIMS Application Integration Platform for Manufacturing Enterprises
Macroeconomics, 344
Macroeconomic developments, 40
Macroeconomic models, 128
MACRO Motion Analyses, 1441–1446
Made-on-demand products, 335
Made-to-stock products, 334–335
Magazines, 1302, 1303
Magazining, 383, 384
Magnetic grippers, 414
Magnetic motion-tracking devices, 1125
"Magnificent Seven" tools, 1857
  cause-and-effect diagram, 1859, 1860
  check sheet, 1858–1859
  control charts, 1861–1875
  defect concentration diagram, 1860, 1861
  histogram, 1856–1857
  Pareto chart, 1859
  scatter diagram, 1860–1862
Maintainability, 1946
  and availability, 1949–1951
  definition of, 1923
  measures of, 1946–1949
  and reliability, 1946–1949
Maintenance, 1586–1622
  assessment of, 1597–1605
    ACE team benchmarking system, 1598–1601
    for benchmarking, 1594–1597
    key performance indicators, 1601–1604

Maintenance (*Continued*)
        maintenance excellence index (MEI) for, 1604–1605
        techniques for, 1597–1598
    benchmarking of, 1593–1597
        assessment for, 1594–1597
        external benchmarking, 1593–1594
        internal benchmarking, 1593, 1597
    best practices for, 1610–1620
        continuous reliability improvement (CRI), 1610–1611
        maintenance repair operations (MRO) parts /materials/supplies, 1615–1616
        operator-based maintenance (OBM), 1620
        planning and scheduling, 1616–1618
        predictive maintenance (PdM), 1612–1615
        preventive maintenance, 1611–1612
        reliability-centered maintenance (RCM), 1618–1619
        total productive maintenance (TPM), 1619–1620
    contract, 1622
    future of, 1620–1622
    and IE principles, 1588
    and information technology, 1605–1610
    and plant engineering, 1566
    predictive, 1606–1608, 1612–1615
    preventive, 1606–1608, 1611–1612
    as profit center, 1588
    requirements for effective, 1588–1593
    scope of, 1587–1588
    test and inspection related to, 1907
    of time standards, 1407
    TPM, 553
Maintenance department, 1550
Maintenance excellence index (MEI), 1604–1605
Maintenance inspection, 1908–1912
Maintenance management:
    enterprise resource planning (ERP) function, 334
    major activities of, 1771
Maintenance operations, human models for, 1121
Maintenance repair operations (MRO), 1615–1616, 1622
Maintenance Steering Group 3 (MSG–3), 1908
Make-to-demand production, 330, 331, 338
Make-to-order manufacturing/service, 330, 1635–1636
Make-to-stock manufacturing/service systems, 330, 331, 1636–1638
Malcolm Baldrige National Quality Award (MBNQA), 8, 645, 950, 1798. *See also* Baldrige criteria
Malcolm Baldrige National Quality Improvement Act of 1987, 1956
Males:
    hand work areas for, 1360
    maximum forces of pull for, 1057, 1058
    reach distances for, 1361
    recommended weight of lift for (industrial workers), 1073–1075
MANs, *see* Metropolitan area networks
Manaco S.A., 609–616
Managed bandwidth services (MBS), 250
Managed business process links, 2118
Management. *See also specific headings*
    benefits of CIM system implementation for, 526
    commitment of, 1705, 2000
    computers and interaction with, 1220–1221
    control decisions by, 111, 112, 136–137
    and enterprise resource planning (ERP), 336
    and health/safety performance, 1161
    and human–computer interaction, 1225–1228
    impact of teams of, 988
    integrated approaches to, 604
    involvement of:
        in employee/management ergonomics committee, 1187
        in safety programs, 1183
    and joint union/management ergonomic committee, 1187
    participatory, 976, 983
    and plant engineering, 1551
    of plant/facilities engineering, 1557–1560
    and process reengineering, 1700, 1701
    quality improvement and actions by, 1032
    quality of, 1796
    responsibility of, 1970
    senior, 1255
    support of:
        for teams, 983, 984
        for teamwork initiatives, 981–982
    of total quality leadership (TQL) process, 1801
Management controls, 46, 48
Management information system (MIS), 491–494
Management systems, 21–22, 1795
    customer relationship, 69
    generic standards for, 1185
    supply chain, 69
Managerial planning, cognitive aid for, 1034–1037
Manager-led teams, 976
Managers:
    perception of hazards in workplace by, 1158
    project managers vs., 1334
    ten-category scheme for interpreting verbal protocols of, 1035
MANDATE (Manufacturing Management Data Exchange), 1782
Mandatory collaboration, 605
Man–Machine Integrated Design and Analysis System (MIDAS), 2413, 2429–2440
    aviation case studies, 2436–2440
        air traffic control, extension of model to, 2439–2440
        flight crew performance, prediction of, 2436–2439
    development of, 2429–2430
    structural architecture of, 2434–2436
    system architecture of, 2431–2434
Manual assembly systems, 356, 358, 359, 363, 416–418
    chained manual assembly stations, 416–417

  criteria for design of, 417–418
  software-based optimization of, 386, 387
  workstations in, 416, 417
Manual inspection, 1892
Manual materials-handling (MMH) tasks, 1070–1082
  biomechanical approach to design of, 1072, 1076
  and job severity index (JSI), 1080, 1081
  and low-back disorders, 1070–1071, 1080–1082
  physiological approach to design of, 1072
  psychophysical approach to design of, 1071–1075
  and revised NIOSH lifting equation, 1076–1080
Manual shift scheduling systems, 1765–1766
Manuals:
  design, 1998–1999
  process platform, 1996–1998
Manufacturers:
  environmental information provided by, 532
  responsibility of, for clean manufacturing, 532
  as service providers, 532
  and supply chain design, 2127–2128
Manufacturing. *See also specific headings*
  activity-based management in, 2317–2329
  advanced planning and scheduling, *see* Advanced planning and scheduling (APS)
  agent-based, 697
  agile, 486, 527
  assembly, *see* Assembly
  CAD/CAM systems for design use in, 1328–1329
  clean, *see* Clean manufacturing
  collaborative, *see* Collaborative manufacturing
  computer integrated, *see* Computer integrated manufacturing
  contract, 263–264
  cost management evolution in, 2318
  designing for, 1311–1330
    assembly, design for, 1328
    CAD/CAM systems for use in, 1328–1329
    and drawings, 1314–1315
    general principles of, 1315–1316
    and hierarchy of design, 1313–1314
    metal, 1316–1323
    and organizational issues, 1329–1330
    plastics, 1324–1328
    processes/materials, selection of, 1316
  and electronic commerce, 347
  ERP tools for, 90–91
  ex post, 276
  flexible production in, 404
  flow management in, 2122
  holonic, 697
  major business functions in, and enterprise resource planning (ERP), 326–327
  for mass customization, 694–701
    and coordination of resource allocation, 697–700
    and shop-floor control, 699–701
    and variant handling, 694–697
  metal, designing for, 1316–1323
    basic processes, 1317–1320
    liquid state, 1316–1318
    secondary processes, 1320–1323
    solid state, 1317, 1319
  nature of, 563
  near-net-shape processes, *see* Near-net-shape processes/production
  operations estimating for, 327–329, 2311–2314
  part, 562–563
  participative bureaucracy and information technology in, 952
  partitioning domain of, 329–331
    customer, nature of, 329
    customer orders, nature of business in terms of, 329–331
    process, nature of, 329–331
    product, nature of, 329
  percent of GDP in, 346
  plastics, designing for, 1324–1328
  and process planning, *see* Process planning
  production planning in, *see* Production planning
  reliability program applications during, 1954
  scope of ERP in, 331–332
  systems approach to, 705–706
  Web-based, 262–263
  work injuries in, 1070
Manufacturing Automation Protocol (MAP), 165
Manufacturing automation system, 496, 497
Manufacturing-based approach to service quality, 626, 638, 639
Manufacturing decision support systems, 348
Manufacturing devices, automated, 500
Manufacturing enterprises, ERP and major business functions in, 326–327
Manufacturing execution systems (MESs):
  for control, 1782–1787
    data in, 1782–1784
    market trends/future directions, 1787
    object models, 1783, 1785–1787
  enterprise resource planning (ERP) interface with, 338–339
Manufacturing facilities, stockrooms serving, 2086
Manufacturing features, 452, 454
Manufacturing hierarchy, 487
Manufacturing information and execution systems (MIES) software, 1772–1774
Manufacturing management:
  enterprise resource planning (ERP) function, 333
  kanban, 549–551
Manufacturing Management Data Exchange (MANDATE), 1782
Manufacturing Message Specification (MMS), 165
Manufacturing progress, 1400

Manufacturing resource planning (MRP II), 85–86, 348, 492, 493
Manufacturing Studies Board (National Research Council), 950
Manufacturing systems:
  flexible, 1633
  queueing models for, 1632–1633
Manugistics Inc., 94, 1738, 2058
MAP (Manufacturing Automation Protocol), 165
Mapping:
  in data models, 117–118
  model, 522–523
  process/feature, 463–466
  service, 641
  of supply chain networks, 2120
  between view of product families, 691, 692
Marginal averages, 2234
Marine Protection, Research, and Sanctuaries Act (MPRSA), 1164
Market(s), 29
  in business model, 40
  capital, 39–40
  changes in, and job design/redesign, 883
  and customers, in business model, 50–51
  customers' relationship to, 50–51
  definition of, 40
  electronic, *see* Electronic commerce
  for enterprise resource planning (ERP), 87–88
  focus, 34
  physical vs. virtual, 262
  potential, 40
  and pricing, 682
  segmentation, 40
  served, 40
Market acceptance, 2130
Market-based pay systems, 910–911
Market change graph, 486
Market concentration, 2130
Market coverage, 2128–2129
Marketing:
  customer access as new paradigm for, 660
  customer-pulled, 276
  of goods vs. services, 624
  and mass customization, 701–705
    customer decision-making process, 703–704
    design by customers, 701–703
    one-to-one marketing, 704–705
  one-to-one, 662, 704–705
  and online advertising, 272–273
  and pricing, 666
  of services/service quality, 623
Marketing channels, 2113, 2115–2116, 2129
Marketing phase (human-centered product planning and design), 1300, 1303–1306
Market interest rate, 2396–2398
Market knowledge, Baldrige criteria for, 1962
Market makers, 271
Market model:
  price mechanism of, 698–699
  for resource allocation coordination, 697–698
Market performance, 49, 50
Market power, in supply chain management, 2127–2128
Market research, 269
Market turbulence, 311–314
Markov chains, 2150–2156
  in continuous time, 2154–2156
  and Markov property, 2150–2151
  queueing model based on, 2153–2154, 2158–2159
  reversible, 2156
  steady-state distributions of, 2152–2153
  transition matrices in, 2151–2152
MARR, *see* Minimum attractive rate of return
Mary Kay Cosmetics, 848
Masculinity (in national cultures), 957
Masking, 2506
MASs (multi-agent systems), 174
Massachusetts Institute of Technology (MIT), 8, 545, 555, 733
Mass balance, estimation of emissions by, 596–597
Mass customization, 684–706
  benefits of, 685
  definition of, 685
  design for, 687–694
    commonality, 688–689
    common bases, 690–691
    customers, design by, 701–703
    derivation processes, 692–694
    modularity, 688–689
    multiple views, synchronization of, 691–692
    and product family concept, 688
    variety, product, 689–690
  and e-commerce, 705–706
  economic implications of, 685–686
  manufacturing for, 694–701
    and coordination of resource allocation, 697–700
    and shop-floor control, 699–701
    and variant handling, 694–697
  and maximization of reusability, 686
  and product life cycle, 687
  and product platform, 686–687
  and retail supply chains, 784
  sales and marketing for, 701–705
    and customer decision-making process, 703–704
    and design by customers, 701–703
    one-to-one marketing, 704–705
  technical challenges presented by, 686
Master production schedule (MPS), 2035, 2039
Master shift rotations, 1760
MAST (mechanized activity sampling technique), 1458
Mast Simulation Environment, 2458
Material analysis (for cost estimating), 2308–2309
Material flow(s):
  and kanban, 549–551
  order-consolidation, 2106–2107
  simulation of, 388
  in warehouse operations, 2084–2085, 2097, 2098

Material handling, 1502–1525
  automated systems for, 500, 1524–1525
  in automated test and inspection, 1902
  and containerization, 1503
  conveyors for, 1513–1520
    belt conveyors, 1513, 1514, 1516
    cart-on-track conveyors, 1518
    chain conveyor, 1516, 1517
    chute conveyors, 1513
    power-and-free conveyor, 1518, 1519
    roller conveyor, 1514, 1516
    skate wheel conveyor, 1515–1517
    slat conveyor, 1515, 1517
    sortation conveyor, 1518–1520
    tow-line conveyor, 1517
    trolley conveyor, 1517–1518
  cost reduction in, 1355, 1356
  definition of, 1502
  equipment used for, 1504, 1505
  industrial trucks for, 1505–1513
    and closest-open-location (COL) rule, 1509–1510
    concurrent vs. sequential travel, 1510–1511
    counterbalanced lift truck, 1506, 1508
    first-come-first-served (FCFS) dispatching, 1511–1513
    number of trucks, determining, 1508–1509
    order picker truck, 1508
    pallet jack, 1505
    shortest-travel-time-first (STTF) rule, 1511, 1513
    sideloader truck, 1507, 1508
    straddle truck, 1506, 1508
    throughput capacity, 1509, 1510
    turret trucks, 1507
    walkie stacker, 1505, 1508
  and information flow, 1503
  in job shops, 1632
  models of, with tool perspective, 171
  scope of, 1502
  and storage systems, 1520–1523
    block stacking, 1520–1521
    cantilever rack, 1523
    pallet flow rack, 1522
    permanent racks, 1521, 1522
    portable racks, 1521
  and unit loads, 1503–1504
  and waste, 1502
Material removal, 456
Material requirements planning (MRP), 85, 348, 492, 2039–2042
  goal of, 2041
  inputs to, 2039–2040
  JIT vs., 545
  logic of, 2041–2042
Materials:
  acquisition of, 332–333
  approximate cost of, 450
  hazard characteristics of, 1160
  selection of, 449, 452, 453, 1316
Materials inventory, 332
Materials management, 968
Material safety data sheets (MSDSs), 1176
Materials specifications, 333
Mathematical programming models, 129
Mathematical programs:
  basic elements of, 2540–2541
  generic form of, 2541
Mathematics:
  of reliability, 1928–1930
  as type of language, 132
Matrix diagrams, 1816, 1819, 1820
Matrix organizational structure, 1265–1266
Maturity levels (of service organizations), 648
MAUT, *see* Multiattribute utility theory
MAW/F, *see* Maximum acceptable weights or forces
Maximax principle, 2379
Maximin principle, 2379
Maximum acceptable weights or forces (MAW/F), 1071–1072
Maximum daily demand constraint (scheduling), 1748
Maximum flow problem, 2572–2575
Maximum-likelihood estimators, 2254–2255
Max-min problems, 2527
MAXREV algorithm, 538
Maynard Operations Sequence Technique (MOST), 1439–1442
MBMSs, *see* Model base management systems
MBNQA, *see* Malcolm Baldrige National Quality Award
Mbps, 232
MBS (managed bandwidth services), 250
MBTI (Myers-Briggs Type Indicator), 937
MC^2-simplex method, 2618–2620
Mean of distribution of system repair time (MTTR), 1947–1951
Mean time between failure (MTBF), 1928, 1950–1951
Mean value, testing:
  with σ^2 known, 2244–2248
  with σ^2 unknown, 2248–2249
Mean value of the time to failure random variable (MTTF), 1928–1929
Mean-variance (MV) analysis, 752–756, 761–769
  combination of, with other techniques, 768–769
  extensions of, 767
  and forecasting, 761–763
  shortcomings of, 756, 757
  and taxation, 764–766
  and time horizon, 766–767
Measurement. *See also* Work measurement
  accuracy of, 1881
  in human-centered product planning and design, 1298, 1301
  in ISO 9001:2000 QMS standard, 1971
  of new product performance, 34
  of product performance, 49–50
  purpose of, 1877
  of reliability, 1941–1946
    estimation, 1944–1946
    test programs, 1942–1944
  of service quality, 640–641, 1963–1964
  variation, measurement, 1986–1987

Measurement (*Continued*)
  of waste-management programs effectiveness, 1570
  when using control charts, 1840
Measurement and test systems (M&TS) variation, 1984–1986
Measurement systems, 20–22, 1877–1885
  accuracy of, in steady state, 1882–1884
  in design and process platform characterization methodology, 1984–1987
  elements of, 1877–1878
  environmental effects in, 1879
  error of, 1883–1884
  hysteresis in, 1879–1880
  linearity/nonlinearity in, 1879–1880, 1883
  noise (interference) in, 1885
  range of values in, 1879
  repeatability in, 1880, 1881
  sensitivity of, 1879
  transfer function of, 1884–1885
Measuring devices:
  assembly of, 392–394
  robots, 376
Mechanical grippers, 414
Mechanical limit switches, 1903
Mechanistic job design, 870, 872, 874, 883–884, 886, 888
Mechanized activity sampling technique (MAST), 1458
Media, instructional use of, 928
Medicaid, 738
Medical instruments manufacturing case study, 2065–2067
Medical robots, 381, 382
Medicare, 738
Medium-complexity network applications, 244
Meetings:
  status, 1347
  three-level, 13
Megabits per second, 232
MEI, *see* Maintenance excellence index
Melt welding, 413
Memory, 930, 1015
Men, *see* Males
Mental assumptions, for effective performance management, 998–999
Mental models, 999–1000, 1210
Mentoring, 857–858, 860, 938
MESs, *see* Manufacturing execution systems
Message-integrity protocols, 733, 734
Metabolic energy requirement, 1118–1119
Meta class level of abstraction, 281, 283
Metadata, 84
Metaknowledge, 223, 1775
Metal, designing for, 1316–1323
  basic processes, 1317–1320
  liquid state, 1316–1317
  secondary processes, 1320–1323
  solid state, 1317–1319
Metallurgical industry, 518
Metaphors, in human–computer interface design, 1213–1214
Method of limits, 1048
Methodology, definition of, 1135
Methods, in object-oriented enterprise modeling, 291, 292
Methods engineering, 740, 1353–1389
  and engineering design steps, 1387–1389
  evolution of, 20
  information gathering/organizing for, 1371–1387
    and arrangement of equipment, 1379–1382
    and balancing flow lines, 1382–1385
    between-operations analysis, 1374–1385
    with flow diagrams, 1374–1376
    with multiactivity charts, 1376–1379
    and SEARCH method, 1373, 1374
    by videotaping, 1371–1373
    within-operation analysis, 1385–1387
  and job design, 1353–1371
    ergonomic criteria, 1354
    error reduction, 1368–1371
    fatigue reduction, 1365–1368
    musculoskeletal disorders, 1362–1365
    workstation organization, 1354–1362
Methods improvement (health care delivery systems), 739–741
Methods–Time Measurement (MTM) systems, 1429–1439
  MTM-1, 1120, 1429–1433
  MTM-2, 1429, 1433–1436
  MTM-3, 1435, 1436
  MTM-C system, 1436–1438
  MTM-M system, 1437, 1438
  MTM-V system, 1436
  specialized MTM systems, 1438–1439
Metropolitan area networks (MANs), 231, 255–256
Mexico, 957–958, 959
M/G/1 queue, 2159–2160
Microassembly, 395–397
Microchip technology, 398
MICRO Motion Analyses, 1441
Micro Saint, 2458–2459
Microsoft, 37, 72, 73, 79, 86, 268, 270, 499, 602, 721
Microsoft Excel, 2535
Microsoft Office for Windows, 2535
Microsoft Project 98, 1257
Microsoft Project 98 Plus, 1257
Microsoft Project 2000, 1261
Microsoft Project Central, 1261
Microsoldering, 431
Microsystem engineering, 365
MIDAS, *see* Man–Machine Integrated Design and Analysis System
Middleware, 251, 661
MIDs, *see* Molded interconnect devices
MIES software, 1774, 1775, *see* Manufacturing information and execution systems software
Military projects, work breakdown structure for, 1273–1275
"Milk runs," 1513
Milling, 1322
  geometric capabilities of, 464
  obtainable accuracy values, 565

  technological capabilities of, 468
MILNet, 238
Miniaturization, 423, 424
Minimax cost/regret, 2177, 2180–2181, 2378–2379, 2381
Minimin principle (decision theory), 2379
Minimum aspiration level, 2180
Minimum attractive rate of return (MARR), 2391, 2396
Minimum cost flow problem, 2569–2570, 2574
Mining industries:
  percent of GDP in, 346
  personnel scheduling for, 1757
“Mini worlds,” 280
MINOPT, 2564
MINOS, 2564
Minuteman Launch Control system, 1936
MIS, *see* Management information system
Mission:
  customer-focused, 654, 656
  as term, 1922
  of test and inspection systems, 1892–1893
Mission reliability, 1923
MIT, *see* Massachusetts Institute of Technology
MIT Commission on Industrial Productivity, 950
Mix-change flexibility, 499
Mixed unit load, 2087
Mix flexibility, 499
MJDQ, *see* Multimethod Job Design Questionnaire
MLR, *see* Multiple linear regression
M/M/1 queue, 2158
M/M/2 queue, 2158
MMH tasks, *see* Manual materials-handling tasks
M/M/∞ queue, 2158
MMS (Manufacturing Message Specification), 165
M/M/s/s, 2158–2159
Mobile Sources Program (emissions), 593
Mobility, range-of-joint, 1043, 1046
Mock-ups, 1303, 1305
Modal qualifiers (logical reasoning), 138
Model(s)/modeling, 1629–1631. *See also specific topics*
  business, *see* Business model
  computer, *see* Computer simulation
  data, 84
  definition of, 281, 2582
  design/process, 1987–1993
    experimental plan for, 1987–1990
    validation of, 1990–1993
  enterprise, *see* Enterprise models/modeling
  enterprise resource planning (ERP) software as tool for, 304
  in flexible manufacturing systems, 503–506
  human performance, *see* Human performance modeling
  of logistics systems, 2008–2011
    network design/configuration, 2008–2010
    supply chain planning, 2009, 2010
    transportation planning, 2011
  mental, 999
  methods of, 281
  queueing models, *see* Queueing models
  scheduling, notation used in, 1719–1722
  for simulation, 207
  tools for, 169–174, 1703
Model base management systems (MBMSs), 113, 115, 124–131
  basic functions of, 145
  database management systems vs., 125
  and DBMS design, 117
  decision processing MBMSs, 125–126
  for groups, 144
  and issue analysis, 127–129
  and issue formulation, 126
  and issue interpretation, 129
  and model base management, 129–131
  model processing MBMSs, 125–126
  objectives for, 125
“Model domain,” 280–281
Modeling languages, LP, 2535–2536
Model mapping, 522–523
Model specification, 2271–2272
Model validation, 2272
Modern Materials Handling, 1505
MODSIM III, 2459
Modular distribution, 1471
Modularity (product families), 688–689
Molded interconnect devices (MIDs), 432–439
  manufacture of, 433, 434
  mounting SMDs onto, 435–438
    optimized MID placement system, 436–438
    six-axis robot system, 435, 436
  soldering
    3D PCBs, 438–439
    conventional, 435
  structure of, 432–433
Molding, plastics, 1324–1327
Momentum, and work postures, 1360
Money, time value of, 2334
Monitored business process links, 2118
Monitoring. *See also* Control(s)
  computerized, 1221, 1225–1227
  project, 1346, 1347
  in retail supply chains, 776
Monitoring/control phase (professional services projects), 1346–1348
Monopolies, 681
Monotony allowances, 1395, 1397
Monte Carlo simulation, 1114, 1115, 2385–2391
  discrete distribution, sampling from, 2385–2386
  general procedure for, 2386–2390
  normal distribution, sampling from, 2386
  numerical example of, 2388–2391
  on project precedence diagram, 1258
  stochastic optimization, application to, 2631–2632, 2634
Montgomery Ward, 654
Morale, and spillover effect, 893
Morality, and leadership, 853
MOST, *see* Maynard Operations Sequence Technique

Most probable future principle (decision theory), 2378
Motion, in digital human modeling, 1116, 1120, 1125–1127
Motion Analysis Corp., 1125
Motion timing, 1120, 1429
Motion tracking:
  technologies for, 1125
  in virtual reality, 1124–1125
Motivation:
  calculative–rational vs. emotional–expressive basis of, 845–846
  and compensation systems, 861
  enhancing, 1961
  and error reduction, 1370
  extrinsic vs. intrinsic, 847–848
  individualistic vs. collectivistic, 846–847
  and leadership, 841, 853
  for localized innovation, 963
  and safety programs, 1182–1184
  in service-driven work systems, 1959
  and spillover effect, 893
  in transactional vs. transformational leadership paradigms, 845–848
  for workers' skill development, 547
Motivational job design, 872–875, 877, 884, 886
Motivational team design, 880
Motorola, 654, 1869
Motors, energy-improvement possibilities for, 1582
Mouse, computer, 1202
Movable magazines, 384
Movements, body, 1047
Movement kanbans, 549
Moving range (control charts), 1842, 1844
MPRSA (Marine Protection, Research, and Sanctuaries Act), 1164
MPS, *see* Master production schedule
MRC (Multiresolution CMAC), 1780
MRO, *see* Maintenance repair operations
MRP, *see* Material requirements planning
MRP-C, *see* Capacitated MRP
MRP II, *see* Manufacturing resource planning
MSDs, *see* Musculoskeletal disorders
MSDSs (material safety data sheets), 1176
MSG–3 (Maintenance Steering Group 3), 1908
MTBF, *see* Mean time between failure
MTM Association, 1120, 1429
MTM-MEK system, 1438
MTM systems, *see* Methods–Time Measurement systems
MTM-TE system, 1438
MTM-UAS system, 1439
M&TS variation, *see* Measurement and test systems variation
MTTF, *see* Mean value of the time to failure random variable
MTTR, *see* Mean of distribution of system repair time
Multiactivity charts, 1376–1379
Multi-agent systems (MASs), 174
Multiattribute utility theory (MAUT), 2177, 2183
Multicast addressing, 242
Multicriteria optimization, 2602–2621
  compromise solutions in, 2610–2614
  domination structures for use in, 2614–2617
    constant cone, 2615–2616
    variable cone, 2616–2617
  fuzzy, 2620
  goal programming solutions in, 2610–2614
  MC^2-simplex method for, 2618–2620
  preferences in, 2603–2605
  satisficing models for, 2608–2610
  value functions in, 2605–2608
Multidimensional search techniques, 2549–2552
Multilayer perceptron, 1778, 1779
Multimedia information:
  on networks, 234, 247
  trends in, 251
  on World Wide Web, 246
Multimethod Job Design Questionnaire (MJDQ), 872–873, 889
Multiple-attribute utility theory, 129
Multiple-class model (flexible machining systems), 1661–1662
Multiple correlation, 2278
Multiple cue probability learning models, 2200
Multiple linear regression (MLR), 2265–2292
  and appropriate use of statistics, 2267
  assumptions in, 2267–2268
  attribute modeling, 2279–2280
  covariates in, 2280
  dangers of, 2266, 2275
  definition of, 2265
  diagnostics in, 2282–2288
    example of, 2286–2288
    internal validation, 2284
    notation for, 2283
    partial plots, 2286
    questions, diagnostic, 2282
    residuals, 2284–2285
    row deletion, 2284
  example of, 2280–2282
  F ratio, 2278–2279
  goals of, 2266
  interactions vs. intercorrelation in, 2279
  intercorrelation effects in, 2275–2277
    ambiguity in assessment of contributions, 2276–2277
    detection of correlation, 2277
    estimates, intercorrelated, 2276
    variances, potentially enlarged, 2275–2276
  multiple correlation in, 2278
  partial correlation in, 2277–2278
  power of, 2266
  practical concerns with, 2291–2292
  ridge regression, 2290–2291
  t ratio, 2278–2279
  two variables, relating, 2268–2275
    coefficient estimation, 2272–2274
    correlation, 2271
    interval estimation for point on the line, 2274
    least-squares method for, 2268–2270
    model specification/validation, 2271–2272
    prediction of future value, 2274–2275

    and residual variance, 2270–2271
  variable selection in, 2289–2290
Multiple-objective approaches (scheduling), 1732
Multiple output control, 161
Multiple-shift scheduling, 1743–1744, 1755–1765
  crew scheduling, 1755–1757
  hierarchical workforce, 1744–1745, 1764–1765
Multiprocess holding, 547
Multiresolution CMAC (MRC), 1780
Multiskilled workers, 547
Multi-spindle automatic machinery, 467
Multi-stage models (production-inventory systems), 1683–1685
Musculoskeletal disorders (MSDs):
  and job design, 1362–1365
  job design/redesign for reduction of
    procedures, 1095
    risk factors, 1093–1094
    surveillance, 1095–1097
  and OSHA Proposed Ergonomic Program Standard, 1166–1167
  of the upper extremities, 1167
  and use of computer technologies, 1224
  work-related, *see* Work-related musculoskeletal disorders (WRMDs)
Mutual funds, 764, 765
MV analysis, *see* Mean-variance analysis
Myers-Briggs Type Indicator (MBTI), 937
MySAP.com, 95, 96

NAAQSs, *see* National Ambient Air Quality Standards
NAIC (North American Industry Classification), 329
Naming system (Internet), 242–243
Nanotechnology, 38
Narrow-range tasks (service systems), 1633
Nasdaq, 277
National Academy of Sciences, 1097
National Ambient Air Quality Standards (NAAQSs), 590, 592, 593, 595
National Center for Manufacturing Sciences, 954
National culture, 956–961
  compatibility of, with organizational culture, 957–958
  definition of, 957
  dimensions of, 957, 958, 960
National Electrical Manufacturers Association (NEMA), 908
National Emission Standards for Hazardous Air Pollutants (NESHAPS), 593
National Health and Nutrition Examination Survey III (NHANES), 1113, 1114
National Health Planning and Resources Development Act, 738
National Institute for Environmental Health Sciences (NIEHS), 1168
National Institute for Occupational Safety and Health (NIOSH), 981, 1082, 1119, 1163–1164, 1167–1170
National Institute of Standards and Technology, 326
National Occupational Research Agenda (NORA), 1168
National Pollution Discharge Elimination System (NPDES) permits, 595, 596
National Research Council, 950, 1097
National Software Testing Laboratories (NTSL), 1260
Natural gas systems, 1580
Naturalistic decision making, 2177, 2205–2209
  contingent decision making, 2207
  and dominance structuring, 2207
  explanation-based decision making, 2207–2208
  and image theory, 2207
  recognition-primed decision making, 2205
  and shared mental models, 2208
  and team leadership, 2208
Naturalist phase (human-centered product planning and design), 1299, 1301–1304
Navigator programs, 240
NDI (nondestructive inspection), 1909
Near-net-shape (nns) processes/production, 562–587
  benefits of, 564
  bulk metal forming techniques, 567–570
  casting, 566–568, 571–573
  cold-formed components, 575–580
    extrusion, 575–577
    orbital pressing, 579–580
    swaging, 577–579
  definition of, 564
  goals of, 564
  hot-formed components, 581–585
    axial die rolling, 584
    extrusion, 582–584
    precision forging, 581–583
  and nature of manufacturing, 563
  as philosophy, 565
  powder metallurgy, 566, 567, 572–576
    forging, powder, 574–576
    hot isostatic pressing (HIP), 572–574
  preconditions for, 564–565
  primary shaping, 566–567
  for prototyping, 586, 587
  semihot formed components, 580–582
    extrusion, 580, 581
    forging, 581
  special applications of, 568
  thixocasting, 568
  thixoforging, 568, 584–586
  trends in, 586, 587
Necessary conditions, 2546–2547
Neck:
  posture checklist for, 1366
  and work posture, 1358
Needs analysis (needs assessment), 926
Needs (of customers), 327
NEMA (National Electrical Manufacturers Association), 908
NESHAPS (National Emission Standards for Hazardous Air Pollutants), 593
Nested control loops, 161

Net present value (NPV), 98–99
Netscape, 244–245
Networks/networking, 228–257. *See also* Event trees; Internet
  of agents, 174
  applications, network-based, 243–244
  applications of, 250–251
  capability requirements of, 233
  for CIMS, 498–499
  classification of, 252–253
  collaboration, networked, 234
  company, types of, 314
  components of, 229
  content generation/provision via, 246–249, 251–252
  content on, 247–248
  cost of, 230
  and distance of users, 233
  and electronic communication, 232, 233
  extranets, 256
  history of, 235–236
  implementation issues, 252, 253
  importance of, 49
  and information access, 230–232
  and information provision, 232, 233
  infrastructure for, 236–237, 249–250
  intelligence of, 229, 230, 234
  intranets, 255–256
  LANs, 255, 256
  logistics models for, 2008–2010
  MANs, 255–256
  multimedia information on, 247
  in production and service industries, 253–254
  protocols for, 237
  rating/filtering of content of, 248–249
  reliability of, 229
  role of, 229–230
  services, networking, 243, 250
  speed of, 229
  supply chain, 2114, 2116–2120
    business process links, 2118–2120
    mapping, 2120
    and members of supply chain, 2117
    structural dimensions, 2117–2118
  and timing of information, 233
  top-down vs. bottom-up approaches to, 254
  transmission speeds in, 236–237
  trends in, 249–252
    applications, 250–251
    content generation/provision, 251–252
    infrastructure, network, 249–250
    services, 250
  and virtual environments, 234–235
  virtual private networks, 237
  WANs, 255–256
Network analysis (for site selection), 1470–1475
Network capability, corporate, 314–317
Network data model, 121
Network economy, 107, 262
Network flow models, 2568–2580
  applications of, 2572–2580
    equipment replacement, 2578–2579
    material-handling system, 2576
    personnel assignment, 2576–2578
    system reliability, 2579, 2580
  assignment problem, 2572
  computer software for, 2572
  features of, 2568–2569
  longest path problem, 2572
  maximum flow problem, 2572–2575
  minimum cost flow problem, 2569–2570, 2574
  shortest path problem, 2572, 2574
  transportation problem, 2570–2571, 2574
  traveling salesman problem, 2573
Networking technologies, 165–166
Network-management protocols, 730–732
Network organizations, 107, 260
Network planning (transportation), 803–812
  definition of problem, 804
  modeling for, 804–806
  network design formulation, 806–807
  package-routing problem, 807–810
  subgradient optimization algorithm, 811–812
  trailer-assignment problem, 810–811
Neural networks. *See also* Artificial neural networks
  for artificial intelligence approaches to control, 1777–1780
  for automated test and inspection, 1906
  in shop floor scheduling, 1777–1780
Neuron C programs, 167
Neurotoxic disorders, 1170
New Economy, 36, 37
  core business processes in, 43
  customers in, 38
  key trends in, 38
  and product/services/customer linkages, 50
  strategic management principles in, 42
New multiple-shift scheduling, 1743–1744
New organizational wealth, 148
New products:
  and competitive ability of company, 486
  lean product development, 556
  management of, with iCollaboration, 968
  measuring performance of, 34
  pricing of, 674, 675
New Source Performance Standards (NSPSs), 590–593
Newspapers, as human-centered product planning/design tool, 1302, 1303
Newsvendor model, 1670
Newsvendor problem, 2626–2627
Newton's method, 2530–2531, 2550–2551
New York Stock Exchange (NYSE), 277
The New York Times, 266
NGT, *see* Nominal group technique
NHANES, *see* National Health and Nutrition Examination Survey III
NIEHS (National Institute for Environmental Health Sciences), 1168
NIOSH, *see* National Institute of Occupational Safety and Health
NIOSH lifting guide, 1076–1080, 1121
Nissan, 212

NLPQL, 2564
NLPQLB, 2564
Nns processes/production, *see* Near-net-shape processes/production
Nodes, 2591
Noise, 1170
  allowances for, 1399
  at computer workstations, 1205
  and human–computer interaction, 1200
  levels of, 1133, 1177
  in measurement systems, 1885
Nominal group technique (NGT), 127, 2213
Nominal interest rates, 2337
Nomogram method for determining sample size, 1453, 1454
Nonbinding constraints, 2541
Nonchipping shaping, *see* Near-net-shape processes/production
Nonconformity, control of, 1971
Nondestructive inspection (NDI), 1909
Nonengineered time estimates, 1392–1393
Nonfinancial outcomes, 1002
Nonlinearities:
  linear programming for handling, 2526–2527
    absolute value functions, 2527
    max-min problems, 2527
    piecewise linear functions, 2526–2527
  in measurement systems, 1879–1880, 1883
Nonlinear programs, 2541
Nonlinear programming, 2540–2565
  computer software for, 2563–2565
  constrained optimization, 2553–2562
    feasible directions, methods of, 2559–2560
    geometric programming problems, 2558–2559
    Karush–Kuhn–Tucker conditions for, 2554–2555
    Lagrange multipliers for, 2553–2554
    and nonsmooth optimization, 2562
    quadratic programming problems, 2555, 2562
    separable programming problems, 2556–2558
    sequential unconstrained minimization techniques for, 2560–2562
    successive linear programming, 2562
    successive quadratic programming, 2562
  convexity in, 2543–2546
  Hessian matrix, 2546
  online resources on, 2563
  solutions in, 2541–2542
  unconstrained optimization, 2546–2553
    classical methods, 2546–2547
    conjugate gradient methods, 2552–2553
    golden section method for, 2547–2549
    line search techniques for, 2547
    multidimensional search techniques for, 2549–2552
Nonmember business process links, 2118, 2119
Nonparametric tests, 2256
Nonproduction test and inspection, 1907–1908
Non-progressive assembly lines, 1355
Nonsmooth optimization, 2562
Nonuniform service lives, 2350
NORA (National Occupational Research Agenda), 1168
Nordstrom's, 656
Normal distribution:
  of reliability, 1931, 1932
  standard, 2386
Normal time, 1394
Normative decision theory, *see* Decision analysis
North American Industry Classification System (NAIC), 329
North Carolina Ergonomic Standard, 1166–1167
Norton, 7
Norway:
  quality standards in, 1968
  social democracy in, 1186
Not-managed business process links, 2118
Nottingham University, 1112
NP charts, 1844
NPDES permits, *see* National Pollution Discharge Elimination System permits
NP-hard problems, 1722
NPSOL, 2564
NPV (net present value), 98–99
NSPSs, *see* New Source Performance Standards
NTSL (National Software Testing Laboratories), 1260
Nuclear power industry, 959–960
Null hypothesis, 2245
Number of defective units, control chart for, 1874–1875
NUMBUS, 2564
Numerical representation, 2582
NYSE (New York Stock Exchange), 277

OB1 software, 2535
Object classes, 291–293
Objective forecasting models, 793
Objective functions (mathematical programs), 2540, 2541
Objective rating method (time study), 1424, 1425
Objectives:
  of business processes, 43, 44
  and job evaluations, 911–912
  for professional services projects, 1336
  profit, and pricing, 682
Objectives document, 1306–1307
Object Management Architecture (OMA), 1773
Object Management Group (OMG), 714, 720, 721, 1772–1774, 1782
Object modeling technique (OMT), 1774
Object orientation, 166
Object-oriented database management systems design (OODBMS), 122–124
Object-oriented database models, 122–124
Object-oriented databases (OODBs), 82–83
Object-oriented data models (OODMs), 82, 121
Object-oriented enterprise modeling, 291–293
Object-oriented modeling methods, 507
Object-oriented programming (OOP), 71, 1328

Object request broker (ORB), 720, 721, 732
Objects:
  in human–computer interface design, 1213
  in object-oriented enterprise modeling, 291, 292
OBM (operator-based maintenance), 1620
OBS, *see* Organizational breakdown structure
Observation(s):
  as task-analysis technique, 1209
  in work sampling:
    frequency of, 1453–1456
    methods for conducting, 1456–1457
    number needed, 1451–1454
Observed time, 1394
Occupational biomechanics, 1068–1070
Occupational disease, 1082–1084
Occupational risk factors, 1086
Occupational safety and health, 1157–1188
  agencies/organizations involved with, 1162–1165
  balance model of, 1159–1162
  and definition of occupational injuries/diseases, 1168–1170
  and employees:
    on employee/management ergonomics committee, 1187
    hazard information for, 1176–1177
    involvement of, 1186–1187
    role of, 1159–1160
  engineering controls for, 1175–1176
  hazards, workplace, 1168, 1171–1187
    engineering controls, 1175–1176
    human factors controls, 1176–1179
    and illness/injury statistics, 1173–1174
    improved work practices for reducing, 1181–1183
    and incident reporting, 1174
    informing employees about, 1176–1177
    inspection programs, 1171–1173
    measurement of potential for, 1171–1174
    new technologies, hazard control for, 1184–1187
    and safety programming, 1183–1184
    safety training for reducing, 1180–1181
    workplace/job design for reducing, 1177–1179
  hazard survey for, 1186–1187
  human factors controls for, 1176–1179
  and illness/injury statistics, 1173–1174
  improved work practices for, 1181–1183
  and incident reporting, 1174
  inspection programs, 1171–1173
  interdisciplinary nature of, 1157
  and job design, 883
  and measurement of hazard potential, 1171–1174
  new technologies, hazard control for, 1184–1187
  and new technology, 1160
  and organizational design, 1179–1180
  and organizational structure, 1161–1162
  public health approach to, 1157–1159
  and quality improvement, 1184–1185
  and safety programming, 1183–1184
  safety programs for, 1183–1184
  and safety training, 1180–1181
  Scandinavian approach to, 1186
  standards for, 1165–1168, 1185–1186
  task factors in, 1160–1161
  and technology/materials, 1160
  and work environment, 1161
  and workplace/job design, 1177–1179
Occupational Safety and Health Act (OSHA), 593, 1162, 1173–1174
Occupational Safety and Health Administration (OSHA), 980, 1097–1100, 1162–1163, 1165–1166, 1176, 1592
Occupational Safety and Health Review Commission, 1162
Occupational strength testing, 1052
OCE, *see* Overall cost effectiveness
OD, *see* Organizational development
OEMs (original equipment manufacturers), 329
Off-highway vehicles, 1470
Office of Information Technology and Applications (NIST), 326
Off-site recycling, 533
Oil analysis, 1613–1614
OJT, *see* On-the-job training
OLAP, *see* Online analytical processing
Oligarchy, perceived, 958
OLTP (online transactional processing), 84
OMA (Object Management Architecture), 1773
OMG, *see* Object Management Group
OMT (object modeling technique), 1774
One-dimensional arrays, 1904
One-factor experiments, 2260
One-handed force magnitudes, 1056
100% rule:
  for objective function coefficients, 2537
  for RHS constants, 2538
One-piece flow production, 547
One-tailed hypothesis tests, 2247
One-to-one marketing, 662, 704–705
Online analytical processing (OLAP), 84, 2013
Online auctions, *see* Auctions, online
Online retailing, 265–267. *See also* Electronic commerce
  digital products, 266, 267, 270–271
  physical products, 266
  services, 266, 267
  storefronts, Web, 265–266
Online transactional processing (OLTP), 84
On/off control models, 160
Onsale.com, 273, 275
On-site recycling, 533
On-the-job training (OJT), 1180, 1181, 1556
OODBMS, *see* Object-oriented database management systems design
OODBs, *see* Object-oriented databases
OODMs, *see* Object-oriented data models
OOP, *see* Object-oriented programming
Opalescent globe lighting, 1198
Open system interconnection (OSI):
  layers of, 165
  management framework, OSI, 729–730
Open systems architecture:
  and client/server (C/S) systems, 714

  and enterprise resource planning (ERP), 88–89
  ERP and, 88–89
Open View (Hewlett-Packard), 732
Operating characteristics, 157
Operating effectiveness and efficiency, 11
Operating permit program, Clean Air Act, 593
Operation(s):
  as dimension of competitive advantage, 327
  in ISO 9001:2000 product realization clause, 1971
  major activities of, 1770
  and plant engineering, 1550–1551, 1565–1569
    automation systems, 1566
    buildings and grounds department, 1566–1567
    design/construction, 1565–1566
    loss control, 1568
    maintenance, 1566
    safety management, 1567–1568
    security function, 1568–1569
  in transformable structures, 317–322
Operational change management, 968
Operational control decisions, 111, 112, 135–137
Operational feasibility (IS systems), 98
Operational information systems, 83
Operational performance decisions, 111, 112, 135–137
Operational planning (warehouse operations), 2088
Operational readiness, 1924
Operational status, 334
Operations analysis, 740
Operations audits, 1544–1547
Operations estimating, 2311–2314
Operations improvement, 18–22
  business processes, 18–20
  ISE's role in, 6
  measurement systems, 20–22
  performance measurement, 21–22
Operations planning, 327–329, 2125
Operations sheet, 2312, 2313
Operations work, 1253
Operator(s):
  choice of, for time study, 1417–1418
  rating performance of, in time study, 1422–1425
Operator-based maintenance (OBM), 1620
Opportunistic thinking, 1024
Opportunities, discovery of, 1705, 1706
Opposing environmental input, 1883, 1884
Optical inspection systems, 431, 432
Optical motion-tracking devices, 1125
Optical transmitter design model, 1979–1980
Optics, 365
OPTIMA Library, 2564–2565
Optimality, principle of, 1726–1727
Optimal solution, 2528, 2536
Optimal value (of linear program), 2528
Optimization:
  direct vs. indirect methods of, 2541
  discrete, *see* Discrete optimization
  linear, *see* Linear programming (LP)
  multicriteria, *see* Multicriteria optimization
  nonlinear, *see* Nonlinear programming
  nonsmooth, 2562
  under uncertainty, 2625–2628
Optimization models, 1630
  for facility locations, 2067
  in health care delivery systems scheduling, 746
  for production planning, 2043–2044
Optional collaboration, 605
Option-to-order production, 330, 331
  ERP configurator in, 338
  and warehousing, 335
OptQuest, 2447, 2459, 2461
Optum Inc., 2058
Oracle, 87, 90, 95, 304, 306, 1002
ORB, *see* Object request broker
Orbital pressing, 570, 579–580
Order(s):
  definition of, 2087
  management of, with iCollaboration, 968
  master table for, 2097
  processing of, 2046, 2093–2095
  retrieval of, 2093–2095
  tracking, 333
Order class, 2087
Order consolidation, 2106–2107
Order cycle, 2131
Order entry:
  enterprise resource planning (ERP) function, 333
  and transportation management software, 2065
Order flows (warehouse operations), 2084–2085
Order fulfillment, 2121–2122
Order picker trucks, 1508
Order picking, 2104–2106
Order processing, 2104
Organization(s):
  combining formal/informal, 1005–1006
  definition of, 284
  functional/project/matrix structures of, 1265–1266
  hierarchical, 284–285
  high-performance, 1000–1001
  impermanent, 1001
  influences of, on job design/redesign, 869
  of plant engineering, 1557–1560
  process, 284–285
  structural, 284
  traditional vs. customer-driven, 1797
Organizational ambiguity, 140
Organizational analysis, 925–926
Organizational breakdown structure (OBS), 1247
  and cost accounts, 1273
  use of, 1266
  and WBS, 1267, 1269
Organizational change, technology and. *see under* Technology
Organizational culture:
  and compatibility with national culture, 957–958

Organizational culture (*Continued*)
  definition of, 956
  dimensions of, 958
  and new technology, 956–961
  process-control-minded, 2002, 2003
  propensity towards change in, 1705
  and supply chain management, 2126, 2127
  for sustaining knowledge management, 216–217
Organizational decision support systems, *see* Group decision support systems
Organizational design, 1179–1180
Organizational development (OD), 938, 939
Organizational effectiveness, 976
Organizational flexibility, 262
Organizational informatics, 146
Organizational intelligence, 146
Organizational issues, in manufacturing design, 1329–1330
Organizational leadership, Baldrige criteria for, 1958
Organizational learning:
  time for, 1400–1406
  and work breakdown structure (WBS), 1276–1277
Organizational performance:
  indicators of, 21
  keys to, in current economy, 147
  measurement of, 21–22
Organizational plan, 1247
Organizational processes, 284–285
Organizational structure(s):
  and occupational safety and health, 1161–1162
  in supply chain management, 2125
  and work breakdown structure (WBS), 1264–1268
Organization charts, 1559
Organization for Industrial Research, 949
Organization view(s), 286–287, 510
Original equipment manufacturers (OEMs), 329
Origin 2000 (Silicon Graphics), 606
OR/MS Today, 2535
Orthogonal arrays, 2232
OS–3 Plus Event Recorder stopwatch, 1412, 1414
OSHA, *see* Occupational Safety and Health Act; Occupational Safety and Health Administration
OshKosh B'Gosh, 846
OSI, *see* Open system interconnection
OSL, 2535, 2575
Outcomes:
  human factors audits, 1145
  nonfinancial, 1002
  and team effectiveness, 987
Out-of-control project conditions, 1347–1348
Output(s):
  from business processes, 45
  definitions of, 287
  GDP as metric for, 344
  in process-oriented enterprise modeling, 287–289
Outsourcing, 263–264, 2051
  coordination mechanisms for, 2134
  and project management, 1249–1250
    in prefabrication field, 404
    in retail supply chains, 778
  in supply chain design, 2115
Ovako Working Posture Analyzing System (OWAS), 1061–1062
Overall cost effectiveness (OCE), 1616–1617
Overhead, 1344, 2300
Overload protector, assembly of, 392
Overpack, 2087
Overtime work, 1178–1179
Ovum model (knowledge management), 222
OWAS, *see* Ovako Working Posture Analyzing System
OWASCA, 1061–1062
Owens-Corning, 953
Owners, as external force, 39
Oxford Health Plans, 950

Pace, work, 1223
Paced systems, models of, 1638–1639
Pacific Area Standards Congress (PASC), 1968
Pacific Environmental Services, Inc., 596
Pacific Gas & Electric, 965
Package software, 68
Packaging industry, scheduling in, 1733
Packet-based transmission, 231
Packet filter, 735
Packet switching, 239
PADER scoring system, 1800–1801
Pallet flow rack, 1522
Pallet jacks, 1505
Pallet magazines, 383, 384
Pallets, 2087
Paper industry, 518
PAQ, *see* Position Analysis Questionnaire
Paragon (Intel), 606
Parallel assembly cells, 409
Parallel machines, 1721
Parallel pricing, 681
Parallel robots, 374, 375
Parameter designs, 2237–2238
Parametrical modeling, 178, 183–185
Pareto charts, 1371–1372, 1815, 1818, 1821–1823, 1832, 1833, 1859
Partial correlation, 2277–2278
Partial plots, 2286
Participative bureaucracy, 952
Participative design, 964
Participative planning, 320, 321
Participatory ergonomics (PE), 980–981, 1184–1185
Participatory management, 976, 983
Part manufacturing, 562–563
Partnerships:
  drivers of, 2135, 2136, 2137
  facilitators of, 2136–2138
  as TQL success factor, 1805
Parts handlers, 373
Part-time workers, 1363, 1744–1745
PASC (Pacific Area Standards Congress), 1968
Passenger vehicles, 2063
Passive redundancy, 1933

PASTA, 2163
Payback period method (cost estimating), 2349
Pay rates, *see* Compensation
PCA, *see* Process capability analysis
PCBs, *see* Printed circuit boards
P charts, 1844–1847, 1872–1874
PCS (Permit Compliance System) database, 596
PDCA cycle, 980–981
PDM, *see* Product data management
PdM, *see* Predictive maintenance
PD (proportional-derivative) control models, 160
PDR, *see* Process design and reengineering
PDSA cycle, 10, 11, 12, 22, 1808–1809
PE, *see* Participatory ergonomics
Peapod groceries, 783
PE control models, *see* Proportional control models
Penetration pricing, 675
PeopleSoft, 87, 89, 95, 1002, 1738
Pepsi, 2135
PERA, *see* Purdue Enterprise Reference Architecture
Perceived change, 958
Perceived oligarchy, 958
Perceived quality, 1247
Perceived tradition, 958
Percent deviation graph, 2363–2364
Percentile method (of estimating), 2305
Percent nonconforming, control charts for, 1872–1874
Perception, 669–671, 1015
Perceptual/motor job design, 873, 875–876, 884, 888
Perceptual/motor skills, 1160
Performance:
  analysis of, 1770
  Baldrige criteria for excellence in, 1957
  as dimension of quality, 1246
  drivers of, 55
  electronic monitoring of, 1225–1227
  feedback to influence, 933–934
  human performance modeling, *see* Human performance modeling
  individual, 937–938
  of inspection systems, 1890
  job design and problems with, 883
  knowledge-based, 2206
  levels of, for decision making, 2205–2206
  and national culture, 957
  organizational, *see* Organizational performance
  and organizational culture, 958
  as outcome of leadership, 851–852
  predictors of, 924
  of a process, analyzing, 1828, 1830–1831
  rule-based, 2205, 2206
  setting goals for, 1708
  skill-based, 2205, 2206
  standard performance, 1422
Performance appraisal, 858
Performance assessment, 1348
Performance evaluation, 728–729
Performance Excellence Framework, 8, 9
Performance improvement:
  and JIT, 545
  total quality leadership (TQL), *see* Total quality leadership
Performance improvement objectives (PIOs), 13
Performance logic, 925–926
Performance management, 995–1010
  and achievement of alignment, 1005–1010
    integrated system, creation of, 1009–1010
    leadership of both formal and informal organization, 1008–1009
    working arenas, identification/alignment of, 1006–1007
  balanced scorecard for, 997–998
  in business model, 48–49
  and change, 996–997
  choice of metrics for, 1003–1005
    cost, 1005
    positive yields, 1005
    quality, on-spec/expected, 1005
    SMART performance goals, 1005
    speed-time, 1004–1005
  and continuous improvement, 1000
  in high-performance organizations, 1000–1001
  in impermanent organization, 1001
  and innovation, 1000
  and leadership, 858–859
  new mental assumptions required for effective, 998–999
  new mental models required for effective, 999–1000
  obstacles to, 1001–1003
    anxieties, 1002
    complexity of megaproject/megaprogram, 1002–1003
    financial management, legacy of, 1002
    flawed assumptions, 1002
    nonfinancial outcomes, expression of, 1002
  and quality, 1000
  time as metric for, 1004–1005
Performance measurement, 20–22
  in alliance management, 49
  in business model, 54–56
  financial vs. nonfinancial, 55, 56
  process, basis for, 30
  product, 49–50
  supply chain, 2131–2132
  for total quality leadership (TQL) process, 1803
Performance measures:
  for design and process platform characterization methodology, 2002–2003
  determined by queueing models, 1631–1632
  of quality, 1246–1247
Performance models, 1126
Performance objectives/criteria, 726–727
Performance ratings:
  for time studies, 1422–1425
  videotape cameras for, 1414
Performance reviews, 1349
Periodic inflation, 2395
Perishability, product, 2130

Permanent racks, 1521, 1522
Permit Compliance System (PCS) database, 596
Perpetuities, 2350–2351
Personality, and workplace accidents, 1160
Personalization of products, 267, 268
Personalized Web Assistant, 269
Personal mastery, 999
Personal quality, 1796
Personnel. *See also* Employees
  assignment of:
    as network flow problem, 2576–2578
    for professional services projects, 1339–1341
  required, 1560–1561
  selection of, 921–924
  warehouse, 2100
Personnel costs, determination of, 1343–1344
Personnel management, *see* Human resource management
Personnel scheduling, 1741–1766
  computerization of, 1765
  crew scheduling, 1743–1744, 1755–1757
  in health care delivery systems, 744
  in health care delivery systems scheduling, 744
  manual shift scheduling systems, 1765–1766
  multiple-shift scheduling, 1743–1744, 1755–1765
    crew scheduling, 1755–1757
    hierarchical workforce, 1744–1745, 1764–1765
    for individuals, 1744, 1757–1764
    new scheduling, 1743–1744
  of part-time workers, 1744–1745
  single-shift scheduling, 1743, 1746–1755
  steps in, 1741–1743
PERT, *see* Project Evaluation and Review Technique
PERT diagrams, *see* Program Evaluation and Review Technique diagrams
Perturbation analysis, 2632–2633
Petri nets, 166, 173, 503–505
Petroleum industry, 518
Petsmart, 781
Pfizer, 911
Pharmaceutical industry:
  assembly in, 398
  as process industry, 518
PHC (productive hour cost), 2314
Photoelectric sensors, 1902
Physical anthropometry, 1043
Physical automation technology, 156
Physical data independence, 116
Physical markets, 262
Physical models, 1630
Physical products, online retailing of, 266
Physical prototyping, 1288
Physical tasks, 1042–1100. *See also* Ergonomics; Work-related musculoskeletal disorders (WRMDs)
  and anthropometry, 1043–1050
    alternative design, 1049
    body position, description of, 1043
    computer-aided models of man, 1050
    design criteria, 1048, 1049
    method of limits, 1048
    physical vs. functional anthropometry, 1043
    range-of-joint mobility, 1043, 1046
    statistical descriptions, 1043, 1048
  human strength, design for, 1050–1058
    computer-simulation, use of, 1054
    joint strengths, maximum voluntary, 1052
    occupational strength testing, 1052
    and push–pull force limits, 1054–1058
    and static vs. dynamic strengths, 1052, 1053
  manual materials-handling (MMH) tasks, 1070–1082
    biomechanical approach to design of, 1072, 1076
    and job severity index (JSI), 1080, 1081
    and low-back disorders, 1070–1071, 1080–1082
    physiological approach to design of, 1072
    psychophysical approach to design of, 1071–1072
    and revised NIOSH lifting equation, 1076–1080
  occupational biomechanics for analysis of, 1068–1070
  static efforts/work, 1052, 1053, 1056–1061
    arm, static efforts of, 1058–1062
    design limits for, 1056, 1057
    intermittent, 1057, 1058
    push–pull force limits, 1055
  and workplace analysis/design, 1061–1068
    international standards, 1065–1068
    postural analysis systems, use of, 1061–1063
    tolerability of working postures, 1063–1064, 1066–1067
Pick-and-place principle, 425, 1525
Pickers, multistop routing of, 2105–2106
Pickett, 653, 654
Picking, batch, 2093–2095
Pick-to-light systems, 2107
Pickup and delivery operations (transportation), 793–803, 2058, 2059
  heuristic construction algorithms for modeling, 795–801
  preassigned routes/territories, 801–803
  VRPTW modeling of, 794–795
Pick wave, 2087, 2095
PID (proportional-integral-derivative) control models, 160
Piecewise linear functions, 2526–2527
Pinch–pull force magnitudes, 1056
PIOs (performance improvement objectives), 13
Pipe design, 187, 189
PI (proportional-integral) control models, 160
Placement systems:
  electronic components, 425–429
  3D PCBs, 435–438
Planes of reference (body position), 1043
Planing, 1322

  - geometric capabilities of, 464
  - technological capabilities of, 470

Planned experimentation, 1820, 1822

Planning, 2034–2035
  - advanced planning and scheduling, *see* Advanced planning and scheduling (APS)
  - algorithms for, 2038–2045
    - finite capacity algorithms, 2042–2045
    - material requirements planning (MRP), 2039–2042
  - cognitive probes for, 1026
  - collaborative customer/demand, 968
  - communications, 1248
  - of control charts, 1839–1841
  - decentralized, in rapid product development, 1288
  - for demand, *see* Demand
  - for experiments, 2226–2227
  - in Flexible Manufacturing Systems, 501–503
  - maintenance, 1592–1593
  - organizational, 1247
  - quality, 1247
  - scheduling vs., 2036–2038
  - shipment, 2063–2067
  - for transformable structures, 317–322
  - transportation, 792–793
  - for warehouse operations, 2088–2095
    - contingency planning, 1530
    - equipment planning, 1541–1544
    - forward-reserve allocation, 2093
    - layout planning, 1538–1541
    - pick wave planning, 2095
    - space planning, 1532–1538
    - strategic master planning, 1530–1532
  - for work sampling, 1451–1457
    - collection methods, determining, 1456–1457
    - frequency of observations, determining, 1454–1456
    - necessary observations, determining, 1451–1454

Planning cycle, 954

Planning and design stage (project life cycle), 1242

Planning system, 11–15
  - change leadership, 14–15
  - ISE's role in, 6
  - policy deployment, 13
  - relationship management, 13–14

Plant, lean, 555

Plant engineering, 1550–1582
  - applying IE techniques to, 1560–1562
  - definition of, 1550
  - and energy management, 1572–1582
    - assessment, energy, 1578–1579
    - demand and power factor charges, 1757
    - environmental issues, 1577
    - financial issues, 1576–1577
    - process, energy, 1574
    - productivity, energy, 1573
    - programs, energy-management, 1578
    - strategies and tactics, 1577–1582
    - system, energy, 1574–1575
  - and enterprise asset management, 1550
  - facilities engineering vs., 1550
  - and facility surveys, 1564–1565
  - financial aspects of, 1562–1564
  - industry characteristics affecting, 1552–1553
  - integration of industrial engineers into, 1553–1557
  - and maintenance department, 1550
  - operational issues with, 1565–1569
    - automation systems, 1566
    - buildings and grounds department, 1566–1567
    - design/construction, 1565–1566
    - loss control, 1568
    - maintenance, 1566
    - safety management, 1567–1568
    - security function, 1568–1569
  - organization/management of, 1557–1560
  - and production/operations, 1550–1551
  - and product/process design and planning, 1551
  - and roles of plant/facilities engineers, 1551–1552
  - strategy for, 1557
  - technological concepts for, 1572
  - and upper management, 1551
  - and waste management, 1569–1572
    - methodology for, 1571–1572
    - solid wastes, 1571
    - streams, waste, 1570–1571
  - work measurement in, 1562

Plant engineers:
  - concurrent engineering involvement of, 1551
  - management functions of, 1552

Plant and facilities engineering, 1586–1588. *See also* Maintenance

Plastics processing, 365

Platform for Privacy Preferences (P3P), 269

Platforms (product), 49

PLATO-Z, 1777

"Playing field" of enterprise, 35–36

PMBOK, see Project Management Body of Knowledge

PM injection molding, 565

PMI (Project Management Institute), 1242

PMISs (predictive management information systems), 112

PMnet, 1260

PMP (Project Management Professional) certificate, 1242

Pneumatic grippers, 414

P-NUT (software package), 173–174

Point estimators, 2475–2477
  - and mean estimation, 2475–2476
  - and probability estimation, 2476, 2477
  - and quantile estimation, 2476, 2477
  - standard error of, 2483–2485

Point method (job evaluation), 907–910

Point processes, 2149–2150

Poisson process, 2149

Poka-yoke, 548, 559

Polhemus FastTrak, 1125

Police, personnel scheduling for, 1744
Policies and procedures:
  customer orientation in, 656–657
  dynamic decision problems, 2639–2640
  and health/safety performance, 1161, 1179
Policy Capture Theory, 129
Policy-capturing models, 2195, 2200
Policy deployment, 13
Political changes, 38
Political trends, 38
Pollution:
  from energy production/waste heat recovery, 1577
  prevention of, 533
Pollution Prevention Act (PPA), 1164
Polymorphism (OOP), 1328
Polynomial time algorithms, 1722
Pools, cost, 2319
Popularity philosophy (warehouse layout), 1540
Portable racks, 1521
Portals, 271, 272
Portfolio-based transactions, 277
Ports (computer), 240
Position Analysis Questionnaire (PAQ), 1137–1139
Positioning, 11, 13
Positive yields, 1005
Post, 653, 654
Postponement (in channel structure theory), 2115–2116
Postprocessing finite element methods (FEM), 201–203
Postural information, 1121
Posture(s):
  postural analysis systems, 1061–1063
  standard, 1062, 1063
  working, 1061, 1063–1064, 1066–1067
Posture allowances, 1396
Posturing, figure, 1115–1116, 1122, 1123
Potential, full, *see* Full potential
Potential market, 40
Powdered metallurgy, 1319
Powder forging, 565
Powder metallurgy, 566, 567, 572–576, 1321
  forging, powder, 574–576
  hot isostatic pressing (HIP), 572–574
Power-and-free conveyors, 1518, 1519
Power distance (in national cultures), 958, 960, 975
Power factors (electricity), 1575–1576
Power law technique, 2303, 2304
Power structure (supply chain management), 2126
PPA (Pollution Prevention Act), 1164
PPOs (preferred provider organizations), 738
Practice-based team training, 934
Precision forging, 581–583, 1317
Precision mechanics, 365
Predatory pricing, 681
Predetermined time standards (PTS), 1413, 1427–1446
  definition/uses of, 1412
  MACRO Motion Analyses, 1441–1446
  Maynard Operations Sequence Technique (MOST), 1439–1442
  Methods–Time Measurement (MTM) systems, 1429–1439
    MTM–1 data, 1429–1433
    MTM–2 data, 1429, 1433–1436
    MTM–3 data, 1435, 1436
    MTM-C system, 1436–1438
    MTM-M system, 1437, 1438
    MTM-V system, 1436
    specialized MTM systems, 1438–1439
  scope of application of, 1428–1429
Predictions, event, 137
Prediction validity, 1134
Predictive maintenance (PdM), 1606–1608, 1612–1615
  advantages of, 1612
  and ferrographic oil analysis, 1614
  and infrared thermography, 1614
  and shock pulse, 1613
  and spectrometric oil analysis, 1613
  and standard oil analysis, 1614
  and ultrasonic detection, 1614–1615
  and vibration analysis, 1613
Predictive management information systems (PMISs), 112
Predictors, 921–924, 2265
  aptitude and ability tests, 921–922
  biodata, 922–923
  drug testing, 923
  references, 923, 924
Preference(s), 2603–2605
  assessment of, in decision analysis, 2194–2195
  in behavioral decision theory, 2201–2205
    and framing of decisions, 2202–2203
    labile preferences, 2204–2205
    and prospect theory, 2203–2204
    and subjective expected utility, 2202
  labile preferences, 2204–2205
  and presentation of decision, 2202–2203
  reversal of, 2202–2203
Preferred provider organizations (PPOs), 738
Preplanning phase (process design and reengineering), 1704, 1705
Preprocessing finite element methods (FEM), 200, 201
Prescriptive approaches for group decision making, 2212–2214
Presentation language, 131–132
Presentation support software, 142
Present worth, probability distribution for, 2367–2369, 2371–2376
  discrete distribution, 2372–2373
  expected present worth, 2367–2368
  mean and variance, using only, 2373–2374
  normal distribution, assumption of, 2374–2376
  variance of present worth, 2368–2369
Present worth factor (interest), 2338–2339
Present worth method (cost estimating), 2346–2347
PRESS, 2284
Press-fitting, 372
Press forming, 1320, 1321
Pressure transducers, 1903
Pressure welding, 413

Prevention-based management, 1805
Prevention of Significant Deterioration (PSD) standards, 592
Preventive maintenance, 1606–1608, 1611–1612
Preventive scheduling, 1734
Price(s):
  definition of, 666–667
  inflation as metric for, 344
  in market model, 698–699
  reference, 671
  relative, 666
  selling, 2298
Price discrimination, 681–682
Price elasticity of demand, 668
Price fixing, 680–681
Price index, 2395
Price information, exchanging, 680
Priceline.com, 275–276
Price promotions, 678–680
Price signaling, 680, 681
Pricing, 666–683
  bundling, price, 676
  and buyer behavior, 668
  buyers' response to, 669–671
  and costs, 672–674
  during declining phase, 675–676
  and demand, 668–672
  demonstration of, 677–678
  and e-commerce, 671–672
  economic concepts related to, 668–669
  with and electronic commerce, 267, 269–271
    of digital product, 270–271
    real-time, 269
  factors involved in, 667
  during growth stage, 675
  guidelines for, 682–683
  in Internet economy, 267, 269–271
    digital products, 270–271
    real-time, 269
  legal issues in, 680–682
  during maturity stage, 675
  of new products, 674, 675
  objectives of, 667–668
    competitive strategies, 668
    profit, 667
    sales volume, 667
  parallel, 681
  predatory, 681
  proactive, 667
  and sales promotions, 678–680
  segmentation, 676–677, 678
  and yield management, 676–677
PRIDE, 1589
Primal feasibility, 2554
Primary hypothesis, 2245
Primary shaping (near-net-shape processing), 566–567
Principle of optimality, 1726–1727
Principles, business, 32
Printed circuit boards (PCBs):
  placement of, 426–428
  process steps in production of, 423
  3D:
    placement systems for, 423
    soldering, 438–439
PRISM (Production Robotics and Integration Software for Manufacturing Group), 607
Prisons, personnel scheduling for, 1744, 1760
Privacy issues:
  and anonymity, 267–268
  current measures to protect, 269
  with networks, 232
  tools for, 268–269
  visual/acoustical requirements, 1205
Private equity, 757, 759–761
Private networks, 237, 238, 243–244
Privatization, 38
Proactive pricing, 667
Probability, 2146–2149. *See also* Hypothesis testing; Monte Carlo simulation
  assessment of, 138–139
  assessment of, in decision analysis, 2191–2193
  estimation of, 2476, 2477
  of loss, 2367
  for present worth, *see* Present worth, probability distribution for
  in risk assessment, 1258–1259
Probability density function, 2147
Probability tree, 2372
Probes, cognitive, 1026
Problem solving, decision making vs., 2173
Procedural knowledge, 67, 1775
Process(es). *See also* Design and process platform characterization methodology
  analyzing performance of, 1828
  definition of, 34, 456, 1243, 1696
  existing, documentation/analysis of, 1708, 1709
  geometric capabilities of, 457, 463–465
  nature of, as industry categorizer, 329–331
  new. *See also* Process design and reengineering
    definition/design of, 1709
    in total quality leadership (TQL), 1803
  PMBOK classification of, 1244
  project management, 1242–1244
  of rapid product development (RPD), 1286
  selection of, 456, 457, 1316
  service, 643–644
  stable/unstable, 1829–1831
  technological capabilities of, 457, 465, 468–471
  tools for viewing, 1809, 1810
  variation in, 1861–1863
  variation in management of, 1830–1831
  variation in operation of, 1830
Process analysis tools, 1703
Process-based approach to service quality, 639
Process capability analysis (PCA), 1869–1871
Process cells/units, 1772
Process chains, 203–205
  classification of, 203
  with common data management, 204
Process charts, 1374–1378
Process design and reengineering (PDR), 1696–1715
  assumptions underlying, 1697
  case studies involving, 1712–1714

Process design and reengineering (PDR) (*Continued*)
  definitions related to, 1696–1697
  development and evolution of, 1699–1702
  diagnose phase of, 1697, 1708, 1709
  envision phase of, 1697, 1705–1707
  evaluate phase of, 1698, 1711, 1712
  failures in application of, 1700, 1701
  future of, 1714–1715
  guidelines for, 1698, 1699
  with health care delivery systems, 747
  implementation of, 1704–1712
    checklists for, 1705, 1708–1712
    customer requirements, determining, 1708
    evaluation, 1711–1712
    existing process, documentation/analysis of, 1708, 1709
    human resource structure, design of, 1710
    information technology levers, identification of, 1706
    management commitment and vision, establishment of, 1705
    new process, definition/design of, 1709
    opportunities, discovery of, 1705, 1706
    performance goals, setting, 1708
    and preplanning, 1704, 1705
    project planning, 1707
    reconstruction, 1710–1711
    stakeholders, informing of, 1706–1707
    teams, organization of, 1707
    training, 1711
  initiate phase of, 1697, 1706–1708
  preplanning for, 1704, 1705
  reconstruct phase of, 1698, 1710, 1711
  redesign phase of, 1698, 1709, 1710
  and second wave of reengineering, 1701
  steps in, 1697–1698
  tools for, 1702–1704
    activity-based costing (ABC), 1704
    analysis tools, 1703
    benchmarking, 1703
    modeling tools, 1703
    simulation tools, 1703–1704
Process equipment, 1580
Process excellence, 42
Process flexibility, 499
Process improvements, 30
Process industry, computer integrated manufacturing (CIM) in, 518–526
  and architecture structure model, 520–521
  definitions, related, 518–519
  and hierarchical structure model, 521–522
  and information integration, 522–523
  key technologies, development of, 519
  refinery enterprise example, 523–526
Processing:
  automated, 156
  of retail goods, 777
Process integration, 490
Process management:
  in EPEM model, 1798
  for increased transformability, 317–319
  major activities of, 1770
Process measures (service quality), 1964
Process modeling, 318, 1703
Process organizations, 284–285
Process-oriented enterprise modeling, 286–291
  data views, 288–290
  function views, 287, 288
  organization views, 286–287
  output views, 287–289
  process view, 290–291
Process performance, 49, 50
Process planning, 448–482
  analysis/evaluation of plan, 458–460
    quality estimation, 460
    time/cost estimation, 459–460
  capability analysis in, 465, 468–471
  computer-aided, 460–482
    advantages of, 473
    capability analysis, 465, 468–471
    CAPP, 474–475
    cost model, 465–467, 472
    development of, 473–474
    generative approach to, 477–478
    group technology, 461–463
    mapping, 463–466
    selection criteria for, 478–482
    tolerance charting, 472–473
    variant approach to, 475–477
  cost model for use in, 465–467, 472
  definition of, 448
  and desired quantity, 452
  detailing, process, 457–459
    optimization, process, 458
    parameters, determination of, 458
    and tool selection, 457–459
  as element of rapid product development (RPD), 1287–1288
  environmental considerations in, 536
  fixture planning, 455, 457
  generative approach to, 477–478
  geometry analysis in, 452
  green engineering for, 599
  gross planning, 453–455
    casting vs. machining, 453, 454
    strength and cost features, 454–455
  group technology as tool for, 461–463
  integration of computer aided design (CAD) with, 191
  mapping for, 463–466
  and plant engineering, 1551
  and product-realization process, 448–452
    design evaluation in, 450
    function analysis in, 450
    geometry, designed, 449, 450
    materials, selection of, 449
    production planning/scheduling, 451–452
    production quantity, 452
  selection of process, 456, 457
  setup planning, 455, 456
  stock selection for, 452, 453
  tolerance charting for use in, 472–473
  variant approach to, 475–477
Process platform(s):
  capability, 1982

  concept/definition of, 1980–1981
  implementation, 1981–1982
  linkage of product design and, 1996–1999
Process platform manuals, 1996–1998
Process quality, 1796, 1797
Process specification management, 333–334
Process tolerances, 1986–1987
Process variables, 985–987
Process view, 88, 290–291, 507–508
PROCRU, 2429
Procter & Gamble, 779, 780, 2010
Procurement management:
  project, 1249–1250
  and supply chains, 2122
Product(s):
  core, *see* Core products and services
  customization of, 261–262
  defects in, *see* Test and inspection
  derivative, 49
  digital, *see* Digital products
  failure rate of:
    in "infant mortality" period, 1925–1972
    in useful life period, 1927
    in wear-out period, 1927
  item master table for, 2096
  manufacturing, nature of, 329
  nature of, as industry categorizer, 329
  new, *see* New products
  physical, *see* Physical products
  platform, 49
  quality of, 1798
  search, 671
  service, 643
  variation in, 1856–1857
  viscosity index of, 2093
Product assurance, 1922
Product-based approach to service quality, 625, 638, 639
Product characteristics, and supply chain design, 2129–2130
Product-characteristics philosophy (warehouse layout), 1540
Product configuration management, 338
Product cost estimating, *see* Cost estimating
Product data exchange, 192
Product data management (PDM), 195
  engineering data management (EDM) vs., 195
  enterprise resource planning (ERP) interface with, 338, 349
Product design:
  assemblability evaluation, 368–369
  design for assembly (DFA), 367–370, 384
  and digital human modeling, 1121–1124
    accommodation, 1122–1124
    usability, 1123
  environmental considerations in, 534–536
  green engineering in, 598–599
  human-centered, *see* Human-centered product planning and design
  improving ability for, with CIM implementation, 526
  linkage of process platforms and, 1996–1999
  and plant engineering, 1551
  user involvement in, 486
Product development, 206–207
  assembly system planner in process of, 371
  as dimension of competitive advantage, 327
  rapid, *see* Rapid product development
  simultaneous engineering and time required for, 371, 372
  stages in, 1977–1978
  and supply chain management, 2122
Product families, 49, 687–694
  common bases (CBs) in, 690–691
  configuration mechanisms (CMs) in, 691
  design of, 692–694
  differentiation enablers (DEs) in, 691
  modularity/commonality issues with, 688–689
  multiple view, synchronization of, 691–692
  and product family architecture (PFA), 690–691
  and variety, product, 689–690
Product flexibility, 499
Product flow, 2125–2126
Product groups, 2087
Production:
  acceptance tests for, 1943
  information integration in, 522–523
  lean, 555
  one-piece flow, 547
  planning/scheduling, 451–452
  and plant engineering, 1550–1551
  pull method of:
    JIT's use of, 545
    kanban as, 545
  queueing models for coordination of, 1662–1667
    base stock control, 1663–1664
    kanban control, 1664–1667
  schedule-initiated, 551
  sequence-synchronized, 551
Production activity control, 497
Production capacity, 553
Production control, 1392
Production efficiency, 526
Production industries, networks in, 253–254
Production-inventory systems, 1669–1692
  assemble-to-order (ATO) systems, 1685–1689
  base-stock control, 1672–1675
    demand over lead time, 1674–1675
    normal approximations, 1673–1674
  deterministic multiperiod model, 1671
  DRP framework for, 1675–1676
  EOQ model, 1670
  kanban control, 1689–1690
  lot-for-lot policy, 1676–1678
  multi-stage models, 1683–1685
  with network of inventory queues, 1690–1692
  newsvendor model, 1670
  (Q, R) model, 1671
  (S, s) model, 1671, 1678–1683
Production leveling, 545–546
Production lines, 1772

Production message standard, 168
Production networks, 616–617
Production-ordering kanbans, 549
Production planning and scheduling, 2034
  advanced planning and scheduling, *see* Advanced planning and scheduling (APS)
  with environmental considerations, 538
  linear programming applications for, 2056
  major activities of, 1770
  managing variety in, 694–697
Production quotas, 20
Production Robotics and Integration Software for Manufacturing Group (PRISM), 607
Production stage (project life cycle), 1242
Production systems:
  assembly subsystem, *see* Assembly
  and Theory of Constraints, 557–558
Production technologies, 949
Production units, areas of responsibility in, 1772
Productive hour cost (PHC), 2314
Productivity:
  computer technologies for increasing, 1223
  energy, 1573
  in food service kitchens, 826
  and job design, 869
  in plant/facilities engineering, 1561–1562
  and work team cohesiveness, 881, 882
Product life cycle(s):
  assessment of, 533
  changes in, 311, 312
  declining phase, 675–676
  and development cost recovery, 486
  failure rates in, 1925–1927
  growth stage, 675
  and mass customization, 687
  maturity stage, 675
  phases in, 1312
  and rapid product development (RPD), 1284–1285
Product measures (service quality), 1964
Product-mix problem, 2524–2525
Product modeling, 210–212
Product planning. *See also* Product design
  human-centered, *see* Human-centered product planning and design
  manufacturing, designing for, 1311–1330
Product platforms, 49, 686–687
Product quality, 526
Product-realization process, 448–452, 1971
  design evaluation in, 450
  function analysis in, 450
  geometry, designed, 449, 450
  materials, selection of, 449
  production planning/scheduling, 451–452
  production quantity, 452
Product selection/sourcing, 777
Professional services, 1333
Professional services projects, 1333–1350
  avoiding problems with, 1349–1350
  close phase for, 1348–1349
  definition of professional services, 1333
  monitoring/control phase for, 1346–1348
  project-definition phase for, 1335–1338
  project-management process for, 1334–1335
  and project manager, 1334
  project-planning phase for, 1338–1346
    budgeting, 1343–1346
    critical path, 1341, 1342
    objectives/scope, confirmation of, 1338
    personnel, assignment of, 1339–1341
    resource loading, 1341, 1342
    task and deliverables list, 1339–1341
    team, project-planning, 1338
    time estimates, 1341, 1342
    work breakdown structure, 1338–1340
Profile analysis, 2072
Profit, 667, 2300
Profitability analysis, 673–674
Profit center, plant engineering as, 1562
Profit-volume ratio (PV), 673–674
Program Evaluation and Review Technique (PERT) diagrams, 104
Programming, stochastic, *see* Stochastic programming
Programming environments, 304
Programming languages:
  ASP, 79
  for building information systems, 70–79
  C++, 72–73
  CGI, 77–78
  ColdFusion, 78–79
  history of, 70–71
  HTML, 76–77
  Java, 78
  Visual Basic, 73–76
  web-based programming, 76–79
Programs:
  energy-management, 1578
  reliability, 1953–1954
Progressive assembly lines, 1355
Project(s):
  definition of, 1242
  division of labor and failure of, 1267
  financial control of, 1273
  professional services, *see* Professional services projects
  WBS as dictionary for, 1277
Project A, 921
Project communications management, 1248
Project concentric circle model, 1253–1255
Project cost management, 1245–1246
Project data management (PDM), 1290
Project-definition phase (professional services projects), 1335–1338
Projected schedule feasibility (IS systems), 98
Project Evaluation and Review Technique (PERT), 2305–2306
Project historical databases, 1260
Project human resource management, 1247
Project integration management, 1244
Projection virtual environments, 2507
Project life, 1242–1243
  decreasing expected, 2392
  and project integration management, 1244
  uncertain, risk analysis with, 2371

Project-life-cycle-based work breakdown structure, 1269, 1270
Project management. *See also* Professional services projects; Work breakdown structure
  avoiding problems in, 1349–1350
  communications management, 1248
  computer-aided, 1252–1262
    automation of, 1256–1260
    future of, 1261–1262
    history of, 1253
    implementation of, 1260–1261
    and project concentric circle model, 1253–1255
  cost management, 1245–1246
  definition of, 1333
  human resource management, 1247
  integration management, 1244
  in learning organizations, 1250–1251
  processes in, 1243–1244
  procurement management, 1249–1250
  quality management, 1246–1247
  risk management, 1248–1249
  scope management, 1244–1245
  time management, 1245
Project Management Body of Knowledge (PMBOK), 1242–1244, 1254, 1259
Project Management Institute (PMI), 1242
Project Management Professional (PMP) certificate, 1242
Project management software, 142, 1242–1251
Project managers:
  for change process with new technology, 965
  and professional services projects, 1334
Project office, 1346–1347
Project organizational structure, 1265
Project planning, 1707
Project-planning phase (professional services projects), 1338–1346
  budgeting, 1343–1346
  critical path, 1341, 1342
  objectives/scope, confirmation of, 1338
  personnel, assignment of, 1339–1341
  resource loading, 1341, 1342
  task and deliverables list, 1339–1341
  team, project-planning, 1338
  time estimates, 1341, 1342
  work breakdown structure, 1338–1340
Project portfolio management tools, 1260
Project procurement management, 1249–1250
Project quality management, 1246–1247
Project risk management, 1248–1249
Project scope management, 1244–1245
Project steering committee, 1346, 1348
Project teams:
  learning processes for, 1250
  management support for, 983, 984
  performance reviews for, 1349
  for professional services projects, 1334, 1338
Project time management, 1245
Project work, 1254
Project workplan, 1347
ProModel, 2459–2460
Proof Animation, 2446
Proportional-derivative (PD) control models, 160
Proportional-integral-derivative (PID) control models, 160
Proportional-integral (PI) control models, 160
Proportionality, in linear models, 2525
Proportional (PE) control models, 160
PROSPECTOR, 2189
Prospect theory, 2203–2204
Protocols:
  choosing, 252
  HTTP, 244–245
  Internet, 239–240
  IP addressing, 241–242
  layers of, 239–240
  network, 237, 239–240
  TCP/IP as group of, 240
Prototyping:
  digital, 1288–1290
  as human-centered product planning/design tool, 1303, 1305–1306
  for human–computer interface design, 1216
  near-net-shape processing for, 586, 587
  of new process concepts, 1709
  new technologies for, 1283, 1284
  physical, 1288
  rapid, *see* Rapid prototyping
  SLDC, 104, 105
  virtual, 2498, 2501, 2509–2512
Provia Software Inc., 2058
Proximity switches, 1903
Proxy gateways, 735
Proxy servers, 268
Prudential, 654
PSD (Prevention of Significant Deterioration) standards, 592
PSDA cycle, 1811
Pseudo-code, 100
Pseudonymity, 268
Pseudonyms, 269
Pseudorandom number generators, 2472–2473
Psychological disorders, 1170
Psychophysics, 1071–1072
Psychosocial stress, 1170
P3P (Platform for Privacy Preferences), 269
PTS, *see* Predetermined time standards
Public health approach to occupational safety and health, 1157–1159
Public information, 232
Public-key cryptosystems, 733–734
Public network-based applications, 243–244
Public utility industries, work injuries in, 1070
Puerto Rico, 957–958, 959
Pull systems (production):
  JIT's use of, 545
  kanban as, 545
  supply chain management as, 2122
Pulp and paper industries, 1765
Punching rivets, 372, 412
Purchasing:
  in ISO 9001:2000 product realization clause, 1971
  lead times for, 2050

Purchasing (*Continued*)
  and transportation management software, 2065
Purchasing power, inflation and, 2395
Purdue Enterprise Reference Architecture (PERA), 506, 507, 519, 1769–1772
  control hierarchy in, 1769–1771
  equipment organization in, 1771–1772
  and ISA/ISO standards, 1772
Purdue University, 607, 660
Purpose statements (business model), 31
Push–pull force limits, 1054–1058
PV, *see* Profit-volume ratio

(*Q*, *R*) model, 1671
QC, *see* Quality circle
QCD (quality, cost, and delivery), 552
QFD, *see* Quality function deployment
QFP (quad flat pack), 424
QI teams, *see* Quality improvement teams
Q-learning, 1780
QLs (query languages), 119
QMS standards, *see* Quality management systems standards
QoS, *see* Quality of service
QR system, *see* Quick response system
QS 9000 standard, 1973
Quad flat pack (QFP), 424
Quadratic programming problems (constrained optimization), 2555, 2562
Quality. *See also* Reliability
  in advanced planning and scheduling (APS), 2049
  basic requirements of, 1889
  causes of variation in, 1828–1829
  definition of, 1889
  definitions of, 1794
  eight pillars of, 1796–1798
  estimated product, 460
  key approaches to, 1794
  on-spec/expected, 1005
  perceived, 669, 680
  and performance management, 1000
  in plant/facilities engineering, 1562
  price and buyers' perceptions of, 669, 670
  service, 1956–1965
    approaches to defining, 625–626
    conceptual model of, 626–630
    definition, 1956–1957
    leadership for, 1958–1959
    measurement/evaluation of, 1963–1964
    measurement of, 640–641
    models of, 638–640
    process dimension, 641, 642
    SERVQUAL instrument for measuring, 627–630
    strategy for creating/maintaining, 1957–1958
    structure dimension of, 641, 642
    systems for delivery of, 1961–1963
    workforce for, 1959–1961
  of service systems, 645–649
    areas, assessment, 645–647
    and maturity model, 648
    procedure for, 648, 649
  total quality leadership, *see* Total quality leadership
  in transportation, 817–818
Quality, cost, and delivery performance (QCD), 552
Quality assurance:
  definition of, 1967
  in electronic production, 431–432
  poka-yoke for, 548
  TQM, 552
  visual control systems for, 548–549
Quality circle (QC), 978–980
Quality control. *See also* Test and inspection
Quality cost analysis, 432
Quality function deployment (QFD), 1817, 1820
Quality improvement:
  and occupational safety and health, 1184–1185
  and teams, 978–980
  tools for (health care), 747
  variation and, 1831–1832
Quality improvement (QI) teams, 748
  in health care systems, 748
  management support for, 984
  QCs vs., 980
Quality management:
  major activities of, 1770
  project, 1246–1247
  in service organizations, 624
Quality management systems:
  computer-aided, 497–498
  standards for, 1185
Quality Management Systems–Guidelines for Performance Improvements (ISO9004–2000), 1972
Quality management systems (QMS) standards, 1966–1974
  definitions related to, 1966–1967
  European standards, 1968
  international standards, 1968–1969
  ISO 9001:2000 standard, 1969–1972
    continual improvement clause in, 1972
    management responsibility clause in, 1970
    measurement/analysis clauses in, 1971
    product realization clauses in, 1971
    resource management clause in, 1971
    scope of, 1969
  ISO 9004–2000 standard, 1972
  non-ISO standards, 1973
  other ISO 9000 standards, 1972–1973
  registration to, 1973–1974
  U.S. standards, 1967–1968
Quality of service (QoS), 250, 1796
Quality planning, 1247
Quality policy, 1967
Qualysis AB, 1125
Quantiles, 2476, 2477
Quantitative forecasting models, 793
Quantity, production, 452
Quasi-Newton methods, 2551, 2553
Queries, 67, 81–82, 84, 2013
Query languages (QLs), 119

QUEST, 2460
Quest for Quality and Productivity in Health Services Conferences, 747, 748
Questionnaires, 127
  for evaluation of job design, 872–873
  for evaluation of team design, 889–892
  as human-centered product planning/design tool, 1302–1304, 1308, 1309
Queueing models, 1628–1668, 2157–2163
  assumptions of, 1634–1635, 2160–2161
  ASTA/PASTA, 2163
  bottleneck queues, 2162
  for dynamic job shops, 1650–1656
    general service times, 1654–1656
    multiple-job-class open Jackson queueing network model, 1652–1654
    single-job-class open Jackson queueing network model, 1650–1652
  in flexible machining systems, 1656–1662
    dedicated material-handling systems, 1660
    general single-class closed queueing network model, 1660–1661
    multiple-class model, 1661–1662
    single-class closed Jackson queueing network model, 1656–1660
  for flow lines and series systems, 1638–1645
    general service times, 1640, 1643–1645
    multiple-stage flow lines with exponential processing times, 1642–1643
    paced systems, 1638–1639
    three-stage flow lines, 1640–1642
    two-stage flow lines, 1639–1640
    unpaced lines, 1639
  for health care delivery systems, 744–745
  and job/customer, 1629
  long-run behavior, determination of, 2161–2162
  for manufacturing systems, 1632–1633
  Markovian queueing models, 2153–2154, 2158–2159
  *M/G/*1 queue, 2159–2160
  and modeling in general, 1629–1631
  notation for, 2157–2158
  performance measures determined by, 1631–1632
  for production coordination, 1662–1667
    base stock control, 1663–1664
    kanban control, 1664–1667
  for service systems, 1633–1634
  for single-stage systems, 1635–1638
    make-to-order manufacturing/service, 1635–1636
    make-to-stock manufacturing/service systems, 1636–1638
  steady state, rate of convergence to, 2162–2163
  and supply chain, 1634
  for transfer lines, 1645–1650
    infinite inventory banks, 1646
    multiple-stage transfer line, 1648–1650
    no inventory banks, 1645–1646
    two-stage synchronized line, 1646–1648
  value of, 1628–1629, 1632
  variability, effects of, 2162
Queueing networks, 2163–2170
  decomposition methods, 2167–2170
  general product-form networks, 2165–2167
  Jackson networks, 2164–2165
Queuing theory, 128
Quick response (QR) system, 2028–2029

Raad voor Accreditatie (RVA), 1974
Racks (storage), 1521–1523
RAD, *see* Rapid application deployment
Radio-frequency (RF) picking systems, 2108
RAM, *see* Responsibility assignment matrix
RAMSIS CAD manikin, 1122
Random-effect models, fixed-effect vs., 2229–2230
Randomization (experimental design), 2228
Randomized complete block experimental design, 2230
Random-location storage, 1534, 1535
Randomness (in simulations), 2472–2473
Random normal deviates, 2386, 2389
Random numbers, table of, 2386, 2387
Random sampling, 1136
Random utility models, 2204
Range-of-joint mobility, 1043, 1046
Range-of-motion (ROM), 1064
Ranking method (job evaluation), 902–903
RAPID, 167
Rapid application deployment (RAD), 72, 104, 105
Rapid product development (RPD), 1283–1294, 2499, 2501
  digital prototyping as element of, 1288–1290
  and engineering solution center, 1290
  human and technical resources involved in, 1286
  and knowledge engineering, 1291–1293
  new technologies facilitating, 1283, 1284
  objectives of, 1284
  and the organization, 1285–1286
  physical prototyping as element of, 1288
  process of, 1286
  process planning as element of, 1287–1288
  and product life cycle, 1284–1285
  results of, 1286–1287
  simultaneous engineering vs., 1284–1287
Rapid prototyping (RP), 191, 207–210, 586, 1288
Rapid prototyping technologies (RPT), 1288
Rapid setup, 547
Rapid tooling, 1288
Rate(s) of return:
  before-tax, 2334
  differentiating, by risk class, 2391–2392
  minimum, 2335
  risk-free, 2334–2335
Rate of return method (cost estimating), 2348–2349
Rating content (of networks), 248–249
Rating scales (job performance), 1423–1425
Ratio control, 161
Rational choice, axioms of, 2178–2179
Rationality:
  bounded, 139, 140, 1020

Rationality (*Continued*)
  in classical decision theory, 2178
Raw-materials storerooms, 1528
Rayovac, 93
RBDSS (rule-based decision support system), 1777
RCCP (rough-cut capacity planning), 2042
RCM, *see* Reliability-centered maintenance
R control charts, 1850–1855
RCRA, *see* Resource Conservation and Recovery Act
R&D projects, *see* Research and development projects
Reach, 1049
Reach distances, 1361
Reactive scheduling, 1733
Ready-made software, 68
Real estate brokers, 1476
Real-life scheduling, theoretical vs., 1732, 1733
Real output, 344
Real-time multimedia transmission, 234, 237
Real-time pricing (e-commerce), 269
Real-time resource scheduling, 497
Real-time systems, 172–173
Real-time transmission of simple information, 237
Reaming, 1322
  geometric capabilities of, 464
  technological capabilities of, 469
Reasoning knowledge, 67
Receiving operations (warehouses), 2103–2104
Recognition, cognitive probes for, 1026
Recognition-primed decision making, 2206
Recommended weight limit (RWL), 1076–1080
Reconstruction, and process design and reengineering (PDR)implementation, 1710–1711
Reconstruct phase (process design and reengineering), 1698, 1710, 1711
Record-based data models, 120
Record (in database), 80
Record keeping. *See also* Database management systems
  as core function of information systems, 66–67
  in enterprise information systems, 69
  in functional information systems, 68
  in local ISs, 68
  in transorganizational information systems, 69
Recruiting:
  of customer service employees, 659
  and leadership, 856–858
Rectilinear transfer machines, 418
Recycling:
  definitions related to, 533
  designing for, 598–599
  of disassembled components, 440, 443
  of returned products, 784
Red Brick Systems, 83
Redesign phase (process design and reengineering), 1698, 1709, 1710
Reductionist models of human performance, 2413
Redundant components, failure behavior of, 1933
"Reengineered" workers, 1700
Reengineering:
  changes required for, 954
  process design, *see* Process design and reengineering
  second wave of, 1701
  of supply chains/supply chain management, 2132–2133
Reengineering the Corporation (Hammer and Champy), 1700
Reference lotteries (decision making), 2192
Reference prices, 671
References (personnel), 923
Referrals, and electronic commerce, 273
Refinery enterprise, CIM in, 523–526
Reflectance, 1199–1200
Reflection effect, 2203
Reflow soldering, 423–425, 429–431, 438, 439
Refresh rate, 2506
Refueling by robots, 381
Registrar Accreditation Board, 1974
Registration to quality management systems (QMS) standards, 1973–1974
Regression analysis, 128–129, 2265. *See also* Multiple linear regression (MLR)
  in health care systems, 745
  hypothesis testing in, 2262
Regressors, 2265
Regret, minimax, 2177, 2180–2181
Regret matrix, 2381
Regulation(s):
  changes in, 38
  as external force on business, 39
  of health care delivery systems, 738
  inspection, regulatory, 1907
  of online markets, 278
Reinforcement learning, 1780
Rejection region (hypothesis testing), 2243, 2244
Relational database model, 80–81, 120
  entity type, 121–122
  hierarchical vs., 121
  spreadsheet vs. entity types, 121
Relational model base management systems, 130–131
Relations (relationship database model), 80
Relationships. *See also* Partnerships
  in business model, 34, 46, 48–49
  coworker/intragroup, 1220
  customer, 1962–1963. *See also* Customer relationship management (CRM)
  manufacturer/retailer, 781
  quality of, 1797
  and team effectiveness, 986–987
  tools for understanding, 1810, 1821–1822
  worker–management, 1220–1221
Relationship charts, 826–828
Relationship diagrams, 829, 830
Relationship management, 13–14
  among professional groups, 23
  and communication, 49
  in EPEM model, 1798

  as partnership within organization, 23
Relative body position, 1043
Relaxations, 2584–2589
  definition of, 2584
  Lagrangean, 2587–2589
  linear programming, 2585–2587
Relevance of models, 284
Reliability, 1922–1954
  of aircraft structural inspection, 1909
  and allocation of reliability requirements, 1937, 1938
  and availability, 1949–1951
  data, and analysis of common elements, 1443, 1444
  definition of, 1922, 1927
  design for, 1922, 1937, 1939–1940
    failure mode and effects analysis, 1940
    probabilistic approach in, 1940
    review, design, 1939
  growth of, 1951–1953
  in human factors audits, 1134–1135
  human factors in, 1941
  and life characteristics curve, 1925–1927
  and maintainability, 1946–1949
  measures of, 1927–1932, 1941–1946
    estimation, 1944–1946
    exponential distribution, 1930
    gamma distribution, 1932
    lognormal distribution, 1931
    mathematics of, 1928–1930
    normal distribution, 1931
    test programs, 1942–1944
    Weibull distribution, 1931, 1945–1946
  as network flow problem, 2579, 2580
  as performance measure of quality, 1246
  programs, reliability, 1953–1954
  and system effectiveness, 1922
  system life cycle, reliability activities during, 1923–1925
  system reliability models, 1932–1937
    fault tree analysis, 1936–1937
    reliability block diagram, 1933–1936
Reliability block diagram:
  bridge structure, 1934, 1936
  coherent systems, 1935–1936
  k-out-of-n configuration, 1935
  parallel configuration, 1934, 1935
  series configuration, 1933–1935
Reliability-centered maintenance (RCM), 1618–1619
Relief labor pool scheduling, 1745
Remote access (computer), 243
Remote procedure call (RPC), 719
Remote services (tele-services), 266
Renault, 302
Rendering, 2504
Renewal processes, 2150
Reorganization, and job design/redesign, 883
Repairability, 1924
Repeatability (measurement systems), 1880, 1881
Repetitive jobs, 1362
Repetitive motion injury (RMI), 1166
Repetitive processes, 1251
Repetitivness, task, 1092
Replenishment flows (warehouse operations), 2084
Replication, 2228
Reporting:
  data access vs. traditional, 37
  project status, 1347
  of safety/health incidents, 1174
  of workplace hazards, 1176
Reports:
  in enterprise information systems, 69
  feasibility (SDLC), 97
  in functional information systems, 68
  from information systems, 67
  in local ISs, 68
  query facility for, 67
  sales/service, 1308, 1309
  timing of, 67
  in transorganizational information systems, 69
  use of standards for, 1406
Representational appropriateness (of information), 140
Representations, operations, memory aids, and control mechanisms (ROMC), 113, 114
Representativeness heuristic, 2198–2199
Requests for estimate (RFEs), 2299–2300
Requests for proposals (RFPs), 1492
Requests for quotes (RFQs), 968
Requirement document (human-centered product planning/design), 1307
Requisite variety, law of, 958
Rescheduling, 503
Research and development (R&D) projects:
  knowledge integration in, 1293
  process planning for, 1287
  work breakdown structure for, 1273, 1274
Residual errors (residuals), 2269, 2284–2285
Residual variance, 2270–2271
Resource allocation:
  major activities of, 1770
  and mass customization, 697–700
  project, 1246
  in rapid product development (RPD), 1286
  for safety programming, 1183
Resource Conservation and Recovery Act (RCRA), 534, 593–594, 1164
Resource loading, for professional services projects, 1341, 1342
Resource management, 30, 41, 43, 59–60
  automation of, 1256, 1257
  in client/server (C/S) systems, 726
  for effective teams, 984
  in ISO 9001:2000 QMS standard, 1971
  risks with, 46
  for services, 637
  tools for multiproject, 1260
Resource Monitoring System (RMS), 742
Resource performance, 49, 50
Resource sharing (over networks), 243
Resource view, modeling method for, 510
Response (in experimental design), 2225
Response time, in client/server (C/S) systems, 726

Responsibility, and work packages, 1268–1269
Responsibility assignment matrix (RAM), 1267–1268
Restaurants, *see* Hospitality industry
Restoration Hardware, 656
Restricted random sampling (RRS), 1456
Restrictive clothing allowances, 1397
Rest time, 1366
Retail distribution warehouses, 2085–2086
Retailers:
  catalog, 2086
  personnel scheduling for, 1743, 1745, 1765
  and supply chain design, 2128
Retailing, 773–785
  online, 265–267
    digital products, 266, 267, 270–271
    physical products, 266
    services, 266, 267
    storefronts, Web, 265–266
  outsourcing in, 264
  personnel scheduling for, 1752–1755
Retail supply chains, 773–785
  benefits of managing, 774–775
  and competitiveness, 781
  components of, 767–779
    distribution, 777
    inbound transportation, 777
    outbound transportation, 777–778
    outsourcing, 778
    processing, 777
    product selection/sourcing, 777
    storage, 778–779
    warehousing, 777
  and e-commerce, 782–784
  emerging paradigm for, 781–782
  and forecasting, 779–781
  and globalization, 782
  history of management of, 776
  and information technology, 782
  and mass customization, 784
  objectives in management of, 779–781
    channel capability, flexible, 781
    forecasting, improved, 779–780
    replenishment, improved, 780, 781
  statistical tools for managing, 782
  transportation in, 777–778
  trends in management of, 776, 782–785
  warehousing in, 777
Retention (employee):
  in customer service department, 659
  and scheduling, 1745
Retention (of information), 930–931
Return of products, 784, 2104, 2122
Return on investment (ROI), 99
Return policies, guarantees and, 656–657
Reusability, maximization of, 686
Revenues from Internet economy, 260
Reverse auctions, 275–276
Reverse distribution, 1470
Reverse supply chain, 784
Reversible Markov chains, 2156
Reviews, handling, 913
Rewards:
  and computer technology, 1222
  dynamic decision problems, 2638–2639
  and leadership, 861–862
  new processes linked to, 1710
  sharing, in supply chain management, 2126
  team/team level, 881, 882
RFEs, *see* Requests for estimate
RFPs (requests for proposals), 1492
RFQs (requests for quotes), 968
RF (radio-frequency) picking systems, 2108
RHS constants, *see* Right-hand side constants
Ridge regression, 2290–2291
Right-hand side (RHS) constants, 2536–2538
Right-to-know concept (workplace hazards), 1176
Rigid work transfer systems, 357
Risk(s):
  business, 45–48
  controls linked to, 45–48
  and globalization, 28
  identifying/mitigating, 30
  perceived, 2203
  of professional services projects, 1337
  responses to, 1249
  sharing, in supply chain management, 2126
  and uncertainty avoidance, 958
Risk analysis, 2367–2376
  in business model, 56–57
  comparison of risky proposals, 2376
  measures for, 2367
  present worth, probability distribution for, 2367–2369, 2371–2376
    discrete distribution, 2372–2373
    expected present worth, 2367–2368
    mean and variance, using only, 2373–2374
    normal distribution, assumption of, 2374–2376
    variance of present worth, 2368–2369
  with uncertain project life/cash flow, 2371
  uncertain timing, cash flows with, 2369–2371
Risk classes, 2391–2392
Risk events (term), 1248, 1249
Risk factors:
  for musculoskeletal disorders, 1362
  for work-related musculoskeletal disorders, 1093–1094
Risk management, 40
  assessment of, 56
  for professional services projects, 1337–1338
  project, 1248–1249
Risk preferences, 753–755
Risk profiles, 57
Risk Radar, 1258, 1259
Ritz-Carlton, 656, 846
Riveting, 372, 411, 412
RMI (repetitive motion injury), 1166
RMS (Resource Monitoring System), 742
Roadnet 5000 (software), 2064
Roadnet Technologies, 2064
ROBCAD, 167, 171–173
Robinson-Patman Act, 681–682
Robots/robotics, 373–382
  accuracy of, 376, 377
  in assembly systems, 366
  for assistance/entertainment, 381, 382

  in automotive industry, 388–392, 420
    gearboxes, unpacking of, 389, 391–392
    steering components assembly, 389, 391
  classification of, 374, 375
  cleaning, 380
  components of, 374, 376–377
    control system, 376, 377
    gripper, 377, 413–415
    handling systems, 413, 414
    measuring equipment, 376
    power supply, 374, 376
  courier/transportation, 379–380
  in electronic assembly, 392–396, 435, 436
    luminaire wiring, 394–395
    measuring instruments, 392–394
    overload protector, 392
  in flexible assembly systems, 360–362
  as flexible handling equipment, 420
  in food industry, 397–398
  industrial, 360–362, 366
    classification/types of, 374–375
    control systems, 376, 377
    definition of, 373
    gripping systems, 377
    measuring equipment, 376
    power supply for, 374, 376
    programming of, 377–378
    simulations, 378, 379
  joining processes accomplished by, 372, 373
  for material handling, 1525
  medical, 381, 382
  in microassembly, 396, 397
  models of, with tool perspective, 171
  in pharmaceutical/biotechnological industries, 398
  programming of, 377–378
  refueling by, 381
  repeatability of motion in, 376, 377
  SCARA, 413
  service, 379–381
    assistance/entertainment, 381
    cleaning, 380
    courier/transportation, 379–380
    medical, 381
    refueling by, 381
  simulations for construction/planning of, 378, 379
  simulators/emulators, 167
  six-DOF robots, 413
  types of, 374
  welding with, 413
ROI (return on investment), 99
Role concepts (human resources), 642
Roller conveyor, 1514, 1516
Roll forming, 1320, 1321
Rolling, 565
ROMC, *see* Representations, operations, memory aids, and control mechanisms
ROM (range-of-motion), 1064
Roses, automated inspection of, 1900–1901
Rotary indexing turntables, 418, 419
Rotary transducers, 1902
Rotational molding, 1327
Rotational sweeping (solid modeling), 183
Rough-cut capacity planning (RCCP), 2042
Routine decision making, 2176
Routing:
  flexibility in, 499
  linear programming applications for, 2056
  package, 807–810
  of pickers in multi-aisle warehouses, 2105–2106
  for pickup and delivery operations, 801–803
  vehicle, 819–821. *See also* Shipment planning
    problems with, 2062–2063
    time windows (VRPTW) problem with, 794–795
    in transportation management systems (TMS), 2058–2059
Royal Dutch Shell, 7
RP, *see* Rapid prototyping
RPC (remote procedure call), 719
RPD, *see* Rapid product development
RPT (rapid prototyping technologies), 1288
RRS (restricted random sampling), 1456
RSA (cryptosystem), 733
Rule-based behavior, 1020
Rule-based decision support system (RBDSS), 1777
Rule-based performance, 2205, 2206
Rule orientation (as national culture dimension), 957, 960
Rules of order (for groups), 2213
Rules:
  decision, 2177–2178
  fuzzy, 163
Run chart, 1817, 1821
Run charts, 1824, 1825, 1832, 1833
Run tests (for hypothesis testing), 2259
Runtime, 297
RVA (Raad voor Accreditatie), 1974
RWL, *see* Recommended weight limit

(*S*, *s*) model, 1671, 1678–1683, 2471–2472, 2478
  approximations, 1680
  dynamic policy, 1681–1683
  safety stock levels, setting, 1683
  stationary analysis, 1678–1679
Sabre Inc., 2058
SAC, *see* Storage analysis chart
SADT, *see* Structured analysis and design technique
SAE-J standards, 1121, 1122
Safe Drinking Water Act (SDWA), 1164
Safety, 957. *See also* Occupational safety and health
Safety culture, 959–961
Safety programs, 1183–1184, 1567–1568
Safety stock levels, 1683
SAG, 291
Saks', 779
Sales:
  ERP tools for, 90
  lost sales, calculation of, 1637–1638
  for mass customization, 701–705
  mass customization for, 701–705

Sales (*Continued*)
    and customer decision-making process, 703–704
    and design by customers, 701–703
    one-to-one marketing, 704–705
  online, estimates of, 671
  volume of, 667
Sales and service phase (human-centered product planning and design), 1300, 1308–1310
Sales promotions, 678–680
Sales reports, 1308, 1309
Salience bias, 1023
SAMMIE, *see* System for Aiding Man-Machine Interaction Evaluation
Sampling. *See also* Work sampling
  in health care systems, 745–746
  in human factors audits, 1135–1136
  when using control charts, 1840
Sanden Corporation, 555
SAP AG, 87, 95, 96, 304, 306, 492, 1002, 1738
SAP software:
  ERP, 89, 90, 95
  R/3, 89, 90, 92, 95, 304–306, 492
SARA, *see* Superfund Amendment Reauthorization Act
SAS Institute, 83
SAS/OR, 2575
SAS (software company), 861
Satisfaction. *See also* Customer satisfaction
  Baldrige criteria for employee, 1961
  expectations as influence on, 2204
  measures of, 1964
Satisficing decision rule, 2177, 2180
Satisficing models, 2608–2610
Savage principle (decision theory), 2381
Sawing:
  geometric capabilities of, 464
  technological capabilities of, 470
Scales:
  for job performance ratings, 1423–1425
  for test and inspection, 1890–1892
Scaling methods (decision making), 2192
Scandinavia, 1186
Scantron, 966
SCARA robots, 374, 375, 413
Scatter diagrams, 1860–1862
Scatterplots, 1819, 1821–1822, 1826, 1832–1834
SCC (Supply Chain Council), 348
Scenarios:
  as human-centered product planning/design tool, 1303, 1305
  in human–computer interface design, 1214
  in TOPP, 1288
Scene sensors, 1904
Schedule-compression options, 1342
Schedule for analysis (control charts), 1841
Scheduling, 329, 1718–1723, 1725–1732, 1725–1739, 2035–2036. *See also* Dispatching
  advanced planning and scheduling, *see* Advanced planning and scheduling (APS)
  AI approaches to shop floor, 1775–1782
    commercial software, 1782
    fuzzy set theory, 1781–1782
    genetic algorithms (GA), 1780–1781
    knowledge-based systems, 1775–1776
    neural networks, 1777–1780
  in automotive industry, 1734
  in aviation industry, 1734–1735
  branch and bound, 1728–1729
  cyclic, 1746–1747
  decomposition heuristics, 1729–1731
  detailed, 2044–2045
  and development of scheduling systems, 1735–1738
  driver, 812–817
    column-generation methodology, 814–815
    definition of problem, 813
    generation of schedules, 816
    iterative process for optimizing, 815–816
    set-partitioning formulation with side constraints, 813–814
  drum-buffer-rope (DBR), 558
  dynamic programming, 1726–1728
  in Flexible Manufacturing Systems, 501–503
  generation, schedule, 1736, 1737
  in health care delivery systems, 742–744
    optimization models, 746
    personnel scheduling, 744
    work scheduling, 742–744
  job, 497
  learning and, 1405
  linear programming, 1725–1726, 2056
  local search, 1731
  major activities of, 1770
  multiple-objective approaches, 1732
  notation used in modeling of, 1719–1722
  in packaging industry, 1733
  past research areas, 1722–1723
  of personnel, *see* Personnel scheduling
  planning vs., 2036–2038
  and polynomial time algorithms, 1722
  preventive, 1734
  production, 451–452
  reactive, 1733
  real-life vs. theoretical, 1732, 1733
  for resource allocation, 697–698
  rules for, 2050
  in semiconductor industry, 1733–1734
  and software development/implementation, 1737–1738
  techniques for, 1725–1732
    branch and bound, 1728–1729
    decomposition heuristics, 1729–1731
    dynamic programming, 1726–1728
    linear programming, 1725–1726
    local search, 1731
    multiple-objective approaches, 1732
Schedules:
  project, 1245, 1258, 1341, 1342
  work, and occupational injuries, 1178–1179
Schemas:
  definition of, 118
  in single-level data models, 118–119
Science, economic growth and, 602
Science network applications, 250

SCI (supply chain integration), 348
SCIS (supply chain information system), 318
SCM, *see* Supply chain management
Scope, project, 1266
Scope changes, 1259
Scope management:
  automation of, 1256, 1257
  project, 1244–1245
Scope statements, 31, 1336
Scope of work (SOW) document, 1266, 1278–1280
The Scoreboard for Maintenance Excellence, 1593–1597, 1610
Score function method, 2633–2634
Scoring rules (decision making), 2192, 2193
Scrap, 2308
Screening designs, 2235–2236
Screen printing (electronic components), 425
Screens, computer, 1195–1198
Screen sharing software, 142
Screwing, 371–372, 410, 411, 441, 442
Scrubbing (data), 85
SDI Industry, 2460
SDLC, *see* Systems development life cycle
SDWA (Safe Drinking Water Act), 1164
SE, *see* Simultaneous engineering
Sealed-bid auctions, 274
SEARCH method, 1373, 1374
Search process (inspection), 1895–1896
Search products, 671
Search techniques, local, 1731
Sears, 14, 654
Seasonality of products, 2130
Secondary hypothesis, 2245
Second-level domains (SLDs), 243
Second price auctions, 274
Second wave of reengineering, 1701
SECS, *see* SEMI Equipment Communication Standard
SECS-I, 165–166
SECS-II, 165–166, 168
Secure disposal, 533
Secure HTTP (S-HTTP), 734
Secure socket layer (SSL), 734
Securities auctions, 273
Security:
  of applets, 78
  with client/server (C/S) systems, 714, 715, 732–735
  for computer records, 1568
  and distributed denial of service attacks, 278
  with electronic commerce, 267–269
  gateways for network, 238
  in OSI management framework, 729
  and plant engineering, 1568–1569
  with public vs. private networks, 244
  technology for, 733–735
    access control, 734
    authentication protocol, 733, 734
    cryptographic systems, 733
    firewalls, 734–735
    message-integrity protocols, 733, 734
    Web security protocols, 734
  test and inspection related to, 1907
Security services, 732
Segmentation, market, 40
Segmentation pricing, 676–678
Selection, 921–924
  future of, 939–940
  predictors for use in, 921–924
    aptitude and ability tests, 921–922
    biodata, 922–923
    drug testing, 923
    interviews, 922
    references, 923
    simulations, 922
    work samples, 923
  and predictors of future performance, 924
  validity of measures used in, 923–924
Selection constraints, 691
Selective distribution, 2129
Selective processing (of information), 2199–2200
Selective soldering, 431, 438–439
Self-contained cell assembly lines, 547, 548
Self-declaration labels, 532
Self-guided vehicles (SGVs), 1524–1525
Self-managed teams, 976
Self-monitoring, cognitive probes for, 1026
Self-optimization, 315, 316
Self-pierce riveting, 372
Selling price, 2298
SEMATECH, 1772, 1774, 1775
SEMI, *see* Semiconductor Equipment and Materials International
Semiautonomous organizational units, 315–317
Semiconductor Equipment and Materials International (SEMI), 165, 1775, 1782
Semiconductor industry:
  automated test and inspection in, 1907
  scheduling in, 1733–1734
SEMI Equipment Communication Standard (SECS), 165–166
Semihot formed components, 568, 580–582
  extrusion, 565, 580, 581
  forging, 581
Semiquantitative job analysis methodology (SJAM), 1087–1091
Semivariance, 2367
Sensing elements (measurement systems), 1877, 1878
Sensitivity:
  in human factors audits, 1135
  of measurement systems, 1879
Sensitivity analysis, 2361–2367
  definition of, 2361
  in linear programming, 2536–2538
    practical uses of, 2536–2537
    simultaneous variations in parameters, 2537–2538
  more than two parameters, 2366–2367
  numerical example, 2361–2362
  with single variable, 2362–2364
  two parameters considered simultaneously, 2364–2366
Sensors:
  for assembly use, 384–386
    force/torque sensors, 385
    tactile sensors, 385
    ultrasound sensors, 385, 386

Sensors (*Continued*)
    video-optical sensors, 385–386
  in automated test and inspection systems, 1902–1904
Separable programming, 2556–2558
Sequences, use, 1214
Sequences (Structured English), 101
Sequencing, synchronization vs., 2036–2037
Sequencing software, 1738
Sequential sampling models, 2204–2205
Sequential truck travel, concurrent vs., 1510–1511
Sequential unconstrained minimization techniques, 2560–2562
Serial assembly cells, 409
Series systems, queueing models for, 1638–1645
  general service times, 1640, 1643–1645
  multiple-stage flow lines with exponential processing times, 1642–1643
  paced systems, 1638–1639
  three-stage flow lines, 1640–1642
  two-stage flow lines, 1639–1640
  unpaced lines, 1639
ServAs, 645, 647–64
Served in random order (SIRO), 2157
Served market, 40
Servers:
  domain name, 243
  network, 240–241
  proxy, 268
  World Wide Web, 240
Service(s):
  core, *see* Core products and services
  and customer satisfaction, 624–625
  customer service, *see* Customer service
  customization of, 261–262
  definitions of, 287, 637, 638
  goods vs., 624, 636–637
  Internet, 243
  need for systematic engineering of, 635–636
  networking, 243, 250
  online retailing of, 266, 267
  professional, *see* Professional services
  quality of, *see* Quality of service (QoS); Service quality
  queueing models for determining level of, 1632
  types of, 637, 638
  variation in, 1856–1857
Serviceability:
  definition of, 1924
  and human modeling, 1121, 1122
  as performance measure of quality, 1246–1247
Service-based economy, 623
Service blueprinting, 641, 642
Service delivery systems, 1961–1963
Service differentiators, 1957
Service encounters, 624–625, 628, 641
Service engineering, 635–636
Service factory, 559
Service industries:
  networks in, 253–254
  percent of GDP in, 346
  personnel scheduling for, 1743
  Theory of Constraints applications in, 559
  3Ts in, 559
  work injuries in, 1070
Service lives, 2332, 2350
Service mapping, 641
Service operations, 1121
Service processes, 643–644
Service products, 643
Service providers, manufacturers as, 532
Service quality, 1956–1965. *See also* Customer service
  approaches to defining, 625–626
  conceptual model of, 626–630
  and customer satisfaction, 640
  customer satisfaction vs., 628–629
  definition of, 1956–1957
  leadership focus on, 1958–1959
  measurement/evaluation of, 1963–1964
  measurement of, 640–641
  models of, 638–640
  process dimension, 641, 642
  SERVQUAL instrument for measuring, 627–630
  strategy for creating/maintaining, 1957–1958
  structure dimension of, 641, 642
  and systems for service delivery, 1961–1963
  workforce focus on, 1959–1961
Service-quality standards, 657–658
Service reports, 1308, 1309
Service robots, 379–381
  assistance/entertainment, 381
  cleaning, 380
  courier/transportation, 379–380
  medical, 381
  refueling by, 381
Service systems:
  assessment of, 645–649
    areas, assessment, 645–647
    and maturity model, 648
    procedure for, 648, 649
  queueing models for, 1633–1634
  structure of, 642–645
Service time:
  exponential, *see* Exponential service time
  general, *see* General service time
Service vehicles, 2062, 2063
Servomechanism, 157
SERVQUAL instrument, 627–630
Set-based engineering, 556
Set-theoretical operations (CAD), 183
Setup, 2312, 2314
  of automated inspection systems, 1901–1902
  of human inspection systems, 1894–1895
Setup costs, 2021
Setup planning, 455, 456
Setup standards (time study), 1427
SEU, *see* Subjective expected utility
Sewell Cadillac, 656
SGVs, *see* Self-guided vehicles
Shadowing (task-analysis technique), 1209
Shaping, 1322. *See also* Near-net-shape processes/production

  geometric capabilities of, 464
  technological capabilities of, 470
Shared mental models, 2208
Shared storage, 2092
Shared system architectures, 212, 213
Shared vision, 999
Sharp hypotheses, 137
SHEL system, 1135
Sherman Antitrust Act, 680
Sherman–Morrison–Woodbury theorem, 2284
Shewhart, Walter, 1828–1830, 1834, 1861
Shewhart control charts, 1818, 1835–1855, 1861–1875
  attribute data, charts for, 1844–1851
    classification data, P chart for, 1844–1847
    count data, C and U charts for, 1847–1850
  and AT&T runs rules, 1863–1864
  data patterns on, 1863
  for determining causes of variation, 1834–1855
    attribute data, charts for, 1844–1851
    construction, chart, 1839–1841
    groupings of data types, 1836–1837
    individual measurements, X chart for, 1841–1844
    interpretation, 1835–1836
    subgrouping, 1837–1839
    X-bar and R control charts, 1850–1855, 1864–1868
  individual measurements, X chart for, 1841–1844
  interpretation of, 1835–1836
  planning/construction of, 1839–1841
  for process capability analysis, 1869–1871
  and subgrouping/stratification, 1837–1839
  X-bar and R control charts, 1850–1855, 1864–1868
Shift schedules, *see* Personnel scheduling
Shift systems, design recommendations for, 1367
Shipment planning, 2063–2067
  case study involving, 2065–2067
  tactical/operational considerations in, 2064–2065
  and total shipping solution, 2065
Shock pulse, 1613
Shop-floor control:
  and artificial intelligence (AI), 1775–1782
  CIM Framework for, 1774–1775
  high-variety, 699–701
  and management system (MAS), 496, 497
Shortage costs, 2021
Short-cycle allowances, 1396
Short-cycle assembly machines, 359
Shortest path problem, 2572, 2574
Shortest setup time first (SST) rule, 1722, 1724
Shortest-travel-time-first (STTF) rule, 1511, 1513
Short-term memory store (STSS), 1015
Shouldice Hospital, 559
SHS, *see* Society for Health Systems
S-HTTP (Secure HTTP), 734
Shusa system, 556
SI, *see* Strain index
SIC (Standard Industrial Classificaiton), 329
Sideloader trucks, 1507–1509
Siebel Systems, 95
Siemens, 311, 312, 381
Signal conditioning elements (measurement systems), 1878
Signal-detection theory, 2185–2186
Signal generation (AVIS), 1904–1905
Signaling, price, 680, 681
Signal preprocessing (AVIS), 1905
Signal processing elements (measurement systems), 1878
Sign test (nono-parametric technique), 2259
Silicosis, 1169
SIMAN, 2446, 2455, 2456
SIMA Program, 326
SIMDRAW, 2459
SimEngine, 2460
Similarity philosophy (warehouse layout), 1540
Simple interest, 2336
Simple network management protocol (SNMP), 731
Simplex algorithm (for LP problems), 2527–2530
Simplex method:
  interior point method vs., 2534
  MC^2-, 2618–2620
Simplicity, in modeling, 1631
SIMPROCESS, 2461
SimRunner, 2447
SIMSCRIPT II.5, 2455–2456
SIMUL8, 2460
Simulated annealing, 1731, 2591
Simulation(s), 2469–2493. *See also* Computer simulation; Digital human modeling
  for construction/planning of robots, 378, 379
  as decision support tool, 2014
  definition of, 378
  digital computer, 2385
  fidelity of, and transfer of training, 932–933
  in flexible manufacturing systems, 503, 505–506
  in health care systems, 748
  for human strength design, 1054
  initial-condition bias in, 2477–2483
    direction of, 2479–2483
    remedial measures for, 2478–2479
  of material flow, 388
  measures of error in, 2483–2487
    confidence intervals, 2485–2487
    standard error, 2483–2485
  and metamodels, 2490–2492
  Monte Carlo, *see* Monte Carlo simulation
  multiple comparisons in, 2488–2490
  optimization/sensitivity of, 2487–2492
  point estimators in, 2475–2477
  as predictors for use in selection, 922
  for process design and reengineering (PDR), 1703–1704
  in product development, 207
  randomness in, 2472–2473
  of revised NIOSH lifting equation, 1079, 1080
  static, 2471

Simulation(s) (*Continued*)
  steady-state, 2471–2472
  for teamwork training, 934
  terminating, 2471
  as training technology, 929
  and transformability, 320
  variance reduction in, 2492–2493
Simulation Dynamics, 2460
Simulation methods, 128
Simulation models, 1630
  for client/server (C/S) system evaluation, 728–719
  for health care delivery systems, 745
Simulation statistics, 2473–2475
Simultaneous engineering (SE), 206–207, 556
  for efficiency in assembly, 369, 371, 372
  rapid product development vs., 1284–1287, 1285
Single-factor systems (job evaluation), 910
Single-level data models, 118–119
Single-minute exchange of dies (SMED), 547
Single-object process charts, 1374–1376
Single-shift scheduling, 1743, 1746–1755
Single-spindle automatic machinery, 467
Single-stage systems, queueing models for, 1635–1638
  make-to-order manufacturing/service, 1635–1636
  make-to-stock manufacturing/service systems, 1636–1638
Sinking fund factor (interest), 2339–2340
SIPs (state implementation plans), 590
SIRO (served in random order), 2157
Site selection and construction, 1465–1501
  architect, selection of, 1496–1499
  checklist, site selection, 1477–1489
  community, selection of, 1476–1477
  contractor, selection of, 1499
  and customer satisfaction, 1468–1469
  distribution network planning for, 1472–1475
  environmental factors in, 1489
  finalizing process of, 1490
  and free trade zones, 1489–1490
  linear programming applications for, 2056
  network analysis for, 1470–1475
  objectives of, 1465–1466
  pitfalls of, 1466
  and project delivery, 1491–1496
    construction management method, 1493–1494
    design-bid-build, 1492–1493
    design-build, 1494–1495
    team design/construct, 1495–1496
  and project management, 1499–1501
  responsibility, areas of, 1771
  and strategic master plan (SMP), 1469–1470
  and supply chain needs, 1467–1468
  team for, 1475–1476
  visitations, site, 1490
Situation rooms, 134, 135
Six-axis robots, 435, 436
Six-DOF robots, 413
Size philosophy (warehouse layout), 1540
SJAM, *see* Semiquantitative job analysis methodology
SJT, *see* Social judgment theory
Skate wheel conveyors, 1515–1517
Skill-, rule-, and knowledge-based (SRK) model, 1019–1021
Skill acquisition, 929–930
Skill-based behavior, 1019
Skill-based pay systems, 911
Skill-based performance, 2205, 2206
Skimming pricing, 675
Skoda, 212
SKU (stock-keeping unit), 2087
Slackness, complementary, 2554
Slat conveyor, 1515, 1517
SLDs (second-level domains), 243
Sleeping, day, 1367
SLIM-MAUD, 2192
SLP (successive linear programming), 2562
Small and medium-sized enterprises (SMEs), 286
"Small wins," 980
SMART (subjective multiattribute rating technique), 2195
SmartEcon.com, 266, 267
SmartFrog, 273
SMART goals, 1005, 1009
SMDs, *see* Surface-mounted devices
SMED (single-minute exchange of dies), 547
SMEs (small and medium-sized enterprises), 286
SMP, *see* Strategic master plan
Snapback timing, 1420
SNMP (simple network management protocol), 731
SOCAP (Society of Consumer Affairs Professionals), 657
Social democracy, 1186
Social environment, and human–computer interaction, 1217, 1220–1222
Socialization of employees, 857
Social judgment theory (SJT), 129, 2200
Social norms, and group decision making, 2209–2210
Social responsibility, 39
Social system (TQL), 1795
Social trends, 38–38
Societal values, and job evaluation, 901
Society for Health Systems (SHS), 739, 748
Society of Consumer Affairs Professionals (SOCAP), 657
Sociotechnical approach to job design, 874–875
Sociotechnical systems (STS) design, 964, 1889
Software, *see* Computer software
Software/hardware/environment/liveware (SHEL) system, 1135
Soldering, 412
  for 3D PCBs, 438–439
  in electronics production, 423–425, 429–431
  methods of, 423
  MIDs, 435
Solectron, 263, 264
Solicitation process, 1250
Solid freeform manufacturing, 586
Solid modeling, 182–184
Solid state metal, designing for, 1317, 1319
Solid waste management, 1571

Solution(s):
  feasible, 2528
  in nonlinear programming, 2541–2542
  optimal, 2528, 2536
Solution process, 201, 202
SOLVOPT, 2565
Sony, 49
Sortation conveyors, 1518–1520
Source reduction, 533
Source selection, 1250
South Afruca, 1968
Southwest Airlines, 559, 861
SOW document, *see* Scope of work document
Space planning:
  and acoustical control, 1200
  for warehousing, 1532–1538, 2088–2093
    alternative storage method space requirements, determining, 1535–1538
    materials to be stored, 1532–1534
    philosophy, storage, 1534–1535
Space standards (storage), 1537–1538
Space-utilization philosophy (warehouse layout), 1541
Span of control, 1264
Spare components, 1933
SPC, *see* Statistical process control
Specialization, 1264, 1354, 1355
Specialized diagnosis (service systems), 1634
Special-purpose assembly systems, 356
Specification, model, 2271–2272
Specification of General Requirements for a Quality Program (ANSI Standard Z1.8–1971), 1968
Specific predetermined time systems, 1429
Spectrometric oil analysis, 1613
Speculation (in channel structure theory), 2115, 2116
Speed, 147. *See also* Transmission speed
  as performance management metric, 1004–1005
  as source of competitive advantage, 2116
Speed rating (time study), 1423, 1424
SPENBAR, 2565
Spherical robots, 375
Spinning, 1320, 1321
Splitting, 2168–2169
Sponsors, project, 1337
Sponsorship:
  of business model development, 31
  executive, of ISE, 22–23
Spontaneous interaction, computer-supported, 143
Sprinklers, 1568
SQl, *see* Structured Query Language
SQP (successive quadratic programming), 2562
Square-root rule of inventory consolidation, 2071
S/R, 2087
SRCs (State Emergency Response Commissions), 594
SRK model, 1019–1021
SSL (secure socket layer), 734
SST rule, *see* Shortest setup time first rule
Stability, computer screen image, 1197
Stable processes, 1829–1831
Staffing:
  assignment, staff, 1247
  determining requirements for, 1742
  in health care delivery systems, 740, 742
Staging, data, 84
Stakeholders, 39
  identifying/determining needs of, 1301–1304
  informing, in process design and reengineering, 1706–1707
  in personnel scheduling, 1741
Standard deviation, *see* Variance
Standard for the Exchange of Product Model Data (STEP), 193–195, 1286
Standard Industrial Classificaiton (SIC), 329
Standard normal distributions, 2386
Standard Oil, 654
Standard oil analysis, 1614
Standard performance, 1422
Standard posture, 1062, 1063
Standards, 1565–1566
  air pollution:
    National Ambient Air Quality Standards (NAAQSs), 590, 592, 593
    New Source Performance Standards (NSPSs), 590–593
    Prevention of Significant Deterioration (PSD), 592
  for control functions, 1772
  for cryptosystems, 733
  electronic data interchange (EDI) transaction standards, 338
  for enterprise resource planning (ERP), 349–350
  for environmental management systems, 539
  ergonomic, 1166–1167
  for expense work, 1461
  Hazard Communication Standard (OSHA), 593
  for human–computer interface design, 1215–1216
  in human factors audits, 1133–1134
  for indirect work, 1459–1462
  for informing employees about hazards, 1176–1177
  ISO quality assurance, *see* International Standards Organization (ISO)
  job, 1449
  "J-standards," 1121, 1122
  for measurement systems, 1881
  for message transmission, 168–169
  National Ambient Air Quality Standards (NAAQSs), 590, 592, 593
  National Emission Standards for Hazardous Air Pollutants (NESHAPS), 593
  for networking technologies, 165–166
  New Source Performance Standards (NSPSs), 590–593
  for occupational safety and health, 1162, 1165–1168, 1185–1186
  OSHA Ergonomics Standard, 980
  Prevention of Significant Deterioration (PSD) standards, 592
  quality, *see* Quality management systems (QMS) standards
  SAE-J standards, 1121, 1122

Standards (*Continued*)
  SEMI Equipment Communication Standard (SECS), 165–166
  service-quality, 657–658
  setup, 1427
  temporary (time studies), 1427
  time, 1392–1394
    and changes in methods/working conditions, 1410. *See also* Predetermined time standards (PTS)
    documentation of, 1406
    engineering estimates, 1393
    nonengineered estimates, 1392–1393
    standard data, 1393–1394
  for working postures, 1068
  for workplace analysis/design, 1065–1068
Standards International, Inc., 1441
Standard time, 1394
Standard time data (for work measurement), 1413, 1443–1445, 1447–1448
  calculating, 1426–1427
  definition/uses of, 1412
  development of, 1447
  elemental level systems, 1445, 1447
  establishing, 1457, 1458
  motion-level systems, 1444–1446
  reliability of, 1443, 1444
  task-level systems, 1445, 1447
  uses of, 1447–1448
Standard work times, establishment of, 1457, 1458
Standing work postures, 1357, 1358
Stanford Research Institute, 238
Star schema, 85
State CIMS Engineering Research Center of China, 500
State Emergency Response Commissions (SRCs), 594
State health and safety agencies, 1164
State implementation plans (SIPs), 590
StatFit, 2446
Static analysis (biomechanics), 1069
Static efforts/work, 1052, 1053, 1056–1061
  arm, static efforts of, 1058–1062
  design limits for, 1056, 1057
  intermittent, 1057, 1058
  push/pull force limits, 1055
Static magazines, 383
Static scheduling, 497, 502, 503
Static simulations, 2471
Static standing forces, 1055
Static strengths, dynamic vs., 1052, 1053
Stationary points, 2546, 2547
Statistical estimation and inference, 2184–2187, 2242–2243
  Bayesian inference, 2184–2185
  in behavioral decision theory, 2196–2201
    biases, 2198–2199, 2201
    and human judgment models, 2200–2201
    and human limitations, 2196–2198
    selective processing, 2199–2200
  as decision support tool, 2013
  Dempster–Schafer method, 2186–2187
  and signal-detection theory, 2185–2186
Statistical experiments, 2225
Statistical hypothesis testing, *see* Hypothesis testing
Statistical method for determining sample size, 1452–1453
Statistical modeling, 2265–2267
Statistical process control (SPC), 1856–1861, 1856–1875
  control charts for, 1861–1875
    attribute data, charts for, 1871–1875
    and AT&T runs rules, 1863–1868
    data patterns on, 1863
    variables, charts for, 1864–1871
  in design and process platform characterization methodology, 1993–1995
  for M&TS, 1987
  tools for, 1856–1861
    cause-and-effect diagram, 1859, 1860
    check sheet, 1858–1859
    defect concentration diagram, 1860, 1861
    histogram, 1856–1857
    Pareto chart, 1859
    scatter diagram, 1860–1862
Statistical thinking, 20
Statistical tools, 782
Statistics, simulation, 2473–2475
Status meetings, 1347
Status reports, 1347–1349
Steady state, accuracy of measurement systems in, 1882–1884
Steady-state behavior:
  convergence to, 2162–2163
  determination of, 2161–2162
Steady-state simulations, 2471–2472
Steam systems, energy-improvement possibilities for, 1581
Steepest descent, method of, 2550
Steering committees, project, 1346, 1348
Steering components (automotive), assembly of, 389, 391
Steiltjes integral, 2147
Stencil printing, 425, 426
STEP, *see* Standard for the Exchange of Product Model Data
Stepwise method, 2290
Stereo lithography format (STL), 208
Stew Leonard's grocery stores, 656
Sticking (joining), 412–413
Stiff linkage, 415–416
Stimulation, environmental vs. task, 1357, 1358
STL (stereo lithography format), 208
Stochastic approximation, 2634–2635
Stochastic counterpart method, 2635
Stochastic decision trees, 2384, 2385
Stochastic models, 2146–2170
  benefits of mathematical analysis of, 2146
  definition of, 2146, 2150
  Markov chains, 2150–2156
    in continuous time, 2154–2156
    and Markov property, 2150–2151
    queueing model based on, 2153–2154

    - reversible, 2156
    - steady-state distributions of, 2152–2153
    - transition matrices in, 2151–2152
  - point processes, 2149–2150
  - and probability, 2146–2149
  - queueing models, 2157–2163
    - assumptions of, 2160–2161
    - ASTA/PASTA, 2163
    - bottleneck queues, 2162
    - long-run behavior, determination of, 2161–2162
    - Markovian queueing models, 2158–2159
    - $M/G/1$ queue, 2159–2160
    - notation for, 2157–2158
    - steady state, rate of convergence to, 2162–2163
    - variability, effects of, 2162
  - queueing networks, 2163–2170
    - decomposition methods, 2167–2170
    - general product-form networks, 2165–2167
    - Jackson networks, 2164–2165
  - randomness in, 2146
- Stochastic programming, 2628–2636
  - likelihood ratio method of derivative estimation, 2633–2634
  - perturbation analysis in, 2632–2633
  - with recourse, 2629–2631
  - sampling methods for, 2631–2632
  - simulation-based optimization methods in, 2634–2636
- Stock-keeping unit (SKU), 2087
- Stockless production, 545
- Stock levels, safety, 1683
- Stockrooms serving manufacturing facilities, 2086
- Stocks, 764
- Stock selection, 452, 453
- Stopped arrival queues, 1637
- Stopwatches, 1411, 1412, 1414
- Stopwatch time studies, 1393, 2307
- Storage analysis chart (SAC), 1532–1534
- Storage equipment, automated, 156
- Storage and storage systems. *See also* Warehousing; Warehousing operations
  - in food service kitchens, 833–834
  - magazines for, 383, 384
  - and material handling, 1520–1523
    - block stacking, 1520–1521
    - cantilever rack, 1523
    - pallet flow rack, 1522
    - permanent racks, 1521, 1522
    - portable racks, 1521
  - in retail supply chains, 778–779
- Stored knowledge, 215
- Storefronts, Web, 265–266
- Stores, data, 99
- Storyboards, 1215–1216
- Straddle trucks, 1506, 1508, 1509
- Strain index (SI), 1087–1090
- Strategic alliances, 48
- Strategic analysis, 51–52
- Strategic business processes, 30
- Strategic change management, 968
- Strategic controls, 46, 48
- Strategic management process, 41–43, 58
- Strategic master plan (SMP), 1469–1470, 1530–1532
- Strategic planning, 111, 112, 135, 136
  - in EPEM model, 1798
  - quality/usefulness of information for, 136–137
  - and technology design, 954
- Strategy(-ies), 13
  - for advanced planning and scheduling implementation, 2051
  - for energy management in plant engineering, 1577–1582
  - of experiments, 2238–2239
  - for job design/redesign, 884–885
    - and basic decisions, 888–889
    - existing jobs, redesign of, 885
    - initial design, 884–885
  - for plant engineering, 1557
  - service, 1957–1958
  - for team design, 884–885
    - and basic decisions, 888–889
    - initial design, 884–885
  - user, 1024, 1025
- Strategy and positioning, 9–11
- Stratification (control charts), 1838–1839
- Stratified random sampling, 1136, 1456
- Stratospheric Ozone Protection Program, 593
- Strength:
  - digital human modeling for assessment of, 1116, 1118
  - human, *see* Human strength
  - of products, 454–455
- Stress:
  - and computer technologies, 1222–1223
  - and decision making, 2208, 2209
  - electronic monitoring's effect on, 1226, 1227
  - with introduction of computer technologies, 1227
- Strictly convex functions, 2543
- Structural data models, 120–122
- Structural organizations, 284
- Structured analysis and design technique (SADT), 304, 507, 508
- Structured English, 100–102
- Structured programming (computers), 71
- Structured Query Language (SQL), 81–82
- STS design, *see* Sociotechnical systems design
- STSS (short-term memory store), 1015
- STTF rule, *see* Shortest-travel-time-first rule
- Studentized range statistic, 2261
- Subclasses (computer programming), 72, 291
- Subcontracting, 263
- Subcritical instability, 2167
- Subjective belief form (decision making), 2191
- Subjective expected utility (SEU):
  - axioms of, 2178, 2179
  - and behavioral decision theory, 2202
  - decision rule, 2177
  - and human preference/choice, 2201–2202
  - and prospect theory, 2203
  - theory of, 2182–2183

Subjective forecasting models, 793
Subjective multiattribute rating technique (SMART), 2195
Subnetworks, 235, 237
Suborder, 2087
Subprocesses, business, 40, 44
Substitutability, product, 2130
Success factors:
  in new technology implementation, 950
  for teams, 981–983
  in total quality leadership (TQL), 1804–1805
Successive linear programming (SLP), 2562
Successive quadratic programming (SQP), 2562
Sufficient conditions, 2546–2547
Summative evaluation, 934–936
Sun Microsystems, 78, 499
SunNet Manager (Sun Microsystems), 732
Superclasses (computer programming), 72, 291
Superfund, *see* Comprehensive Environmental Response, Compensation and Liability Act
Superfund Amendment Reauthorization Act (SARA), 594–595
Superposition, 2169
Supervised-learning neural networks, 1778–1779
Supervision, variation in, 1832
Supplier networks, lean, 555–556
Supplier performance management, 1799
Supplier relationship management, 337, 2134–2138
Suppliers, power of, 39
Supply chain(s), 1690, 2110, 2115–2125, 2127–2133. *See also* Transportation management; Warehousing
  and business processes, 2118–2125
    customer order-fulfillment process, 2121–2122
    customer relationship management process, 2121
    customer service management process, 2121
    demand management process, 2121
    information flow, 2124
    links, business process, 2118–2120, 2123–2124
    manufacturing flow management process, 2122
    procurement process, 2122
    product development/commercialization, 2122
    returns process, 2122
  and channel structure, 2115–2116
  design of, 2127–2131
    customer service objectives in, 2130–2131
    manufacturer's perspective on, 2127–2128
    market coverage objectives in, 2128–2129
    and product characteristics, 2129–2130
    retailer's perspective on, 2128
    wholesaler's perspective on, 2128
  identifying members of, 2117
  logistics models for planning, 2009, 2010
  mapping of, 2120
  network structure of, 2114, 2116–2120
    business process links, 2118–2120
    mapping, 2120
    and members of supply chain, 2117
    structural dimensions, 2117–2118
  outsourcing pieces of, 2115
  primary/supporting members of, 2117
  and queueing models, 1634
  reengineering of, 2132–2133
  retail, *see* Retail supply chains
  and site selection, 1467–1468
  and transportation, 790–791
Supply Chain Council (SCC), 348
Supply chain information system (SCIS), 318
Supply chain integration (SCI), 348
Supply chain management (SCM), 2111–2115
  and benefits of modeling, 306
  and bullwhip effect, 546
  business processes in, 2120–2123
    customer order-fulfillment process, 2121–2122
    customer relationship management process, 2121
    customer service management process, 2121
    demand management process, 2121
    links, business process, 2118–2120
    manufacturing flow management process, 2122
    procurement process, 2122
    product development/commercialization, 2122
    returns process, 2122
  common understanding of, 348
  definition of, 94, 2111–2112
  electronic data interchange (EDI) transaction standards, 338
  and enterprise resource planning (ERP), 348
  enterprise resource planning (ERP) interface with, 94–95, 350
    execution, 338
    operations interfaces, B2B, 343
    planning, 338
  and ERP, 94–95
  in health care systems, 748
  integrated, 2133–2134
  logistics vs., 2111–2115, 2112–2115
  management components of, 2125–2127
  performance measurement with, 2131–2132
  planning, 327, 328
    at operational level, 329
    at strategic level, 327
    at tactical level, 327, 329
  reengineering of, 2132–2133
  and supplier relationships, 2134–2138
  supply chain integration (SCI) vs., 348
  and transportation management, 2055–2056
  transportation management in, 2057–2058
  and Web-based procurement, 262–263
Supporting systems, 45
Support processes:
  automation of, 1260
  project management, 1254, 1255
Support tasks, budgeting for, 1344
Surface modeling, 180–182

Surface-mounted devices (SMDs), 423–425, 435–438
  optimized MID placement system, 436–438
  six-axis robot system, 435, 436
Surfaces, work, *see* Working surfaces
Surveillance:
  fleet/field tests operational evaluation tests, 1943
  in public health approach to safety/health, 1157–1158
  for reduction of work-related musculoskeletal disorders, 1095–1097
Surveys, 127, 1209, 1811, 1813
Sustainable development, 533
Svenka MTM Grupen, 1436
Swaging, 565, 570, 577–579
Sweden:
  quality management systems in, 1185
  social democracy in, 1186
Swedish Environmental Institute, 531
Sweeping (solid modeling), 183
Swift, 654
Swivel/tilt, of computer screen, 1197–1198
Symantec Visual Caf–126 for Java, 304
Synchronization, sequencing vs., 2036–2037
Synthetic rating scales (time study), 1423
SyRS (systematic random sampling), 1456
System(s). *See also specific systems*
  definition of, 280, 489
  effectiveness of, 1922
  energy, 1574–1575
  quality of, 1797
  reliability of:
    and effectiveness, 1922
    and employee participation, 976
    models for, 1932–1937
  successful performance/failure of, 1927
  tools for viewing, 1809, 1810
Systematic random sampling (SyRS), 1456
Systematic structure of models, 284
System design phase (CIM), 514, 515
System effectiveness (term), 1922
Systems development life cycle (SDLC), 96–106
  analysis phase, 97, 99, 105
  computer aided software engineering, 105
  data dictionaries, 102–103
  data flow diagrams, 99–101
  design phase, 97
  entity relationship diagrams, 102, 103
  feasibility analysis, 98–99
  Gantt charts, 103–104
  implementation phase, 97
  joint application deployment, 105
  PERT diagrams, 104
  planning phase, 96–98, 103
  prototyping, 104, 105
  rapid application deployment, 104, 105
  Structured English, use of, 100–102
  use and maintenance phase, 97
Systems ecology, 146
Systems engineering:
  basic phases/steps in, 126
  models/modeling, 126–129
System for Aiding Man-Machine Interaction Evaluation (SAMMIE), 1049, 1050, 1112
System management (client/server systems), 729–730
System reliability models, 1932–1937
  fault tree analysis, 1936–1937
  reliability block diagram, 1933–1936
Systems for Integrating Manufacturing Applications (SIMA) Program, 326
Systems thinking, 999

Table of random numbers, 2386, 2387
Tables:
  heights of (food service kitchens), 834
  in relationship database model, 80
  work, 1203
Taboo search, 1731
Tabu algorithms, 2591
Tabu search, 800–801
Taco Bell, 559
Tactical naval decision making system (TANDEM), 922
Tactile sensors, 385
Taguchi method, 2232
TAKD, *see* Task analysis for knowledge description
TANDEM (tactical naval decision making system), 922
Tapping:
  geometric capabilities of, 464
  technological capabilities of, 471
Target costing, 207
Task allocation process, 1210–1211
Task analysis:
  cognitive task analysis vs., 1025
  contextual, 1206–1211
  hierarchical task analysis (HTA), 1028, 1029, 1909–1912
  as technique for training, 926
  videotaping for, 1371–1373
Task analysis for knowledge description (TAKD), 1208, 1209
Task and deliverables list, 1339–1341
Task list, 1339–1341
Task network models, 2413–2429
  crew workload, evaluation of, 2420–2427
    future command and control process, modeling workload of, 2421–2425
    other environments, extension to, 2424–2427
  design issues, 2420
  elements of, 2414–2419
  new task environments, extension of findings to, 2427–2429
  process control operator example, 2419–2420
Task/operator/machine/environment (TOME) system, 1135
Task relevance (of information), 140
Task repetitiveness, 1092
Tasks:
  cognitive, *see* Cognitive task(s)
  combining, in job design/redesign, 885–886
  constant, 740

Tasks (*Continued*)
  decoupled, 1356–1357
  definition of, 40
  and ergonomics interventions, 1195
  error reduction by simplification of, 1370
  interdependence of, 887
  jobs vs., 869
  and occupational safety and health, 1160–1161
  physical, *see* Physical tasks
  and safety hazards, 1178
  in service systems, 1633–1634
  size comparability of, 1341
  and team effectiveness, 985
  team members, tasks performed by, 877
  of teams vs. individuals, 987–988
  variable, 740
Task stimulation, environmental vs., 1357, 1358
Taxation issues, 764–766
Taxonomy, 461
Taylor ED, 2460–2461
Tbps, 232
TCP, *see* Transmission Control Protocol
TCP/IP, *see* Transmission Control Protocol/Internet Protocol
TD(δ), 1780
TDKA (trace-driven knowledge-acquisition) methodology, 1776
TDM (time division multiplexing), 231
Teaching methods, 928
Team(s), 975–989
  advantages of using, 976
  characteristics of, 977
  and collective intelligence, 976
  design of, *see* Team design
  for development of business model, 31
  effectiveness of, 983–987
    outcome variables affecting, 987
    process variables affecting, 985–987
    structure variables affecting, 983–985
  for ERP choice/implementation, 92–94
  impact of, 987–898
    on employees, 98–99
    on management, 988
  maintenance, 1590
  negative consequences of, 988–989
  and participatory ergonomics, 980–981
  in plant engineering, 1557
  for professional services projects, 1335
  project, *see* Project teams
  and quality improvement, 978–980
  quality improvement teams (QITs), 748
  redesigning existing, 885
  reengineering, 1707
  site selection, 1475–1476
  success factors for, 981–983
  total quality leadership (TQL), 1802–1803
  training of, 933–934
  types of, 976–977
Team coordination training, 934
Team design, 870, 877–882, 977, 1495–1496
  advantages of, 880–881
  combining tasks in, 885–886
  development of, 877
  disadvantages of, 881–882
  evaluating need for, 882–884
  evaluation of, 899–884
    biases, potential, 893
    and data sources, 892
    example of, 893–894
    long-term effects, 892–893
    with questionnaires, 889–892
  existing teams, redesigning, 885
  and individual differences, 886–888
  input factors in, 878, 879
  output factors in, 880
  process factors in, 879
  strategies for, 884–885
    and basic decisions, 888–889
    initial design, 884–885
Team-development process, 1247
Team leadership, 2208
Team learning, 999
Team meetings, 143
Team mind, 2208
Team-oriented project planning system (TOPP), 1287–1288
Team performance assessment technology (TPAT), 922
Team-performance measurement, 934
Teams incorporating distributed expertise (TIDE), 922
Team task analysis, 934
Teamwork, 975–976, *see* Simultaneous engineering
  and individual rewards, 862
  negative consequences of, 988
  in total quality leadership (TQL) process, 1802
Teamwork integration evaluator (TIE), 606–608
Teamwork simulation exercises, 934
Teamwork (software package), 173
Technical Committee on Musculoskeletal Disorders (IEA), 1067
Technical system (TQL), 1795, 1796
Technical University of Berlin, 441
Technique, definition of, 1135
Technological, organizational, and people (TOP) changes, 949, 953
Technological capability (of process), 457
Technology-based work breakdown structure, 1269
Technology(-ies). *See also specific topics*
  comprehensive solutions needed for implementation of new, 953–954
  context specificity of new, 953
  and culture, 956–961
  decision aids, use of, 965–968
  estimating costs for, 2298
  failures, implementation, 949–952
  and malleability of factors, 952–953
  and occupational safety and health, 1160
  organizational change and implementation of new, 949–969
    agreement on change process for, 963–965
    breadth of factors in, 955–956
    and business purpose, 955
    comprehensive solutions needed for, 953–954
    context specificity of, 953

    - as cross-functional problem, 953
    - and culture, 956–961
    - decision aids, use of, 965–968
    - failures, implementation, 949–952
    - iCollaboration software, 966–968
    - innovation, encouragement of, 961–963
    - and length of planning cycle, 954
    - malleability of factors, 952–953
    - role of compromise in, 954
    - TOP Modeler system, 965–966
    - and unpredictability of new technology, 952
  - for plant engineering, 1572
  - quality of, 1797
  - in training, 928, 929
  - for transportation management, 2056–2057
  - and unpredictability of new technology, 952
- Technology integration and management, 1798, 1799
- "Technology spirit," 952
- Tecnomatix, 167
- Telediagnosis, 432
- Teleimmersion, 251
- Telemedicine, 251
- Telephone workers, 1744
- Telephony extensions, 142
- Telepresence, 251
- Telescoping belt conveyors, 1516
- Teleteaching (teleinstruction), 251
- Teleworking, 235, 250, 1217, 1220, 1222
- Telnet, 240
- Temperature, work area, 1200
- Temperature transducers, 1903
- Template juggling (layout design), 1538–1539
- Temporary (ad hoc) teams, 976, 982
- Temporary workers, 1745
- Ten-category scheme for interpreting verbal protocols of managers, 1035
- Ten Commandments for Experimental Design, 2228–2229
- 10 net addresses, 238
- TERs, *see* Time-estimating relationships
- Terabits per second, 232
- Teriminating simulations, 2471
- Terminology standard ISO/DIS 9000–2000, 1966–1967
- Territory Planner (software), 2064
- Test and inspection, 1887–1916, 1942–1944
  - automation in, 1900–1907
    - equipment, 156
    - image processing, 1904–1907
    - materials handling, 1902
    - sensing, 1902–1904
    - setup, 1901–1902
    - signal processing, 1904
  - and decision making, 1890
  - distribution of, 1889–1890
  - of electronics products, 431, 432
  - and global business environment, 1887–1889
  - for hazard identification, 1171–1173
  - human role in, 1894–1900
    - decision, 1896–1899
    - and job design, 1899–1900
    - present, 1895
    - respond, 1899
    - search, 1895–1896
    - setup, 1894–1895
  - human vs. automated, 1892
  - hypothesis testing, *see* Hypothesis testing
  - and information technology, 1889
  - job design, 1899–1900
  - logical function allocation in, 1912–1916
  - logical structure of, 1892–1893
  - maintenance inspection, 1908–1912
  - as measurement issue for successful design, 1299, 1301
  - mission/function in, 1892–1893
  - nonproduction, 1907–1908
  - OSHA inspections, 1162–1163
  - reliability program applications during, 1953–1954
  - scales for, 1890–1892
  - systems design for, 1914–1916
- "Tests for Instability" (AT&T runs rules), 1863
- Texas Instruments, 654, 966
- Text filtering software, 142
- TFNs, *see* Triangular fuzzy numbers
- THDs, *see* Through-hole devices
- Theoretical expected value, 2390
- Theoretical scheduling, real-life vs., 1732, 1733
- Theory of Constraints (TOC):
  - and JIT, 557–558
  - and just-in-time (JIT), 557–558
- Thermoforming, 1325
- Thinking, opportunistic, 1024
- Third-party intermediaries, 277
- Thixocasting, 568, 584, 585
- Thixoforging, 568, 584–586
- 3D computer aided design (CAD), 180–183
  - solid modeling with, 182–183
  - surface modeling with, 180–182
  - 2D CAD vs., 178
- 3D printed circuit boards (3D PCBs):
  - placement systems for, 423
  - soldering, 438–439
- 3D SSPP, *see* Three Dimensional Static Strength Program
- 3M, 7, 263, 2120, 2122
- 3T's, 552–555, 558
- Three Dimensional Static Strength Program (3D SSPP), 1054, 1118
- Three-level meetings, 13
- Three Mile Island, 875, 883
- Three-stage flow lines, queueing models for, 1640–1642
- Thresholds, price, 669–670
- Through-hole devices (THDs), 423, 425
- Throughput, 726, 1631
- Throughput capacity (trucking), 1509, 1510
- TIDE (teams incorporating distributed expertise), 922
- TIE, *see* Teamwork integration evaluator
- Tier (term), 342
- TIGER/line files (GIS), 2017, 2018
- Time, 1392–1407
  - allowances affecting, 1394–1400
    - delay allowances, 1398, 1400
    - fatigue allowances, 1394–1400
  - for learning, 1400–1406
    - individual learning, 1400

Time (*Continued*)
    organization learning, 1400–1406
  normal, 1394
  observed, 1394
  as performance management metric, 1004–1005
  standard, 1394
  standards for, 1392–1394
    documentation of, 1406
    establishment of, 1392
    maintenance of, 1407
    predetermined standards, 1427–1446, 2307
    using, 1406–1407
Time division multiplexing (TDM), 231
Time estimates, for professional services projects, 1341, 1342
Time-estimating relationships (TERs), 2302, 2308
Time horizons, 766–767, 954
Time limits:
  and team performance, 982
  for working postures, 1063–1064
Time logs, 1393
Time management. *See also* Scheduling
  automation of, 1256, 1257
  project, 1245
Time pressure, 2208–2209
Time-recording equipment, 1411, 1412, 1414
Time required to perform task, 1120
Time series forecasting, 128
Time slotting, 1459–1461
Time span of discretion (TSD), 910
Time study, 1411–1427, 2307
  allowances to basic time, 1426–1427
  definition/uses of, 1412
  equipment for, 1411, 1412, 1414–1416
    board, time study, 1414
    forms, time study, 1414–1416
    requirements for effective, 1414, 1417
    time-recording equipment, 1411, 1412, 1414
  fair day's work, determination of, 1411
  procedure for conducting, 1417–1425
    breakdown of job into elements, 1418–1419
    methods of timing, 1420
    and number of cycles, 1419–1421
    operator, choice of, 1417–1418
    performance rating, 1422–1425
    variations, 1420, 1422
Time study boards, 1414
Time study forms, 1414–1416
Time to failure random variable, 1928
Time-to-market, 2116
Time value of money, 2334. *See also* Inflation
Time-varying demand (scheduling), 1752–1755
Time window, 2087
Timing:
  in advanced planning and scheduling (APS), 2047
  methods of, 1420
  uncertain, cash flows with, 2369–2371
TIPs (Treasury inflation-protected securities), 761
Titanic Auto Production Company case study, 2319–2329
Tivoli Management Environment (TME), 732
TLDs, *see* Top-level domains
TLS (Transport Layer Security), 734
TME (Tivoli Management Environment), 732
TMS, *see* Transportation management systems
TOC, *see* Theory of Constraints
Tolerance charting, 472–473
TOME (task/operator/machine/environment) system, 1135
Tool management, 497
Tool–oriented technologies, 169–174
Tool selection, 457–459
Tool systems, automated, 500
TOP changes, *see* Technological, organizational, and people changes
Top-down networking, 254
Top-level domains (TLDs), 242, 243
TOP Modeler, 965–966
TOPP, *see* Team-oriented project planning system
Torque sensors, 385
Total demand constraint (scheduling), 1748
Total-enclosure concept, 598
Total productive maintenance (TPM), 551–553, 1557, 1619–1620
  and concurrent TQM implementation, 555
  and JIT implementation, 553–555
  just-in-time (JIT) vs., 553
  principle activities of, 553
Total quality improvement process (TQIP), 1793–1795
Total quality leadership (TQL), 1793–1805
  and eight pillars of quality, 1796–1798
  PADER scoring system for use in, 1800–1801
  steps for implementation of, 1801–1803
  success factors in, 1804–1805
  system areas for, 1795–1796
  tools for, 1798–1801
Total quality management (TQM), 19, 551–552, 1699, 1793, 1794, 1889
  and concurrent TPM implementation, 555
  and employee assessment systems, 938
  in health care systems, 748
  and improved safety/health, 1184
  and JIT, 552–555
  maxims expressing, 552
  process design and reengineering within, 1712
Total quality (TQ), 1793
Tote, 2087
Tower Records, 263
Tow-line conveyors, 1517
Toxic Releases Inventory (TRI), 594
Toxic Substances Control Act (TSCA), 1164
Toyota Motor Corporation, 16, 37, 493, 544, 551, 555–557, 783, 976
Toyota Production System (TPS), 544–545, 1502
Toys "R" Us, 778
TPAT (team performance assessment technology), 922

TPM, *see* Total productive maintenance
TPM Excellence Award, 555
TPQM, 553
TPS, *see* Toyota Production System
TP (transaction processing) monitor, 723
TQIP, *see* Total quality improvement process
TQL, *see* Total quality leadership
TQM, *see* Total quality management
TQ (total quality), 1793
Trace-driven knowledge-acquisition (TDKA) methodology, 1776
Trading markets, 275
Tradition, perceived, 958
Traditional organizations, customer-driven vs., 1797
Traffic (on Internet), 232
Traffic flow, in work areas, 1177–1178
Training, 924–937. *See also* Education; Learning
  and acquisition of learned information, 929–930
  appropriateness of, 925
  Baldrige criteria for, 1960
  computer-based, 222
  for computer simulation software, 2448
  for creating service-driven workforce, 1959–1961
  of customer service employees, 659
  design/development of, 926–927
  development of, 926–927
  development vs., 937
  education vs., 924, 925
  evaluation of, 934–937
  future of, 940
  for industrial engineers to become plant engineers, 1553–1557
  instructional methods/techniques for, 928–929
  with introduction of computer technology, 1226
  ISD model for, 926–927
  job analysis as prerequisite to, 926
  of job evaluators, 913
  and leadership, 859–861
  needs analysis for, 926
  organizational analysis as prerequisite to, 925–926
  outcome of, 924
  and process design and reengineering (PDR) implementation, 1711
  and productivity, 1888
  purpose of, 924
  and retention of information, 930–931
  safety, 1180–1183
  simulations, use of, 929
  system approach to, 926
  task analysis as technique for, 926
  of teams, 933–934, 985
  and team success, 982
  technology in, 928, 929
  as TQL success factor, 1804, 1805
  transfer-of-training, 931–933
  videotaping for, 1371–1373
Transactional leadership, 841–844
  calculative–rational basis of, 845
  extrinsic motivation in, 846
  individualistic orientation in, 846–847
Transaction management:
  distributed, 721–723
  and enterprise resource planning (ERP), 332–336
    accounting, 336
    finance/management, 336
    human resource management, 335–336
    maintenance management, 334
    manufacturing management, 333
    materials acquisition, 332–333
    materials inventory, 332
    order entry/tracking, 333
    process specification management, 333–334
    transportation, 335
    warehousing, 334–335
  functions supported by ERP, 332–336
  objective of, 347
Transaction processing (TP) monitor, 723
Transactions, four ACID properties of, 721–723
Transcendent approach to service quality, 625, 638, 639
Transducers, continuous, 1903
Transfer function (of measurement systems), 1884–1885
Transfer lines, 1632–1633, 1645–1650
  infinite inventory banks, 1646
  multiple-stage transfer line, 1648–1650
  no inventory banks, 1645–1646
  two-stage synchronized line, 1646–1648
Transfer molding, 1324, 1326
Transfer-of-training, 931–933
Transfer systems (assembly systems), 375–358
Transformability, 313
Transformable structures, 314–322
  and corporate network capability, 314–317
  and market turbulence, 314
  methods for planning/operating, 317–322
    integrated evaluation tool, 321, 322
    integrated simulation, 320
    participative planning, 320, 321
    process management, 317–319
  and semiautonomous organizational units, 315–317
Transformation, data, 84, 85
Transformational leadership, 843–845
  collectivistic orientation in, 846–847
  emotional–expressive basis of, 845–846
  intrinsic motivation in, 847–848
Transform Technologies, 1050
Transition matrices, 2151–2152
Transmission Control Protocol/Internet Protocol (TCP/IP), 235, 238
  packet switching, 239
  protocol layers in, 240
Transmission Control Protocol (TCP), 240, 245
Transmission speed:
  Internet, 235, 236
  network, 213–232, 234, 236–237, 249
Transorganizational information systems, 69–70, 107

Transparency, of models, 1631
Transport, 788
Transportation, 788–822, 1459
  automated, 156
  definition of, 788
  driver scheduling, 812–817
    column-generation methodology, 814–815
    definition of problem, 813
    generation of schedules, 816
    iterative process for optimizing, 815–816
    set-partitioning formulation with side constraints, 813–814
  and driver scheduling, 812–817
    column-generation methodology, 814–815
    definition of problem, 813
    generation of schedules, 816
    iterative process for optimizing, 815–816
    set-partitioning formulation with side constraints, 813–814
  enterprise resource planning (ERP) function, 335
  functions associated with, 789–790
  of goods, 791–793
  and information, 819
  intelligent transportation systems (ITS), 819, 822
  and large-scale network planning, 803–812
    definition of problem, 804
    modeling for, 804–806
    network design formulation, 806–807
    package-routing problem, 807–810
    subgradient optimization algorithm, 811–812
    trailer-assignment problem, 810–811
  parameters associated with, 789
  pickup and delivery operations, 793–803
    heuristic construction algorithms for modeling, 795–801
    preassigned routes/territories, 801–803
    VRPTW modeling of, 794–795
  planning, transportation, 792–793
  planning for, 789–790, 792
  quality in, 817–818
  in retail supply chains, 777–778
  and role of industrial engineer, 790
  routing, vehicle, 819–821
  and supply chain, 790–791
  as a system, 788–789
  work injuries in, 1070
Transportation industries, 346
Transportation management, 2054–2057
  location problems in, 2067–2068
  optimization problem in, 2054–2055
  and shipment planning, 2063–2067
    case study involving, 2065–2067
    tactical/operational considerations in, 2064–2065
    and total shipping solution, 2065
  and supply chain management, 2055–2056
  technology requirements for, 2056–2057
Transportation management systems (TMS), 2057–2063
  electronics industry case study, 2059–2060
  pickup/delivery/routing in, 2058, 2059
  and traveling salesman problem, 2060–2062
  and vehicle routing problems, 2062–2063
Transportation planning, logistics models for, 2011
Transportation problem (network flow models), 2570–2571, 2574
Transportation resources planning and management (TRPM) systems, 2064
Transportation robots, 379–380
Transport Layer Security (TLS), 734
t ratio, 2278–2279, 2284
Traub AG, 302
Traumatic injuries, 1168, 1169–1170
Traveling salesman problem (TSP), 794, 2060–2062, 2573
Traveling salesman problem with time windows (TSPTW), 794, 2062
Travel time (trucks), 1509–1513
Treasure Chest Model, 9
Treasury inflation-protected securities (TIPs), 761
Treatments (in experimental design), 2226
Tree diagrams, 1817, 1820, 2591–2592
Trend chart, 1817
Trend extrapolation, 128
Treplanning, 1323
Triangular fuzzy numbers (TFNs), 1781–1782
TRI (Toxic Releases Inventory), 594
Trolley conveyors, 1517–1518
TRON, 2565
Troughed belt conveyors, 1516
TRPM (transportation resources planning and management) systems, 2064
Trucks, industrial, *see* Industrial trucks
Trunk, posture checklist for, 1366
TRUSTe, 269
TSCA (Toxic Substances Control Act), 1164
TSD (time span of discretion), 910
TSP, *see* Traveling salesman problem
TSPTW, *see* Traveling salesman problem with time windows
Tuples (relationship database model), 80
Turning, 1322
  cost of machinery for, 467
  geometric capabilities of, 464
  obtainable accuracy values, 565
  technological capabilities of, 469
Turning time, calculation of, 460
Turnover and startup stage (project life cycle), 1242
Turret trucks, 1507, 1509
2D computer aided design (CAD), 178–180, 190
2 1/2 computer-aided design (CAD) systems, 180
Two-dimensional arrays, 1904
Two-stage flow lines, queueing models for, 1639–1640
Two-stage synchronized line, queueing models for, 1646–1648
Two-tailed hypothesis tests, 2247
Two-way tables, 1820, 1822, 1826

Uarco, Inc., 1713
U charts, 1844, 1847–1849, 1851
UILS, *see* Universal indirect labor standards
Ultrasonic detection, 1614–1615
Ultrasonic machining, 1323
Ultrasound sensors, 385, 386
UML, *see* Unified Modeling Language
UMTRCA (Uranium Mill Tailings Radiation Control Act), 1153
Uncertainty:
  interval of, 2548
  optimization under, 2625–2628
Uncertainty avoidance (in national cultures), 957, 958, 960
Unconstrained optimization (nonlinear programming), 2546–2553
  classical methods, 2546–2547
  conjugate gradient methods, 2552–2553
  golden section method, 2547–2549
  line search techniques for, 2547
  multidimensional search techniques for, 2549–2552
Underground storage tanks (USTs), 1489
Understanding:
  business model as aid to, 30
  model development for, 1630
Unemployment, 344
Unicast addressing, 242
Unified Modeling Language (UML), 291–293, 644–645
Unions:
  and job evaluation, 913
  and joint union/management ergonomic committee, 1187
United Kingdom:
  industrial robots in, 373
  quality standards in, 1968
United Parcel Service, 266, 2115
U.S. Army, *see* Anthropometric Survey
United States/Canada MTM Association, 1437
U.S. Department of Commerce, 793
U.S. Department of Defense, 238, 1243, 1967
U.S. Department of Transportation, 1592
U.S. National Academy of Sciences, 1195
U.S. Postal Service (USPS), 1520
U.S. quality management systems (QMS) standards, 1967–1968
U.S. Treasury, 273
U.S. Treasury securities, 273, 274
Unit loads, 1503–1504
Unit method (of cost estimating), 2301
Units:
  of analysis, 18–19
  experimental (in experimental design), 2226
Universal indirect labor standards (UILS), 1459–1460
Universal Resource Locators (URLs), 244, 245
University of Arkansas, 1860, 1861
University of California at Irvine, 174
University of California at Los Angeles, 238
University of California at Santa Barbara, 238
University of Connecticut, 739
University of Michigan, 1062
University of Michigan Center for Ergonomics, 1118
University of Saarland (Germany), 290–291
University of Southern California Information Sciences Institute, 1112
University of St. Gallen, 217–218
University of Utah, 238
Unnecessary work, 1459
Unpaced lines models of, 1639
Unrelated events, 2147
Unstable processes, 1830
Unsupervised neural networks, 1779–1780
Unsustainability, 997
Upper-Extremity Checklist, 1143–1144
Upper-extremity cumulative trauma disorders, 1365
Upper extremity musculoskeletal disorders, 1224
Uranium Mill Tailings Radiation Control Act (UMTRCA), 1153
URLs, *see* Universal Resource Locators
Usability:
  in human digital modeling, 1123
  in human factors audits, 1135
  operational goals for, 1212
Usability engineering, 1193
Usability evaluation of human–computer interaction, 1216–1220
USAF Aerospace Medical Research Laboratory Crew Systems' Interface Division, 1112
User-based approach to service quality, 626, 638, 639
User interface(s). *See also* Human–computer interaction
  consistency in, 133–134
  desirable characteristics in, 134
  DGMS as, 132
  external-to-ERP, 343
  internal-to-ERP, 343
  for models, 1631
  top-level attributes of, 134
User interface languages, 119
User profiles (contextual task analysis), 1207–1208
User strategies, 1024, 1025
Use statements (business model), 31
U-shaped assembly lines, 547, 548
USPS (U.S. Postal Service), 1520
USTs (underground storage tanks), 1489
Utility deregulation, 1577, 1580
Utility function assessment (decision analysis), 2193–2194
Utility models, 2392
Utility theory, 2392
Utilization (in space planning), 2088, 2089

Vacuum grippers, 414
Validation, model, 2272
Validity:
  in human factors audits, 1134
  as measurement issue for successful design, 1299, 1301

Value(s), 15–16
  adding, for customer satisfaction, 653
  buyers' perception of, 669
  conceptualization of, 629
  in context of knowledge management, 214
  dimensions of national, 957
  expected value, maximization of, 2181
  lifetime, of customers, 651, 652, 654
  optimal, of linear program, 2528
  perceived, 671
  product, 2129
  range of, 1879
  societal, 901
  of warehousing, 1528
Value-adding chain:
  holistic view of, 404, 405
  significance of assembly in, 402
Value-based approach to service quality, 626, 639–640
Value chains, 52–54. *See also* Supply chain(s)
  and design by customers, 702
  "value web" vs., 262
Value engineering, 834
Value functions, 2605–2608
Value-in-exchange, 669
Value-in-use, 669
Value system design, 127
Value trees, 2188–2189
"Value web," 262
Vantive, 95
VAR, *see* Vector autoregression
Variability:
  modeling, 1126
  and queueing models, 1628, 2162
Variable cone domination structures, 2616–2617
Variable data, 1837
Variable probability method (utility function assessment), 2193, 2194
Variable state activation theory (VSAT), 2209
Variable tasks, 740
Variance(s), 2367
  homogeneity of, 2255
  and hypothesis testing, 2249–2252
  hypothesis testing for equality of means and, for *k* populations, 2255–2256
  reduction of, 2492–2493
  residual, 2270–2271
  and testing for mean value, 2244–2249
Variant process-planning systems, 475–477
Variation, 1828–1855, 2266. *See also* Statistical process control (SPC)
  common vs. special causes of, 1828–1832
  and improvement of quality, 1831–1832
  in management of processes, 1830–1831
  measurement, process tolerances vs., 1986–1987
  measurement and test systems (M&TS), 1984–1986
  in operation of processes, 1830
  in processes, 1861–1863
  product-to-product, 1856–1857
  service-to-service, 1856–1857
  Shewhart control charts for determining causes of, 1834–1855
    attribute data, charts for, 1844–1851
    construction, chart, 1839–1841
    groupings of data types, 1836–1837
    individual measurements, X chart for, 1841–1844
    interpretation, 1835–1836
    subgrouping, 1837–1839
    X-bar and R control charts, 1850–1855
  in supervision/leadership, 1832
  in time studies, 1420, 1422
  tools for understanding, 1810, 1821, 1832–1834
  X-bar and R control charts for determining causes of, 1850–1855
Variety generation methods, 691, 692
VB, *see* Visual Basic
VBScript, 76, 79
VC, *see* Venture capital
VDA-FS, 192, 193
VE, *see* Virtual environments
Vector autoregression (VAR), 761, 762
Vehicle routing, *see* Routing; Transportation
Vehicle-routing problem (VRP), 2059, 2061–2063
Vehicle-routing problem with time windows (VRPTW), 2061
Vehicles, off-highway, 1470
Velocity encoders, 1903
Vendor-managed inventory (VMI), 779
Ventilation:
  energy-improvement possibilities for, 1580–1581
  work area, 1200
Venture capital (VC), 757, 759–761
Verband der Automobilindustrie-FlächenSchnittstelle (VDA-FS), 192, 193
Verification, as measurement issue for successful design, 1299, 1301
Viability, as measurement issue for successful design, 1299, 1301
Vibration allowances, 1399
Vibration analysis, 1613
Vicon Motion Systems, 1125
Video conferences, computer supported, 142
Video display terminals, *see* Screens, computer
Video-optical sensors, 385–386
Videos, online retailing of, 266
Videotaping:
  for performance rating, 1414
  for task analysis/training, 1371–1373
  for work sampling, 1456–1457
Viewing distance, computer, 1195, 1197
Virtual corporations, 107
Virtual environments (VE), 230, 234–235, 2497–2519
  applications of, 2509–2518
    education/training, 2513
    engineering task analysis for, 2514–2516
    and process integration, 2514–2518
    process simulation, 2512–2513
    scientific visualization, 2514
    virtual prototyping, 2498, 2501, 2509–2512

  definitions of, 2499–2500
  field of use for, 2497–2499
  fully immersive, 2507
  hardware for, 2501–2503
    computation systems, 2503
    display systems, 2502
    interaction/manipulation systems, 2503
    networks, 2503
    position/orientation systems, 2502–2503
  human–computer interaction for, 2504–2509
    combined interaction, 2509
    direct manipulative interaction, 2508
    formal language interaction, 2508
    gesture interaction, 2508–2509
    natural language interaction, 2508
    and visualization, 2504–2507
  projection, 2507
  software for, 2503–2504
Virtual machines, 168–170
  application of LonWorks to, 166
  definition of, 169
Virtual manufacturing, 527–528
Virtual marketplace, 262
Virtual prototyping, 2498, 2501, 2509–2512
Virtual reality (VR), 235, 251, 378, 1124, 2501. *See also* Virtual environments (VE)
Virtual Technology, Inc., 1125
Virtual workforce, 38
Viscosity, product, 2093
VISIO, 304
Vision:
  shared, 999
  for total quality leadership (TQL) process, 1801
Visionary companies, 7, 8
Visionary leadership, *see* Transformational leadership
Vision systems (test and inspection), 1904
Visual aspects of human–computer interaction, 1198–1200
Visual-based instruction, 928
Visual Basic (VB), 73–76
Visual control systems, 548–549
Visual control systems for, 548–549
Visual lobe (inspection systems), 1895
Visual Simulation Environment (VSE), 2461
Visual SLAM, 2446
Visual Thinking International Inc., 2460
VMI, *see* Vendor-managed inventory
VOCs (volatile organic compounds), 597
Voice-guided picking, 2107, 2108
Voice recognition programs, 658
Volatile organic compounds (VOCs), 597
Volume (solid) modeling, 182–183
Volumetric efficiency (in space planning), 2088–2089
Volvo, 531
VR, *see* Virtual reality
VRP, *see* Vehicle-routing problem
VRPTW, *see* Vehicle-routing problem with time windows
VSAT (variable state activation theory), 2209
VSE (Visual Simulation Environment), 2461
VW, 212
Wages, *see* Compensation
Walkie stackers, 1505
Walls, and acoustical control, 1200
Wal-Mart, 37, 263, 656, 659, 773, 775, 778–780, 782, 2115
WANs, *see* Wide area networks
Wants (of customers), 327
Warehouse management systems (WMSs), 2083–2084. *See also* Warehouse operations
Warehouse-network restructuring problem (WNRP), 2072–2079
Warehouse operations, 2083–2108
  daily operational factors in, 2103–2108
    inventory control, 2104
    order consolidation, 2106–2107
    order picking, 2104–2106
    order processing, 2104
    receiving, 2103–2104
    storage, 2104
  databases for, 2095–2103
    backups, data, 2103
    equipment masters, 2097, 2099–2102
    flow control, 2097, 2098
    hardware controllers, links to, 2103
    products and orders, 2096–2097
    protocols, 2102–2103
  functional structure of, 2084–2085
  planning for, 2088–2095
    forward-reserve allocation, 2093
    individual product assignment to storage positions, 2090, 2092–2093
    order retrieval, 2093–2095
    pick wave planning, 2095
    space utilization, 2088–2093
    zone picking, 2093–2095
  strategic factors in, 2087–2088
  tactical factors in, 2088
  terminology related to, 2087
Warehouses, 2085–2087
  catalog retailers, 2086
  data, *see* Data warehouses/warehousing
  factory warehouses, 2085
  retail distribution warehouses, 2085–2086
  stockrooms serving manufacturing facility, 2086
  terminology related to, 2087
Warehousing, 1527–1547, 2070–2081
  centralization strategy, 2071
  and contingency planning, 1530
  decentralization strategy, 2071–2072
  and decision support systems, 2079–2081
  electronic commerce and changes in, 2070–2071
  enterprise resource planning (ERP) function, 334–335
  and equipment planning, 1541–1544
  future research, directions for, 2081
  and layout planning, 1538–1541
  objectives of, 1529–1530
  and operations audits, 1544–1547
  profile analysis for selecting strategy for, 2072
  requirement for successful, 1528–1529
  in retail supply chains, 777

Warehousing (*Continued*)
  scope of, 1527–1528
  and space planning, 1532–1538
    alternative storage method space requirements, determining, 1535–1538
    materials to be stored, 1532–1534
    philosophy, storage, 1534–1535
  and strategic master planning, 1530–1532
  and value, 1528
  and warehouse-network restructuring problem, 2072–2079
    application of model, 2075–2079
    case scenario, 2072–2074
    formulation of model, 2074–2075
Warm standby components, 1933
Warnings (of workplace hazards), 1176–1177
Washington State Proposed Ergonomics Program Rule, 1166
Waste audits, 533, 534
Waste management:
  and material handling, 1502
  and plant engineering, 1569–1572
    methodology for, 1571–1572
    solid wastes, 1571
    streams, waste, 1570–1571
Waste Mangement (company), 654
Waste minimization, 533
Waste streams, 1570–1571
Waste treatment, 533
Wastewater permits, 596
Water pollution control:
  Clean Water Act, 595
  water permits, 596
Water systems, 1581
Wave soldering, 423, 424, 429
WBS, *see* Work breakdown structure
WBT (Web-based training), 940
Wealth, creation vs. preservation of, 755
Web-based procurement, 262–263
Web-based programming, 76–79
Web-based training (WBT), 940
Web browsers, 240, 244–245
Web pages, 245
Web security protocols, 734
Web servers, 256
  Active Server Pages, 79
  application program interfaces, 78
Web stores, 262
Weekend constraint (scheduling), 1747
Weibull distribution:
  of reliability, 1931, 1932, 1945–1946
  reliability estimation, 1945–1947
Weighted average cost of capital, 2334–2335
Weighted shortest processing time first (WSPT) rule, 1722–1724
Welding, 412, 413
  building parts with, 454
  and computer aided design (CAD), 188, 190
Well-being, employee, 1961
Westinghouse Electric Corporation, 1423
Westinghouse Performance-Rating Plan, 1423–1425
WFMC workflow description language, 507
WfMC (Workflow Management Coalition), 350
Wheel conveyors, *see* Skate wheel conveyors
Whirlpool, 950
Whole Food Markets, 861
Whole-population analysis, 1115
Wholesalers, and supply chain design, 2128
WHO (World Health Organization), 1165
Wide area networks (WANs), 212, 213, 231, 255–256
Wilcox signed-rank test, 2259
WISE (Workplace Improvement in Small Enterprises), 1144
Withdrawal-authorizing kanbans, 549
Within-operation analysis, 1385–1387
WMSs, *see* Warehouse management systems
WNRP, *see* Warehouse-network restructuring problem
Women, *see* Females
Words (as types of language), 132
Work, static, *see* Static efforts/work
Work attitudes, and quality-related teamwork, 979
Work breakdown structure (WBS), 1245, 1247, 1264–1280
  advantages/disadvantages of using, 1271, 1272
  applications of, 1273–1275
  change control of, 1274, 1276
  design of, 1268–1272
  and division of labor, 1264, 1266–1267
  function of, 1268
  geography-based, 1269, 1271
  logistics-based, 1271
  need for, 1266–1267
  and organizational learning, 1276–1277
  and organizational structures, 1264–1268
  for professional services projects, 1338–1340
  project-life-cycle-based, 1269, 1270
  technology-based, 1269
  types of, 1269–1271
  and work packages, 1272–1273
Work breaks, 1205
  for fatigue reduction, 1368
  time allowed for, 1394
Work cells, 1772
Work constraints, 1024–1025
Work content, 740, 742
Work cycle(s):
  definition of, 1418
  number of, for time study, 1419–1421
Work environment:
  ergonomic recommendations for, 1196
  and occupational safety and health, 1161
Worker Right to Know laws, 593
Workers:
  categories of, 1742
  deskilling of, 962
Workers' compensation, 1082
Workflow:
  and computer technologies, 1222
  in supply chain management, 2125
  use, in human–computer interface design, 1215
Workflow Management Coalition (WfMC), 350
Workflow-management systems, 221, 1251

Workforce. *See also* Employees; Staffing
  changes in, and job design/redesign, 883
  multiskilled, for just-in-time (JIT), 547–548
  reengineering and reduction in, 1710–1711
  service-driven, creating, 1959–1961
    education/training, 1959–1961
    employee well-being and satisfaction, 1961
    work systems, 1959
  virtual, 38
Work groups, *see* Team(s)
Work hours, increase in, 1888
Working arenas, identification/alignment of, 1006–1007
Working postures, 1061, 1063–1064, 1066–1067
Working surfaces, 1202–1203
Work-in-progress:
  advanced planning/scheduling and accuracy of data on, 2049
  and CIM implementation, 525
Workload modeling, 727–728, 1223
Work measurement, 1410–1413, 1562. *See also* Time study; Work sampling
  basic procedure of, 1410–1411
  common uses of, 1410
  comparison of techniques for, 1412, 1413
  definition of, 1410
  evolution of, 20
  indirect labor operations, 1458–1462
  in plant/facilities engineering, 1562
  predetermined time standards, *see* Predetermined time standards (PTS)
  standard data for, 1413, 1443–1445, 1447–1448
    calculating, 1426–1427
    definition/uses of, 1412
    development of, 1447
    elemental level systems, 1445, 1447
    establishing, 1457, 1458
    motion-level systems, 1444–1446
    reliability of, 1443, 1444
    task-level systems, 1445, 1447
    uses of, 1447–1448
  techniques for, 1411–1413
Work orders, 2051
Work pace, 1223
Work packages (WPs), 1272–1273
  and accountability, 1268
  authority for, 1268
  definition of, 1272
  for professional services projects, 1338, 1349
  responsibilities for, 1267–1268
Workplace analysis/design:
  and digital human modeling, 1120–1121
  physical tasks, 1061–1068
    international standards, 1065–1068
    postural analysis systems, use of, 1061–1063
    working postures, 1063–1064, 1066–1067
  for safe/healthful workplaces, 1177–1178
Workplace Improvement in Small Enterprises (WISE), 1144
Work practices:
  and human–computer interaction, 1205
  models of, 1210
Work-related diseases, 1082–1084
Work-related injuries, 1070, 1082. *See also* Occupational safety and health
  definition of, 1168–1170
  descriptions of, 1167–1170
  statistics related to, 1157, 1173–1174
Work-related musculoskeletal disorders (WRMDs), 1082–1086, 1166
  conceptual models for development of, 1083–1086
  definition of, 1082
  job analysis and design for reduction of, 1093–1097
    procedures, 1095
    risk factors, 1093–1094
    surveillance, 1095–1097
  management of, 1091–1093
    with administrative/engineering controls, 1092–1093
    with ergonomic guidelines, 1091–1092
Work-related upper-extremity disorders (WUEDs), 1082
  causal mechanism for development of, 1086
  definition of, 1082
  ergonomic design for reduction of, 1086–1091
    programs, ergonomics, 1097
    proposed OSHA regulations, 1097–1100
    and quantitative models, 1087
    wrist/hand disorders, 1087–1091
  occupational risk factors for, 1086
Work sampling, 1393, 1413, 1448–1458
  computerized, 1458
  control chart techniques with, 1457
  definition/uses of, 1412
  methodology for, 1449–1451
  objectives of, 1448–1449
  planning for, 1451–1457
    collection methods, determining, 1456–1457
    frequency of observations, determining, 1454–1456
    necessary observations, determining, 1451–1454
  and selection, 923
  standard times, establishment of, 1457, 1458
Work schedules/scheduling, *see* Personnel scheduling
Work simplification, 740
Work statements, *see* Work packages
Workstations:
  design of:
    ergonomic recommendations for, 1196
    and human–computer interaction, 1202–1205
  in food service kitchens, 834
  groups of, 1354–1358
  individual, 1357–1362
Work systems:
  Baldrige criteria for, 1960
  service-driven, creating, 1959
Work teams, *see* Team(s)
Work team support software, 143

Work to be done, determining quantity of, 1742
Work week, definition of, 1746–1747
World Health Organization (WHO), 1165
World Wide Web Consortium (W3C), 77, 244–245, 269
World Wide Web (WWW), 236, 238, 244–246, 256–257
  architecture of, 245
  and client/server (C/S) systems, 712–713
  collaboration/integration supported by, 246
  content provision services for, 250
  history of, 244–245
  HTML as language of, 76
  and HTTP/HTML, 245–246
  multimedia elements of, 246
  servers, 240
  storefronts on, 265–266
  as tool for communication/exchange of information, 246
  and universal information access, 246
WPs, *see* Work packages
Wrappers, 764, 765
Wrist/hand disorders, 1087–1091
Wrist rests, computer keyboards, 1202
WRMDs, *see* Work-related musculoskeletal disorders
WSPT rule, *see* Weighted shortest processing time first rule
W3C, *see* World Wide Web Consortium
WUEDs, *see* Work-related upper-extremity disorders
WWW, *see* World Wide Web

X-bar control charts, 1841–1844, 1850–1855
Xerox, 784, 859, 2122
XHTML (programming language), 77
XML, *see* Extensible Markup Language
X Mosaic, 244
XRP, *see* Extended Resource Planning
XSL (programming language), 77

Yahoo, 266, 272
Yamaha, 964
Yerkes-Dodson law, 933, 2208
YHAT, 2268, 2272
Yield-enhanced cash substitutes, 761
Yield management, 676–677
Y-YHAT, 2268, 2272

Zachman enterprise framework, 302–303
Zero inventory, 545
Zero Knowledge systems, 269
Zone(s):
  definition of, 2087
  warehouse, 2101
Zone picking, 2093–2095